



# Chemogenomics

*Edited by: Sandro Castro*



# Chemogenomics





# Chemogenomics

Editor:

---

**Sandro Castro**

**Chemogenomics**

**Editor: Sandro Castro**

**www.bibliotex.com**

**email: info@bibliotex.com**

**e-book Edition 2024**

**ISBN: 978-1-98467-787-7 (e-book)**

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated. Copyright for individual articles remains with the authors as indicated and published under Creative Commons License. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data and views articulated in the chapters are those of the individual contributors, and not necessarily those of the editors or publishers. Editors or publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

**Notice:** Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

**© 2024 Intelliz Press**

In Collaboration with Intelliz Press. Originally Published in printed book format by Intelliz Press with ISBN 978-1-68251-847-2



# TABLE OF CONTENTS

|                      |           |
|----------------------|-----------|
| <i>Preface</i> ..... | <i>xi</i> |
|----------------------|-----------|

|                  |                                      |          |
|------------------|--------------------------------------|----------|
| <b>Chapter 1</b> | <b>Introduction to Chemogenomics</b> | <b>1</b> |
|------------------|--------------------------------------|----------|

|  |    |
|--|----|
| Introduction.....  | 1  |
| 1.1 Chemogenomics in Drug Discovery .....  | 2  |
| 1.1.1 Chemical Biology.....  | 3  |
| 1.1.2 Chemical Genetics .....  | 4  |
| 1.1.3 Chemogenomics .....  | 5  |
| 1.1.4 Enzyme Inhibitors .....  | 9  |
| 1.1.5 Receptor Ligands .....   | 12 |
| 1.2 Chemogenomics Strategy .....   | 16 |
| 1.2.1 Forward Chemogenomics .....  | 19 |
| 1.2.2 Reverse Chemogenomics .....  | 20 |
| 1.2.3 Predictive Chemogenomics.....  | 21 |
| 1.2.4 Ligand and target selection.....   | 23 |
| 1.2.5 Reverse-chemogenomics case study: designing a<br>focused library of ligands for monoamine-related<br>GPCRs ..... | 24 |
| 1.3 Chemogenomics Application.....   | 26 |
| 1.3.1 The Definition of the Mode of Action.....  | 29 |
| 1.3.2 The Identification of New Drugs.....   | 30 |
| 1.3.3 Identifying Genes in Biological Reactions .....  | 31 |

|  |    |
|--|----|
| 1.4 Challenges and Limitations.....      | 31 |
| 1.4.1 Hit Selection and Validation ..... | 32 |
| 1.4.2 Data Integration.....              | 33 |
| References .....                         | 35 |

|                  |   |           |
|------------------|---|-----------|
| <b>Chapter 2</b> | <b>Chemogenomics Analysis of Drug Targets</b> | <b>41</b> |
|------------------|---|-----------|

|   |    |
|---|----|
| Introduction.....   | 41 |
| 2.1 Comparative Chemogenic Analysis For Predicting Drug-Target.....                               | 42 |
| 2.1.1 Identification of Chemogenomic Features from Drug-Target Interaction.....                   | 45 |
| 2.1.2 Materials .....   | 47 |
| 2.1.3 Model .....   | 48 |
| 2.1.4 Binary classifiers .....  | 48 |
| 2.1.5 Extraction of Chemogenomic Features .....   | 49 |
| 2.2 Open-source Chemogenomic data-driven algorithms for predicting drug-target interactions ..... | 52 |
| 2.2.1 Data set transformation .....   | 55 |
| 2.2.2 Cross-validation and Evaluation Metric .....  | 55 |
| 2.2.3 Open-source chemogenomic data-driven DTI prediction algorithms .....                        | 56 |
| 2.2.4 Algorithm Comparison Procedure .....  | 57 |
| 2.2.5 Funding.....  | 62 |
| 2.3 Chemogenomic Approaches to Rational Drug Design.....  | 63 |
| 2.3.1 Ligand space.....   | 64 |
| 2.3.2 Target space .....  | 65 |
| 2.3.3 Target-ligand space.....  | 66 |
| 2.4 Ligand-Based Chemogenomic Approaches.....   | 67 |
| 2.4.1 Ligand-based <i>in silico</i> screening.....  | 69 |
| 2.5 Target-Based Chemogenomic Approaches.....   | 70 |
| 2.5.1 Sequence-based comparisons .....  | 71 |
| 2.5.2 Structure-based comparisons .....   | 71 |
| 2.5.3 Chemical annotation of target binding sites.....  | 72 |
| 2.5.4 2-D searches.....   | 72 |
| 2.5.5 3-D searches.....   | 73 |

|   |    |
|---|----|
| 2.5.6 Concluding remarks.....   | 74 |
| 2.6 Gene Knockout Technology .....  | 74 |
| 2.6.1 Utility and Importance of Gene Knockout Animals...                  | 74 |
| 2.6.2 Experimental approaches to modulate gene<br>expression in vivo..... | 76 |
| References .....  | 78 |

## **Chapter 3    The Value of Chemical Genetics in Drug Discovery    79**

|  |     |
|--|-----|
| Introduction.....  | 79  |
| 3.1 Knowledge Management in Drug Discovery .....   | 80  |
| 3.2 Knowledge Gaps, Their Importance, and How to<br>Address Them.....  | 81  |
| 3.3 Target Validation: The Foundation of Drug Discovery .....  | 83  |
| 3.4 Chemical Genetics – How Chemistry Can Contribute<br>to Target Identification and Validation.....                               | 84  |
| 3.5 Integration of Chemistry and Biology: Importance<br>and Issues.....  | 87  |
| 3.6 Finding New Chemical Tools and Leads .....   | 87  |
| 3.7 Is Biological Selectivity an Illusion? .....   | 102 |
| 3.8 Synthesis of Chemical Genetics Libraries: New Organic<br>Synthesis Approaches to the Discovery of Biological<br>Activity ..... | 106 |
| 3.9 Information and Knowledge Management Issues.....   | 109 |
| 3.10 Annotation of Small Molecules.....  | 110 |
| References .....   | 113 |

## **Chapter 4    Structural Informatics: Chemogenomics 115**

|  |     |
|--|-----|
| Introduction.....  | 115 |
| 4.1 Structural Informatics .....                                   | 117 |
| 4.1.1 Calculating the Structural Informatics Universe .....        | 118 |
| 4.1.2 Structural Relationships .....                               | 120 |
| 4.1.3 Binding Site Relationships .....                             | 121 |
| 4.1.4 Ligand Binding Mode Relationships .....                      | 125 |
| 4.2 Tools for Ligand Based Drug Design.....                        | 125 |
| 4.2.1 Quantitative Structure–activity Relationship<br>(QSAR) ..... | 126 |
| 4.2.2 Pharmacophore.....   | 129 |

|   |     |
|---|-----|
| 4.2.3 Target Fishing.....                 | 132 |
| 4.2.4 Reverse Docking.....                | 135 |
| 4.3 Bioinformatics .....                  | 138 |
| 4.3.1 Data of Bioinformatics .....        | 139 |
| 4.3.2 Storage and Retrieval of Data ..... | 140 |
| 4.3.3 Goals.....                          | 141 |
| 4.3.4 Relation to Other Fields.....       | 143 |
| References .....                          | 144 |

## **Chapter 5      A Chemical Genomics Approach for Ion Channel Modulators      147**

|   |     |
|---|-----|
| Introduction.....   | 147 |
| 5.1 Structural Information on Ion Channels: Ion Channel Families..... | 150 |
| 5.2 Lead-finding Strategies for Ion Channel Modulators .....          | 157 |
| 5.2.1 Ligand-based Lead Finding .....                                 | 157 |
| 5.2.2. Structure-based Lead Finding .....                             | 160 |
| 5.3 Design of Ion Channel Focused Libraries: Chemical Genomics .....  | 165 |
| 5.3.1 Design Principles.....  | 165 |
| 5.3.2 Example: Building the Aventis Ion Channel Library.....          | 169 |
| References .....  | 175 |

## **Chapter 6      Phosphodiesterase Inhibitors: A Chemogenomic View      179**

|   |     |
|---|-----|
| Introduction.....   | 179 |
| 6.1 Basics of Phosphodiesterase Inhibitors.....                       | 180 |
| 6.1.1 How do phosphodiesterase inhibitors work? .....                 | 181 |
| 6.1.2 Uses.....   | 181 |
| 6.1.3 Some Examples of Common Phosphodiesterase Inhibitors.....       | 183 |
| 6.1.4 Side Effects of Phosphodiesterase Inhibitors.....               | 184 |
| 6.1.5 Important Facts to Know About Phosphodiesterase Inhibitors..... | 185 |
| 6.2 Modular Structure of PDEs .....                                   | 185 |
| 6.2.1 The PDE Superfamily.....  | 185 |
| 6.2.2 Structure/Function of PDEs .....                                | 186 |

|  |     |
|--|-----|
| 6.2.3 Significance of the PDE Complexity .....                                   | 188 |
| 6.3 Mechanisms of Regulation of PDEs .....                                       | 189 |
| 6.3.1 Protein-Protein Interaction and PDE Function .....                         | 190 |
| 6.3.2 Cyclic Nucleotide and Other Allosteric<br>Regulations of PDEs.....         | 191 |
| 6.3.3 Posttranslational Modification.....  | 192 |
| 6.3.4 Signaling Cascades Involving PDEs in<br>Endocrine Cells.....               | 192 |
| 6.4 General Pharmacology of<br>cAMP-Dependent Phosphodiesterase Inhibitors.....  | 199 |
| 6.4.1 In PDE3 Inhibitors .....   | 199 |
| 6.4.2 In PDE5 inhibitors .....   | 202 |
| 6.4.3 Therapeutic Indications .....  | 202 |
| 6.4.4 Specific Drugs .....   | 203 |
| 6.4.5 Side Effects and Contraindications .....                                   | 204 |
| 6.5 PDE6 Inhibitors.....   | 205 |
| 6.5.1 PDE6, the central effector of visualtransduction<br>in rods and cones..... | 205 |
| 6.5.2 Subunit composition and structureof the PDE6<br>holoenzyme .....           | 207 |
| 6.5.3 Similarities and differences between PDE5<br>and PDE6 .....                | 210 |
| 6.5.4 Regulation of PDE5 and PDE6 by post-translational<br>modifications.....    | 211 |
| 6.5.5 Drug selectivity for PDE5 and PDE6 .....                                   | 212 |
| 6.6 Inhibitors of Other<br>Phosphodiesterases .....                              | 212 |
| 6.6.1 PDE1 .....   | 212 |
| 6.6.2 PDE2 .....   | 216 |
| 6.6.3 PDE7 .....   | 218 |
| References .....   | 219 |

## Chapter 7 Computational Chemogenomics 221

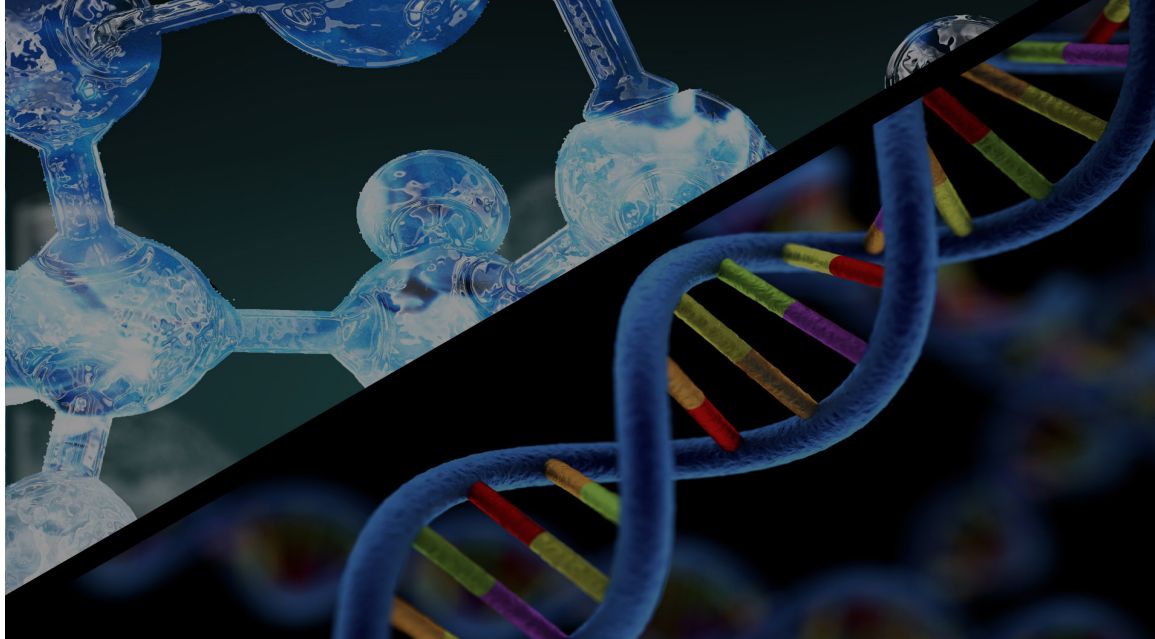
|  |     |
|--|-----|
| Introduction.....  | 221 |
| 7.1. Computational chemogenomics: Is it more than<br>inductive Transfer.....                 | 222 |
| 7.1.1. New Insights in Protein Kinase Conformational<br>Dynamics .....                       | 229 |
| 7.2. Chemogenomics Approaches for the Quantitative<br>Comparison of Biological Targets ..... | 241 |

|   |     |
|---|-----|
| 7.2.1. Target-Based Similarity Methods.....                 | 244 |
| 7.2.2. Ligand-Based Target Comparison .....                 | 249 |
| 7.2.3. The Impact of Data Quality on<br>Chemogenomics ..... | 260 |
| References .....  | 267 |

## INDEX

271





## PREFACE

Until the recent sequencing of the human genome, drug discovery has long been a multidisciplinary effort to optimize ligands properties (potency, selectivity, and pharmacokinetics) towards a single macromolecular target. A robust knowledge of the interactions between small molecules and specific proteins aids the development of new biotechnological tools and the identification of new drug targets, and can lead to specific biological insights. Chemogenomics is a complementary strategy for the investigation of chemically related compounds and libraries against various members of a target family. It is largely based on the intelligent application of automated parallel synthesis. Chemogenomics is a new strategy in drug discovery which, in principle, searches for all molecules that are capable of interacting with any biological target. Because of the almost infinite number of drug-like organic molecules, this is an impossible task. Therefore Chemogenomics has been defined as the investigation of classes of compounds (libraries) against families of functionally related proteins.

This thorough book provides a collection of techniques used in the emerging field of Chemogenomics. Chemogenomic modeling is concerned with the application of techniques to extract patterns in ligand-target binding, aiming to exploit similarity of bioactivity between similar molecules which is then expected to contribute to the identification of

bioactive pairs more efficiently than random molecule selection and activity measurement. In this book, the concepts and core elements to build a computational Chemogenomic platform are presented, with special emphasis on adaptive instance selection by the active learning technique. Despite its adaptation to drug discovery almost two decades prior, it is not until recently that active learning has been investigated in Chemogenomic contexts; nonetheless, the technique is demonstrating the ability to build models of ligand-target binding that are also predictive on external prediction challenges.



# CHAPTER 1

## INTRODUCTION TO CHEMOGENOMICS

### INTRODUCTION

Chemogenomics, or chemical genomics, is the systematic screening of targeted chemical libraries of small molecules on specific target families of drugs, with the ultimate goal of identifying new drugs and medicines. Usually some members of a target library have been well characterized, where the function has been determined and compounds that modulate the function of these goals have been identified. Other members of the target family may have an unknown function with no known ligands and are therefore classified as orphan receptors. By identifying screening hits that modulate the activity of the less well characterized members of the family, the functions of these new tasks can be clarified. In addition, it is ideal for these purposes can be used as a starting point for drug discovery. The completion of the project “human Genome” has provided an abundance of potential targets for

therapeutic intervention. Chemogenomics aims to explore the intersection of all possible drugs to all these potential targets.

The General method of constructing target chemical library should include known ligands of at least one and preferably several members of the family. Because some of the ligands that have been designed and synthesized to bind to one family member are linked to additional family members, compounds contained in the target chemical library must collectively associated with a high percentage of the target family.

## 1.1 CHEMOGENOMICS IN DRUG DISCOVERY

Chemogenomics is a new strategy in drug discovery which, in principle, searches for all molecules that are capable of interacting with any biological target. Because of the almost infinite number of drug-like organic molecules, this is an impossible task. Therefore chemogenomics has been defined as the investigation of classes of compounds (libraries) against families of functionally related proteins. In this definition, chemogenomics deals with the systematic analysis of chemical–biological interactions. Congeneric series of chemical analogs are probes to investigate their action on specific target classes, e.g., GPCRs, kinases, phosphodiesterases, ion channels, serine proteases, and others. Whereas such a strategy developed in pharmaceutical industry almost 20 years ago, it is now more systematically applied in the search for target- and subtype-specific ligands. The term “privileged structures” has been defined for scaffolds, such as the benzodiazepines, which very often produce biologically active analogs in a target family, in this case in the class of G-protein-coupled receptors. The SOSA approach is a strategy to modify the selectivity of biologically active compounds, generating new drug candidates from the side activities of therapeutically used drugs.

Chemical biology, chemical genetics, and chemogenomics are recent strategies in drug discovery. Although definitions in the literature are somehow diffuse and inconsistent, a differentiation

of the terms will be attempted here: Chemical biology may be defined as the study of biological systems, e.g., whole cells, under the influence of chemical libraries. If a new phenotype is discovered by the action of a certain substance, the next step is the identification of the responsible target.

Chemical genetics is the dedicated study of protein function, e.g., signaling chains, under the influence of ligands which bind to certain proteins or interfere with protein–protein interaction; sometimes orthogonal ligand–protein pairs are generated to achieve selectivity for a certain protein. Chemogenomics defines, in principle, the screening of the chemical universe, i.e., all possible chemical compounds, against the target universe, i.e., all proteins and other potential drug targets. Whereas this task can never be achieved, due to the almost infinite size of the chemical universe, the systematic screening of libraries of congeneric compounds against members of a target family offers unprecedented chances in the search for compounds with significant target or subtype specificity

### 1.1.1 Chemical Biology

In classical drug discovery, research was often based on vague hypotheses on structure–activity relationships. Compounds were synthesized and tested in whole animals. If a biological effect was observed, a medicinal chemistry project started to optimize chemical structures with respect to activity, pharmacokinetic properties, and lack of toxic side effects. Later on, this approach was replaced by in vitro screening on defined targets, most often human proteins. Only in recent years have we experienced a more systematic investigation of drug-like compounds in biological systems, called chemical biology.

One illustrative example of the chemical biology approach is the discovery of monastrol, a molecule that prevents spindle formation in mitotic cells by inhibiting the kinesin Eg5, a motor protein required for spindle bipolarity. In this manner, monastrol stops cell division by mitotic arrest. Another example of the concept of chemical biology is the discovery of synthetic small molecules

that influence embryonic stem (ES) cell fate. A high-throughput phenotypic cell-based screen identified a 4,6-disubstituted pyrrolo-pyrimidine, which induces the differentiation of ES cells to neurons. Glycogen synthase kinase-3 $\beta$  (GSK-3 $\beta$ ) has been identified as the target of this compound.

On the other hand, screening of any compounds may not result in the desired output of results. The production of a 2.18 million-compound natural product library by diversity-oriented synthesis generated much hype but, so far, not the anticipated results with respect to biological activities. In a later comment, the author Stu Schreiber had to admit that the chemical diversity of his library was seemingly too narrow – “disappointingly similar” by molecular descriptors; the compounds “tend to cluster in discrete regions of multidimensional descriptor space”. This goes hand in hand with another problem: biologically active compounds seem to be distributed only in certain areas of chemical space, by their physicochemical properties and their structural features. If we consider the chemical universe as a huge ocean, with small islands or groups of islands of biologically active compounds, we have to understand and accept that most chemistry-driven approaches will end up in water, instead of discovering new islands. For the broad exploration of biology with small organic molecules, the National Institutes of Health (NIH) has started an initiative to provide a repository of chemically diverse molecules for the public and private sector

### 1.1.2 Chemical Genetics

Classical genetics sets a (random) mutation, e.g., by irradiation, and tries to conclude from a new phenotype to the genotype. “Chemical genetics” is another new term for a strategy that has also been used since long ago, in a less systematic manner; it describes the purposeful investigation of proteins by small molecules or libraries, for target identification (forward chemical genetics) or target validation (reverse chemical genetics). Sometimes, orthogonal ligand-receptor pairs are constructed if selective ligands are not available. Selective kinase inhibition



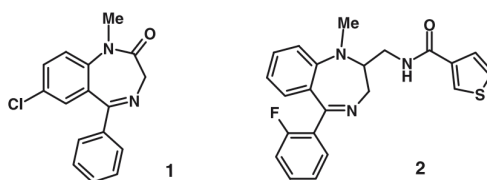
has been achieved by specifically converting nonspecific, low-affinity inhibitors into larger analogs and to construct certain kinase mutants (e.g., v-Src I338G or Cdk II F80G) that specifically accommodate these originally less well-fitting ligands by their larger binding pocket. In this manner, the specific inhibition of a certain kinase can be studied without having developed an inhibitor of comparable specificity against the wild-type kinase.

### 1.1.3 Chemogenomics

As well as in the other two cases, chemogenomics defines an approach that has also been used earlier, but less systematically. Since a screening of the chemical universe against the target universe is practically impossible, due to the almost infinite number of potential drug-like compounds, the method defines the screening of congeneric chemical libraries against certain target families, e.g., the G protein-coupled receptors, nuclear receptors, different protease families, kinases, phosphodiesterases, ion channels, transporters, etc.; this systematic strategy aims to discover highly potent, selective ligands against functionally and evolutionarily related targets, with the least effort.

#### *Privileged Structures*

Many drugs have been derived from certain chemotypes, e.g., phenethylamines, tricyclics, steroids, or benzodiazepines, whereas others have certain structural features in common, e.g., diphenylmethane, diphenylamine, or arylpiperazine groups.

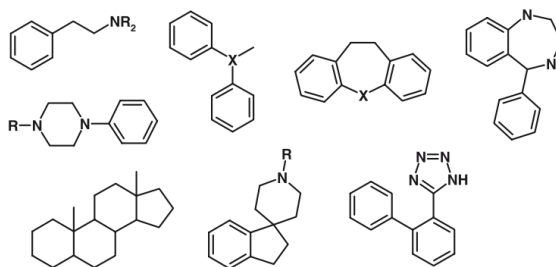


**Figure 1:** Diazepam 1 (Valium) was one of the first tranquilizers and the prototype of a series of other GABA receptor agonists, antagonists, and inverse agonists. The chemically closely related benzodiazepine Tiflua-

dom 2 is a  $\kappa$ -opiate receptor agonist and a nanomolar cholecystokinin receptor antagonist.

The systematic chemical variation of benzodiazepines, e.g., the GABA-agonist diazepam 1 produced not only tranquilizers but also GABA antagonists, inverse agonists, and the strong  $\kappa$ -opiate receptor agonist tifluadom 2 (Fig. 1)

When Evans discovered that tifluadom is also a nanomolar cholecystokinin receptor antagonist, he concluded that “these structures appear to contain common features which facilitate binding to various . . . receptor surfaces, perhaps through binding elements different from those employed for binding of the natural ligands . . . ” and formulated “. . . what is clear is that certain ‘privileged structures’ are capable of providing useful ligands for more than one receptor and that judicious modification of such structures could be a viable alternative in the search for new receptor agonists and antagonists”. Minor chemical modifications of such privileged structures (Fig. 2) may result in highly selective ligands or drugs, e.g., the estrogenic, gestagenic, androgenic, glucocorticoid, and mineralocorticoid steroids, or the  $\alpha$ -adrenergic,  $\beta$ -adrenergic, and  $\beta$ -antiadrenergic phenethylamines. Others lack such target selectivity: the atypical neuroleptic olanzapine is a highly promiscuous tricyclic ligand, with nanomolar affinities at various GPCRs, including 5-HT<sub>2A</sub>, 5-HT<sub>2B</sub>, 5-HT<sub>2C</sub>, dopaminergic D<sub>1</sub>, D<sub>2</sub>, D<sub>4</sub>, muscarinic M<sub>1</sub>, M<sub>2</sub>, M<sub>3</sub>, M<sub>4</sub>, M<sub>5</sub>, adrenergic  $\alpha_1$ , and histaminic H<sub>1</sub> receptors, as well as the 5-HT<sub>3</sub> ion channel.



**Figure 2:** Privileged structures are scaffolds or substituents that often produce biologically active compounds, e.g., phenethylamines, diphenylmethyl and diphenylamine compounds (X = C or N, respectively),



tricyclic compounds (X = C or N), benzodiazepines, arylpiperidines, steroids, spiropiperidines, and tetrazolobiphenyls (from the upper left to the lower right).

Privileged structures, even if they are promiscuous ligands, should not be confused with some structural classes, which seemingly bind with micromolar affinity to various enzymes.

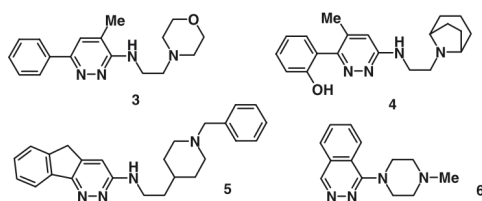
This unspecific binding behavior is caused by an aggregation of the ligands and clumping of these aggregates to the protein.

### ***Drugs from Side Effects – The SOSA Approach***

Many drugs of the past resulted from the experimental or clinical observation of side effects. Diuretic, antihypertonic, antiglaucoma, and antidiabetic drugs were derived from the bacteriostatic sulfonamides; the mood-improving effect of iproniazid was discovered when it was tested as an antituberculous drug; antidepressant inhibitors of neurotransmitter re-uptake, like imipramine and desipramine, stem from the antipsychotic dopamine antagonist chlorpromazine, which itself was derived from H1 antihistaminics; there are many other stories of this kind. Only recently, Camille Wermuth proposed to investigate the side effects of drugs more systematically, by his “selective optimization of side activities” (SOSA) approach.

Whenever a side effect of a drug is observed, it might be possible to optimize the candidate to a selective drug with this other biological activity, following a statement by Sir James Black that “the most fruitful basis for discovery of a new drug is to start with an old drug”.

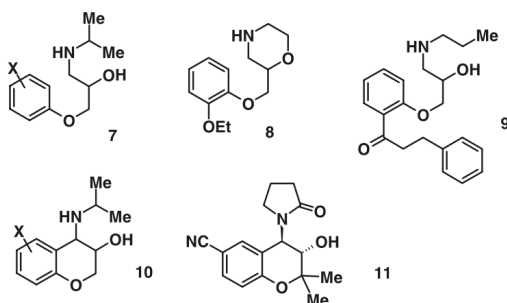
Among several other examples, Wermuth demonstrated by his own research the optimization of different weak side effects of the antidepressant minaprine 3 to the nanomolar muscarinic M1 receptor ligand 4 and the reversible acetylcholinesterase inhibitor 5; a closely related analog of minaprine was optimized to the nanomolar 5-HT<sub>3</sub> antagonist 6 (Fig. 3)



**Figure 3:** The antidepressant minaprine 3 is also a weak muscarinic M1 receptor antagonist ( $K_i = 17 \mu\text{M}$ ) and an acetylcholinesterase inhibitor ( $K_i = 600 \mu\text{M}$ ). By systematic structural variation, these activities could be enhanced to the nanomolar M1 receptor antagonist 4 ( $K_i = 3 \text{ nM}$ ) and the acetylcholinesterase inhibitor 5 ( $K_i = 10 \text{ nM}$ ). A closely related analog of minaprine was optimized to the nanomolar 5-HT3 receptor antagonist 6 ( $\text{IC}_{50} = 10 \text{ nM}$ )

### *From Target Family-Directed Masterkeys to Selective Drugs*

Chemogenomics is mainly based on the masterkey concept of tailor-made privileged structures. Starting from such masterkeys, selective ligands can be derived, either by classical medicinal chemistry or by systematic structural variation in combinatorial libraries. The masterkey concept will be illustrated by just one example: selective  $\beta_1$  and  $\beta_2$  agonists, as well as  $\beta$  antagonists ( $\beta$ -blockers) were derived from the mixed  $\alpha/\beta$  agonist epinephrine.



**Figure 4:** The  $\beta$ -blocker prototype structure 7, Phenyl-O-CH<sub>2</sub>-CH(OR<sub>1</sub>)-CH<sub>2</sub>NHR<sub>2</sub> is also the key structural element of the antidepressant viloxazine 8 and the class Ic antiarrhythmic propafenone 9. Structural variation of a cyclic  $\beta$ -blocker analog 10 yielded the potassium channel opener levcromakalim 11.

Further chemical variation of the typical  $\beta$ -blocker phenoxypropanolamine structure 7 yielded the antidepressant viloxazine 8 and the class Ic antiarrhythmic propafenone 9. The optimization of a cyclic  $\beta$ -blocker prototype 10 indeed produced an antihypertensive drug; however, levocromakalim 11 is no longer a  $\beta$ -blocker, it is a vasodilatory potassium channel opener (Fig. 4).

### 1.1.4 Enzyme Inhibitors

Protease inhibitors are most often derived from the sequence of the amino acids in the positions next to the bond that is cleaved by the enzyme. A simple strategy for a first inhibitor is a conversion of the amide bond of the cleavage site into a noncleavable analog or a group that reacts or coordinates with the catalytic center of the enzyme; the P1, P2, ... and/or P1' , P2' , ... amino acids are kept constant.

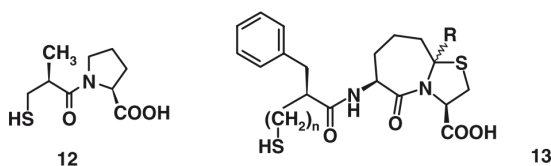
The structural requirements of the individual protease classes are different:

- For aspartyl protease inhibitors, it is necessary to attach some amino and carboxy-terminal amino acid side chains to a group that mimics the transition state of the enzymatic cleavage.
- For metalloprotease inhibitors, a metal-coordinating group is introduced at the amino-terminal side of the peptide.
- For serine and cysteine protease inhibitors, the groups that interact with the catalytic center are not necessarily but most often at the carboxy-terminal end of the peptide.

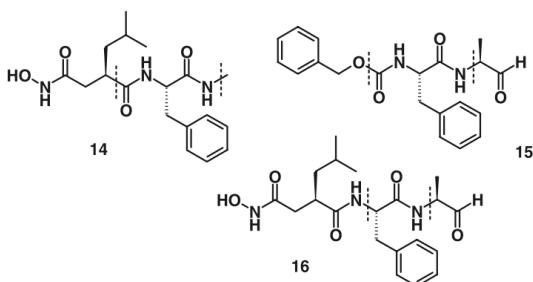
The chemogenomics strategy in the design of protease inhibitors will be illustrated by four examples: the design of HIV protease inhibitors, thrombin and factor Xa inhibitors, selective ACE and dual zinc protease inhibitors, and “dual warhead” MMP/cathepsin inhibitors. Renin is an aspartyl protease, which is involved in blood pressure regulation by converting angiotensinogen into angiotensin I, the substrate of angiotensin-converting enzyme

(ACE). Hundreds of person years of research were invested to arrive at orally active peptidomimetics, without much success. When it became known that HIV protease is also an aspartyl protease, the accumulated experience on the design of transition state inhibitors could be transferred to this new project.

The same situation applies to inhibitors of the serine protease thrombin; here also all efforts to arrive at orally active analogs had only limited success. However, structural elements from inhibitors of another serine protease, elastase, e.g., the pyrimidone ring system as a substitute for a flexible amino acid, could also be applied to thrombin inhibitors. Later on, the search for inhibitors shifted from thrombin to factor Xa, a serine protease with similar specificity as thrombin.



**Figure 5:** Captopril 12 was the very first marketed angiotensin-converting enzyme (ACE) inhibitor. The specific ACE inhibitor 13a ( $n = 0$ ,  $R = \beta\text{-H}$ ;  $K_i$  ACE = 11.5 nM,  $K_i$  NEP24.11 = 2,820 nM) resulted from structural variation, as well as the dual zinc protease inhibitors 13b ( $n = 0$ ,  $R = \alpha\text{-H}$ ;  $K_i$  ACE = 16 nM,  $K_i$  NEP24.11 = 11.5 nM) and 13c ( $n = 1$ ,  $R = \alpha\text{-H}$ ;  $K_i$  ACE = 5.5 nM,  $K_i$  NEP24.11 = 1.1 nM)



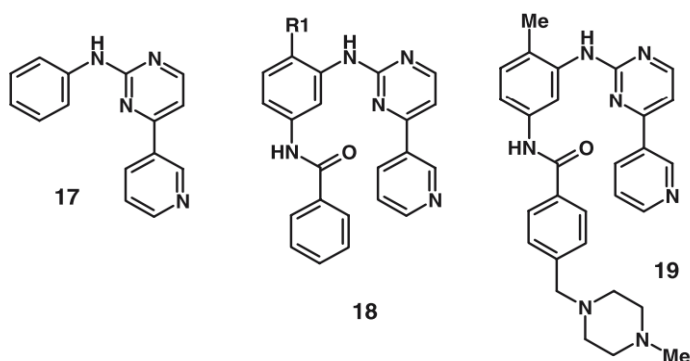
**Figure 6:** Compound 14 is a nanomolar metalloprotease inhibitor ( $IC_{50}$  MMP1 = 3 nM;  $IC_{50}$  Cat L > 1,000 nM), whereas compound 15 is a nanomolar cysteine protease inhibitor ( $IC_{50}$  MMP-1 > 1,000 nM;  $IC_{50}$  Cat L = 3

nM). Crossover of the two structures produces the dual inhibitor 16 ( $IC_{50}$  MMP1 = 25 nM;  $IC_{50}$  Cat L = 15 nM); the dashed lines indicate the common center part of all three molecules

Captopril 12 was the very first ACE inhibitor that was introduced into human therapy. A multitude of ACE-inhibiting analogs resulted from this drug, e.g., the ACE-specific inhibitor 13a and the dual ACE/NEP24.11 inhibitors 13b and 13c (Fig. 5).

A dual warhead inhibitor resulted from a merger of the structures of a selective matrix metalloprotease (MMP) inhibitor 14 with a cathepsin L inhibitor 15. Although MMP-1 is a zinc protease and cathepsin L is a cysteine protease, the resulting inhibitor 16, which bears both “warheads,” inhibits both enzymes with nanomolar activity (Fig. 6).

Kinases play a most important role in cell signaling. More than 500 different kinases are coded by the human genome; after activation, they phosphorylate either a tyrosine hydroxyl group (tyrosine kinases) or a serine or threonine hydroxyl group (serine/threonine kinases). Some kinase mutants are constitutionally active: they activate a signaling cascade without any external stimulus.

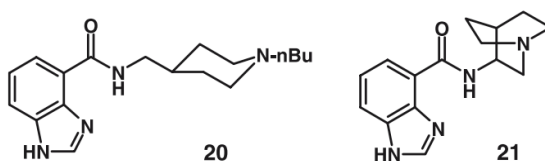


**Figure 7:** Structural variation of the protein kinase C (PKC) inhibitor 17 produced the dual PKC/bcr-abl inhibitor 18a (R = H). A minor structural modification to 18b (R = CH<sub>3</sub>) abolished the undesired PKC activity. After introduction of a methylpiperazine residue, to enhance the aqueous solubility, the bcr-abl inhibitor imatinib 19 (Glivec, Gleevec) resulted

Chronic myelogenous leukemia is caused by such a constitutionally active kinase. The coding regions of an abl tyrosine kinase at chromosome 9 and a bcr serine/threonine kinase at chromosome 22 form after reciprocal translocation a bcr-abl coding region at the new, shorter version of the chromosome 9, the so-called Philadelphia chromosome. The resulting bcr-abl tyrosine kinase is constitutionally active. At Novartis, a class of protein kinase C (PKC) inhibitors were optimized to the PKC inhibitor 17. Amide analogs 18a of this compound showed activity against PKC and bcr-abl kinase; surprisingly, the methyl analog 18b inhibited only bcr-abl kinase; finally, an N-methyl-piperazine residue was added to increase solubility (Fig. 7). Imatinib (Gleevec, Glivec), 19, was clinically developed and is successfully used for the treatment of chronic myelogenous leukemia.

### 1.1.5 Receptor Ligands

G protein-coupled receptors (GPCRs) are a large group of evolutionarily related seven-transmembrane proteins. They are activated by such different agents as light, ions, odorants, neurotransmitters, peptides, and proteins and transfer the stimulus by the G protein complex. Serotonin receptors are made up of 14 subtypes, 13 of which are GPCRs, whereas the 5-HT<sub>3</sub> subtype is a ligand-controlled ion channel.



**Figure 8:** Compound 20 is a highly selective 5-HT<sub>3</sub> antagonist ( $K_i$  5-HT<sub>3</sub> = 3.7 nM,  $K_i$  5-HT<sub>4</sub> > 1,000 nM), whereas the chemically closely related compound 21 is a selective 5-HT<sub>4</sub> antagonist ( $K_i$  5-HT<sub>3</sub> > 10,000 nM,  $K_i$  5-HT<sub>4</sub> = 13.7 nM).

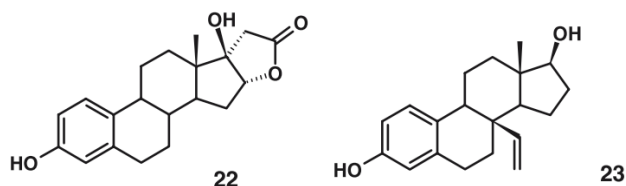
From pharmacophore models, Lopez-Rodriguez et al. designed the structure of a highly selective 5-HT<sub>4</sub> receptor ligand 20, which shows a selectivity difference of more than five orders of

magnitude to its closely related, 5-HT<sub>3</sub>-selective analog 21 (Fig. 8)

Somatostatin receptors are made up of five subtypes: sst1–sst5. In their attempt to obtain selective, peptidomimetic ligands for each subtype, Rohrer et al. synthesized four  $\beta$ -turn-mimicking combinatorial libraries, with up to 350,000 compounds per library. Highly specific ligands resulted for all five subtypes.

Nuclear receptors are another important receptor family. They are made up of a ligand-binding domain and a DNA-binding domain. After activation by their specific ligands, e.g., the steroid hormones, the thyroid hormone or retinoic acid, receptor dimers bind to DNA and activate the expression of certain proteins.

Estrogen receptors exist as two distinct subtypes, ER $\alpha$  and ER $\beta$ , which are relatively abundant in several tissues. As their function in all those organs and potential interaction, forming ER $\alpha$ /ER $\beta$  heterodimers, has not been completely elucidated so far, it is most important to find selective ligands for both receptors. By homology modeling of the ligandbinding domain of the ER $\beta$  receptor, based on the corresponding 3D structure of the ER $\alpha$  receptor, Hillisch et al. inspected the minor differences in the estradiol binding site: in human ER $\beta$ , the leucine of ER $\alpha$  at the “top” of the binding site (“top” refers to the  $\beta$  side of the steroid ring) is replaced by a flexible, sterically less demanding methionine, whereas at the “bottom” of the binding site, close to ring D, a methionine in ER $\alpha$  is replaced by an isoleucine in ER $\beta$ .



**Figure 9:** The estradiol analogs 22 (40% of estradiol activity, ER $\alpha$ -selective) and 23 (50% of estradiol activity, ER $\beta$ -selective) have been designed as selective ER $\alpha$  and ER $\beta$  receptor ligands. Even though they are less active than estradiol, they show 300-fold and 190-fold selectivity for the different receptor subtypes.

Using this information on the narrower binding pocket above and below the estradiol binding sites of ER $\alpha$  and ER $\beta$ , respectively, the selective ligands 22 and 23 could be designed (Fig. 9).

Whereas 22 has only about 40% of the activity of estradiol at ER $\alpha$ , it shows a 300-fold selectivity against ER $\beta$ ; on the other hand, compound 23 has only 50% of the activity of estradiol at ER $\beta$  but a 190-fold selectivity against ER $\alpha$ .

The thyroid hormone T3 and its less active storage form T4 are iodinated phenoxy-phenylalanines, which bind to two nuclear receptor subtypes TR $\alpha$  and TR $\beta$ . Unfortunately, the affinity of T3 to TR $\alpha$  is higher than to TR $\beta$ , which causes cardiac side effects, if hypothyroid patients are treated with T3.

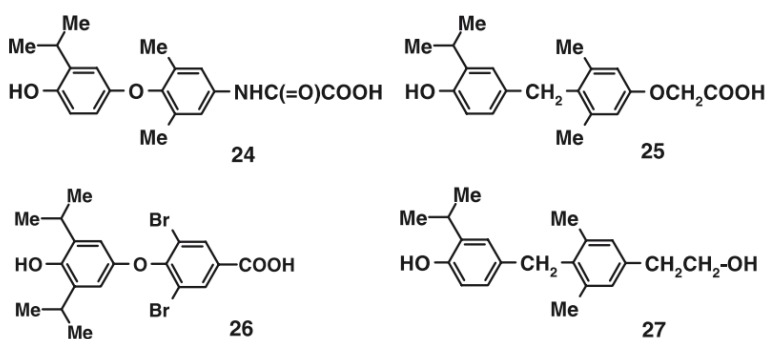
The alkyl analogs 24 and 25 are less active at TR $\alpha$  than at TR $\beta$  (Fig. 10). Compound 26 binds to both receptor subtypes but has no agonistic activity at TR $\alpha$  and is only a weak partial agonist at TR $\beta$ ; correspondingly, this compound might be used to treat hyperthyroid patients.

Other patients suffer from a R320C mutant of TR $\beta$ ; due to the exchange of the strongly basic arginine side chain against the neutral cysteine, T3 binds with much lower affinity to this receptor, causing a hypothyroid condition.

Treatment with T3 or compound 25 is impossible, due to the high affinity of these compounds to the TR $\alpha$  receptor. Conversion of the acid 25 into the neutral analog 27 solved the problem: 27 has a higher affinity to the TR $\beta$  mutant than to TR $\alpha$  (Fig. 10).

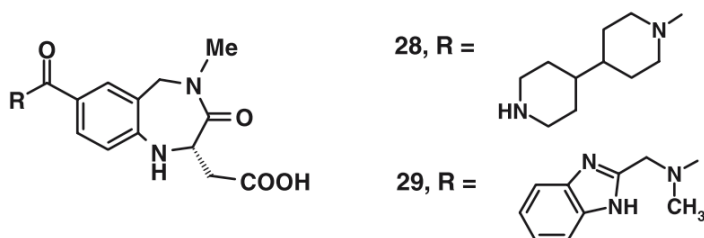
Integrins are another group of receptors. They are expressed at cell surfaces and their endogenous ligands, e.g., fibrinogen at the GP IIb/IIIa integrin (also called fibrinogen receptor) or vitronectin at the  $\alpha_v\beta_3$  integrin (also called vitronectin receptor), mediate cell-cell contacts.





**Figure 10:** Compounds 24 (CGS-23425) and 25 (GC-1, UCSF) are alkyl analogs of the thyroid hormone T3; in contrast to T3, which has a higher activity at TR $\alpha$ , these analogs have a higher activity at the TR $\beta$ . Compound 26 is a thyroid hormone antagonist at TR $\alpha$  and a weak partial agonist at TR $\beta$ . Neither T3 ( $EC_{50}$  hTR $\alpha$  = 0.14 nM,  $EC_{50}$  hTR $\beta$  = 0.66 nM,  $EC_{50}$  hTR $\beta$  R320C mutant = 4.3 nM) nor compound 25 ( $EC_{50}$  hTR $\alpha$  = 6.6 nM,  $EC_{50}$  hTR $\beta$  = 3.7 nM,  $EC_{50}$  hTR $\beta$  R320C mutant = 38 nM) has sufficient activity at a hTR $\beta$  R320C mutant. Compound 27 is a neutral, weakly active but TR $\beta$  R320C mutant-selective thyromimetic ( $EC_{50}$  hTR $\alpha$  = 38 nM,  $EC_{50}$  hTR $\beta$  = 32 nM,  $EC_{50}$  hTR $\beta$  R320C mutant = 7.0 nM)

The recognition motif of these two receptors is the Arg-Gly-Asp (RGD) sequence of the ligands, obviously in different conformations. Research at SmithKline Beecham led to the discovery of ligands that showed, after minor chemical modification of a basic side chain, some selectivity for each of these two receptors.



**Figure 11:** Compound 28 (lotrafiban,  $K_i$  GP IIb/IIIa = 2.5 nM,  $K_i$   $\alpha_v\beta_3$  = 10,340 nM; failed in phase III clinical trials) is a specific fibrinogen receptor antagonist, whereas compound 29 ( $K_i$  GP IIb/IIIa = 30,000 nM,  $K_i$   $\alpha_v\beta_3$  = 2 nM) is a specific vitronectin receptor antagonist.

After extensive structural modification, the highly selective ligands 28 (SB 214 857, lotrafiban) and 29 (SB 223 245) resulted in their selectivity (Fig. 11) (Samanen et al. 1996; Keenan et al. 1997; Miller et al. 2000). They differ by more than seven orders of magnitude.

## 1.2 CHEMOGENOMICS STRATEGY

Chemogenomics integrates target and drug discovery by using active compounds that function as ligands as probes to characterize proteome functions. The interaction between small compound and protein that causes the phenotype.

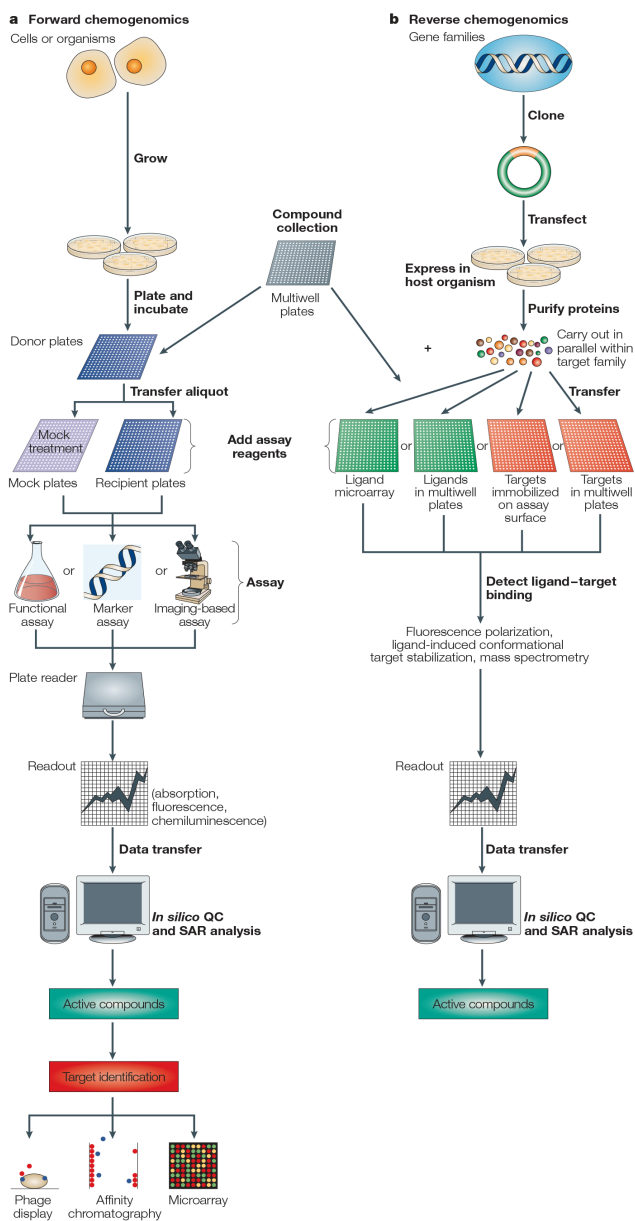
Once the phenotype is characterized, we could link the protein to a molecular event. Compared to genetics, chemogenomics methods can change the function of the protein, not the gene. In addition, chemogenomics is able to observe the interaction and reversibility in real time.

For example, a change of the phenotype can be observed only after adding a special component and can be interrupted after its withdrawal from the environment.

Currently there are two experimental chemogenomic approaches: classical forward and reverse chemogenomics. Forward chemogenomics attempt to identify drug targets in the search for molecules which give a certain phenotype on cells or animals, while reverse chemogenomics to validate phenotypes by searching for molecules that interact specifically with this protein.

Both of these approaches require a suitable collection of compounds and an appropriate model system for screening substances and looking for the parallel identification of biological targets and biologically active compounds.

Biologically active compounds that can be detected by direct or reverse chemogenomics approaches are called the modulators because they bind to and modulate specific molecular targets, so they can be used as a target therapy'.



**Figure 12:** Forward- and reverse-experimental chemogenomic approaches. The initial step in forward and reverse chemogenomics is to select a suitable collection of compounds and an appropriate model system in which to screen them. In both approaches, the sequential steps

of the assay — the transfer of ligands, cells, assay reagents and plates — can be fully or semiautomated. IT integration is a key element in industrial setups. A. In forward chemogenomics, the cell or organismal model system is typically dispensed in multiwell microtitre or nanowell plates. Solutions of single ligands are added from the stock plates to different wells. After incubation, an aliquot is transferred from the donor plate to a new recipient plate, in which the ligand–target binding assay is carried out. The effects of a compound are assayed by one of several methods: functional assays directly measure cellular activities such as cell division; marker assays, such as reporter-gene assays and whole-culture CYTOBLOTS, identify specific molecular events that act as surrogate transcriptional and post-transcriptional markers for phenotypic changes of interest; automated microscopy or imaging-based screening are innovative approaches that attempt to capture further morphological changes. The end point of most cell-based high-throughput assays is a spectroscopic readout; readout data are automatically transferred to a microprocessor for final data calculation, including *in silico* quality control and structure–activity relationship (SAR) analysis. Active compounds that achieve the desired phenotypic change are then selected to identify their molecular targets. This can be done in several ways, of which affinity matrix purification, phage display or transcriptional or proteomic profiling are the most commonly used approaches. In profiling experiments, protein or RNA isolates of the treated model system are analysed in reference to mock treatment for global molecular drug signature assessment. B. In reverse chemogenomics, emphasis is especially placed on the parallel exploration of gene and protein families. Here, target gene sequences that show a certain degree of homology are expressed in a host cell: these family target proteins are purified, collected and subjected to assay design. Based on the SAR homology concept<sup>116</sup> that the degree of similarity of ligands determines similarities in target binding, candidate ligands that show a desirable similarity to a ligand that is known to interact with one member of a target family are selected. The ultimate goal is to identify new ligands that hit either the same target or analogous target-family members. Reverse chemogenomics is normally carried out using a cell-free binding system, in which either the target protein or the libraries are immobilized on assay plates or dispensed in multiwell plates, and the study compounds or target proteins, respectively, are added in solution. Various technologies are used to detect ligand–target binding. In fluorescence-based detection, an imaging camera automatically captures the fluorescence signal that corresponds to ligand–target binding. As in forward chemogenom-

ics, readout data are transferred to a data analysis system, which uses sophisticated computational algorithms to mine the large amounts of — and, to some extent, noisy and error-prone — data.

### 1.2.1 Forward Chemogenomics

Live chemogenomics, which is also known as classic chemogenomics, the phenotype studied and the small coupling of interacting with this function defined. The molecular basis of this desired phenotype is unknown. Once the modulators have been identified, they will be used as tools to search protein responsible for the phenotype. For example, the loss of function phenotype could be the inhibition of tumor growth. After connecting that lead to a target phenotype have been identified, the identification of gene and protein targets should be the next step. The main challenge of forward chemogenomics strategy lies in designing phenotypic assays that lead immediately from screening to target identification.

In ‘forward chemogenomics’ (FIG. 12a), the molecular basis of a desired phenotype is unknown. Here, a so-called ‘phenotypic screen’ is performed in single-cell organisms or cells from multicellular organisms using a panel of ligands. The biological systems can consist of prokaryotic and eukaryotic single cell organisms (bacteria and fungi; for example, the yeast *Saccharomyces cerevisiae*), physiological or pathological cells from complex multicellular vertebrate or mammalian organisms, or even whole higher organisms, such as fly, worm, zebrafish or mouse. Subsequently, high-throughput cell-based or organismal phenotypic assays are used to identify biologically active compounds. In other words, in this approach, compounds are identified on the basis of their conditional phenotypic effect on a whole biological system rather than on the basis of their inhibition of a specific protein target. The phenotypic screen is designed to reveal a novel conditional phenotype (either a loss-of-function or a gain of-function phenotype) and the affected protein or pathway. For example, a loss-of-function phenotype

could be an arrest of tumour growth. Once biologically active compounds that lead to a target phenotype have been identified, efforts are directed towards the study of the mechanistic basis of the phenotype by identifying the gene and protein targets using various high throughput methods (FIG. 12a). The biological and structural information on the target can, in turn, be used in reverse chemogenomics to identify and develop, through HTS, new and more potent compounds that disrupt the function of the target. On the other hand, active ligands that are identified using reverse chemogenomics can be biologically validated by examining the phenotypic effect of altering the function of their protein targets in a forward chemogenomics setting. The main challenge of this chemogenomics strategy lies in designing phenotypic assays that lead immediately from screening to target identification.

### 1.2.2 Reverse Chemogenomics

In reverse chemogenomics, small compounds that disturb the function of the enzyme in terms of enzymatic in vitro test to be determined. Once the modulators have been identified, the phenotype caused by the molecules to be analyzed in test cells or on the entire body. This method will allow you to identify or confirm the enzyme role in biological reactions. Reverse chemogenomics used almost identical target sets that were used in the development of drugs and molecular pharmacology over the past decade. This strategy is now enhanced by parallel screening and the ability to perform lead optimization on many targets that belong to one target family.

In 'reverse chemogenomics', gene sequences of interest are first cloned and expressed as target proteins, which are then screened in a high throughput, 'target-based' manner by the compound library (FIG. 12b). Such a HIGH-THROUGHPUT SCREENING (HTS) method can involve many different bioassays, which monitor the effects of different compounds on specific targets (such as the ability to bind a protein), on specific cellular pathways (for example, the capacity to inhibit the mitogenic pathway of a tumour cell) or on the phenotype of a whole cell or organism.

The assays can be generally divided into cell-free, cell-based and organismal assays. Cell-free, universal binding assays — in which several compounds are simultaneously tested for their binding affinity to a wide panel of specific targets — are usually simple, precise, highly automated and compatible with a very high throughput approach. Target–ligand interactions (so-called ‘hits’) are unambiguously identified in the absence of confounding variables. For example, fluorescent-based methods for detecting the LIGAND-INDUCED CONFORMATIONAL STABILIZATION of proteins or MASS-SPECTROMETRY-based detection systems have been described as a means to examine the effect of bound ligands. By contrast, cell-based and organismal assays — in which selected compounds are delivered directly to cells or organisms *in vitro* — identify hits within a relevant cellular context, but, because of the interaction with multiple targets, hits require additional mechanistic characterization. Whereas cell-free assays are primarily used in reverse chemogenomics, if target information is available, cell-based and organismal assays are predominantly used in forward chemogenomics (see below) to examine broad compound effects on intact biological systems.

The hits that are revealed in this way are used to generate lead compounds. These are then optimized by the careful selection of the most promising candidates and through the synthesis and testing of chemical ANALOGUES with similar and, it is to be hoped, improved properties. Reverse chemogenomics is therefore virtually identical to the target-based approaches that have been applied in drug discovery and molecular pharmacology over the past decade; however, these are now enhanced by parallel screening and by the ability to perform lead optimization on many targets that belong to one target family.

### 1.2.3 Predictive Chemogenomics

Whereas the principal goal of chemogenomics is to identify new therapeutic targets and drugs, ‘predictive chemogenomics’ strategies primarily attempt to holistically characterize treatment responses, coupled with the secondary aim of identifying novel



therapeutic molecules. The central approach of predictive chemogenomics is to initially collect the genomic responses (for example, through microarray analysis) and the pharmacological responses (for example, through growth inhibition assay) of a cell type or tissue to treatment with various drugs.

**Table 1:** Key *in silico* methods that are used to support chemogenomics approaches

| <i>In silico</i> method                          | Application setting and objective  |
|--|--|
| Similarity and pharmacophore* searching          | Virtual screening or design of compound libraries to provide focused subsets, on the basis of the knowledge of a known active ligand. Homology-based searching aims to identify compounds that are active on a target for which there are no known active compounds but that are related by homology to one or more targets for which active compounds are known (using the 'structure–activity relationship homology concept') <sup>38</sup> .  |
| High-throughput docking† (HTD)                   | Virtual screening or design of compound libraries to provide focused subsets, on the basis of the knowledge of the 3D structure of the target. Identification or design of selective compounds: docking a focused library against a comprehensive panel of 3D structures of one protein family. Identification of potential targets and mechanisms of action: docking of selected compounds against the entire PDB‡ database.  |
| 3D bioinformatics target-binding site comparison | Identification of potential targets and mechanisms of action of compounds on the basis of the structure-based comparison of the ligand-binding sites. Design of selective compounds within a target family with conserved molecular recognition.   |
| Target and ligand annotation and ontologies      | By linking the sequence information of targets to their ligands, ligand–target classification schemes allow ligands to be matched to targets on the basis of their sequence similarity and provide reference sets for homology-based similarity searching. The automatic build-up of compound annotations on the basis of Medline literature reports provides the knowledge basis for generating annotated compound libraries. These can then be used to guide experiments to determine the components of signalling pathways. |

Each drug profile represents the drug’s own signature at the transcriptional and molecular pharmacological level. Biostatistical integration of the genomic and pharmacological data then reveals predicted gene–drug relationships. This approach can be extended to look at families of drugs to extract the signatures that are common to a class of molecules (that is, the effect that is linked to a chemical structure) and those that are drug specific. This strategy will not necessarily reveal drug targets but might identify molecules that significantly influence the effect of a drug. Computational or *in silico* methods can complement experimental chemogenomics strategies in the search for such predictive molecules (TABLE 1). Predictive chemogenomics has considerable overlap with ‘PHAR-



MACOGENOMICS'. In contrast to pharmacogenomics, however, predictive chemogenomics strategies generate gene–ligand response associations by concurrently considering the response profiles of thousands of drugs, rather than those of one molecule at a time.

### 1.2.4 Ligand and target selection

From a cost perspective, it is important that biologically active chemical candidates are identified quickly, efficiently and accurately. The original combinatorial chemistry approach to ligand and target identification involved the ULTRA HIGH-THROUGHPUT SCREENING (uHTS) of diverse-compound collections; however, it soon became clear that such massive screening is hardly applicable in the above experimental chemogenomics settings because of the necessary high level of financial investment and the substantial efforts needed for data handling (IT logistics and automation) and interpretation. In experimental chemogenomics, emphasis has therefore now been placed on the pre-selection of potential ligands to allow the study of smaller, focused, libraries of compounds. Several strategies have been applied to generate so-called 'targeted libraries', in which the design and selection of ligands is based on the information that is available on either the target itself (for example, through three-dimensional X-ray/nuclear magnetic resonance (NMR) structure) or on ligands that are known to interact with the target. For targets with limited or no biostructural information, PRIVILEGED STRUCTURES are often considered in library design. Today, typical chemogenomics screening libraries contain several types of designed subset, including annotated known biologically active compounds, target-family focused libraries (for example, kinases, proteases), peptide mimetics (for example,  $\beta$ -strand,  $\beta$ -turn and  $\alpha$ -helix structural mimetics), natural products and derivatives thereof, and DIVERSITY SETS of drug-like compounds.

As well as using focused library design, cost-efficient chemogenomics requires genetic sequences of relevance to be dissected from those that do not contribute to the ligand–target

interaction. For this purpose, sequence homology alignments and biostructural algorithms can be used to narrow down the number of targets to be tested.

### 1.2.5 Reverse-chemogenomics case study: designing a focused library of ligands for monoamine-related GPCRs

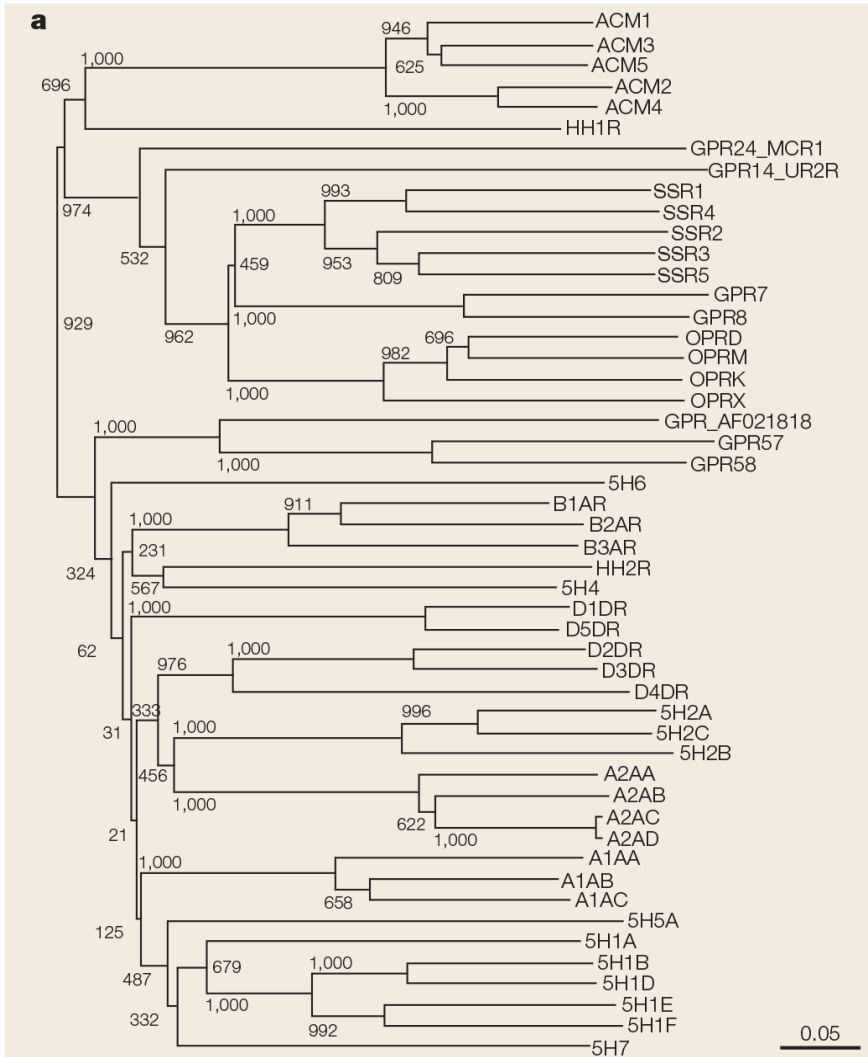
The successful design of family-focused target libraries depends on how similar the ligands and binding sites are among the members of a gene family. We recently proposed that the monoamine-related G-protein coupled receptor (GPCR) subfamily, for which motif-based sequence searches identified 50 human GPCR members, recognizes its ligands through 3 binding sites.

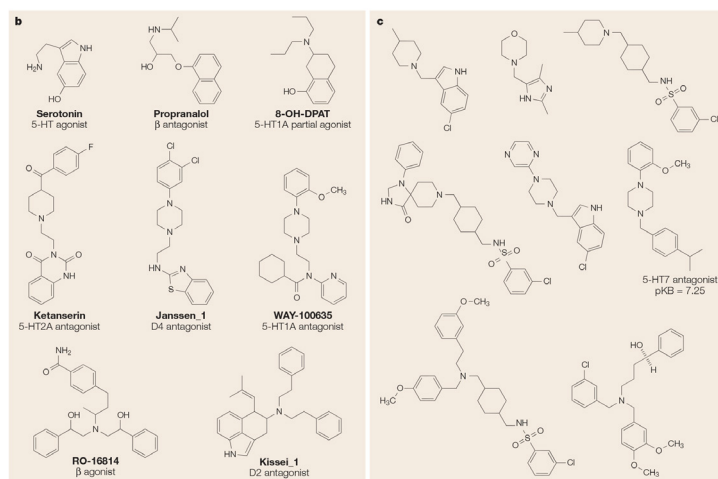
The sequence comparisons of the 50 identified GPCRs, based on the 7-transmembrane (7TM) domain, are shown in the dendrogram in panel a (the scale of sequence identity is indicated by the 5% distance bar and the numbers on the branches are BOOTSTRAP values out of 1,000 replicates).

For the serotonin 5HT<sub>1A</sub>-receptor subtype, each of the three spatially distinct binding regions allows the receptor to bind a different ligand (panel b, top). These regions are located within the highly conserved 7TM domain of the receptor and overlap at the residue D3.32 in TM3, which is responsible for the recognition of the basic amino group of the ligands. This information motivated the design of the Novartis tertiary amine (TAM) combinatorial ligand library. The TAM structures, for which prototypes are shown in panel c, were designed to be similar in architecture and properties to known monoamine-related GPCR ligands, for which examples are shown in panel b.

The successful search for antagonists for the 5HT<sub>7</sub> GPCR, which has the 5HT<sub>1A</sub> receptors as next neighbor in the sequence dendrogram, illustrates the use of the TAM library. By searching with 5HT<sub>1A</sub> reference compounds in the TAM library (using the Similog method), we were able to identify a 10% hit rate ( $p_{KB} < 5 \mu\text{M}$ , where  $p_{KB}$  = the negative logarithm of the binding constant) when only

a biological assay with limited capacity was available — that is, when only a limited number of compounds could be screened. The hits corresponded to arylpiperazines(see panel c), which, in follow-up studies, were also active on other monoamine-related GPCRs.





## 1.3 CHEMOGENOMICS APPLICATION

Chemogenomics strategies are increasingly being harnessed by various fields of medical research – for example, those related to cancer or immune, inflammatory and hormone disease – in attempts to develop new targeted therapies as rapidly as possible.

Several advances have been made by chemogenomics in understanding the molecular biology of various diseases and in identifying potential pharmacological therapies for them. TABLE 2 summarizes some recent results of chemogenomics research into human diseases.

In general, chemogenomics approaches can be used for three different purposes in disease research.

First, chemogenomics can be used to identify new drug targets and might allow their biological functions to be understood. In this context, forward chemogenomics strategies are used to initially examine the phenotypic effect of a compound or a panel of compounds on a

**Table 2: Chemogenomics applications**

| Focus/approach                 | Ligands   | Target                           | Disease                           | Model   | Validation model   |
|--------------------------------|---|----------------------------------|-----------------------------------|---|--|
| Target identification/F        | FK228, Trichostatin A, HDAC Depudecin           | HDAC                             | Cancer/angiogenesis               | Hras-Ras1, vras-NIH3T3 cells, mammalian cell lines  | Proliferation/enzyme assay   |
| Target identification/F        | TNP-470   | MetAP2                           | Cancer/angiogenesis               | W303 yeast strain, endothelial cells  | Yeast-deletion model   |
| Target identification/F        | Dihydroepone-mycin                              | 20S proteasome (LMP2, LMP7)      | Cancer/angiogenesis               | EL4 murine thymoma cells  | 2D-gel electrophoresis, immunoblotting                               |
| Target & drug identification/F | PNRI-299  | Ref1                             | Asthma                            | A549 lung epithelial cells  | BALB/c mice  |
| Target & drug identification/F | Radical-derivatives BR-1, BR-6, KT8529, KF25706 | Hsp90, ACL                       | Cancer/angiogenesis               | NIH3T3, RAS-373, SRC-373 mouse fibroblasts, HeLa cells, normal rat and human tumour cells | BALB/c mice, ex vivo western blot, immunoblotting, enzyme inhibition |
| Target & drug identification/F | FK506, cyclosporine A                           | Calcineurin, CPH1 & FPR1         | Immune disease                    | Yeast strains   | <i>In vitro</i> and mice   |
| Drug & target identification/F | K252a   | CaMPKs                           | Neuroinflammation                 | BV-2 microglia cells  | Western blot   |
| Drug identification/F          | Compound A                                      | pRB                              | Cancer                            | HT29, HCT116, C33A cancer cells (cytoblot)  | Western blot   |
| Drug identification/F          | Compounds 5a-5h                                 | Ras-Raf                          | Cancer                            | MDCK epithelial cells, MDCK-F3 (H-Ras)  | Phenotype reversal   |
| Drug identification/R          | CDK-731   | MetAP-2                          | Cancer/angiogenesis               | Endothelial cells   | Phenotype reversal (proliferation inhibition)                        |
| Drug identification/R          | Peptide 18 & analogues                          | CaMPKs/MLCK                      | Various                           | Enzyme inhibition assay   | Kinetic analysis of target inhibition                                |
| Drug identification/R          | Isatin oximes 30, 50, 51                        | UCH-L1                           | Cancer                            | High-throughput screening   | Phenotype induction in H1299 lung cancer cells                       |
| Drug identification/R and IS   | Compounds 1a-c and 11l                          | HDAC                             | Cancer/angiogenesis               | <i>In vitro</i> enzyme assay, mouse A20 cells, virtual mutation                           | Phenotype reversal in murine erythro-leukemia cells                  |
| Drug identification/R and IS   | NSC-65828, R5A, H8A, N68A, des (121-123)        | AGN                              | Cancer/angiogenesis               | Cell-free high-throughput screening, VHTS   | Athymic mice (s.c. PC-3 prostate and HT-29 colon cancer cells), HPLC |
| Drug identification/IS         | 1-850, D1-D4                                    | TR                               | Hyper-thyroidism                  | VHTS  | HeLa, GH4 rat pituitary cells  |
| Drug identification/IS         | Indoloquina-zolinones 3a, 4, 5                  | CK2                              | Cancer                            | VHTS  | Rat liver kinase inhibition assay                                    |
| Drug mechanism/F               | Wortmannin                                      | 1,067 on/off targets             | Immune/inflammatory disease, etc. | 6,025-strain homozygous and heterozygous yeast-deletion collection                        | Phenotype complementation by ORF reintroduction                      |
| Drug mechanism/F               | Rapamycin                                       | TOR1p, TOR2p, 106 gene mutations | Cancer, immune disease            | 2,216 homozygous and 50 heterozygous yeast-deletion strains                               | Transfection   |
| Drug mechanism/F               | Artesunate                                      | 54 genes                         | Cancer                            | 55 human tumour cell lines  | Transduction ( $\Delta$ EGFR, GCS, CDC25A)                           |
| Response factor/P              | >70,000 anti-cancer drugs                       | Many on/off targets              | Cancer                            | 60 human cancer cell lines (NCI60)  | 71   |

Biological system, followed by an investigation into how these compounds interact with the drug targets. As a result of these approaches, the number of newly identified targets (and compounds) is steadily increasing (TABLE 2). One of the earliest successes of this method was the discovery of the immunosuppressant FK-506 (which has entered clinical practice as tacrolimus) in 1987 and the subsequent identification of calcineurin as its molecular target. The application of forward chemogenomics to cancer and angiogenesis research, for example, has identified the heat-shock protein 90 (HSP90) molecular chaperone and ATP citrate lyase

(ACL) as targets of radicicol (*Hemicolli fuscoatra*, an antifungal antibiotic that has anti-tumour and anti-neoplastic activity) and its analogues KT8529, KF25706, BR-1 and BR-6 (REFS 32–34). Second, chemogenomics approaches are applied to discover, in a high-throughput fashion, new chemical candidates for molecular targets and phenotypes of interest. Hundreds of novel drug-like ligands have been identified for various molecular targets in the past few years through reverse, in silico and forward chemogenomics (TABLE 2). These targets relate to different disease areas, such as cancer, asthma, neuroinflammation and hyperthyroidism. Finally, chemogenomics can be used to understand the mechanism of drug action, which also includes finding genetic markers of drug susceptibility (TABLE 2). For example, the complex molecular effects of rapamycin (an antibiotic derived from *Streptomyces hygroscopicus* that has antifungal, anti-inflammatory, anti-tumour and immunosuppressive properties) and wortmannin (a fungal metabolite isolated from *Penicillium wortmanni* that has anti-neoplastic and radio sensitizing effects) were recently examined in a chemogenomics fashion using homozygous and heterozygous collections of the yeast *S.cerevisiae*: defined and known regions of the yeast genome had been deleted and targets were identified on the grounds that a yeast strain bearing a deletion in a gene that encodes a protein target is more resistant to treatment. Similarly, the heterozygous yeast-deletion model has been recently used to identify gene products that functionally interact with various compounds, including anticancer, antifungal and anticholesterol agents. The published data show signs of promise for all three aspects of chemogenomics.

***Combining forward and reverse chemogenomics.*** Several studies stress the power of combining forward and reverse chemogenomics in concurrent target and drug discovery. As such, targets identified for lead compounds by phenotypic screening have been subsequently used to develop more potent synthetic analogues by HTS. To take a specific example, sequential forward and reverse chemogenomics have led to the initial identification of histone deacetylase (HDAC) and the subsequent development of HDAC inhibitors, such as depsipeptide, sodium phenylbutyrate, CI-994 and suberoylanilide hydroxamic acid (SAHA). Clinical testing of

these compounds has already begun in patients with advanced solid cancers, peripheral and cutaneous T-cell lymphoma, acute myeloid leukaemia and MYELODYSPLASTIC SYNDROME. In turn, phenotypic screening of compounds that have been recently discovered by reverse chemogenomics to inhibit a certain biological pathway can be used to identify the immediate target of compound action in this pathway. For example, the redox effector factor-1 (Ref1) gene has been recently identified as a therapeutic target for asthma, following HTS for small-molecule inhibitors of activator protein-1 (AP1) transcription. The potential of chemogenomics to identify novel compounds for many targets has stimulated particular interest in its application to cancer research. Compared with other diseases, cancers normally demonstrate complex genetic changes. These changes involve multiple alterations at the genetic and gene-expression levels that lead to aberrant protein (target) abundance. The genetic and epigenetic profile is highly variable even among cancers that seem to be histologically similar. In addition, genetic instability can cause substantial intratumour genetic heterogeneity, which further increases the number of molecular aberrations. This presents two difficult challenges: how to identify the enormous quantity of pathogenic changes (and therefore potential targets) that are present in tumours and how to identify the many targeted therapeutics that are necessary to address as many genetic alterations as possible. Chemogenomics might successfully rise to both these challenges and it is in fact in the field of cancer research that currently most chemogenomics studies have been done.

### 1.3.1 The Definition of the Mode of Action

Chemogenomics have been used to determine the MOA mode of action of traditional Chinese medicine TCM and Ayurveda. Substance contained in traditional medicine is usually more soluble than synthetic compounds, "privileged structures" chemical structure, which are more often used for binding in various living organisms, and more fully known safety factors and tolerance. Thus, this makes them particularly attractive as a resource for



leadership structures in the development of new molecular entities. Databases containing chemical structures of compounds used in alternative medicine along with their phenotypic effects, in Silico analysis can be used to assist in determining the MOA for example, to predict indicators of the ligand, which are related to known phenotypes for traditional medicines. In the study, the case for CPR, therapeutic grade a tonic and restorative medicine" was estimated. Therapeutic actions or phenotypes for this class include anti-inflammatory, antioxidant, neuroprotective, hypoglycemic activity, immunomodulatory, anti-metastatic and anti-hypertensive. Sodium-glucose transport proteins and PTP1B to insulin signaling regulator has been identified as goals that are associated with the hypoglycemic phenotype suggested. Case study for Ayurveda participates anti-cancer drugs. In this case, the target forecasting software is enriched for the purposes directly related to the progression of cancer, such as steroid-5-alpha-reductase and synergistic goals efflux pump P-GP. These target-phenotype links can help to identify new MoA.

In TCM and Ayurveda, chemogenomics can be applied at an early stage of drug discovery to determine the mechanism of action of compounds and use of genomic biomarkers of toxicity and efficacy for use in Phase I and II clinical trials.

### 1.3.2 The Identification of New Drugs

Chemogenomics profiling can be used to identify entirely new therapeutic targets, such as new antibacterial drugs. The study is based on an existing library of ligand to enzyme, called murD, which is used in the synthesis of peptidoglycan the way. Based on the principle of similarity chemogenomics, the researchers made a map of the murD ligand library to other family members of the Mur ligase, to identify new targets for known ligands. Identified ligands would be expected to be broad-spectrum gram-negative inhibitors in experimental tests because the synthesis of peptidoglycan is exclusive to bacteria. Structural and molecular studies have shown the docking of the ligands of candidate for Goro and Ligas Muir.



### 1.3.3 Identifying Genes in Biological Reactions

Thirty years after post translationally modified derivative of histidine diphthamide, it was determined chemogenomics have been used to detect the enzyme responsible for the last stage of its synthesis. Diphthamide is post translationally modified histidine residue found in elongation factor 2 EEF-2. The first two steps of the biosynthesis leading to diphthine was known, but the enzyme responsible for amidation diphthine in diphthamide remains a mystery. The researchers used *Saccharomyces* cofitness data. Data Cofitness data representing the similarity of growth fitness under different conditions between any two different strains of the deletion. Under the assumption that strains lacking diphthamide synthetase gene should have a high cofitness strain with no other diphthamide biosynthesis genes, they identified *ylr143w* as a strain with high cofitness for all other strains lacking known diphthamide biosynthesis genes. Subsequent experimental assays confirmed that YLR143W was required for diphthamide synthesis and was the missing diphthamide synthase.

## 1.4 CHALLENGES AND LIMITATIONS

Chemogenomics is one example of the many innovative platforms that pharmaceutical and biotechnology research have generated to accelerate the pace of drug discovery. Such combined efforts have been motivated by the assumption that the judicious application of genomics can help to improve efficiencies throughout the drug-discovery process — for example, by identifying candidate failures early in the process — before moving into expensive later-phase trials.

However, the decreased output of commercialized drugs during the past few years indicates that genomics might currently be doing more to hinder drug-discovery programmes than to stimulate them. The sheer volume of data that are generated by chemogenomic analyses creates a gap between drug discovery and drug development. Considering that an important driving

concept of chemogenomics is to reduce the time taken for drug development, it is under tremendous pressure to deliver some valid results quickly. Therefore, current chemogenomics programmes are anxiously seeking to accelerate their throughput at every stage of the drug-development process.

### 1.4.1 Hit Selection and Validation

HTS often generates an enormous number of hits. In addition, HTS data are noisy and error prone and include a substantial portion of false-positive and false-negative hits. This presents two main challenges: how to mine the large data sets to identify those hits with the greatest potential as leads and how to weed out false-positive hits. It is generally agreed that the bottleneck that is created by genomics-based drug discovery occurs not because the process fails to identify hits but because of the slow process of optimizing and validating them — that is, it is improving their ADMET PROPERTIES and determining whether they can convincingly reverse or ameliorate a disease state<sup>80</sup>. Negligent hit validation often occurs at the expense of taking compounds through development, only to later fall short of success. Validation is not always straight forward, however, as sufficient biological information about many targets is often lacking and, in turn, for many new targets with true disease impact, leads cannot be identified or optimized because of the increasing sizes and concurrent stagnating diversities and complexities of chemical libraries.

The design, adaptation and refinement of sophisticated ‘front-end’ computational approaches that can assist in making an informed, quick decision as to which hits should be pursued will be crucial for the success of chemogenomics. Statistical and chemoinformatics approaches for assessing the quality control of HTS data and for mining their chemical and biological information have been developed, for example, by incorporating pattern-detection methods for the identification of pipetting artefacts or for the detection of chemical-class-related effects. The development of chemoinformatics methods and procedures,

such as recursive partitioning, phylogenetic-like tree algorithms or binary quantitative structure–activity relationships (QSARS), which support the automatic identification of hits that are frequently identified by HTS, false positives and negatives, as well as structure–activity relationship (SAR) information, is essential for generating knowledge from HTS data. The myriad efforts that surround the design of appropriate analytic tools have to cope with the difficulty of integrating disparate types of information, especially PARSING and assimilating both chemical and genomics information data (tools such as Scitegic Pipeline pilot or Kensington Inforsense provide the required integration concepts).

### 1.4.2 Data Integration

Chemogenomics-orientated drug discovery programmes face many data-management challenges, which have partly been met by the recent development of innovative ‘biochemoinformatics’ (or pharmacoinformatics) platforms. These integration platforms aim to collect, store and disseminate diverse data sets — such as combinatorial library diversity data, small-molecule chemical structure and biomolecule protein structure data, annotation information, HTS bioassay data and imaging data — in as efficient and productive a manner as possible and to maximize the potential of these data sets.

In particular, integrating information about ligands and targets is complex. This process is related to the current bioinformatics project to create ontologies for proteomics, the aim of which is to generate a systematic definition of the structure and function of all proteins in a genome. The key difficulty lies in integrating two ontologies—one for protein structure and the other for protein function — that have been developed separately and remain largely isolated. The description of active sites and binding sites in protein structures is recognized here as one potential connection point that describes the protein function. Classifications that are based on molecular interactions, in which each protein is associated with a vector that consists of the probability of binding to various

ligands — the central chemogenomics idea—might therefore become prominent in the future. The emphasis on protein-structure similarity is recognized here as a guiding principle. The establishment of standardized molecular informatics platforms and real drug-discovery ontologies at the genome level — that integrate the relevant chemical and biological knowledge—is therefore pursued by the academic and industrial drug-discovery organizations and by companies that are involved in informatics-based discovery.

Knowledge-based chemogenomics companies are currently developing comprehensive molecular information systems for several target classes, including G-protein-coupled receptors, kinases, ion channels and proteases. Their main contribution has been to integrate, in a comprehensive manner, data from patents and selected literature, including two-dimensional structures of the ligands, target sequence and classification, mechanisms of action, structure–activity data, assay results and bibliographic information, together with chemical and biological search engines. Further academic and commercial programmes are gathering information on other types of target, such as those involved in adverse reactions, in ADMET mechanisms and those that define metabolic and signalling pathways.

## REFERENCES

1. Access and patient autonomy are also open to discussion. Genomics Chemogenomics Clinomics Genetic engineering Toxicogenomics Metabolomics Pharmacovigilance
2. Advantages for physicians and the military. Pharmacology Metabolite Chemogenomics Neurotransmitter Peptidomimetic Macielag MJ 2012 Chemical properties
3. Analysis of protein networks and systems biology. Chemical genetics Chemogenomics Cox Ju, Mann M 2007 Is Proteomics the New Genomics? Cell. 130
4. Austin CP, Brady LS, Insel TR, Collins FS (2004) NIH molecular libraries initiative. Science 306:1138–1139
5. Award from The Wellcome Trust, resulting in the creation of the ChEMBL chemogenomics group at EMBL - EBI, led by John Overington. The ChEMBL database contains
6. Baxter JD, Goede P, Apriletti JW, West BL, Feng W, Mellstrom K, Fletterick RJ, Wagner RL, Kushner PJ, Ribeiro RC, Webb P, Scanlan TS, Nilsson S (2002) Structure-based design and synthesis of a thyroid hormone receptor (TR) antagonist. Endocrinology 143:517–524
7. Bierbach Guri Giaever Corey Nislow August 28, 2012 Comparative Chemogenomics To Examine the Mechanism of Action of DNA - Targeted Platinum - Acridine
8. Bioinformatics Chemical file format Chemicalize.org Cheminformatics toolkits Chemogenomics Computational chemistry Information engineering Journal of Chemical
9. biomedical topics RIKEN integrated database of mammals TDR Targets: a chemogenomics database focused on drug discovery in tropical diseases TRANSFAC: a
10. Bishop AC, Ubersax JA, Petsch DT, Matheos DP, Gray NS, Blethrow J, Shimizu E, Tsien JZ, Schultz PG, Rose MD, Wood JL, Morgan DO, Shokat KM (2000) A chemical switch for inhibitor-sensitive alleles of any protein kinase. Nature 407:395–401

11. Bleicher KH (2002) Chemogenomics: bridging a drug discovery gap. *Curr Med Chem* 9:2077–2084
12. Capdeville R, Buchdunger E, Zimmermann J, Matter A (2002) Glivec (ST571, Imatinib), a rationally developed, targeted anticancer drug. *Nature Rev Drug Discov* 1:493–502
13. Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko MA (2001) Chemogenomic approaches to drug discovery. *Curr Opin Chem Biol* 5:464–470
14. Ding S, Wu TY, Brinker A, Peters EC, Hur W, Gray NS, Schultz PG (2003) Synthetic small molecules that control stem cell fate. *Proc Natl Acad Sci U S A* 100:7632–7637
15. Eastwood EL, Wojtovich AP, Elliot SJ, Schaus SE, Collins JJ (March 2005) Chemogenomic profiling on a genome - wide scale using reverse - engineered gene networks
16. Elliott, Sean J. Schaus, Scott E. Collins, James J. March (2005) Chemogenomic profiling on a genome - wide scale using reverse - engineered gene networks
17. forward transfection library screening can entail reverse transfection or chemogenomics Pharmacy compounding is yet another application of personalised medicine
18. Gagna, Claude E. Clark Lambert, W. (1 March 2007) Cell biology, chemogenomics and chemoproteomics - application to drug discovery Expert Opinion
19. Hillisch A, Peters O, Kosemund D, Müller G, Walter A, Elger W, Fritzemeier KH (2004a) Protein structure-based design, synthesis strategy and in vitro pharmacological characterization of estrogen receptor  $\alpha$  and  $\beta$  selective compounds. *Ernst Schering Res Found Workshop* 46:47–62
20. Hillisch A, Peters O, Kosemund D, Müller G, Walter A, Schneider B, Reddersen G, Elger W, Fritzemeier KH (2004b) Dissecting physiological roles of estrogen receptor  $\alpha$  and  $\beta$  with potent selective ligands from structure-based design. *Mol Endocrinol* 18:1599–1609
21. Hillisch A, Pineda LF, Hilgenfeld R (2004c) Utility of homology models in the drug discovery process. *Drug Discov Today* 9:659–669

22. include drug repositioning, polypharmacy, high - throughput screening and chemogenomics While these research approaches have proved effective in helping scientists
23. ISSN 0022 - 3565. PMID 15075359. Hugo Kubinyi Gerhard Muller 6 March 2006 Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective. John Wiley Sons
24. Jacoby E, Schuffenhauer A, Floersheim P (2003) Chemogenomics knowledgebased strategies in drug discovery. *Drug News Perspect* 16:93–102
25. Kubinyi H (2004) Drug discovery from side effects. In: Kubinyi H, Müller G, (eds) Chemogenomics in drug discovery – a medicinal chemistry perspective, vol. 22 of Methods and principles in medicinal chemistry. Mannhold R, Kubinyi H, Folkers G (eds) Wiley-VCH, Weinheim, pp 43–67
26. Kubinyi H, Müller G (eds) (2004) Chemogenomics in drug discovery – a medicinal chemistry perspective, vol. 22, Methods and principles in medicinal chemistry. Mannhold R, Kubinyi H, Folkers G (eds) Wiley-VCH, Weinheim
27. Ligands. She also researches analytical methods for metabolomics and chemogenomics for the reaction of plants to pesticides and hypoxia using NMR and mass
28. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. *Nature* 432:855–861
29. McGovern SL, Caselli E, Grigorieff N, Shoichet BK (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem* 45:1712–1722
30. McGovern SL, Shoichet BK (2003) Kinase inhibitors: not just for kinases anymore. *J Med Chem* 46:1478–1483
31. Miller WH, Keenan RM, Willette RN, Lark MW (2000) Identification and in vivo efficacy of small-molecule antagonists of integrin  $\alpha v \beta 3$  (the vitronectin receptor). *Drug Discov Today* 5:397–408
32. Müller G (2004) Target family-directed masterkeys in chemogenomics. In: Kubinyi H, Müller G (eds) Chemogenomics in drug discovery – a medicinal chemistry



- perspective, vol. 22 of Methods and principles in medicinal chemistry. Mannhold R, Kubinyi H, Folkers G (eds) Wiley-VCH, Weinheim, pp 7–41
33. Other biological areas include chemical biology, chemical ecology, chemogenomics systems biology, molecular modeling, chemometrics, and chemoinformatics
  34. PMID 21935381. Botet J, Mateos L, Revuelta JL, Santos MA 2007 A chemogenomic screening of sulfanilamide - hypersensitive *Saccharomyces cerevisiae* mutants
  35. R. Heman - Ackahc, S. Martin, S. Youle, R.J. Inglese, J. 2015 Chemogenomic profiling of endogenous PARK2 expression using a genome - edited coincidence
  36. Russell K, Michne WF (2004) The value of chemical genetics in drug discovery. In: Kubinyi H, Müller G (eds) Chemogenetics in drug discovery. WileyVCH, Weinheim, pp 69–96
  37. Seidler J, McGovern SL, Doman TN, Shoichet BK (2003) Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem* 46:4477–4486
  38. Shanmugam D, Van Voorhis WC, Agüero F January 2012 TDR Targets: a chemogenomics resource for neglected diseases *Nucleic Acids Research*. 40 Database
  39. Target Family - directed Masterkeys in Chemogenomics In Kubinyi H, Muller G, Mannhold R, Folkers G eds. *Chemogenomics in Drug Discovery: A Medicinal Chemistry*
  40. Targeting a particular gene of interest. Chemical biology Chemogenomics Kubinyi H 2006 *Chemogenomics in drug discovery* In Weinmann H, Jaroch S eds.
  41. Targets 6: Driving drug discovery for human pathogens through intensive chemogenomic data integration *Nucleic Acids Research*. doi: 10.1093 nar gkz999. TDR
  42. universally accepted, and others have offered alternative terms such as chemogenomics to describe essentially the same field of study. The nature and complexity



43. was the Chief Operating Officer at Iconix Pharmaceuticals, a leading chemogenomics company, where he led the research, development and informatics operations
44. Wermuth CG (2001) The “SOSA” approach: an alternative to high-throughput screening. *Med Chem Res* 10:431–439
45. Wermuth CG (2004) Selective optimization of side activities: another way for drug discovery. *J Med Chem* 47:1303–1314
46. Ye HF, O’Reilly KE, Koh JT (2001) A subtype-selective thyromimetic designed to bind a mutant thyroid hormone receptor implicated in resistance to thyroid hormone. *J Am Chem Soc* 123:1521–1522



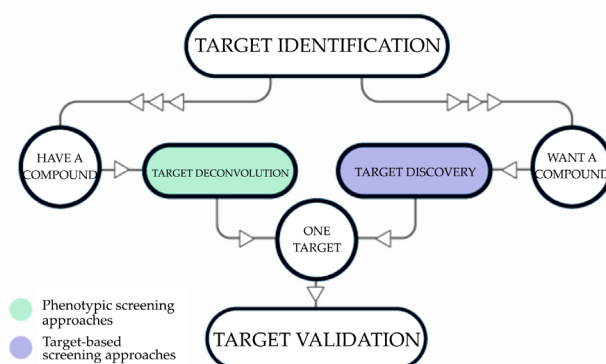


## CHAPTER 2

# CHEMOGENOMICS ANALYSIS OF DRUG TARGETS

### INTRODUCTION

Target-identification and mechanism-of-action studies have important roles in small-molecule probe and drug discovery. Biological and technological advances have resulted in the increasing use of cell-based assays to discover new biologically active small molecules. Such studies allow small-molecule action to be tested in a more disease-relevant setting at the outset, but they require follow-up studies to determine the precise protein target or targets responsible for the observed phenotype. Target identification can be approached by direct biochemical methods, genetic interactions or computational inference. In many cases, however, combinations of approaches may be required to fully characterize on-target and off-target effects and to understand mechanisms of small-molecule action.



## 2.1 COMPARATIVE CHEMOGENIC ANALYSIS FOR PREDICTING DRUG-TARGET

A computational technique for predicting the DTIs has now turned out to be an indispensable job during the process of drug finding. It tapers the exploration room for interactions by propounding possible interaction contenders for authentication through experiments of wet-lab which are known for their expensiveness and time consumption. Chemogenomics, an emerging research area focused on the systematic examination of the biological impact of a broad series of minute molecular-weighting ligands on a broad raiment of macromolecular target spots.

Additionally, with the advancement in time, the complexity of the algorithms is increasing which may result in the entry of big data technologies like Spark in this field soon.

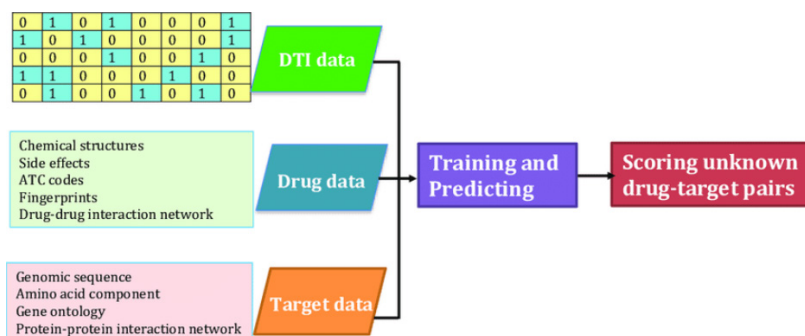
In the presented work, we intend to offer an inclusive idea and realistic evaluation of the computational Drug Target Interaction projection approaches, to perform as a guide and reference for researchers who are carrying out work in a similar direction. Precisely, we first explain the data utilized in computational Drug Target Interaction prediction attempts like this. We then sort and explain the best and most modern techniques for the prediction of DTIs. Then, a realistic assessment is executed to show the projection performance of several illustrative approaches in

various situations. Ultimately, we underline possible opportunities for additional improvement of Drug Target Interaction projection enactment and also linked study objectives.



The accurate prediction of interactions formed between a drug and its targeted protein via computational approaches is highly demanding because it is an efficient analog to the wet-lab experiments that cost heavily and requires additional efforts. Drug–target interactions (DTIs) which are newly discovered are critical for discovering novel targets that can interact with the existing drugs, as well as new drugs that can target some specific genes causing diseases. Drug repositioning is one of the efficient methods for the recovery of existing drugs for a novel cause, i.e. drugs which are developed for some particular purposes can be used to treat other biological conditions, meaning a single drug can be applied to many targets. There is already massive research going on the existing drugs based on the bioavailability and their safe use. Repositioning can limit drug costs and may enhance the process of drug discovery, making drug repositioning an eminent method for drug discovery. Some major techniques employed for the drug repurposing involve network-based approach, network-based cluster approach, network-based propagation approach, text mining-based approach, and semantics-based approach. Drug repositioning is different from the traditional drug development that involves five stages, however, this method requires only 4 stages which include compound recognition, obtaining a compound, production and FDA based safety monitoring. The

Gleevec (imatinib mesylate) is a well-known example of drug repositioning which was initially thought to interact only with the Bcr-Abl fusion gene related to leukemia. But later on, it was found that interaction of the Gleevec with PDGF and KIT can also be achieved, with an added advantage as a repositioned drug for the treatment of gastrointestinal stromal tumours. The success of Gleevec as a repositioned drug is one of the admired stories reported in the literature. As drug repositioning is already revealed by the example of Gleevec, it opens new doors for scientists to reposition other drugs as well. A drug's feasibility (i.e. interaction of a single drug with multiple targets) may enrich its polypharmacology (i.e. having multiple beneficial effects), which motivates the scientists to discover more about drug repositioning.

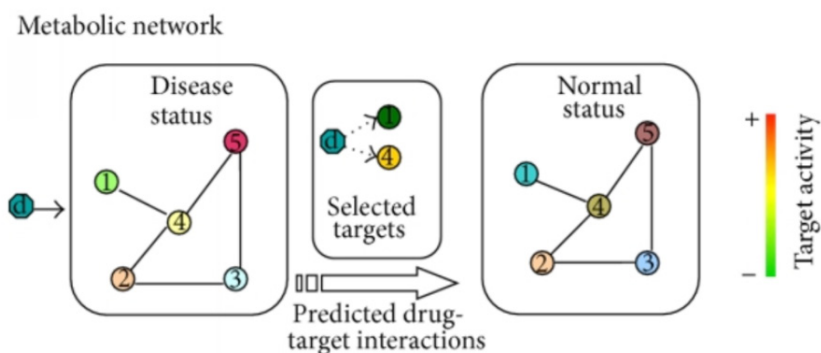


On the other side, there still exist a lot of small molecules that can be used as drugs but because of their interaction profiles, they cannot be used. For example, more than 90 million compounds are stored in the PubChem database whose interaction profiles are still unknown. Thus, by knowing the interactions between the disease-causing genes and the target proteins for these compounds may help in the discovery of new drugs as it can help the drug candidates with low potential to work within the drug discovery field. Therefore, for drug repositioning, the discovery of DTIs is very useful, as it aids with the drug candidate selection and predicts the side effects of these drugs in advance. Definitely, the experimental wet-lab techniques are more helpful in predicting such types of interactions but this job is much tiresome and also consumes a lot of time. Thus, from here, the computational methods take over as they are proven to be highly useful and may

prove efficient in predicting potential interacting candidates with satisfactory accuracy, hence reducing the DTIs to be inspected via *in-vitro* correspondent.

### 2.1.1 Identification of Chemogenomic Features from Drug-Target Interaction

Drug phenotypic effects are caused by the interactions between drug molecules and their target proteins including their primary targets and off-targets. Polypharmacology, the idea that drug phenotypic effects are not due only to its primary target, but rather to its whole spectrum of interactions, tends to become a new paradigm in drug design. It is important to identify the molecular mechanisms behind overall drug–target interactions or more generally compound–protein interactions, leading to many applications at different levels of the drug design process. There is a hypothesis that polypharmacology is strongly involved in both drug chemical substructures and protein functional sites, so there is a strong incentive to develop new methods to explore the association between drug chemical substructures and protein functional sites in terms of drug–target interactions.



Docking or ligand–based approach (e.g. QSAR) have been proposed to analyze and predict interactions with respect to a single protein, so these methods cannot be applied to mine ligand–protein pairs across many different proteins. Chemogenomics is an emerging research area that attempts to associate the chemical space of

possible ligands with the genomic space of possible proteins. These methods are purely predictive and do not provide any further understanding of molecular mechanisms behind ligand–protein interactions. Drug–target interactions are due to drug chemical substructures and protein functional sites. Beyond the ligand–protein interaction prediction problem, a variety of methods have been proposed to investigate the correlation between chemical substructures, biological activities and phenotypic effects.

Several methods based on binding pockets comparison have been proposed, but they require the knowledge of protein 3D structures, which is not genome-wide available. However, most previous works have been performed from the viewpoint of either chemical substructures or protein functional sites.

One of the most challenging issues in recent chemogenomic research is to identify the underlying associations between drug chemical substructures and protein functional sites which are involved in drug–target interaction networks.

Recently, a variant of sparse canonical correspondence analysis (SCCA) has been proposed to extract sets of chemical substructures and protein domains governing drug–target interactions, but the variation of detectable protein domains is very limited.

The use of both graph mining and sequence mining has been proposed to extract drug substructures and protein subsequences which tend to appear in known drug–target interactions. However, the size of extracted subsequences is very small (e.g. two or three amino acids), which makes biological interpretation difficult, and any prediction framework for new interactions based on the extracted features was not provided.

We develop a classifier-based approach to identify chemogenomic features (the underlying associations between drug chemical substructures and protein domains) which are strongly involved in drug–target interaction networks. We propose a novel algorithm for extracting informative chemogenomic features by using  $L_1$  regularized classifiers over the tensor product space of possible drug–target pairs. In the results, we show



that the proposed method can extract a very limited number of chemogenomic features without losing the performance of predicting drug–target interactions. We underline that the extracted chemogenomic features are biologically meaningful and discuss how the method can help the drug development process. The extracted substructure–domain association network enables us to suggest ligand chemical fragments specific for each protein domain and ligand core substructures important for a wide range of protein families.

### 2.1.2 Materials

Drug–target interactions involving human proteins were obtained from the DrugBank database. Target proteins belong to many different classes such as enzymes, ion channels, G protein-coupled receptors (GPCRs) or nuclear receptors. The dataset consists of 4809 drug–target interactions involving 1862 drugs and 1554 target proteins.

Chemically identical drugs with the same structures (duplicates) are removed, so structures of all drugs in the above interaction data are unique.

Each drug was represented by an 881 dimensional binary vector whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Among the 881 substructures used to represent the chemical structures, 663 are actually used, because 218 do not appear in our drug set.

Genomic information about target proteins was obtained from the UniProt database, and associated protein domains were obtained from the PFAM database. Target proteins in our dataset were associated with 876 PFAM domains. Each target protein was represented by a 876 dimensional binary vector whose elements encode for the presence or absence of each of the retained PFAM domain by 1 or 0, respectively.

### 2.1.3 Model

Linear model is a useful tool for classification and regression. Generally, a linear model represents each example  $E$  by a feature vector representation  $\Phi(E) \in \mathbb{R}^D$  and then estimates a linear function  $f(E) = \mathbf{w}^T \Phi(E)$  whose sign is used to predict whether the example  $E$  is classified into positive or negative. The weight vector  $\mathbf{w} \in \mathbb{R}^D$  is estimated based on its ability to correctly predict the classes of examples in the training set. In addition to its classification ability, linear models have an interpretability of features. Since each element of a feature vector  $\Phi(E)$  corresponds to an element of its weight vector  $\mathbf{w}$ , we can extract effective features contributing to the prediction by sorting elements of  $\Phi(E)$  according to the values of the corresponding elements of  $\mathbf{w}$ .

The prediction of drug–target interactions or compound–protein interactions is more complicated because the dataset consist of drug–target pairs or compound–protein pairs. Let  $C$  be a drug (or drug candidate compound) and  $P$  be a target (or target candidate protein). To apply the previous machine learning approach to this problem, we need to represent a pair of a compound  $C$  and a protein  $P$  by a feature vector  $\Phi(C, P)$  and then estimate a linear function  $f(C, P) = \mathbf{w}^T \Phi(C, P)$  whose sign can be used to predict whether a pair of  $C$  and  $P$  interacts or not. The weight vector  $\mathbf{w}$  is estimated based on its ability to correctly predict interactions of drug–target pairs or compound–protein pairs.

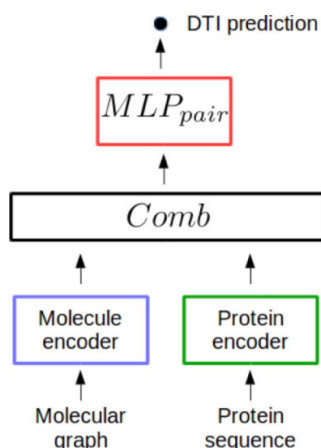
### 2.1.4 Binary classifiers

We apply two popular binary linear classifiers: logistic regression and linear support vector machine (SVM). Models are typically learned to minimize objective functions with a regularization for both classifiers. It is well known that the use of regularization is necessary to achieve a model that generalizes well to unseen data, particularly if the dimension of features is very high relative to

the amount of training data. One common regularization is  $L_2$ -regularization which keeps most elements in the weight vector to be non-zeros. Therefore, one can suffer from interpreting features from learned weights. We shall, respectively, refer to  $L_2$ -regularized logistic regression and linear SVM as L2LOG and L2SVM. Another possible regularization is  $L_1$ -regularization that makes most elements in the weight vector to be zeros. In this study, we introduce logistic regression and linear SVM with  $L_1$ -regularization for its high interpretability.

### 2.1.5 Extraction of Chemogenomic Features

We tested the feature extraction ability of five feature extraction methods: L1LOG, L1SVM, L2LOG, L2SVM and SCCA. Note that L1LOG and L1SVM are the proposed methods with L1-regularization, L2LOG and L2SVM are the proposed methods with L2-regularization and SCCA is the previous method. We extracted chemogenomic features that were positively weighted in each method. The parameters in each method (e.g. regularization parameters, sparsity parameters and number of components) were optimized by performing cross-validation.

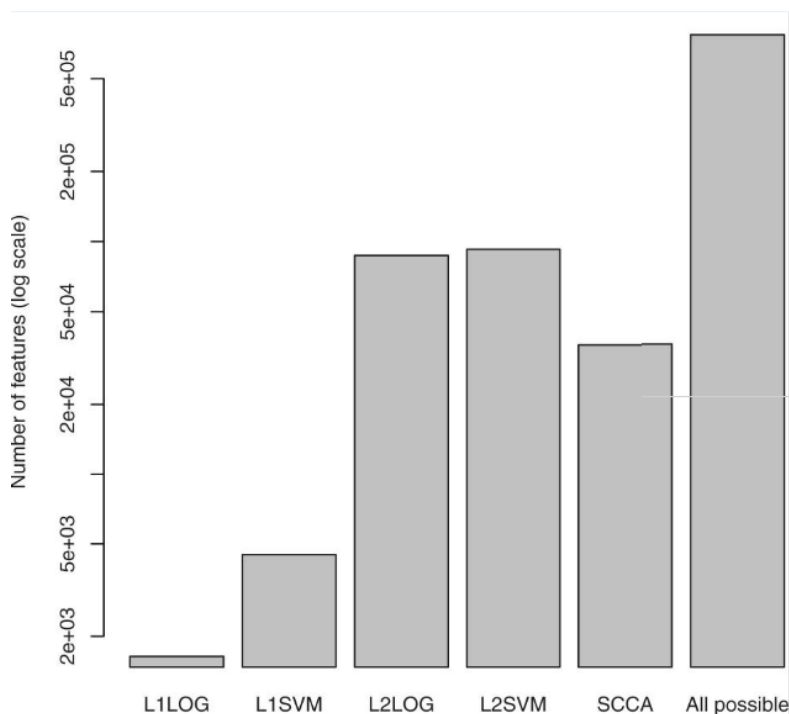


**Table 1:** Examples of extracted chemogenomic features by the L1LOG method

| Rank | Weight | Substructure ID<br>Domain ID | PubChem substructure definition<br>PFAM domain definition |
|------|--------|------------------------------|---|
| 1    | 2.1468 | SUB158                       | $\geq 3$ any ring size 5                                  |
| 1    | 2.1468 | PF00106                      | short chain dehydrogenase                                 |
| 2    | 2.1118 | SUB414                       | $S(-C)(-H)$   |
| 2    | 2.1118 | PF00255                      | Glutathione peroxidase                                    |
| 3    | 1.9413 | SUB158                       | $\geq 3$ any ring size 5                                  |
| 3    | 1.9413 | PF01126                      | Heme oxygenase  |
| 4    | 1.8035 | SUB686                       | $O = C-C-C-C-N$   |
| 4    | 1.8035 | PF01094                      | Receptor family ligand binding region                     |
| 5    | 1.7707 | SUB687                       | $O = C-C-C-C-O$   |
| 5    | 1.7707 | PF03171                      | 2OG-Fe(II), oxygenase superfamily                         |
| 6    | 1.7514 | SUB348                       | $C(-C)(-H)(-O)(-O)$                                       |
| 6    | 1.7514 | PF03414                      | Glycosyltransferase family 6                              |
| 7    | 1.6343 | SUB387                       | $C(:C)(:C)(:N)$   |
| 7    | 1.6343 | PF00042                      | Globin  |
| 8    | 1.6299 | SUB409                       | $O(-H)(-S)$   |
| 8    | 1.6299 | PF00167                      | Fibroblast growth factor                                  |
| 9    | 1.5807 | SUB32                        | $\geq 2 P$  |
| 9    | 1.5807 | PF00348                      | Polyprenyl synthetase                                     |
| 10   | 1.5797 | SUB567                       | $O-C-C-N$   |
| 10   | 1.5797 | PF00464                      | Serine hydroxymethyltransferase                           |
| 11   | 1.5105 | SUB309                       | $O-H$   |
| 11   | 1.5105 | PF00102                      | Protein-tyrosine phosphatase                              |
| 12   | 1.5065 | SUB433                       | $C(-C)(-C)(=O)$   |
| 12   | 1.5065 | PF02518                      | Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase   |
| 13   | 1.5033 | SUB449                       | $C(-H)(=O)$   |

|    |        |         |                                    |
|----|--------|---------|------------------------------------|
| 13 | 1.5033 | PF00107 | Zinc-binding dehydrogenase         |
| 14 | 1.4956 | SUB695  | $O = C-C-C-C-C = O$                |
| 14 | 1.4956 | PF00551 | Formyl transferase                 |
| 15 | 1.4784 | SUB433  | $C(-C)(-C)(=O)$                    |
| 15 | 1.4784 | PF07884 | Vitamin K epoxide reductase family |

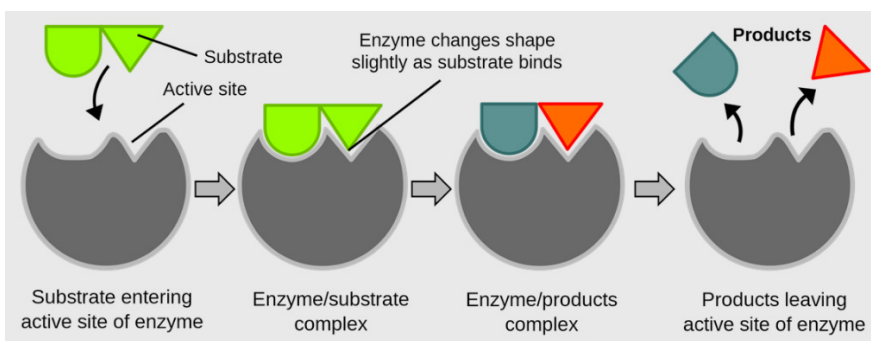
Figure 1 shows a comparison of the number of extracted features between the five different feature extraction methods. In the case of SCCA, we evaluated the association between chemical substructures and protein domains by computing the product of their weight elements between chemical substructures and protein domains within each canonical component, and we took unique combinations as chemogenomic features if they were present in different canonical components.



**Figure 1:** Comparison of the number of extracted features between different methods.

## 2.2 OPEN-SOURCE CHEMOGENOMIC DATA-DRIVEN ALGORITHMS FOR PREDICTING DRUG-TARGET INTERACTIONS

Drug development is a complex and expensive process. Over the past decades, despite technological advances in drug discovery and increase of investments in pharmaceutical research and development, the number of new drug approvals has remained stagnant. The most significant causes of drug failures are toxicity and a lack of efficacy. Thus, there is an urgent need to develop effective drugs to overcome these limitations. Drug repositioning, the process of finding new uses outside the scope of the original medical indications for existing drugs, is considered to be a promising strategy with the benefit of providing a rapid route to clinic than through the traditional drug discovery approaches because of the use of existing knowledge about drugs. The new indication-driven discovery by using repositioning methods has already yielded several successes. For example, HIV protease inhibitors such as nelfinavir can be used as a new class of anticancer drugs. Sunitinib, originally developed for treating renal cell carcinoma, was found to be effective for patients with pancreatic neuroendocrine tumors. Imatinib, developed originally for chronic myeloid leukemia, has shown clinical benefits to the treatment of gastrointestinal stromal tumor.



One of the necessary steps of drug repositioning is to accurately identify the drug–target interactions (DTIs). However, experimental determination of such associations is time-

consuming and costly. Thus, computational methods have been proposed alternatively to infer potential DTIs in effective ways. Traditionally, computational methods for DTI predictions include molecular docking simulation, quantitative structure–activity relationship (QSAR) and so forth. However, these methods possess inherent limitations. For example, docking simulation requires 3D crystal structure of the drug target, which is difficult to obtain for membrane proteins. Traditional QSAR often handles compound analogs targeting a single molecular target, which is less efficient for processing chemogenomic data with a large library of compounds and many targets.

Unlike QSAR (chemical data-based) and molecular docking (genomic data-based) approaches, chemogenomic data-driven DTI prediction methods simultaneously consider both chemical information and genomic information (often from large-scale screenings of small molecule libraries against a panel of drug target, which may or may not be biologically related). For example, Yamanishi *et al.* proposed a bipartite graph learning method to infer the relationship between chemical/genomic space and pharmacological space. Kim and coworkers explored the effect of drug–drug interactions (DDIs) on DTI predictions. They used two machine learning algorithms, including support vector machine (SVM) and kernel-based L1-norm regularized logistic regression (KL1LR) to build prediction models. As a result, they concluded that DDI from pharmacological information is a promising feature in predicting DTIs when compared with other data sources such as chemical structures of drugs, and KL1LR is useful for investigating the contributing features. In the work by Wang *et al.*, a two-layer graphical model, called restricted Boltzmann machine, was proposed to predict not only the direct and indirect drug–target relationships but also the drug modes of action, including binding, activation and inhibition, which extended the conventional binary DTI predictions. Most recently, Meng *et al.* proposed a novel feature-based approach, called predicting drug targets with protein sequence (PDTPS), to infer potential DTIs. In PDTPS, for each protein sequence, position-specific score matrix (PSSM) was first constructed, and the bigram probability feature

extraction method was used to represent a given protein sequence based on the calculated PSSM. After this, principal component analysis (PCA) was adopted to reduce the protein sequence feature vector. For each drug compound, the structural features were calculated. As a result, the feature representation of each drug–target pair was obtained by concatenating both protein vector and drug vector. Finally, relevance vector machine was used to predict potential DTIs. Another feature-based approach proposed by Li *et al.* adopted local binary pattern operator to compute the histogram descriptors for protein sequences. For drug molecules, they calculated the fingerprints, and used PCA to extract the low-dimensional features for both proteins and drugs. Finally, they used the discriminative vector machine classifier to identify DTIs.

Among these algorithms, many of them are made publicly available. Researchers often compared different algorithms based on the benchmark data set, and they adopted two commonly used metrics [i.e. area under the curve (AUC) and area under precision–recall curve (AUPR)] as the evaluation criteria. However, the comparison may be suboptimal and less objective because of differences in program parameter setting and details of cross-validation methods. In this work, we first review chemogenomic data-driven and open-source algorithms published in recent years, and then we compare five representative algorithms based on a new recall-based evaluation metric in the same framework. We hope the reviewed algorithms can be continuously improved to make stronger prediction, and can be optimized to ease reuse and ensure result replication.

**Table 2:** Benchmark data set for DTI prediction algorithms

| Data set | Number of drugs | Number of targets | Number of interactions | Sparsity value |
|----------|-----------------|-------------------|------------------------|----------------|
| Enzyme   | 445             | 664               | 2926                   | 0.010          |
| IC       | 210             | 204               | 1476                   | 0.034          |
| GPCR     | 223             | 95                | 635                    | 0.030          |
| NR       | 54              | 26                | 90                     | 0.064          |



## 2.2.1 Data set transformation

It should be emphasized that the benchmark data set may be transformed slightly based on the individual prediction algorithms. For algorithms such as bipartite local models (BLMs), the zero components in matrix  $Y$  may be transformed to  $-1$ , while for algorithms such as dual-network integrated logistic matrix factorization (DNILMF), the original zero elements remain unchanged. Besides the interaction matrix  $Y$ , the drug similarity matrix  $S_c$  and target similarity matrix  $S_g$  may be transformed to corresponding kernel matrices,  $K_c$  and  $K_g$ , respectively. For example, BLM requires a kernel matrix as input to build a model. A general transformation procedure can be performed in the following way: taking the conversion from  $S_c$  to  $K_c$  as an example,  $S_c$  was first converted to a symmetrical matrix by adding its transposed matrix and then divided by 2. The obtained symmetrical matrix was finally converted to a positive semi-definite matrix by adding an identity matrix with a small value (0.1 in the work) in the main diagonal line for multiple times. A similar procedure was applied to  $S_g$  for generating  $K_g$ .

## 2.2.2 Cross-validation and Evaluation Metric

Stringent cross-validation is important for model evaluation. Different from previous methods, which put both positive and negative interaction pairs (considering unknown interactions as negative ones) into the test set. We, in this work, only include the positive interaction pairs in the test set in the process of cross-validation. Specifically, in each split, we removed a random subset of 10% of the known entries in the drug–target adjacency matrix  $Y$  as the test set and trained on the remaining 90% of the known DTI. In addition, we ensured each drug has at least one interaction with a target (and vice versa, each target has at least one interaction with a drug as well) in the training matrix as reported by a previous DTI prediction work. Then, we used a ranking-based statistical metric to evaluate different DTI prediction algorithms.

### 2.2.3 Open-source chemogenomic data-driven DTI prediction algorithms

We review several chemogenomic data-driven and open-source DTI prediction algorithms focusing on model properties and model evolutionary relationships. Table 3 lists the reviewed algorithms with the corresponding Web links.

**Table 3:** Open-source chemogenomic data-driven DTI prediction algorithms based on the benchmark data set

| No. | Algorithm    | Open-access link   | Year |
|-----|--------------|--|------|
| 1   | BLM          | <a href="http://cbio.mines-paristech.fr/~yyamanishi/bipartitelocal/">http://cbio.mines-paristech.fr/~yyamanishi/bipartitelocal/</a>  | 2009 |
| 2   | KronRLS      | <a href="http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/">http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/</a>  | 2011 |
| 3   | KBMF2K       | <a href="http://users.ics.aalto.fi/gonen/bioinfo12.php">http://users.ics.aalto.fi/gonen/bioinfo12.php</a>  | 2012 |
| 4   | DTHybrid     | <a href="http://alpha.dmi.unict.it/dtweb/dthybrid.php">http://alpha.dmi.unict.it/dtweb/dthybrid.php</a>  | 2013 |
| 5   | KronRLS-WNN  | <a href="http://cs.ru.nl/~tvanlaarhoven/drugtarget2013/">http://cs.ru.nl/~tvanlaarhoven/drugtarget2013/</a>  | 2013 |
| 6   | SC-MLKNN     | <a href="http://web.hku.hk/~liym1018/projects/drug/drug.html">http://web.hku.hk/~liym1018/projects/drug/drug.html</a> or <a href="http://www.bmln-wpu.org/us/tools/PredictingDTI_S2/METHODS.html">http://www.bmln-wpu.org/us/tools/PredictingDTI_S2/METHODS.html</a> | 2015 |
| 7   | RLSKF        | <a href="https://github.com/minghao2016/RLS-KF">https://github.com/minghao2016/RLS-KF</a>  | 2016 |
| 8   | KronRLSMKL   | <a href="http://www.cin.ufpe.br/~acan/kronrlsmkl/">http://www.cin.ufpe.br/~acan/kronrlsmkl/</a>  | 2016 |
| 9   | KronRLS-WNNS | <a href="https://github.com/hkmztrk/SMILES-basedSimilarityKernels">https://github.com/hkmztrk/SMILES-basedSimilarityKernels</a>  | 2016 |
| 10  | NRLMF        | <a href="https://github.com/stephenliu0423/Py-DTI">https://github.com/stephenliu0423/Py-DTI</a>  | 2016 |
| 11  | COSINE       | <a href="http://bioinfo.cs.uni.edu/COSINE.html">http://bioinfo.cs.uni.edu/COSINE.html</a>  | 2016 |
| 12  | DNILMF       | <a href="https://github.com/minghao2016/DNILMF">https://github.com/minghao2016/DNILMF</a>  |      |

### 2.2.4 Algorithm Comparison Procedure

We performed the following evaluation procedures for a more rigorous comparison of the reviewed algorithms based on the recall-based statistical metric. Step 1:

The adjacency matrix,  $Y$ , was first split for 10-fold cross-validation in the abovementioned approach. Briefly, only positive pairs were used to perform the subset splitting to compare the recall-based evaluation metric.

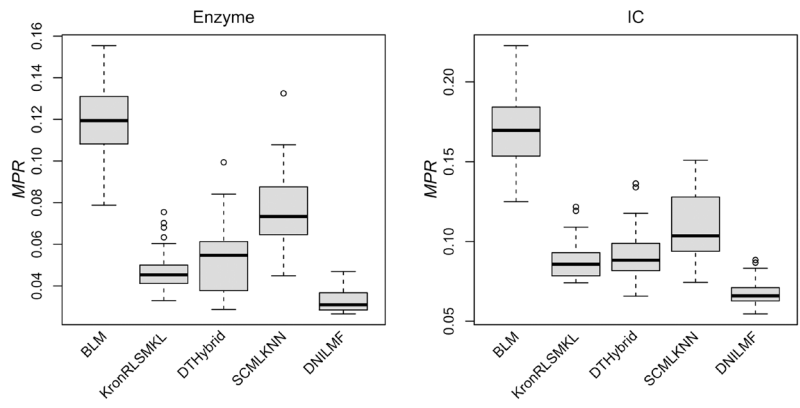
In each fold, at least one link (known interaction) was kept in each row and each column of  $Y$ , respectively. Five trials of 10-fold cross-validation processes were performed to yield 50 fold matrices with test set data points included in each matrix. Each fold matrix was used to build the model and predict the data points in the test set. Step 2:

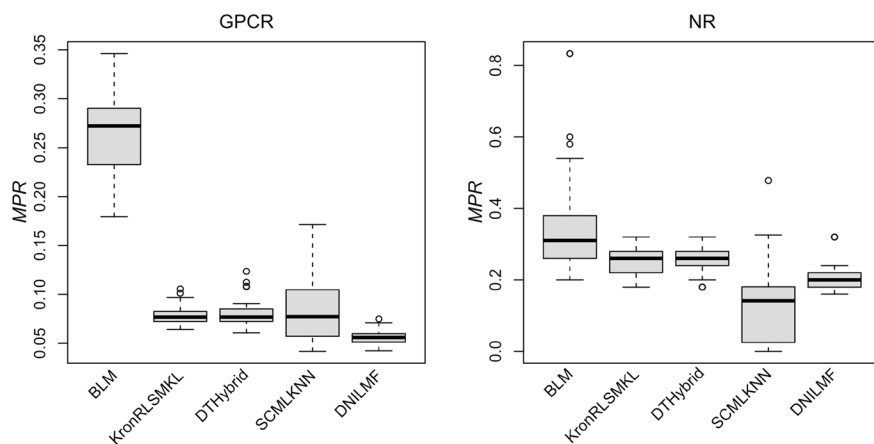
Optionally, the similarity matrices were converted to the corresponding kernel matrices as shown in the data set transformation section for the kernel-based algorithms such as BLM. Step 3: Multiple models were built based on similarity matrix either from targets or from drugs (or based on matrices from both targets and drugs), as well as each fold matrix (generated from Step 1). Step 4: The recall-based evaluation metric, MPR, was calculated from the test set data points in each fold matrix.

**Table 4:** Comparison of open-source algorithms based on MPR, AUC and AUPR for the benchmark data set

| Data   | Method     | MPR<br>(mean $\pm$ SE) | AUC<br>(mean $\pm$ SE) | AUPR<br>(mean $\pm$ SE) |
|--------|------------|------------------------|------------------------|-------------------------|
| Enzyme | BLM        | 0.119 $\pm$ 0.002      | 0.923 $\pm$ 0.003      | 0.750 $\pm$ 0.003       |
|        | KronRLSMKL | 0.047 $\pm$ 0.001      | 0.993 $\pm$ 0.000      | 0.963 $\pm$ 0.001       |
|        | DTHybrid   | 0.053 $\pm$ 0.002      | 0.986 $\pm$ 0.001      | 0.939 $\pm$ 0.001       |
|        | SCMLKNN    | 0.076 $\pm$ 0.002      | 0.986 $\pm$ 0.000      | 0.839 $\pm$ 0.002       |
|        | DNILMF     | 0.033 $\pm$ 0.001      | 0.996 $\pm$ 0.000      | 0.951 $\pm$ 0.001       |

|      |            |                   |                   |                   |
|------|------------|-------------------|-------------------|-------------------|
| IC   | BLM        | $0.169 \pm 0.003$ | $0.899 \pm 0.002$ | $0.684 \pm 0.009$ |
|      | KronRLSMKL | $0.088 \pm 0.002$ | $0.990 \pm 0.001$ | $0.953 \pm 0.003$ |
|      | DTHybrid   | $0.090 \pm 0.002$ | $0.989 \pm 0.002$ | $0.918 \pm 0.002$ |
|      | SCMLKNN    | $0.109 \pm 0.003$ | $0.975 \pm 0.000$ | $0.823 \pm 0.004$ |
|      | DNILMF     | $0.068 \pm 0.001$ | $0.996 \pm 0.000$ | $0.947 \pm 0.002$ |
| GPCR | BLM        | $0.266 \pm 0.006$ | $0.752 \pm 0.010$ | $0.326 \pm 0.009$ |
|      | KronRLSMKL | $0.079 \pm 0.001$ | $0.987 \pm 0.003$ | $0.833 \pm 0.005$ |
|      | DTHybrid   | $0.080 \pm 0.002$ | $0.969 \pm 0.002$ | $0.768 \pm 0.014$ |
|      | SCMLKNN    | $0.086 \pm 0.005$ | $0.968 \pm 0.003$ | $0.650 \pm 0.011$ |
|      | DNILMF     | $0.056 \pm 0.001$ | $0.987 \pm 0.001$ | $0.826 \pm 0.008$ |
| NR   | BLM        | $0.349 \pm 0.020$ | $0.777 \pm 0.050$ | $0.211 \pm 0.091$ |
|      | KronRLSMKL | $0.254 \pm 0.006$ | $0.979 \pm 0.001$ | $0.613 \pm 0.060$ |
|      | DTHybrid   | $0.257 \pm 0.005$ | $0.917 \pm 0.003$ | $0.566 \pm 0.087$ |
|      | SCMLKNN    | $0.135 \pm 0.016$ | $0.951 \pm 0.002$ | $0.342 \pm 0.009$ |
|      | DNILMF     | $0.205 \pm 0.005$ | $0.952 \pm 0.011$ | $0.605 \pm 0.063$ |
| kd   | BLM        | $0.320 \pm 0.003$ | $0.755 \pm 0.009$ | $0.233 \pm 0.015$ |
|      | KronRLSMKL | $0.166 \pm 0.003$ | $0.817 \pm 0.004$ | $0.200 \pm 0.002$ |
|      | DTHybrid   | $0.126 \pm 0.002$ | $0.957 \pm 0.001$ | $0.686 \pm 0.004$ |
|      | SCMLKNN    | $0.181 \pm 0.005$ | $0.908 \pm 0.002$ | $0.526 \pm 0.008$ |
|      | DNILMF     | $0.122 \pm 0.002$ | $0.966 \pm 0.001$ | $0.721 \pm 0.004$ |





**Figure 2:** MPR of five representative DTI prediction algorithms based on the benchmark data set. For Enzyme, IC and NR, all results differ significantly except KronRLSMKL VS. DTHybrid ( $P < 0.01$ ,  $t$ -test).

For three relatively larger data sets (i.e. Enzyme, IC and GPCR), all of the five representative algorithms keep the same trend of prediction performance, where DNILMF outperforms the other four algorithms consistently, and BLM shows large MPR values compared with others.

Both KronRLSMKL and DTHybrid show comparable results, which are both (slightly) better than the ones from SCMLKNN. Interestingly, for NR, which is the smallest data set, SCMLKNN exhibits the best MPR value. However, for all four data sets, these algorithms already gave a large improvement over a purely random model with MPR expected as of 50% (especially for the larger data sets).

We emphasize that the current results of MPR were calculated based on the original parameter settings from the reported algorithms with slight differences, and the parameters were fixed during the cross-validations. Therefore, it is anticipated that the performance might be improved by fully exploiting the parameter space. As reported in the previous work, the nested cross-validation can be used to perform parameter tuning from the inner loop, and the outer loop is used to evaluate the model. In this work, we took

the DTHybrid algorithm as an example to perform nested cross-validation. As a result, the model with optimal parameters (i.e. lambda and alpha used by DTHybrid) derived from the nested cross-validation exhibits similar or slightly better results than the one using fixed parameters.

Most models would show improved performance with the increase of samples. Interestingly, though IC includes more data points compared with GPCR, algorithms from KronRLSMKL, DTHybrid, SCMLKNN and DNILMF consistently give relative better results for GPCR. One of the possible reasons for this result may be that the ratio of the number of targets to the number of drugs in  $Y$  for GPCR is much less than the ratio in the IC group. To further investigate the influence of data size, we subsampled three larger data sets (i.e. Enzyme, IC and GPCR in descending order of size) to the sizes approximate to that of the smaller data sets and calculated the MPR values. For computational efficiency, we took DTHybrid as the tested algorithm. To subsample a larger data set, for example, for the Enzyme data set (i.e. 664 targets and 445 drugs), three subsamples were generated with the approximated size of IC (i.e. 204 targets and 221 drugs), GPCR (i.e. 95 targets and 125 drugs) and NR (i.e. 30 targets and 55 drugs). Similarly, two subsamples for the IC data set and one subsample for the GPCR data sets were generated.

It is evident that DNILMF shows better performance for all the subsets except NR. However, we notice that the enhanced performance of DNILMF is not only derived from the proposed algorithm itself but also from the KF method, which is an important but understudied approach in the DTI prediction field. In fact, the KF method can be applied to any algorithm, as it is independent of the model itself. In this work, we combined two kinds of kernels (in the DNILMF algorithm) including the drug kernel from structural information (or target kernel from sequence information) and the drug GIP kernel from the interaction matrix  $Y$  (or target GIP kernel of  $Y$ ). However, multiple kernels are allowed to KF.

Besides the KF technology, many other methods were also proposed to improve the data set itself, which are independent

of algorithms. For example, in the SCMLKNN algorithm, the authors proposed the super-target technique, which first clusters the targets into protein families on the basis of sequence similarity. By performing such operation, the data sparsity problem can be solved to some extent. In the KronRLSWNN algorithm, the authors proposed to use a WNN to infer the interaction profiles for new drugs, which have no interaction data with any targets.

In both RLSKF and DNILMF, a similar process was also used to infer those profiles for both new drugs and new targets. Technologies such as KF, super-target clustering and WNN, which are unsupervised methods, are more straightforward and flexible to combine with other algorithms, as they are obtained before the model building step.

Therefore, such unsupervised technologies are often adopted by researchers who do not have statistical/mathematical background because of the simplicity and easy implementation compared with supervised algorithms, which often require optimization process with complex mathematical knowledge.

Besides the mentioned algorithms above, in fact, there are many algorithms from other scientific disciplines such as implicit feedback, which can be smoothly transformed and applied to tackle DTI predictions.

Indeed, progress for one scientific field may be accelerated by 'borrowing' ideas, concepts or theories from a different discipline. For example, NRLMF borrows the logistic matrix factorization technique used by collaborative filtering with enhanced objective function, and DNILMF borrows the 'trust ensemble' idea from the recommender systems field by adding similarity fusion technique and extending the original one to dual integration. Thus, with the development of algorithms in various research fields, it is beneficial to transfer across-discipline methods into the DTI prediction field.

In this work, we assessed five open-access DTI prediction algorithms based on the experimental setting where both training and test sets share common drugs/targets, and proposed to use a recall-based metric to evaluate the models. Algorithms, which can

handle new drug/target scenarios, will be studied in the future, and additional and multiple evaluation metrics for estimating one-class classification problems may be taken into consideration. Despite of the many applications in previous work, the benchmark data sets used are rather limited. In fact, as more DTI data becomes available in the public domain, it will be beneficial to apply the DTI algorithms to diverse data sets for comparison. In summary, the current work compares and analyzes the performance on DTI predictions using the commonly used benchmark data set and DTI data in the kd data set. Such review and comparative work may provide insights for advancing the state of art for DTI predictions by developing new methods to improve scalability and gain stronger generalization abilities, as well as to effectively incorporate negative samples and better handle regression-format data.

### 2.2.5 Funding

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

**Ming Hao** is a Postdoctoral Researcher at National Center for Biotechnology Information, National Institutes of Health. His research focuses on bioinformatics and chemoinformatics related to drug design.

**Stephen H. Bryant** is a Senior Investigator at National Center for Biotechnology Information, National Institutes of Health. His structure group conducts basic research in bioinformatics and cheminformatics.

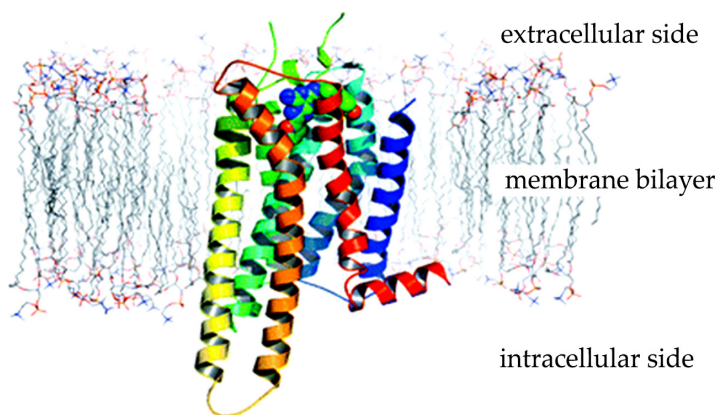
**Yanli Wang** is a Lead Staff Scientist at National Center for Biotechnology Information, National Institutes of Health. Her research interests include computational methods for drug discovery and data mining chemogenomic data.



## 2.3 CHEMOGENOMIC APPROACHES TO RATIONAL DRUG DESIGN

Remarkably, medicinal chemistry followed a parallel boost with the miniaturization and parallelization of compound synthesis, such that over 10 million non-redundant chemical structures covers the actual chemical space, out of which ca. 1000 have been approved as drugs. Therefore, only a small fraction of compounds describing the current chemical space has been tested on a fraction of the entire target space. Chemogenomics is the new interdisciplinary field, which attempts to fully match target and ligand space, and ultimately identify all ligands of all targets. Various definitions of overlapping fields (chemical genetics, chemical genomics) have been proposed. We will herein consider a broad definition of chemogenomics encompassing chemoproteomics, namely the study of small-molecular-weight drug candidates on gene/protein function. From the definition of the field, one easily understands that chemogenomics will be at the interface of chemistry, biology and consequently informatics since data mining is required to extract reliable information. Furthermore, methodologies at the border of chemistry and biology (medicinal chemistry), chemistry and informatics (chemoinformatics), biology and informatics (bioinformatics) will also play a major role in bringing these major disciplines together. Chemogenomic approaches to drug discovery rely on at least three components, each necessitating hard experimental work: (1) a compound library, (2) a representative biological system (target library, single cell and whole organism), and (3) a reliable readout (for example, gene/protein expression, high-throughput binding or functional assay). By definition, analysing chemogenomic data is a never-ending learning process aimed at completing a two-dimensional (2-D) matrix, where targets/genes are usually reported as columns and compounds as rows, and where reported values are usually binding constants ( $K_i$ ,  $IC_{50}$ ) or functional effects (for example,  $EC_{50}$ ). This matrix is sparse as far as all possible compounds have not been tested on all possible genes/proteins. Predictive chemogenomics will thus attempt to fill existing holes by predicting compounds–genes/

proteins relationships. *In silico* approaches to predict such data (target selectivity for various ligands and ligand selectivity for various targets) will span pure ligand-based approaches (comparison of known ligands to predict their most probable targets), pure target-based approaches (comparison of targets or ligand-binding sites to predict their most likely ligands) or ultimately target-ligand based approaches (using experimental and predicted binding affinity matrices).



### 2.3.1 Ligand space

To efficiently navigate in ligand space, one first needs to describe the compound using appropriate properties (descriptors) and then to use a master equation to measure a distance between two compounds (similarity metric).

**Table 5:** Ligand descriptors

| <i>Dimension</i> | <i>Nature</i> | <i>Examples</i>  |
|------------------|---------------|--|
| 1-D              | Global        | Molecular weight, atom and bond counts (for example, number of H-bond donors, number of rings), polar surface area, polarizability, $\log P$ |

|     |                |   |
|-----|----------------|---|
| 2-D | Topological    | Topological and connectivity indices, fragments, substructures (for example, maximum common substructures), topological fingerprints (for example, structural keys) |
| 3-D | Conformational | <i>n</i> -points pharmacophore, shape, field, spectra and fingerprints  |

### 2.3.2 Target space

However, sequence lengths may considerably vary within a protein family (for example, sequence lengths of human GPCRs range from 290 to 6200 residues), such that analysing similarities and differences first requires an alignment of amino-acid sequences which can be tricky in case of large insertions/deletions. Therefore, one may focus on specific motifs which are a collection of continuous residues specific of a protein family (for example, DRY motif in TM III of rhodopsin-like GPCRs). To take into account the structural organization of the target, it can be of interest to look at the 2-D structure (mapping of  $\alpha$ -helices,  $\beta$ -sheets, coils and random structures) and even better at the 3-D structure (atomic coordinates provided by X-ray diffraction, NMR or molecular modelling) and/or the corresponding fold. In chemogenomics-related approaches, one usually focuses on the ligand-binding site, where structural similarities among related targets are usually much higher than when considering the full 1-D sequence or 3-D structure.

**Table 6:** Structural classification of proteins

| <i>Dimension</i> | <i>Classification scheme</i> | <i>Databases</i>  |
|------------------|------------------------------|---|
| 1-D              | Sequence                     | UniProt (Wu <i>et al.</i> , 2006) and Pfam (Finn <i>et al.</i> , 2006)        |
|                  | Patterns                     | PRINTS (Attwood <i>et al.</i> , 2003) and PROSITE (Hulo <i>et al.</i> , 2006) |
|                  |                              |   |
| 2-D              | Secondary structure fold     | SCOP (Casbon and Saqi, 2005) and CATH (Reeves <i>et al.</i> , 2006)           |

|     |                    |   |
|-----|--------------------|---|
|     |                    |   |
| 3-D | Atomic coordinates | PDB (Berman <i>et al.</i> , 2000) and MODBASE (Pieper <i>et al.</i> , 2006)           |
|     | binding site       | Binding MOAD (Hu <i>et al.</i> , 2005) and sc-PDB (Kellenberger <i>et al.</i> , 2006) |

Targets may also be classified according to their pharmacological profile (binding affinity for a panel of ligands) which means according to the nature of ligands they recognize. Of course, there is a considerable overlap between sequence- and ligand-based classifications, since ligands generally bind to a subset of the protein universe. However, relationships across protein subfamilies are particularly interesting in drug design for predicting/modifying the pharmacological profile of a drug.

### 2.3.3 Target–ligand space

It is possible to directly navigate in the protein–ligand space by browsing full matrices in which either affinity or structural information is stored.

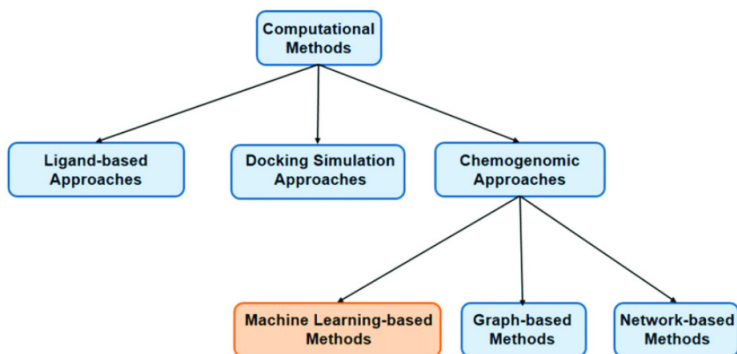
Experimental evaluation of  $x$  compounds on  $y$  targets (for example, *in vitro* binding affinity assay) leads to a matrix of  $xy$  numbers (for example,  $IC_{50}$  values), which can be used to predict the affinity of a new compound to an existing target by multivariate linear regression measure a structure–activity relationships distance between two targets and predict a global pharmacological profile.

A clear advantage of this approach is that it relies on true binding affinity values and that experimentally derived descriptors will usually outperform computed descriptors.

A clear drawback is the enormous amount of data required to derive true information such that similar approaches are not realistic, for example, in an academic environment.

## 2.4 LIGAND-BASED CHEMOGENOMIC APPROACHES

The basic paradigm underlying ligand-based chemogenomic approaches is that molecules sharing enough similarity to existing biologically annotated ligands have enhanced probability to share the same biological profile. It is therefore very important to annotate chemical libraries with biological information (targets, *in vitro* affinity data and ADMET properties). Over recent years, there has been a huge effort mainly from small biotech companies to compile such data by an exhaustive survey of literature and patent data. Since chemogenomic approaches usually focus on target families, most of these archives are related to the most pharmaceutically important target families (GPCRs, kinases, nuclear hormone receptors (NHRs), proteases and phosphodiesterases).



**Table 7:** Biologically annotated compound libraries

| Database | Description   | Website   |
|----------|---|---|
| AurSCOPE | Target family-oriented knowledge database containing pharmacological and pharmacokinetic data for 160 000 GPCR ligands and 77 000 kinase inhibitors | <a href="http://www.aureus-pharma.com">http://www.aureus-pharma.com</a> |

|                       |   |   |
|-----------------------|---|---|
| Bioprint              | Biological profile ( <i>in vitro</i> and clinical data) of 2400 small-molecular-weight drugs and drug-like compounds                | <a href="http://www.cerep.fr/">http://www.cerep.fr/</a>                             |
| ChemBank              | Storage of 50 000 compounds and related biological properties in 441 high-throughput screening and small molecule microarray assays | <a href="http://chembank.broad.harvard.edu/">http://chembank.broad.harvard.edu/</a> |
| ChemBioBase           | Target centric ligand databases (GPCRs, kinases, PDE)   | <a href="http://www.jubilant-biosys.com/">http://www.jubilant-biosys.com/</a>       |
| Kinase knowledge base | kinase structure–activity and chemical synthesis data   | <a href="http://www.eidogen-sertanty.com/">http://www.eidogen-sertanty.com/</a>     |
| MDL Drug Data Report  | 132 000 biologically relevant compounds and well-defined derivatives  | <a href="http://www.mdli.com/">http://www.mdli.com/</a>                             |
| MedChem database      | 650 000 compounds with biological and pharmacological information   | <a href="http://www.gvkbio.com">http://www.gvkbio.com</a>                           |
| StARLite              | Highly curated target-compound SAR relationships  | <a href="http://www.inpharmatica.co.uk/">http://www.inpharmatica.co.uk/</a>         |
| Wombat                | 154 236 entries over 307 700 biological activities on 1320 unique targets   | <a href="http://sunsetmolecular.com/">http://sunsetmolecular.com/</a>               |

On the other hand, annotation of targets was based on existing classifications for enzymes and receptors. Linking MDDR ‘activity keys’ to the target classification scheme enabled the annotation of 53 000 compounds totalling 799 different activity keys and related targets. Since the target’s sequence is linkable to the ligand, sequence-based similarity searches of ligands for protein homologues of liganded targets are therefore feasible. Annotated reference ligands for a particular GPCR were used as starting points to recover either new receptor ligands or ligands of receptors

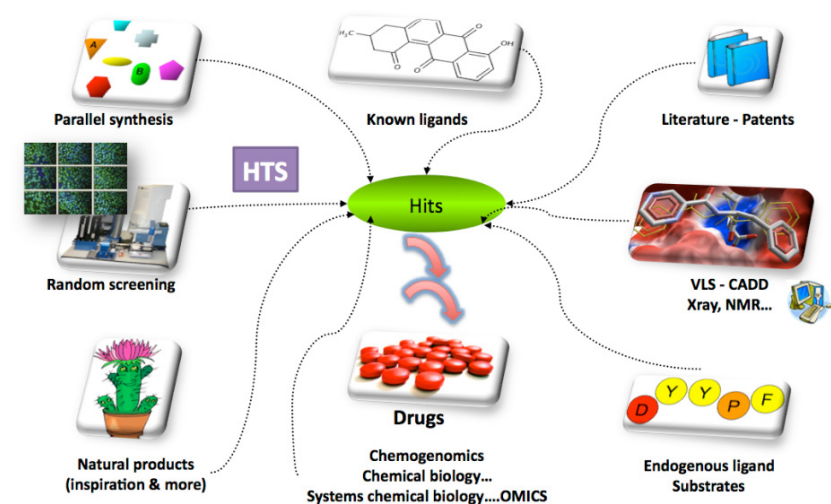
close to the reference GPCR. Interestingly, the efficiency of the virtual screening approach was dependent on the phylogenetic distance between the reference and the query targets. Another straightforward application of biologically annotated compound libraries is the design of target-directed combinatorial libraries focusing on chemotypes preferred by a family of targets.

Natural products also cover a very interesting chemical space of biological relevance because of the evolutionary pressure put on these compounds to bind, usually through highly specific mechanisms, to particular targets. The chemical space spanned by biologically annotated natural products was described recently as a structural and hierarchical scaffold tree which can be browsed to design natural product-oriented chemical libraries.

Biologically annotated compound libraries are a direct source of potentially new biological mechanisms to correct a phenotype. It designed a library of 2036 biologically active compounds covering 169 different biochemical mechanisms, which was shown to be structurally diverse and able to provide 85 hits in a cell viability and proliferation assay. Among the 85 hits, 27 were supposed to be active by new biochemical mechanisms.

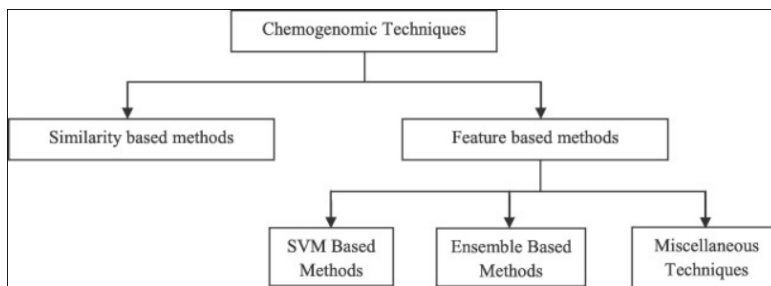
### 2.4.1 Ligand-based *in silico* screening

Main target families can be distinguished by a simple look at physicochemical properties (molecular weight,  $\log P$ , polar surface area, H-bond donor and acceptor counts) of their cognate ligands. One can thus easily imagine that more sophisticated descriptors can be used to predict a global target profile for any given compound, provided that targets to be predicted are sufficiently well described by existing ligands. Ligand-based *in silico* approaches to target fishing begin to appear in the literature.



## 2.5 TARGET-BASED CHEMOGENOMIC APPROACHES

Controlling the selectivity of ligands towards related targets from the same family is crucial information in early drug-discovery stages. There is therefore a growing interest in comparing all targets from the same family especially those for which there is enough structural data (X-ray or NMR structures) to enable a proteome-wide comparative modelling of targets of still unknown structure (for example, protein kinases). Target-based chemogenomic approaches can be classified in two categories depending on whether the amino-acid sequence or the 3-D structure of targets is compared.





### 2.5.1 Sequence-based comparisons

Sequence-based approaches are intended to be used for any kind of target family, provided that a multiple alignment of all targets to compare is reachable. They are generally used for target families where a lack of high-resolution structural data hampers target comparison. A simple one consists in target hopping, which means discovering receptor ligands for a particular receptor by considering first the known ligands of closely related receptors. For example, CRTH2 receptor antagonists could have been identified from existing angiotensin II type 1 receptor antagonists, because both receptors were found close in the GPCR cavity-biased tree. In addition, the design of targeted libraries towards a particular area of the tree is facilitated by addressing those residues responsible for selectivity/promiscuity

### 2.5.2 Structure-based comparisons

Structure-based comparisons are only possible for target families where there are enough good structural templates (X-ray structures) to afford the homology modelling of other related targets. In general, only ligand-binding sites are compared, since the basic aim of such comparisons is to understand the selectivity/permissivity features of related targets of known ligands.

Starting from a structural alignment of all targets, interaction energies generated by rolling several probe atoms (for example, sp<sup>3</sup> carbon atom) at each point of 3-D grid encompassing the ligand-binding site are then concatenated into a MIF vector, which can be placed in a global matrix where rows describe targets and columns interaction energies at a given 3-D grid point.

A nice example of binding site similarities for distant proteins has been exemplified by, who detected cross-reactivity of arylsulfonamide-based COX-2 inhibitors with human carbonic anhydrase (HCA) based on the similarity of COX-2 and HCA binding pockets. A problem with these matching techniques is that the computed similarity score (usually dependent on the number

of atom/pseudocenter/triangle matches) is not always easy to interpret, notably for active sites of different dimensions, because large active sites will have a tendency to present more matches than small ones even if the latter are more similar. Therefore, normalized distance metrics similar to those used for comparing ligands are needed. A promising approach is proposed by, who discretizes an active site by a dimensionless 80-triangle sphere and projects, from  $c\beta$  atoms of cavity-lining residues to the sphere centre, various topological and physicochemical descriptors.

### 2.5.3 Chemical annotation of target binding sites

However, as far as information about the binding site is missing, there is a potential risk to compare compounds sharing the same target but not the same binding site (for example, orthosteric and allosteric ligands). It is therefore important to rigorously annotate protein sequences and/or binding site by the chemotype of the ligands they can recognize. The SMID (Small Molecule Interaction Database) archive is an interesting initiative to annotate protein amino-acid sequences by domain-specific ligands. A total of 6300 ligands covering 230 000 experimentally observed domain/small molecule interactions have been stored in a relational database, which can be browsed to predict the most likely ligand of proteins of unknown 3-D structures by comparison of their domains to known protein structures using a reverse position-specific basic local alignment search (BLAST) procedures.

### 2.5.4 2-D searches

To browse and predict protein–ligand complexes, one needs to set up simple descriptors for both ligands and proteins from knowledge databases and concatenate them into a single protein–ligand description. A machine-learning algorithm was trained from 5319 non-redundant known complexes and applied to a set of 1 911 415 virtual complexes (55 orphan receptors and 34 753 drug-like compounds from the NCI database) to predict the most likely associations. Out-of-sample validations (finding the receptors of

a promiscuous ligand and the ligands of a single target) were in general agreement with literature data and some predictions still awaiting experimental validations have been made.

### 2.5.5 3-D searches

A straightforward way to predict putative targets of ligands is to dock each of the ligands of the compound library into each of the active site of the target library. This strategy has been validated by several groups and proved able to recover the known ligands of known targets and predict their off-targets and thus some potential side effects. Up to now, there is a single successful target fishing application described in the literature utilizing a docking approach. Hence, inverse docking requires first a high-quality 3-D dataset of binding sites whose automated set-up is quite difficult, and second an accurate scoring function to properly rank targets. A problem is that energy-based scoring functions are not very good at quantifying very heterogeneous protein–ligand complexes by decreasing binding-free energies and that alternative ways of scoring are requested for efficient target selection.

Usage of IFPs have shown several promising features: (i) enhancing the quality of pose prediction in docking experiments; (ii) clustering protein–ligand interactions for a panel of related inhibitors according to the diversity of their interaction with a target subfamily; (iii) assisting target-biased library design.

However, docking-independent 3-D methods may also constitute an interesting approach to predict protein–ligand complexes. A significant problem is to encode protein and ligand properties with similar descriptors such that one partner can be retrieved by using the second one as a query. A promising solution is proposed with the CoLiBRI (Complementary Ligands Based on Receptor Information) method in which both ligand and active site atoms are described by a same vector of molecular descriptors derived from shape and electronic properties of isolated atoms. Therefore, it is possible to directly correlate chemical similarities between active site and their ligands by mapping patterns of active sites onto

patterns of their complementary ligands. When applied to a test data set of 800 high-resolution PDB complexes, the complementary ligand was ranked among the top 1% of a large library in 90% of tested active sites. Accuracy dropped significantly for active sites very different from those in the test set but still usable as a prefiltering step for removing the most improbable ligands

### 2.5.6 Concluding remarks

Chemogenomic approaches to rational drug discovery have been exploding in the last years as high-throughput data (structure, binding affinity and functional effects) become available for both targets and ligands of pharmaceutical interest. Numerous ways to link those data have been proposed focusing on either ligand or target neighbourhood. A clear data organization and storage is necessary to foster such applications and begins to emerge for the most interesting target families (kinases, GPCRs and NHRs). In a near future, an earlier and better control of ligand selectivity can be anticipated by using chemogenomic data. This does not mean that more selective ligands are going to be designed, but simply that the observed selectivity profile of the compound will be compatible with a therapeutical usage. In addition, novel genomic targets could be better addressed after locating them in the target space and exploiting the associated chemical information.

## 2.6 GENE KNOCKOUT TECHNOLOGY

A gene knockout (abbreviation: KO) is a genetic technique in which one of an organism's genes is made inoperative ("knocked out" of the organism). ... Knockout organisms or simply knockouts are used to study gene function, usually by investigating the effect of gene loss.

### 2.6.1 Utility and Importance of Gene Knockout Animals

The importance of many dietary constituents in maintenance of health is obvious. While deficiencies in dietary intake of specific

nutrients may be detrimental to growth, reproduction and immunity, excessive amounts of specific nutrients in the diet can also lead to disease states. For example, increased intake of dietary fat and cholesterol is associated with hyperlipidemia and an increased risk of coronary heart disease. Excessive intake of vitamin D has also been shown to result in soft-tissue calcification and renal calculi. However, it must be noted that while the correlation exists between increase nutrient uptake and specific diseases, the response of a given individual is quite variable. These individual variations in dietary responsiveness are likely due to the different genetic composition of each individual. Numerous genetic factors are involved in determining responsiveness to specific dietary nutrients. These include genes important for nutrient absorption as well as those important for the metabolism and processing of the nutrient in the diet. Additionally, the amount of each nutrient in the diet also has an impact on the level of specific gene expression. Such regulatory mechanisms may also account for individual differences in susceptibility to diet-induced diseases.

Advances in molecular biology techniques during the past decade have led to an explosion of research aimed at understanding diet and gene interactions in health and diseases. While most of the earlier work focused on nutrient regulation of gene expression, transgenic technology has also been used to study the effect of overexpression of specific genes in modulation of dietary nutrient effects and on the metabolism of dietary components as they relate to normal health and pathogenesis of diseases. More recently, targeted gene inactivation, commonly known as gene knockout, has been employed to study the functional importance of specific genes and the impact of specific genetic mutations and deletions on complex metabolic processes which ultimately lead to various diseases. This review provides an overview of the utility and importance of the gene ablation technology to nutritional and metabolic research

## 2.6.2 Experimental approaches to modulate gene expression in vivo

Three different approaches have been used to inhibit specific gene expression in mammalian systems. The most common approach is specific gene ablation by homologous recombination in embryonic stem cells and then the production of animals with defects in expression of the specific gene. However, technical difficulties in obtaining high level stable expression of antisense nucleotides and ribozymes have limited the usefulness of these approaches. Most of the research with antisense nucleotides and ribozymes were restricted to in vitro cell culture systems and, thus, were of limited value to nutritional and metabolic studies. In contrast, targeted gene disruption by homologous recombination in embryonic stem cells have been employed widely to produce animal models with specific gene deletions. Many of these models are quite useful for nutritional and metabolic research. Therefore, this review will focus on the use of this technique.

Targeted gene disruption by homologous recombination takes advantage of the observation that pluripotent embryonic stem (ES) cells obtained from mouse blastocysts can be cultured in vitro and remain viable for differentiation after their injection into a different embryo and reimplantation into a foster mother. In a typical experiment, the ES cells and the recipient embryo are obtained from animals carrying genes of different coat colors, such that the initial selection of chimeric mice can be based on the coat color of the offspring. The most commonly used ES cells to date are those derived from the mouse strain 129, which has an agouti coat color. These ES cells can then be microinjected into embryos obtained from C57BL/6J mice, which have a black coat color. Offspring with a high degree of agouti coat color, indicating the transmission of ES cell-derived genes, can then be crossbred with each other to obtain animals with a genetic background identical to that of the ES cells. Using this approach, mice with specific gene modifications can be obtained by manipulation of the ES cell genome.

Modification of specific genes in the ES cell genome depends on the ability of transfected DNA to recombine with the homologous gene in the chromosome. Although such targeted recombination is a rare event in comparison with nonhomologous integration of the transfected DNA, methodology has been developed to optimize the chance of homologous recombination and to rapidly screen and select ES cells in which such event has taken place. Most of the experiments to date utilized isogenic DNA for the targeting construct to maximize hybridization of the targeting DNA to the proper gene locus in the chromosome.

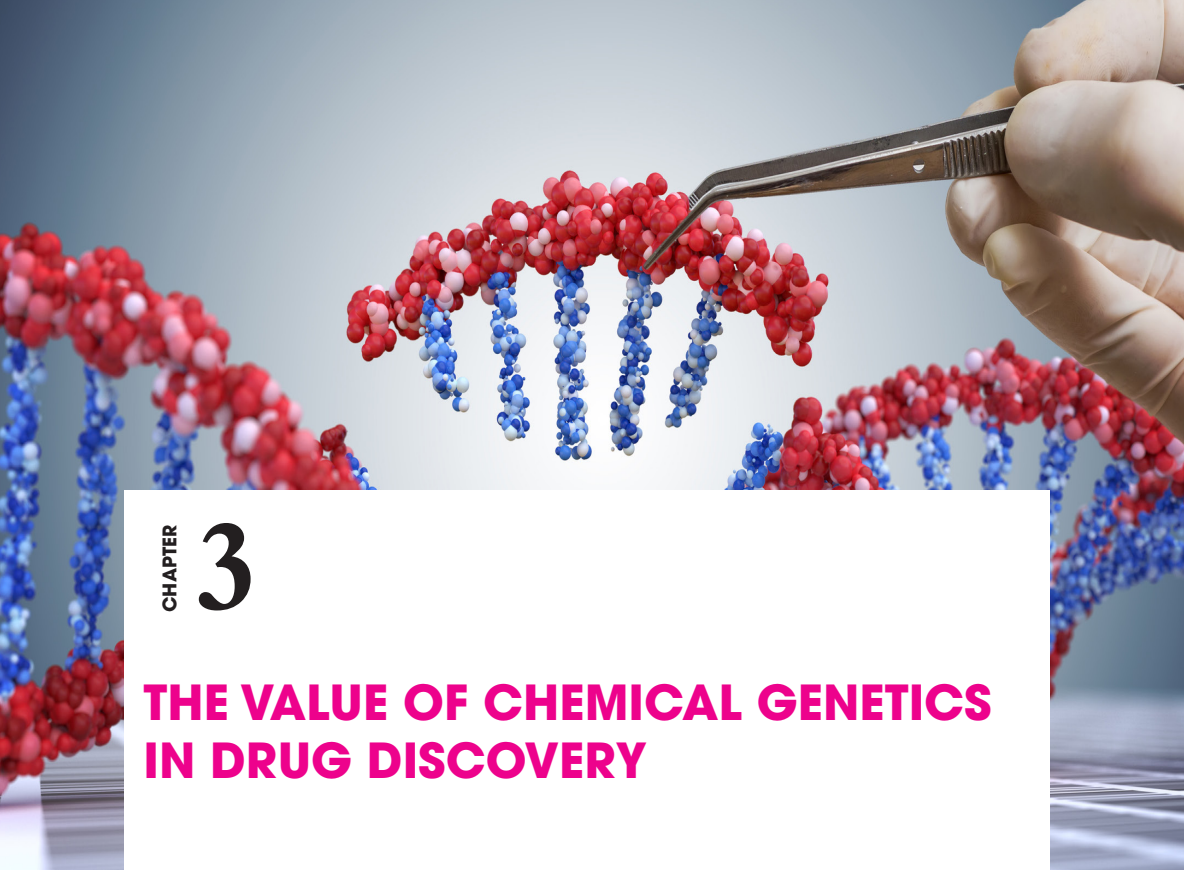
Homologous recombination of the transfected DNA with chromosomal DNA at the target locus will result in the replacement of a portion of the endogenous gene with the targeting construct, thus disrupting the coding sequence and inactivation of the endogenous gene. The use of a selectable gene marker allows the selection for cells that have taken up and expressed the transfected DNA. Growth of the ES cells in the presence of antibiotic selection indicates the integration of the transfected DNA into the ES cell genome. In a successful experiment, approximately 0.01–0.001 of the antibiotic-resistant cells would have the transfected DNA targeted to the proper gene locus, while the remaining cells would have incorporated the DNA in a nonhomologous site. In some cases, investigators have included a negative selectable gene marker at the 5' or 3' end of the targeted construct to allow for selection against random insertion events. Homologous recombination at the targeted gene locus would result in deletion of the negative selectable gene marker, while integration at nonhomologous sites would have included this marker in the genome. The inclusion of a negative selection marker usually results in an additional 10-fold enrichment of homologous recombination clones.



## REFERENCES

1. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics*. 2005;4:752–761.
2. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*. 1994;243:327–344.
3. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*. 2003;31:400–402.
4. Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem*. 2004;2:3204–3218.
5. Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW. ‘Bayes affinity fingerprints’ improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept. *J Chem In Model*. 2006;46:2445–2456.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res*. 2000;28:235–242.
7. Gashaw, Isabella, et al. “What makes a good drug target?”. *Drug discovery today* 17 (2012): S24-S30.
8. Hughes, James P., et al. “Principles of early drug discovery.” *British journal of pharmacology* 162.6 (2011): 1239-1249.
9. Kuhn, Max, Ian Peers, and Stan Altan. *Nonclinical statistics for pharmaceutical and biotechnology industries*. Springer, 2016.





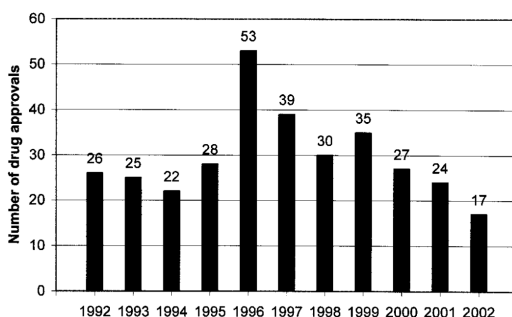
## CHAPTER 3

# THE VALUE OF CHEMICAL GENETICS IN DRUG DISCOVERY

## INTRODUCTION

To understand what chemical genetics is and how it can add value to the drug discovery process, we must first consider some of the challenges and needs of the pharmaceutical industry. The process of discovering new drugs is a highly complex multidisciplinary activity requiring very large investments of time, intellectual capital, and money. Today the average cost of bringing an NCE to market is on the order of \$ 900 million. For every 5000 compounds synthesized, only one makes it to the market. Only three of ten drugs generate revenue that meets or exceeds average R&D costs, and 70% of total returns are generated by only 20% of the products. Given this gloomy backdrop it is even more disturbing to learn that, despite the proliferation of many new technologies of great potential (and great cost), pharmaceutical productivity levels

have not increased in the past ten years (as shown graphically in Figure 1).



**Figure 1.** US drug approvals during the past ten years.

Pharmaceutical R&D costs continue to grow exponentially, driven in part by investments in new technologies, but the return on this investment remains elusive. There are many reasons for these disturbing trends. If we consider the pharmaceutical industry as primarily a generator of knowledge (defining knowledge as compiled and interpreted information that can be acted upon) and focus on the knowledge creation process, we can shed some light on how the current situation, a productivity gap, emerged. Working harder is not likely to overcome this productivity gap to deliver more drugs. Working smarter, doing things differently, and focusing on what we actually need to deliver, i.e., knowledge, may be a new way to approach the problem. Ultimately, spanning the ‘knowledge gap’ will lead us to the efficient exploitation of the human genome to discover new drugs to meet major medical needs.

### 3.1 KNOWLEDGE MANAGEMENT IN DRUG DISCOVERY

Pharmaceutical companies create and sell knowledge, e.g., knowledge that a drug product will rid patients of the symptoms of their disease while not causing serious side effects. The

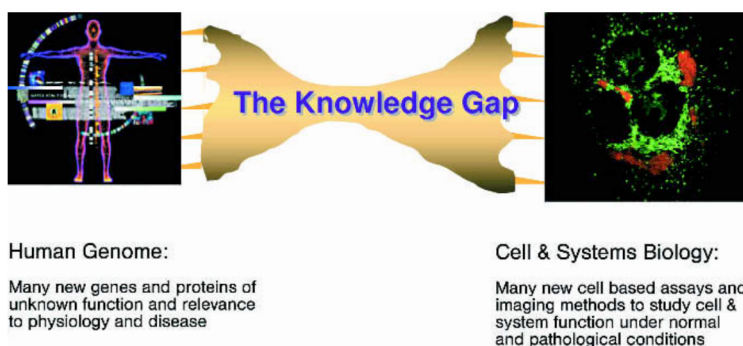
resources that go into the production of the drug pale alongside the resources needed to discover the knowledge of what the drug will do when administered to a patient. In the early years of drug discovery it was often true that the literature provided a significant knowledge base for our efforts. Two approaches were taken: (1) function-based screening, where one did not know what the target was but could easily screen for small molecules that possessed the right biology [3]; and (2) ‘rational drug discovery’, where one has knowledge of the target and its function. What was needed were small molecules that would interact with the target in the right way before being optimized for in vivo activity and safety.

The existing and evolving chemistry and biology literature fueled these efforts. It is probably also true to say that the medical problems addressed in these early days of drug discovery represented the more accessible opportunities. Often the biology was not only reasonably well understood, but it was reasonably easy to study and measure. Examples of biological effects that were tackled include blood pressure, acid secretion, and cytotoxicity. The situation today is very different. We now face many new targets we know little about and biology that is complex to study and understand. In addition to these issues, advances in our knowledge of distribution, metabolism, and pharmacokinetics, as well as toxicology and pharmacogenetics, have led to the introduction of discovery processes that front-load measurement of such small-molecule properties. This also raises the bar for passage of compounds through the process – making the process more difficult and slower. While this may lead to lower output of development candidates, it should also lead to lower failure rates later in development, i.e., improvements in quality.

## **3.2 KNOWLEDGE GAPS, THEIR IMPORTANCE, AND HOW TO ADDRESS THEM**

The human genome has been solved and optimistic promises have been made. It is clear that the human genome did not deliver knowledge (i.e., something immediately useful); rather, it

delivered a massive amount of data. Significant advances have also been made in cell biology and systems biology. The relationship between genes/proteins derived from the human genome and their function as a part of a biological system constitutes the knowledge gap, and our appreciation of the extent of this void is still emerging. The human genome is thought to consist of ca. 30 000 genes. Each gene can potentially produce several proteins via alternative splicing and post-translational modification, and every protein can potentially combine with other proteins to form many different protein complexes. Clearly, the number of different proteins and protein complexes is much larger than 30 000. To add further complexity, small molecules (that we hope will become drugs) can interact with different sites on a protein or via different mechanisms to further expand the diversity of possible outcomes from the interaction of small molecules with a protein target. We do not know what many gene products (proteins) do, either physiologically or pathologically, and we do not really know how many of these proteins can interact with small-molecule ligands. There are many genes about which we know nothing at all. In summary, there is clearly a vast knowledge gap between knowing a gene and knowing the function (physiology and pathology) of its protein product (Figure 2). The enormity of this knowledge gap has been underestimated by the pharmaceutical industry.



**Figure 2.** The knowledge gap represents the large gap in understanding that exists between genetic information from the human genome project and information regarding biological function from cell and sys-

tems biology To illustrate the size of the knowledge gap, consider the following (admittedly approximate) analysis from the area of substance P. Substance P antagonists have emerged in recent years as potential new treatments for depression, although none have yet been approved for this use. Substance P has been known since 1937, and since that time (67 years!) there have been over 6500 papers published providing significant new information on substance P. Thousands of scientists have worked on generating this information during this period. It is sobering that our understanding of Substance P's role in depression is still in its infancy. No one pharmaceutical company can generate this volume of information. New faster and more efficient methods must be developed to fill these knowledge gaps. Partnership with the academic community will become increasingly important as the number of druggable targets expands.

### 3.3 TARGET VALIDATION: THE FOUNDATION OF DRUG DISCOVERY

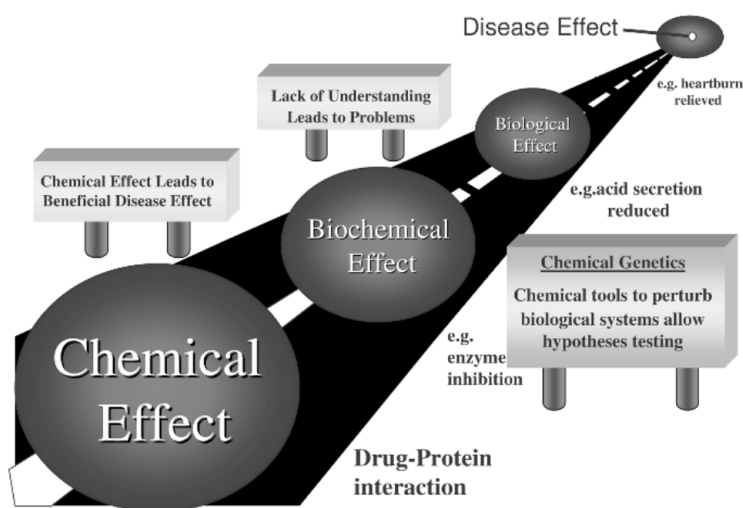
One critical piece of knowledge to the pharmaceutical industry relates to knowledge of a drug target and its link to a disease process. In the context of small-molecule drug discovery, we define target validation in a broader sense as including knowledge of the protein target and its specific interaction with small molecules, and the consequences of this interaction in terms of modifying a disease process. In fact, drug discovery is primarily focused on the biology of a target in the presence of a drug, i.e., drug-induced biology. It begins with a chemical effect – the interaction of a ligand with a protein at a specific site in a specific manner – and ends in patients' gaining benefit from taking a drug derived from the application and exploitation of this knowledge. Target validation that simply links a specific protein and its function to a disease state does not include reference to whether a small molecule can modulate the function of the protein. The protein may not therefore constitute a true target since it is not a target for a small-molecule ligand and efforts to do target validation on

such a protein will ultimately lead to a negative outcome. We can (and do) proceed to work on drug discovery before we have all the knowledge we need. The absence of this knowledge constitutes the major risk of drug discovery. One way to proceed is to focus on obtaining the most critical knowledge first. This is the knowledge that modulation of a protein target by a small molecule can ultimately lead to a clinical benefit in patients.

### **3.4 CHEMICAL GENETICS – HOW CHEMISTRY CAN CONTRIBUTE TO TARGET IDENTIFICATION AND VALIDATION**

Target validation (TV) is the foundation of drug discovery and requires greater attention if we are to reduce the risk of failure after significant investment. Traditionally, target validation has been thought of as a biology problem. Thinking in terms of what knowledge we need makes it clear that the problem does not neatly fall into any particular discipline and is better characterized as an integrated biology and chemistry problem. A schematic target validation roadmap is shown in Figure 3, where the entire validation path from a chemical effect through various levels of biological effects to a clinical effect is outlined. To begin with, an understanding of the function of a particular gene product can often be achieved through the methods of classical genetics. However, the process can be slow and tedious. For example, developing a mouse carrying the mutation of interest could take months or years. Indeed, if the gene product is essential, the organism may not survive long enough to be studied. On the other hand, the situation wherein a molecule is available that alters the function of the gene product has a number of advantages. However, we should recognize that significant chemical effort is often required. The phenotype of interest is conditional, in that it is present only when the molecule is present, allowing the study of essential gene products. It is also tunable, i.e., the intensity of the phenotype can be adjusted by controlling the concentration of the molecule.



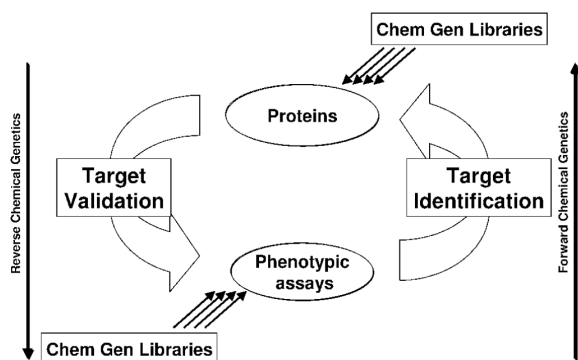


**Figure 3.** The knowledge roadmap for target validation, beginning with a chemical effect between a small molecule and a protein target and ending with a beneficial clinical effect on a person with a disease. Chemical genetics approaches provide some assistance in pursuing this path.

Chemical genetics is the purposeful modulation of protein function through its interaction with a small molecule. The principles of chemical genetics were established in the rich history of using small molecules to explore biological function and, in this sense, chemical genetics is not new. What is new is the development of a systematic approach to studying biological function with small molecules – this is the emerging field of chemical genetics. Just as genetic changes can alter protein function, so can small molecule–protein interactions. It is important to appreciate that, by interaction of a ligand with a protein, we mean interaction of a small molecule at a specific site on a protein causing a specific protein change, conformational or otherwise, ultimately leading to a specific biological effect. Small molecules can often interact with multiple sites on proteins and cause a multitude of consequences such as agonism, antagonism, partial agonism, modulation, competitive and noncompetitive inhibition, etc. They can also interact at junctions between protein subunits. The sophistication of small molecule–protein interactions and their biological consequences cannot be easily reproduced by techniques such as gene knockin/

out or the use of siRNA, by which genes/proteins are simply removed or increased in concentration in a biological system. Having said that, knockout models have certainly contributed significantly to drug discovery and will continue to do so. The power of chemical genetics resides in this sophistication of the small molecule–protein interaction and the precise way we can (in principle) modulate the function of a protein. As a precursor to drug discovery it serves the purpose of focusing us on where small molecule drug discovery really begins – with the chemical interaction between a small molecule and a protein.

The knowledge gap outlined above can be thought of as a cycle linking the target (a protein or protein complex) with a function ultimately linked to an effect important in a disease process (Figure 4). Going from target to function represents the knowledge path of target validation. Going from a function to a target represents the knowledge path of target identification (TI) or deconvolution. Chemical genetics approaches can be applied to both knowledge paths. Application to the target validation path is called reverse chemical genetics. Application to the target identification/deconvolution path is referred to as forward chemical genetics. At the heart of this approach to knowledge generation in TI/TV is the simple concept that small molecules are used to perturb biological systems. Manipulation of a biological system in a controlled manner by small molecules allows us to study these systems more systematically.



**Figure 4.** Chemical genetics tools (libraries) can help uncover the function of proteins (target validation) and the protein target responsible for



biological function (target identification) in a phenotype assay.

### 3.5 INTEGRATION OF CHEMISTRY AND BIOLOGY: IMPORTANCE AND ISSUES

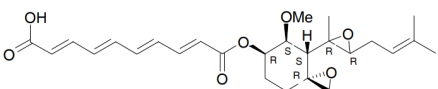
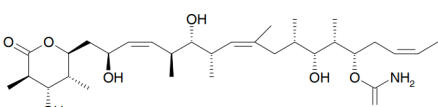
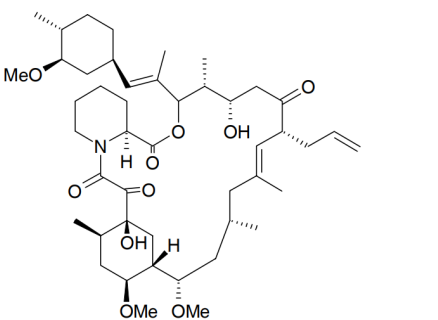
Given that the foundation of target validation is a ligand–protein interaction (a chemical effect) and its consequence (a biochemical/biological effect), we can expect that advances in this area will come from a close integration of chemistry and biology. Some key questions at the interface of chemistry and biology that are fundamental to chemical genetics include – why are some molecules biologically active while others are not? What is the biological profile of a small molecule’s structure and how do we dissect this into what each part (fragment) of the small molecule is doing to each protein target? Is there a protein ‘code’ for recognition of small molecules that is used by every protein in the proteome? The following sections begin to address these questions.

### 3.6 FINDING NEW CHEMICAL TOOLS AND LEADS

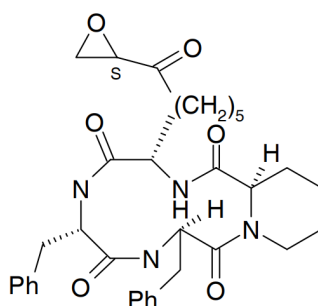
A chemical tool is small molecule that is sufficiently potent and selective for a protein target to be used in the identification and validation of that target. It could, although it need not, meet the rigorous absorption, distribution, metabolism, excretion, and toxicology criteria required of a lead to start an optimization project. How do we find such tools? The total number of ‘reasonable’ drug-like molecules has been estimated as approximately  $10^{63}$  discrete molecules, a number so large that synthesizing all of them is simply impossible. Natural products were designed by nature to bind to proteins and other macromolecular targets and represent powerful chemical tools for use in chemical genetics. Numerous examples exist in which natural products have been identified that modulate biological function. The natural products are then used to identify proteins that they interact with and so to begin deconvolution (forward chemical genetics) of the target

responsible for the biological effect. For example, fumagillin inhibits new blood vessel growth (angiogenesis), and analogs of this compound are now in Phase 3 trials. Using fumagillin as a starting point, chemical tools (e.g., biotinylated analogs) were constructed to bind and tag cellular proteins. One of these proteins, methionine aminopeptidase, has been identified as the likely target for this class of molecules. Some other examples of natural products used in forward chemical genetics are shown in Table 1. Cases in which these natural products were then used to deconvolute the target protein are noted. Interestingly, some of the top-selling drug classes originated from a forward chemical genetics approach, e.g., the gastric acid secretion inhibitors omeprazole and esomeprazole were discovered by a process that began with screening for antisecretory agents that lowered stomach acid.

**Table 1.** Natural products used to identify targets.

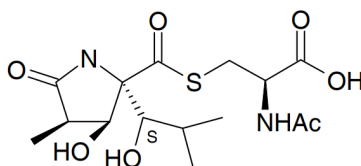
| <i>Biological effect</i>                       | <i>Protein target</i>                          | <i>Natural product</i>  |
|--|--|---|
| Angiogenesis                                   | Methionine aminopeptidase                      | <br>fumagillin       |
| Immuno-suppressive, anticancer                 | Microtubule binder/stabilizer                  | <br>discodermolide |
| Immuno-suppression, IL-2 production inhibition | Calcineurin (a protein phosphatase) inhibition | <br>FK506          |

Histone  
deacetylase  
inhibitor



trapoxin

Proteasome  
inhibition



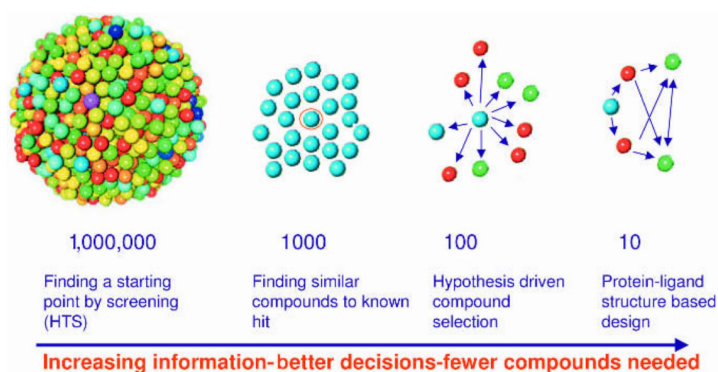
lactacystin

Interestingly, given the discussion of the importance of understanding small molecule–target protein interactions early in drug discovery, there is renewed interest in reexamining many older drugs to more fully understand how they work.

An advantage of the chemical genetics approach is that the small molecules identified in biological screens can act both as conditional switches for inducing phenotypic changes and as probes/chemical tools for identifying protein targets implicated in those phenotypic changes. However, identifying the molecular target and mechanisms by which the small molecules affect biological systems (target deconvolution) can sometimes be difficult. Classical deconvolution approaches, such as affinity chromatography and biochemical fractionation using photoactivatable and other affinity ligands to pull out the target protein, often work well. More recently, genomics-based techniques have been added to the deconvolution toolset.

Beyond natural products, finding chemical tools to modulate biological systems is a difficult step and shares many of the risks associated with finding leads in a drug discovery program. Strategies for finding small-molecule tools representing two

poles on a continuum of approaches are illustrated by structure-based design and the high-throughput screening approach. Given our focus on knowledge generation, it is interesting to note that molecules at either end of this spectrum also reflect different levels of information content. Individual molecules used in high-throughput screening teach us (if we are fortunate) about a simple  $IC_{50}$  or  $EC_{50}$ . Molecules that additionally teach us how they bind to their molecular target provide us with much more useful information, especially when we consider what to do next to improve or change the biology of the molecule (Figure 5).



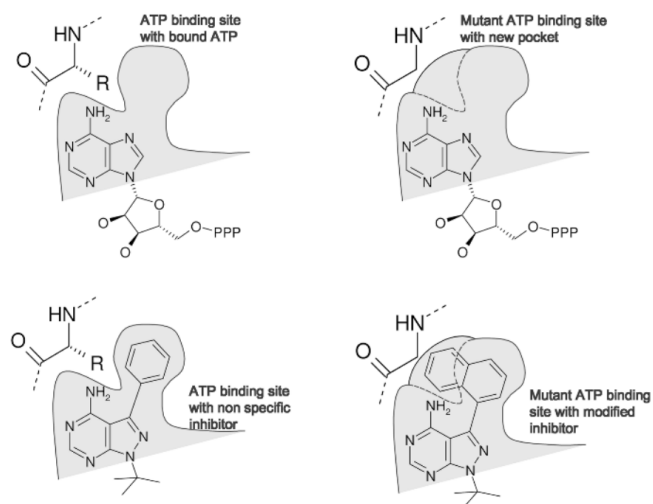
**Figure 5.** The spectrum of approaches to finding chemical tools or leads, illustrating the inverse relationship between information content and number of compounds needed.

Schreiber has been a pioneer in this rapidly developing area of chemical biology. He has constructed several structurally complex screening libraries using a diversity-oriented synthesis approach and has used these libraries to uncover chemical tools to begin to unravel complex biology. Using this approach, Schreiber discovered a small-molecule chemical tool that he named uretupamine, which interacts with the protein Ure2p. Ure2p represses the transcription factors Gln3p and Nil1p.

Uretupamine was found to specifically modulate a subset of glucose-sensitive genes downstream of Ure2p. As noted earlier, this type of behavior, modulating a subset of the function of Ure2p, cannot be replicated by gene knockouts (e.g., knockout of the

URE2 gene) or siRNA approaches and represents a real challenge for proteomics to identify and control all inputs and outputs of a protein. He used the natural product FK506 to uncover its target FKBP12 and then went on to design specific molecular probes derived from FK506, guided by crystal structures of FKBP, FK506, and calcineurin to uncover its mechanism of action as a 'small-molecule dimerizer' of FKBP12 and calcineurin. The formation of this ternary complex led to inhibition of the protein phosphatase activity of calcineurin. This discovery, together with the discovery by Gerald Crabtree of NFAT proteins, helped define the calcium-calcineurin-NFAT signaling pathway, now known to be essential for immune function, heart development, and the acquisition of memory in the hippocampus.

Peter Schultz's team used a combinatorial library of purines to identify agents that could disassemble multinucleated myotubes into mononucleated fragments (a morphological differentiation screen). A new microtubule-binding molecule, mysosoverin, was identified in this way.



**Figure 6.** Replacing a bulky amino acid with glycine in the ATP-binding site of a kinase enlarges the site. ATP binding and catalytic activity are unaffected. The nonselective kinase inhibitor can now be modified to create a molecule that selectively blocks the mutant enzyme.

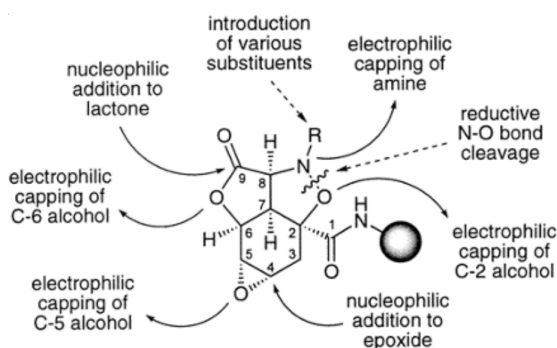
Structure-based design has been employed in some powerful examples of chemical genetics by teams led by Kevan Shokat and by John Koh. Shokat's team has studied the function of kinases by engineering designed modifications into both the kinase and kinase inhibitor ligand to create highly selective chemical tools that can then be used to probe the function of individual kinases in complex kinase cascades (for an explanation of the basic concept see Figure 6).

John Koh's team have focused their efforts on nuclear hormone receptors, including the vitamin D receptor, in an effort to target specific clinical problems. Koh studied a mutant version of the vitamin D receptor (an arginine located in the binding pocket is mutated to a leucine) that binds vitamin D with only one thousandth the affinity of the normal receptor. Analogs of vitamin D were synthesized, based on computer modeling of their interaction in the mutant vitamin D receptor. Some of these compounds were found to bind 500 times better than vitamin D to the mutant receptor. This work may ultimately lead to drugs to treat a disease known as vitamin D resistant rickets. Koh previously demonstrated the feasibility of this approach with other nuclear hormone members, including thyroid hormone receptor.

David Corey's team has also employed this approach, termed 'engineered orthogonal ligand-receptor pairs', in studies of retinoid x receptor to find 'near drugs'. These near drugs are chemical tools used to discern the biology of the retinoid x receptors.

To date, the results of efforts to find new biologically active molecules through preparation of large libraries based solely on diversity considerations have been disappointing. On the other hand, the collective experience of the global bioorganic and medicinal chemistry community indicates that biological activity is not uniformly distributed in chemistry space; rather, it is found within discrete regions. Since we cannot know the locations of these regions a priori, we might look to known biologically active molecules to guide our search. There have been several approaches to doing this. Many natural products derived

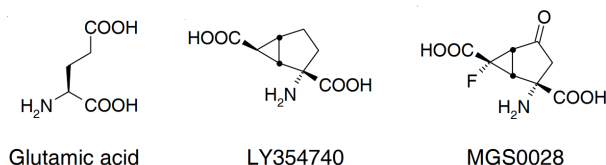
from plants and animals have evolved over time to have specific biological effects on either the parent organism or an unrelated one. The pool of natural products is extremely large with respect to both numbers and structural diversity. Not surprisingly, a number of methods to produce natural product libraries have emerged. Some companies provide prefractionated extracts of unknown structures for screening. Structures are determined after a hit is found. Many companies have established libraries of single pure natural products. Yet another approach is the assembly of libraries of derivatized natural products. Finally, one can develop syntheses of natural product core structures and, using combinatorial techniques, decorate the cores with diverse elements. In this way it is possible to prepare large libraries of peripherally diverse compounds related to natural products for general screening. The following library (Fig. 7) is illustrative. It contains over 2 million compounds that are both sterically and functionally complex. Little biological activity was observed; for the purposes of the pharmaceutical industry this result might be viewed as somewhat disappointing, given the size of the library and the effort invested in preparing it. Why were more active compounds not found?



**Figure 7.** Potential coupling sites on a natural product-related core-based diversity library.

One reason might be related to the high degree of overall molecular complexity of the library. Hann and coworkers reported an in-depth analysis of the relationship between molecular complexity

and the probability of finding leads. They derived a model system in which ligand complexity and ability to bind to a protein target could be studied statistically. They found that, as systems became more complex, the chance of observing a useful interaction for a randomly chosen ligand fell dramatically. Thus, there may be an optimal complexity for molecules in a screening library. Smaller libraries of less-complex molecules are likely to be more productive in terms of finding relevant chemistry space, with enhancements in potency and selectivity resulting from iterative rounds of synthesis and testing to increase complexity. Although the compounds were not derived from a library, a comparison of glutamic acid to LY354740 and MGS0028 is illustrative (Figure 8).



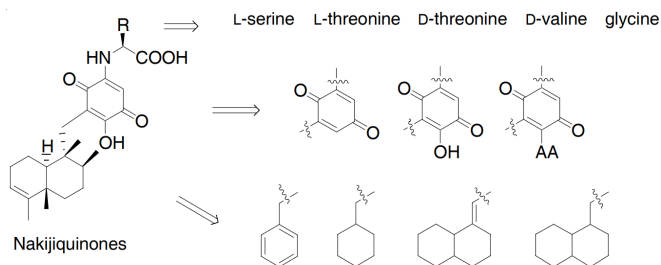
**Figure 8.** Increasing structural complexity of glutamate analogs.

Glutamic acid is a relatively simple molecule with several degrees of rotational freedom, and obviously interacts with all glutamate receptors, both ionotropic and metabotropic. LY354740 is arguably more complex with respect to stereochemistry and rigidity, is much more potent than glutamate at Group 2 mGluR's, and has no activity at iGluR's. MGS0028 is even more complex with respect to functionality and heteroatoms and, although no more selective than LY354740, it is about 20 times more potent. Most chemists would no doubt agree that molecular complexity increases from glutamic acid to LY354740 to MGS0028, but there have been few attempts to quantify molecular complexity. Bertz developed a general index of molecular complexity based on concepts of graph theory and information theory and included features such as branching, rings, multiple bonds, heteroatoms, and symmetry. In the work reported by Hann, the number of bits set in the Daylight 2D structure representation was taken as an indication of the internal bond complexity, but the method does not capture notions of stereochemistry and rigidity.



A rather different approach to natural product-based libraries is being promoted by Waldmann and coworkers. Recent results in structural biology and bioinformatics indicate that the number of distinct protein families and folds is fairly limited. Often, many proteins use the same structural domain in a more or less modified form created by divergent evolution. Protein families can have similar folds, even though they at first seem to have completely different sequences and/or catalyze quite different chemical reactions with a different arrangement of activesite residues. However, proteins in these families evolved from the same ancestors and can still bind similar ligands. If ligand types or frameworks for certain domain families are already known from the investigation of evolutionarily related proteins, the underlying structure of this ligand may be employed as the guiding principle for library development. Such ligands would provide targeted, biologically validated starting points in structural space for the development of relatively small compound libraries, which should yield significantly higher hit rates than much larger libraries designed exclusively on the basis of available and proven chemical transformations.

Accordingly, they synthesized a library of nakijiquinone analogs (Figure 9), the only natural products known to be inhibitors of the Her-2/Neu receptor tyrosine kinase, and investigated them as possible inhibitors of the receptor tyrosine kinases involved in angiogenesis. This led to the identification of inhibitors of IGF1R, Tie-2, and VEGFR-3, with IC<sub>50</sub>'s in the range of 0.5–18  $\mu$ M.



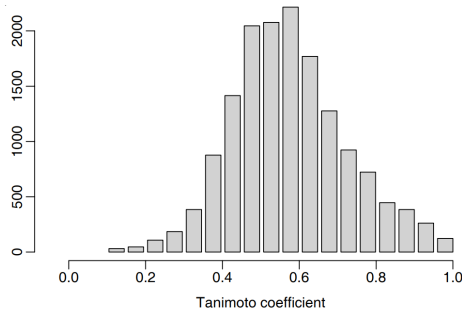
**Figure 9.** Molecular composition of the nakijiquinone library.

The growing awareness that biological activity is not uniformly distributed throughout chemistry space has led to a number of efforts to determine those molecular attributes that are drivers of that activity. At an elementary level, Ghose and coworkers carried out quantitative and qualitative characterization of known drug databases with respect to computed physicochemical property profiles, such as log P, molar refractivity, molecular weight, and number of atoms, as well as characterization based on the occurrence of functional groups and important substructures. For many parameters, they defined a qualifying range covering  $\geq 80\%$  of the compounds. They also found that the benzene ring is the most abundant substructure, slightly more abundant than all heterocyclic rings combined, and that nonaromatic heterocycles were twice as abundant as aromatic heterocycles. The most abundant functional groups were tertiary aliphatic amines, alcohols, and carboxamides.

Bemis and Murcko carried out an extensive structure-based analysis using shape description methods to analyze a database of commercially available drugs and prepare a list of common drug shapes. A useful way of organizing this structural data is to group the atoms of each drug molecule into ring, linker, framework, and side-chain atoms. On the basis of the 2D molecular structures (without regard to atom type, hybridization, or bond order), there were 1179 different frameworks among the 5120 compounds analyzed. However, the shapes of half of the drugs in the database were described by the 32 most frequently occurring frameworks. This suggests that the diversity of shapes in the set of known drugs is extremely low. Within the set of 32 frameworks, 23 contained at least two six-membered rings linked or fused together, and only three had more than five rotatable bonds. In a second method of analysis, in which atom type, hybridization, and bond order were considered, more diversity was seen: there were 2506 different frameworks among the 5120 compounds in the database, and the most frequently occurring 42 frameworks accounted for only one-fourth of the drugs. Subsequently, the same workers analyzed the side chains of the same set of drugs. On the basis of the atom pair shape descriptor (taking into account atom type, hybridization,

and bond order), there were 1246 different side chains among the 5090 compounds analyzed. The average number of side chains per molecule was 4, and the average number of heavy atoms per side chain was 2. Ignoring the carbonyl side chain, there were approximately 15 000 occurrences of side chains. Of these 15 000, approximately 11 000 were from the 'top-20' group of side chains. This suggests that the diversity that side chains provide to drug molecules is also quite low. The authors have combined this information to generate new structures that are likely to be druglike and synthetically accessible. They used this approach to generate a set of molecules optimized for blood–brain barrier penetration.

Ajay and coworkers used a Bayesian neural network to distinguish between drugs and non-drugs. They evaluated commercial databases of drug (Comprehensive Medicinal Chemistry, CMC) and nondrug (Available Chemicals Directory, ACD) molecules with respect to 1D and 2D parameters. The former contain information about the entire molecule, like molecular weight, and the latter contain information about specific functional groups. Their results correctly predicted over 90% of the compounds in the drug database while classifying about 10% of the molecules in the nondrug database as drug-like. The neighborhoods defined by their model are not similar to those generated by standard Tanimoto similarity calculations, and thus new and different information is being generated by these models, as shown in Figure 10.



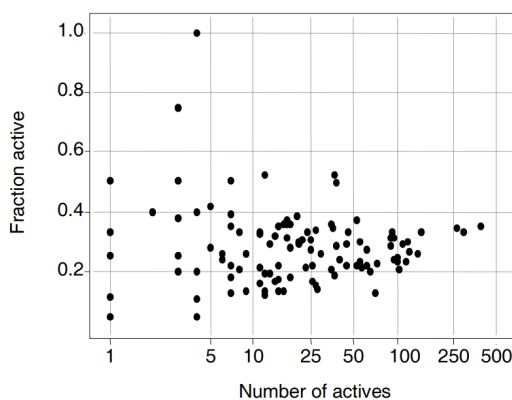
**Figure 10.** Histogram of Tanimoto coefficients based on topological tor-

sions of the most similar CMC molecule for each of the drug-like molecules from the ACD.

Further efforts have been made to distinguish between drugs and non-drugs. Sadowski and Kubinyi developed a scoring scheme for rapid and automatic classification of molecules into drugs and nondrugs. The method was set up by using atom type descriptors for encoding the molecular structures and by training a feed-forward neural network for classifying the molecules. It was parameterized and validated by using large databases of drugs (World Drug Index, WDI) and non-drugs (ACD). The method revealed features in the molecular descriptors that either qualify or disqualify a molecule for being a drug and classified 83% of the ACD and 77% of the WDI appropriately.

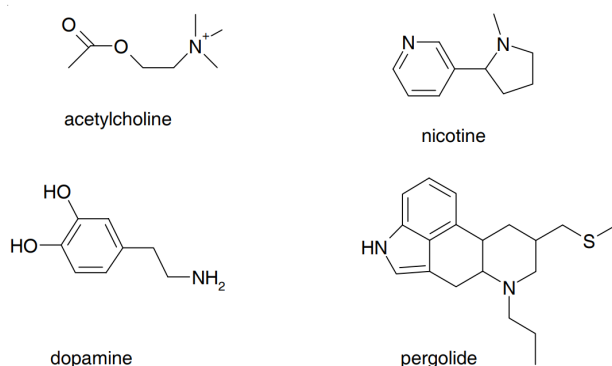
Clark and coworkers investigated techniques for distinguishing between drugs and non-drugs using a set of molecular descriptors derived from semiempirical molecular orbital (AM1) calculations. These descriptors had been used successfully to build absorption, distribution, metabolism, and excretion-related QSPR models. A principal-components analysis was carried out for the descriptors in property space. The third-most significant principal component of this set of descriptors served as a useful numerical index of drug-likeness, but no others were able to distinguish between drugs and non-drugs. The set of descriptors was extended, and ultimately three descriptors were used to train a Kohonen artificial neural net for the entire Maybridge dataset. Projecting the drug database onto the map so obtained resulted in clear distinction between drugs and non-drugs.

Figure 11 demonstrates that there is no simple relationship between druglikeness and standard 2D similarity measures of molecules. Martin and coworkers [40] addressed this question in a study using Daylight fingerprints. They showed that, for IC<sub>50</sub> values determined as a follow-up to 115 high-throughput screening assays, there is only a 30% chance that a compound that is  $\geq 0.85$  Tanimoto similar to an active is itself active.



**Figure 11.** The fraction of molecules that are similar to any active that are themselves active, as a function of the number of actives with similars.

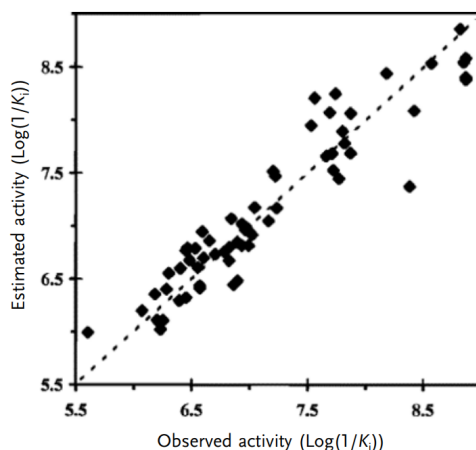
These workers also asked whether biologically similar compounds have similar chemical structures. Considering such classic example pairs as the nicotinic agonists acetylcholine and nicotine or the dopaminergic agonists dopamine and pergolide (Figure 12), the expected answer is no. In fact, the highest Tanimoto similarity within this group of four compounds is between nicotine and pergolide, and the second-highest is between nicotine and dopamine. Nevertheless, in general, the Daylight and Unity fingerprints are more similar for compounds with the same biological properties than for compounds with different biological activities. What might at first be perceived as a disappointing level of similarity-predicted actives might be the result of compounds binding in subtly different ways to the same receptor or to different but related populations of receptors.



**Figure 12.** Pairs of cholinergic and dopaminergic agonists.

Pearlman and Smith indicated that such distance-based algorithms are quite satisfactory for simple subset selection, but are considerably less useful for all other diversity-related tasks. In their view, traditional descriptors make rather poor chemistry space metrics for three reasons: many of the traditional descriptors are highly correlated, some traditional descriptors are strongly related to pharmacokinetics but only weakly related to receptor affinity, and traditional descriptors convey very little information about substructural differences that are the basis of structural diversity. They defined BCUT metrics in a manner that incorporates both connectivity information and atomic properties relevant to intermolecular interaction, i.e., atomic charge, polarizability, and H-bond donor and acceptor abilities. Given a set of active compounds that all bind to a given receptor in the same way, it is certainly reasonable to expect that these active compounds should be positioned near each other in a small region of chemistry space if the chemistry space metrics are valid. They developed the Activity-Seeded Structure-Based clustering algorithm, which provides a method for directly testing that expectation in the typical case in which the chemistry space dimensionality is greater than three and, thus, simple visual inspection of the distribution of active compounds is difficult or impossible. Given a number of compounds for which a particular receptor has significant affinity, they can then identify the receptor-relevant subspace for that receptor by identifying the axes along which compounds are

tightly clustered. The algorithm also accounts for the possibility of multiple receptor binding modes by allowing more than one cluster of actives per relevant axis. In addition to their own application to ACE inhibitors as an illustration of the method, Stanton independently applied this method to a QSAR study of dihydrofolate reductase inhibitors. The resulting model was highly predictive, as shown in Figure 13. It is apparent that the BCUT metrics are measuring particular structural features that can be related to the observed properties of a variety of molecules. They appear to perform quite well in capturing structural information important for understanding polar intermolecular interactions.



**Figure 13.** Comparison of estimated and observed DHFR inhibitor activity values using a BCUT-based model.

BCUT metrics are being used increasingly in QSAR studies and library design. A particularly interesting study was done by Pirard and Pickett, who presented studies with BCUTs for the classification of ATP site-directed kinase inhibitors active against five different protein kinases, three from the serine/threonine family and two from the tyrosine kinase family. In combination with a chemometric method, the BCUTs were able to correctly classify the ligands according to their target. The authors concluded that BCUTs are indeed a useful set of descriptors for design tasks, extracting information in a manner relevant to describing ligand–

receptor interactions. They are particularly suited to the design of targeted libraries and virtual screening of compound collections, as they are quick to calculate while containing more information than a standard 2D fingerprint type descriptor.

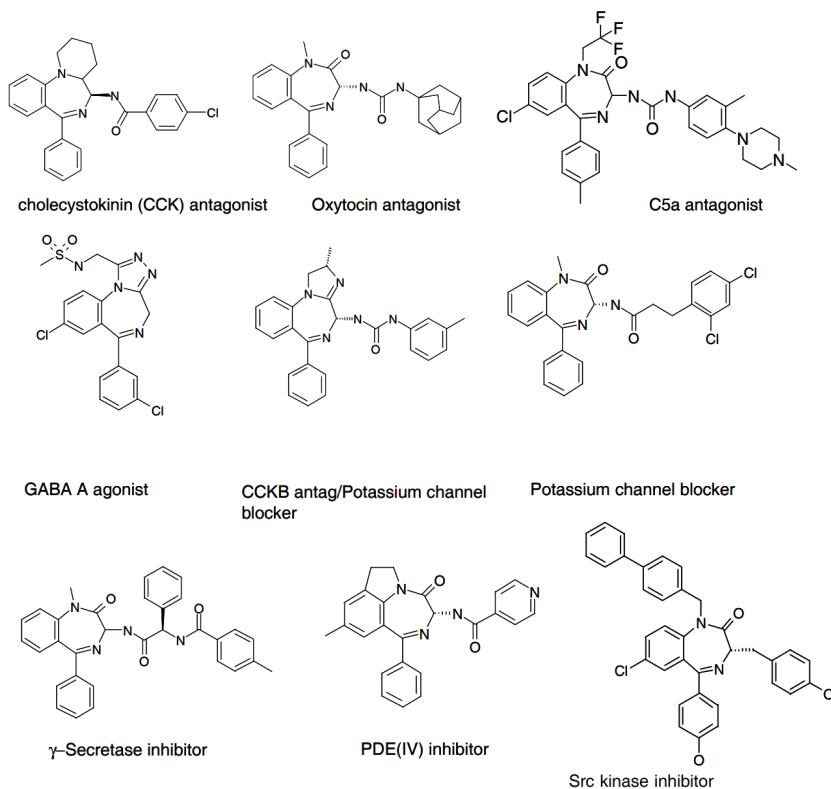
### 3.7 IS BIOLOGICAL SELECTIVITY AN ILLUSION?

We have illustrated the enormity of chemistry space and the focus on biologically relevant chemistry space, but what about biology space itself? How many biologically relevant targets are there? Although this number has been estimated to be around 3000, it may well be much larger than this if we extrapolate from what we know about particular target classes, e.g., GPCRs, where there are many potential druggable targets and many potential pharmacologies, from agonists to antagonists to modulators to inverse agonists. In a typical drug discovery program, selectivity of potential development candidates is often assessed against a panel of 50–100 biologies. Clearly, this does not cover a very large fraction of available biology space. In fact, many compounds originally thought to be very selective were later found to have effects against many other targets. For example, cholesterol-lowering HMGCoA reductase Inhibitors (statins) are among the world's top-selling drugs. It was recognized recently that statins possess additional biology. e.g., anti-inflammatory activity that is not explained by their interaction with this enzyme. High-throughput screening of large chemical libraries has identified lovastatin (a statin) as an extracellular inhibitor of LFA-1. Lovastatin was shown to decrease LFA-1-mediated leukocyte adhesion to ICAM-1 and T-cell co-stimulation. Unexpectedly, lovastatin was found to bind to a hitherto unknown site in the LFA-1 I (inserted) domain, as documented by nuclear magnetic resonance spectroscopy and crystallography.

Some structural classes, e.g., benzodiazepines, are well known to exhibit diverse biology depending on the precise substituent pattern and conformation. Selective ligands with common cores have been obtained against many protein targets (Figure 14).



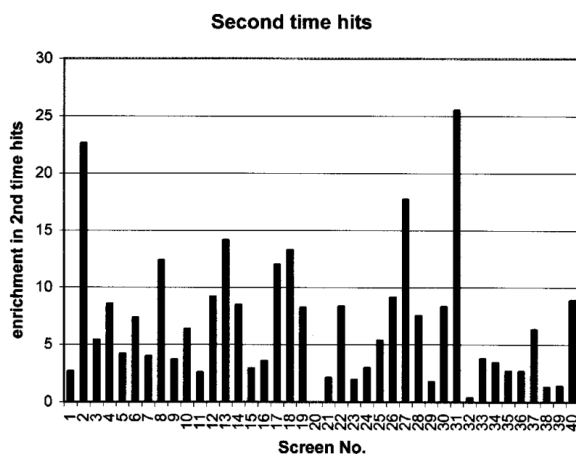
The existence of such privileged structures suggests that some common structural binding motifs on proteins are reused across many different protein families. It is widely accepted that few if any of the known biologically active molecules are exclusively selective for a single biological target. This forms the basis for the discovery of new uses for existing drugs and the explanation of side effects observed for all drugs. Indeed, in a commentary on the molecular basis for the binding promiscuity of antagonist drugs, LaBella stated that it is unlikely that binding-site dimensions, geometry, charge environments, hydrophobic surfaces, and other features will ever be known to the extent that drug design technology will yield a compound with absolute specificity for one species of functional protein. On a molecular level this may well be a consequence of there being a relatively small number of protein families and folding motifs. These considerations are being applied in interesting ways to quickly find new biologically active compounds. For example, Kauvar and Dixon have developed a method called affinity fingerprinting, for predicting ligand binding to proteins. In this method, the binding potency of a small molecule is measured against a panel of reference proteins, in which the panel members have been empirically selected to provide binding sites that are well diversified with regard to interactions with small molecules. The resulting set of pIC<sub>50</sub>'s constitutes the molecule's molecular fingerprint. Libraries of compounds can be evaluated and the collection of corresponding fingerprints entered into a database. From this large set, a subset is then chosen to represent the diversity of the set. The subset is then screened against a new target protein. Those compounds with the best pIC<sub>50</sub>'s against the new protein are used to query the database to find other compounds with the same or similar fingerprint. Repetition of the cycle quickly finds the bestbinding compounds in the collection. These can then serve as seeds for combinatorial expansion, presumably accelerating the lead discovery process.



**Figure 14.** The classic privileged structure – the benzodiazepine nucleus with small structural modifications – is capable of many different biologicals.

We have used a related strategy to analyze the performance of our corporate collection in high-throughput screening over the past several years. Our panel of proteins consists of drug targets of interest and spans several target classes, including GPCRs, several classes of enzymes, ion channels, etc. Our thesis is that a compound that exhibits biological activity in any target class is more likely to exhibit activity in another unrelated class than is a compound that has never exhibited biological activity of any kind. We initially used a relatively small set of assays and screened compounds and identified about 3500 compounds that were biologically active in at least one assay and met our internal criteria with respect to molecular weight, cLogP, polar surface area, and other chemistry-

based filters. About 10% of these compounds were found to exhibit activity in other assays. The number of active compounds was then expanded about 10 000, and the number of assays to 40. The hit rate of the general corporate collection was normalized to a frequency of 1 and compared to the hit rate of the 10 000 known biologically active set. The results are shown in Figure 15.



**Figure 15.** Observed hit rates for a biology-based library on a scale in which the hit rate of the general collection was normalized to 1.

Clearly, the hit rate exceeds that of the general collection in the majority of screens. However, recent publications have sounded a cautionary note. Roche and coworkers reported the development of a virtual screening method for the identification of 'frequent hitters'. These compounds appear as hits in many different biological assays covering a wide range of targets for two main reasons: (1) the activity of the compound is not specific for the target; and (2) the compound perturbs the assay or the detection method. They found that, with an increasing drug-likeness of the database, a decreasing fraction of frequent hitters is predicted. Sheridan reported finding multiactivity substructures by mining databases of drug-like compounds. Shoichet and coworkers described a common mechanism underlying this phenomenon. In their study they observed that several nonspecific inhibitors formed aggregates 30–400 nm in diameter and that these aggregates

were likely responsible for the inhibition. With these two reports in mind, we returned to our corporate database and identified, again after suitable filtering, a set of 72 000 biologically active compounds. We then selected a subset of about 25 000 compounds based on the following criteria: (1) compounds with confirmed activity in at least two assays, (2) compounds with confirmed activity in no more than five assays, (3) compounds tested in at least ten assays. We felt that this simple approach would give us a set of information-rich compounds largely free of frequent hitters. Using Daylight 2D fingerprints and a Tanimoto distance of 0.3, the set consists of 9200 clusters, of which there are almost 5100 singletons. We propose that this richly diverse subset is an ideal starting platform for the design of screening libraries and for the discovery of new privileged structures. Interestingly, with respect to physical properties, the subset is slightly more lipophilic and has slightly larger polar surface area than the general collection, but the distribution of molecular weights and the numbers of hydrogen-bond donors and acceptors is the same. We conclude that the currently accepted drug-like physical properties boundary conditions are necessary but not sufficient to define biological activity and that other, poorly understood, factors are the true drivers of such activity. We continue to explore just what those factors might be.

### **3.8 SYNTHESIS OF CHEMICAL GENETICS LIBRARIES: NEW ORGANIC SYNTHESIS APPROACHES TO THE DISCOVERY OF BIOLOGICAL ACTIVITY**

The recognition that the intersection of biology space is limited within chemistry space has encouraged the development of new strategies in organic synthesis for the discovery of biological activity. For example, Ellman and coworkers have developed combinatorial target-guided ligand assembly. In this method, a set of potential binding elements is prepared in which each molecule incorporates a common chemical linkage group. The set of potential binding elements is screened to identify all binding

elements that interact even weakly with the biological target. A combinatorial library of linked binding elements is prepared in which the binding elements are connected through a set of flexible linkers. The library is then screened to identify the tightest-binding ligands. Using this approach they identified a potent ( $IC_{50} = 64$  nM) inhibitor of the non-receptor tyrosine kinase c-Src. An extension of this strategy has been developed by Lehn and others. So-called dynamic combinatorial chemistry uses self-assembly processes to generate libraries. In contrast to the stepwise assembly of molecules in the library, this method allows for the generation of libraries based on continuous inter-conversion among the library constituents. Addition of the target ligand or receptor creates a driving force that favors formation of the best-binding constituent. Sharpless and coworkers have investigated a slightly different approach. Rather than using a set of interconverting constituents, they allow the target to select building blocks and synthesize its own inhibitor. Dubbed 'click chemistry,' it depends on the simultaneous binding of two ligands, decorated with complementary reactive groups, to adjacent sites on the protein. Their colocalization is then likely to accelerate the reaction that connects them. The reaction of course must be selected so as to not take place in undesired ways within biochemical systems. One such reaction is the cycloaddition of azides to acetylenes to yield 1,2,3-triazoles. As a proof of principle, AChE was used to select and synthesize a triazole-linked bivalent inhibitor by using known site-specific ligands as building blocks. This resulted in the discovery of an inhibitor with a  $K_d$  in the range of 77–410 fM (femtomolar), depending on the species. This is the most potent noncovalent AChE inhibitor known to date, by approximately two orders of magnitude.

The standard approach to parallel synthesis of libraries is to start with a polyfunctional common core and elaborate those functions with diversity elements. With just a few diversity locations and the large number of commercially available diversity reactants, this can result in libraries consisting of tens or hundreds of thousands, or even more, members. Nevertheless, such libraries retain the common core for all members, which necessarily limits the total

diversity of the library. Far more challenging, and arguably more valuable to the efficient exploration of chemistry space, would be the synthesis of libraries whose members are based on disparate cores. Schreiber is addressing the problem of skeletal diversity by using a synthesis strategy that involves transforming substrates with different appendages that pre-encode skeletal information into products that have different skeletons, with the use of common reaction conditions.

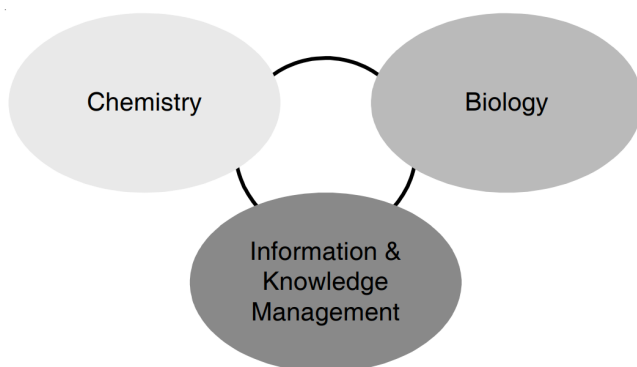
Our own interest in this problem was the result of our work on the biology-based collections discussed above. We found that roughly only half the compounds were available as solid samples for further study, and the remainder were dropped from consideration for that reason. The efficient resynthesis of hundreds or thousands of disparate compounds was simply not practical. Or was it? Perhaps there was an easy way to sort multiple syntheses into common starting materials and reactions and to carry them out in parallel. To that end, we used LeadScope software as our management tool. Normally, LeadScope links chemical and biological data, allowing chemists to explore large sets of compounds by a systematic substructural analysis using a predefined set of 27 000 structural features. More importantly for our purposes, two sets can be compared with respect to these features. We chose the ACD database as our second set. We could then easily select those starting materials that would give rise to many products via different routes. We then ran as many reactions as possible using parallel synthesis methods. We have used this method for syntheses of up to four steps and have been able to maintain a productivity level of one compound per chemist per day, 25 mg scale, purified  $\geq 85\%$ , and characterized by LC/MS and NMR.

We are developing an approach to true simultaneous synthesis of disparate core compounds. Most molecules of the size and complexity we are interested in would likely be prepared in no more than five steps. The actual transformations are usually limited to the chemistry background and experience of the chemist(s) involved in the project. However, the routes need not be so limited. Indeed, consider the generation of tens or hundreds

of routes to each compound of interest. The problem then becomes one of how to prepare the maximum number of compounds using the minimum set of common chemistries, staging the routes as necessary so as to maximize the overlap of reagents and conditions. The generation of syntheses is software based. Two or three decades ago there was a lot of effort to develop software to predict the most efficient syntheses of complex organic molecules; most have been abandoned. We chose to use the SynGen program for the very reason that it usually produces several routes to a molecule, each of which begins with a commercially available starting material and whose transformations usually have a literature precedent. Common chemistries can be grouped at three levels: (1) reaction type, e.g., acylation of amines; (2) reagent type, e.g., acylation of secondary amines; and (3) specific reagents, e.g., acylation of diethyl amine. Each level is specifically encoded by the program, making searching, sorting, and matching fairly easy. We will not necessarily choose the shortest route to each molecule, since it is entirely possible that some longer routes would give rise to additional commonalities, thereby allowing the preparation of a larger total number of compounds. We are in the process of testing this concept using a set of 100 very different structures and will report the results in due course.

### 3.9 INFORMATION AND KNOWLEDGE MANAGEMENT ISSUES

The integration of chemistry and biology that constitutes the engine for chemical genetics presents a major challenge for existing models of information and knowledge management. The management of information and knowledge is so critical as to deserve a place as one of the three critical components necessary to truly enable chemical genetics (Figure 16). Linking chemical structures with biology in a systematic way has challenged pharmaceutical companies and software vendors for many years, and several proprietary and off-the-shelf solutions now exist. Typically, these products are not scaleable or flexible enough to deal with the problems exposed by chemical genetics.



**Figure 16.** Chemical genetics requires the integration of the three critical elements of chemistry, biology, and information/knowledge management.

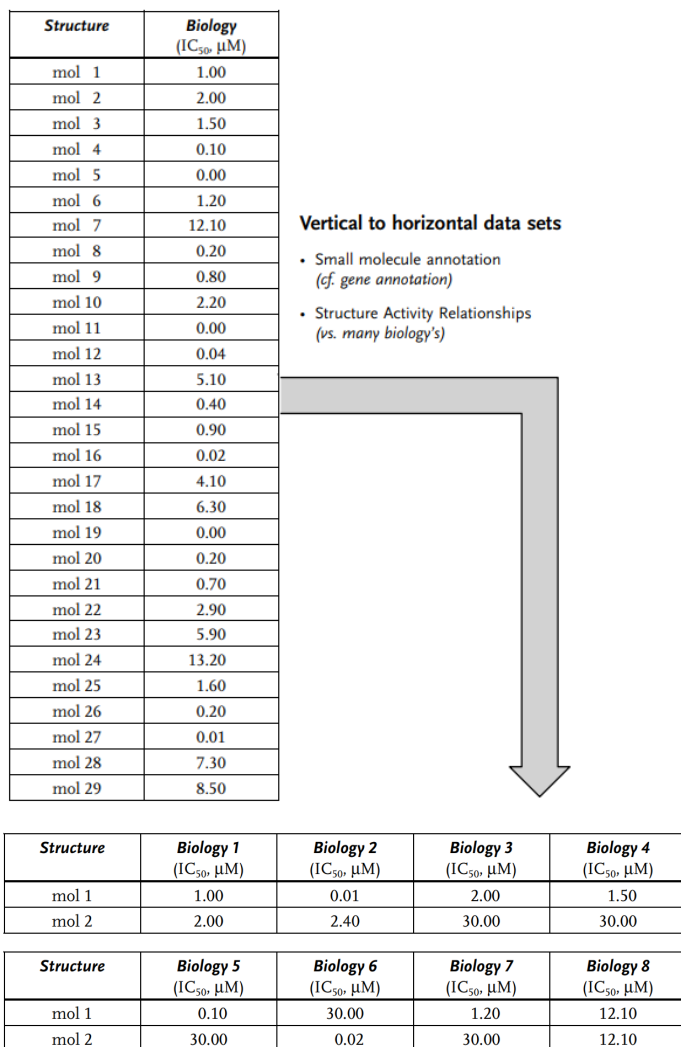
### 3.10 ANNOTATION OF SMALL MOLECULES

Several groups have realized the information management challenges posed by chemical genetics. The US National Cancer Institute is developing a powerful openaccess database called ChemBank that will link small-molecule structure and associated effects on proteins, cell pathways, and tissue formation. Additionally the effect of small molecules on an organism's phenotype will also be captured. ChemBank is a chemical genetics database, which has been described as a chemical version of GenBank, the online repository of genetic data. The NCI plans to synthesize and screen thousands of molecules for their biological activity. Annotation of small molecules should allow for much closer integration of chemical structure and biological activity. Use of such annotated compounds (sometimes referred to as information-rich compounds) as chemical tools for probing biological systems promises to be a fruitful area of future research.

The central informatics issue in chemical genetics is annotation of chemical structures in the same way as annotation of genes, i.e., annotation of the biology and other properties of a chemical structure. In a typical single-drug discovery project, it is common



for many structures to be profiled by a single biological screen generating a simple vertical data format (Figure 17). In chemical genetics we focus on single compounds annotated with many biologies – a horizontal data format (Figure 17).



**Figure 17.** Chemical genetics databases require the annotation of individual compounds with many biologies, in contrast to the more traditional way of capturing the assay results of many compounds against a single biology

NCI is asking scientists from all over the world to deposit information on the effects of small molecules on cells on the micro (gene expression) and macro levels in ChemBank. One of the hopes here is to link phenotypic changes with structures and to use this information in predicting the mechanism of action of drugs.

## REFERENCES

1. A. L. Hopkins, C. R. Groom, *Nature Rev. Drug Discov.* 2002, 1, 727–736.
2. A. Nakazato, T. Kumagai, K. Sakagami, R. Yoshikawa, Y. Suzuki, S. Chaki, H. Ito, T. Taguchi, S. Nakanishi, S. Okuyama, *J. Med. Chem.* 2000, 43, 4893–4909.
3. B. P. Zambrowicz, A. T. Sands, *Nature Reviews* 2003, 2, 38–51.
4. B. Pirard, S. D. Pickett, *J. Chem. Inf. Comput. Sci.* 2000, 40, 1431–1440.
5. D. F. Doyle, D. A. Braasch, L. K. Jackson, H. E. Weiss, M. F. Boehm, D. J. Mangelsdorf, D. R. Corey, *J. Am Chem. Soc.* 2001, 123, 11367–11371.
6. D. J. Maly, I. C. Choong, J. A. Ellman, *Proc. Nat. Acad. Sci. USA* 2000, 97, 2419–2424.
7. E. Shorter, *Nature Rev. Drug Discov.* 2002, 1, 1003–1006.
8. F. G. Kuruvilla, A. F. Shamji, S. M. Sternson, P. J. Hergenrother, S. L. Schreiber, *Nature* 2002, 416, 653–656.
9. G. R. Rosania, Y.-T. Chang, D. Sutherlin, H. Dong, D. J. Lockhart, P. G. Schultz, *Nature Biotechnol.* 2000, 18, 304–308.
10. G. Roberts, G. J. Myatt, W. P. Johnson, K. P. Cross, P. E. Blower, *J. Chem. Inf. Comput. Sci.* 2000, 40, 1302–1314.
11. G. Weitz-Schmidt, *Trends Pharmacol. Sci.* 2002, 23, 482–486.
12. H. J. Kwon, *Curr. Med. Chem.* 2003, 10, 717–726.
13. L. Kissau, P. Stahl, R. Mazitschek, A. Giannis, H. Waldmann, *J. Med. Chem.* 2003, 46, 2917–2931.
14. L. Olbe, E. Carlsson, P. Lindberg, *Nature Rev. Drug Discov.* 2003, 2, 132–139.
15. M. Brüstle, B. Beck, T. Schindler, W. King, T. Mitchell, T. Clark, *J. Med. Chem.* 2002, 45, 3345–3355.
16. M. D. Burke, E. M. Berger, S. L. Schreiber, *Science* 2003, 302, 613–618.
17. M. M. Hann, A. R. Leach, G. Harper, *J. Chem. Inf. Comput. Sci.* 2001, 41, 856–864.

18. O. Ramström, J. Lehn, *Nature Rev. Drug Discov.* 2002, 1, 26–36.
19. O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmerman, G. Schneider, *J. Med. Chem.* 2002, 45, 137–142.
20. R. Breinbauer, I. R. Vetter, H. Waldmann, *Angew. Chem. Int. Ed.* 2002, 41, 2879–2890.
21. R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* 2003, 43, 1037–1050.
22. S. L. McGovern, E. Caselli, N. Grigorieff, B. K. Shoichet, *J. Med. Chem.* 2002, 45, 1712–1722.
23. T. Gura, *Nature* 2000, 407, 282–284.
24. U. Abel, C. Koch, M. Speitling, F. G. Hansske, *Curr. Opin. Chem. Biol.* 2002, 6, 453–458.
25. W. G. Lewis, L. G. Green, F. Grynszpan, Z. RadiÆ, P. R. Carlier, P. Taylor, M. G. Finn, K. B. Sharpless, *Angew. Chem. Int. Ed.* 2002, 41, 1053–1057.
26. Y. C. Martin, J. L. Kofron, L. M. Traphagen, *J. Med. Chem.* 2002, 45, 4350–4358.



## CHAPTER 4

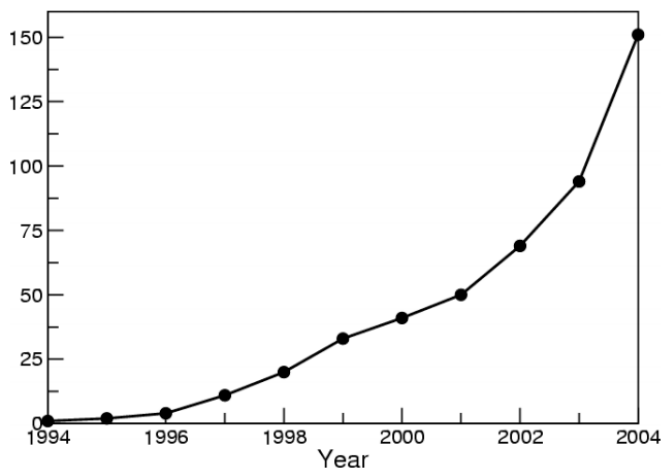
# STRUCTURAL INFORMATICS: CHEMOGENOMICS

## INTRODUCTION

The goal of chemogenomics is a systematic understanding of how various chemical compounds modulate the function or activity of each and every gene product (protein) in the human body. Before an approach qualifies as chemogenomic, it must provide knowledge about multiple targets or pathways in a holistic manner that represents a departure from drug discovery's historic target-by-target approach. Hence, technologies such as cell-based screening and expression profiling are commonly labeled chemogenomic by virtue of their ability to furnish data that spans multiple targets and pathways. Naturally, as these chemogenomic technologies begin to yield new types of data, new chemogenomic informatics approaches must be developed to convert this data into knowledge relevant to drug discovery. The goal of human structural genomics is a systematic determination of all of the protein structures in the

human body. Current experimental efforts have been instrumental in establishing high-throughput structural genomics platforms that employ automated protein expression, crystallization, data acquisition, and model refinement technologies.

The number of protein-ligand co-crystal structures available for drug targets has also increased substantially over the last decade (Fig. 1). Historically, structure-based drug design (SBDD) approaches have utilized cocrystal information to rationally optimize the activity and ADME properties of lead compounds. Since the first successful structure-based drugs were developed to target HIV protease and influenza neuraminidase in the early nineties, inhibitors for more than 40 distinct targets have been developed using SBDD approaches.



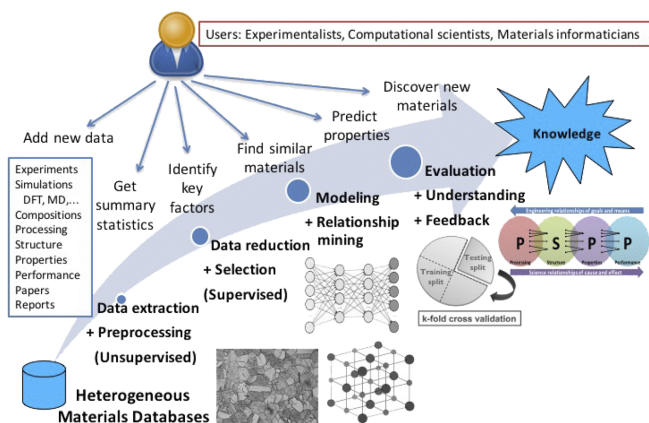
**Figure 1:** High throughput technologies have rapidly expanded the number of targets that have been expressed and successfully co-crystallized.

As the amount of protein and protein-ligand complex structural data increases, structural coverage across many important gene families is becoming much more complete. That is, rather than having the structure for just one target within a gene family, structures are becoming available for many or all of the targets within a gene family. This increase in structural coverage offers the possibility of replacing the historic target-by-target utilization of

structural data with a holistic, chemogenomic approach. Structural informatics is an important branch of chemogenomic informatics whose goal is to utilize the rapidly expanding structural database in new ways to enhance the discovery and optimization of small molecule protein modulators on a genomic scale.

## 4.1 STRUCTURAL INFORMATICS

Informatics is defined as, “the collection, classification, storage, retrieval, and dissemination of recorded knowledge.” In bioinformatics, gene and protein sequences are classified according to their similarity to infer function for genes and proteins whose functions have not been verified experimentally. For example, many of the proteins that we refer to as kinases have never been assayed for kinase activity, we infer that they are kinases since their sequences are similar to verified kinases.



This well-established process of inference via similarity is not without error; every once in a while the bioinformatics-based inference of protein function will be incorrect. Much more frequently, the cheminformatics-based inference of small molecule activity is in error, since slight changes in a molecule can dramatically affect its ability to bind. Hence, in cheminformatics, inference of function from similarity classification is less reliable than in bioinformatics. Because of this lack of reliability, inference in cheminformatics is thought of as an imperfect screening

process, whose less than ideal performance is analyzed in terms of an enrichment factor (a measure of how much better the cheminformatic inference performs than random inference).

Structural informatics utilizes the same process of inference through classification as its well established informatics cousins. In structural informatics, the data being compared and classified are protein structures, binding sites, and ligand binding modes. Correspondingly, the types of algorithms used for the purposes of classification are structure alignments, site alignments, and binding mode alignments.

| <i>Field of Informatics</i> | <i>Primary Data</i>            | <i>Similarity Relationships</i> | <i>Key Applications</i>   |
|-----------------------------|--------------------------------|---------------------------------|---|
| <b>Bio-</b>                 | Protein Sequences              | Sequence Alignments             | Function Inference  |
|                             | <b>Structure Determination</b> |                                 |   |
| <b>Structural</b>           | Protein Structures             | Structure Alignments            | Function Inference  |
|                             | <b>Site Annotation</b>         |                                 |   |
|                             | Binding Sites                  | Site Alignments                 | Target Hopping, Cross Reactivity, Selectivity, Opportunity Mining |
|                             | Sites+Ligands                  | Binding Mode Alignments         | Enhanced Screening, Scaffold Hopping, Novel Scaffolds             |
|                             | <b>Small Molecule Docking</b>  |                                 |   |
| <b>Chem-</b>                | Small Molecules                | Molecular Similarities          | Screening   |

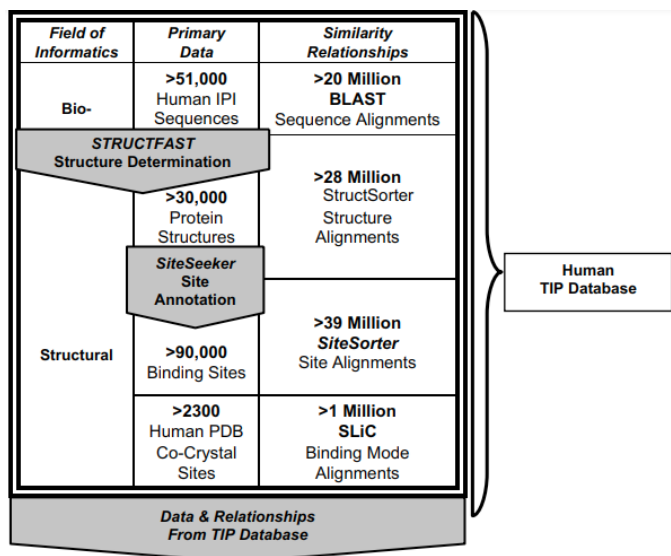
**Figure 2:** The relationship between bio-, structural, and cheminformatics.

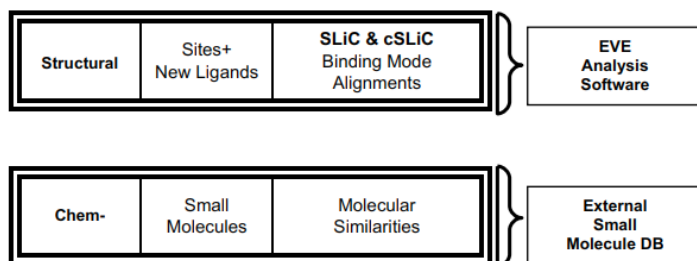
#### 4.1.1 Calculating the Structural Informatics Universe

While the number of compounds that can potentially modulate the activity of human proteins is infinite, the number of human



proteins is finite. Hence, it is theoretically feasible to construct bioinformatics and structural informatics databases that contain the sequences, structures, and binding sites for all human proteins. Since experimental structure determination plays a key role in expanding the amount of reliable structural and ligand binding mode data, the foundation of such a database is robust, updatable knowledge management of experimental protein structure information. At Eidogen-Sertanty, we have developed a structural informatics database that utilizes predictive algorithms to amplify the existing experimental protein structure and binding site data and to classify the resulting structures, sites, and ligand binding modes according to their respective similarity relationships. We call this database the Target Informatics Platform (TIP). Figure 3 shows the data and algorithms in the TIP database, and how data can be extracted from the database to interface with selected compounds from the infinite space of potential molecules. Figure 4 shows a snapshot of the protein structure and binding site data in TIP for those targets recently assigned to the human druggable genome.



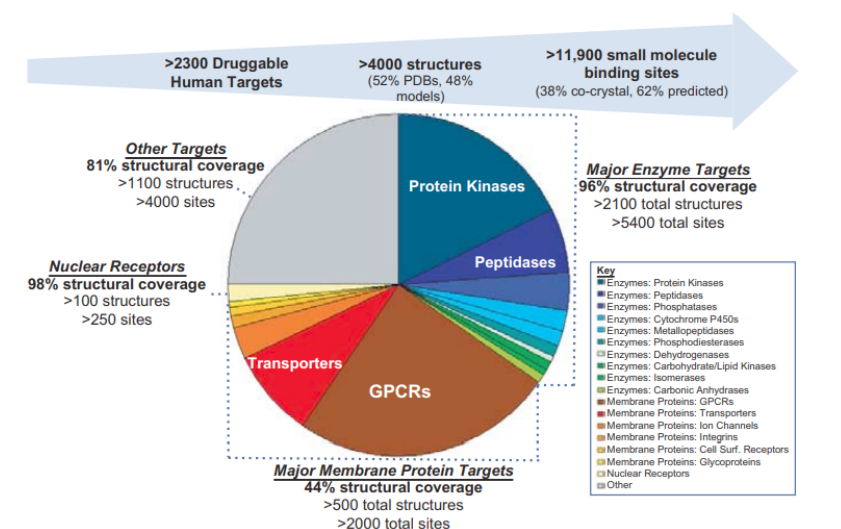


**Figure 3:** The data and algorithms in Eidogen-Sertanty's Target Informatics Platform human database.

### 4.1.2 Structural Relationships

Due to divergent or convergent evolution, structural homology can be conserved between proteins with undetectable sequence homology. In such instances, protein structure alignment algorithms, such as DALI and CE, can be used to find structural similarities and potential functional relationships that cannot be found using sequence alignment approaches. The well known structural classification databases SCOP, CATH, and FSSP store the results of structure alignments for protein structures from the PDB, and the Gene3D database goes a step further by providing the CATH structural classification for gene and protein sequences from completed genomes. The TIP database goes an additional step beyond Gene3D, not only providing the structural classification for all protein structures in a genome, but also the explicit structural alignments between each of the structures.

While protein structure alignment is certainly an important tool for functional genomics, the knowledge gained from structural classification is of limited value for chemogenomics applications. Inferring whether a compound is likely to bind to a target protein requires an understanding of the relationships at the level of the binding site.



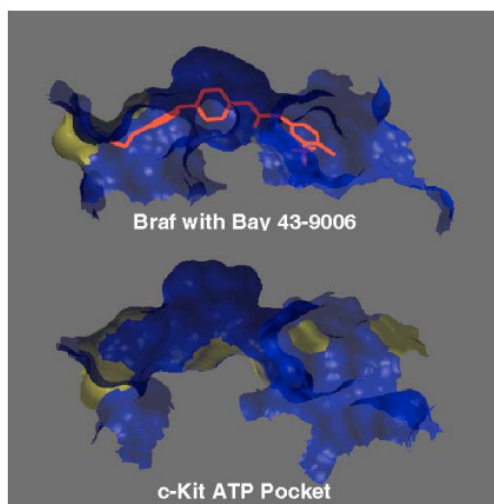
**Figure 4:** TIP's structure and binding site coverage for the major drug target families that comprise the druggable human genome.

### 4.1.3 Binding Site Relationships

While there are many resources available for obtaining protein structure relationships, there are comparatively few resources available for understanding binding site relationships. Sali and co-workers developed the first resource of this kind, LigBase. The LigBase database was created by coupling site annotations from the co-crystal record in the PDB along with the CE structure alignment algorithm, yielding multiple alignments for known binding sites. A distinguishing feature of CavBase is that it contains additional similarities between binding sites from proteins that do not share any structural homology, since the binding sites are directly aligned using a clique detection algorithm, not a structure alignment algorithm. At Eidogen-Sertanty, we have developed a site alignment algorithm, SiteSorter, which uses a weighted-clique detection approach to directly overlay binding sites and avoid the requirement for structure homology. By integrating SiteSorter with fully automated homology modeling (STRUCTFAST) and site annotation (SiteSeeker), TIP goes an additional step beyond

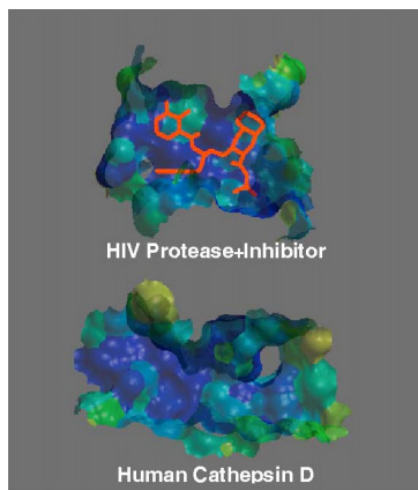
LigBase and CavBase, providing intra- and inter family binding site for the entire proteome, not just for those proteins whose structures have been resolved experimentally.

Since closely related binding sites are more likely to bind to the same small molecules, binding site similarity analysis allows us to infer important cross-reactivity information. During lead discovery for a new target, finding a cross-reactivity to a target for which there are already leads enables the fast discovery of new leads via target-hopping. With the potential of short circuiting the lead discovery process on a genomic scale, target hopping is an important chemogenomic application of structural informatics. Figure 5 shows an example of intra-family target hopping, while Fig. 6 shows an example of inter-family target hopping.



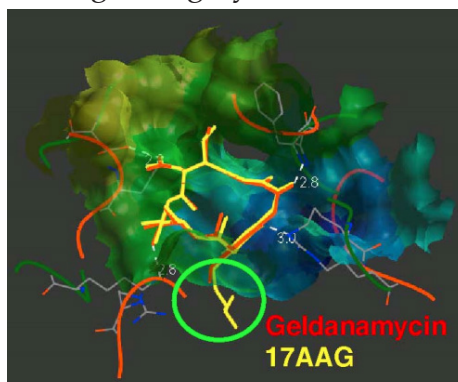
**Figure 5:** An example of intra-family target hopping within kinases.

While the potential for target hopping exists when two binding pockets are highly similar, a second set of applications emerges from a detailed understanding of the differences between two similar binding pockets. During lead optimization, where the goal is a highly selective binder, understanding the detailed mechanism of cross-reactivity between targets is critical for modifying existing leads to enhance their affinity for the desired target.



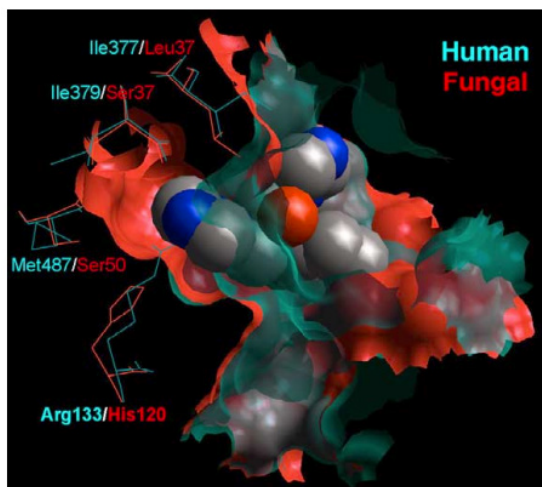
**Figure 6:** An example of inter-family target hopping between human and viral aspartyl proteases.

Figure 7 shows an example of an undesirable inter-family crossreactivity found in the TIP database, and proposes a mechanism for an optimized lead series to avoid the undesirable off-target. In addition to enabling the optimization of known leads, structural informatics offers the possibility of mining the proteome for interesting drug discovery opportunities that are likely to succeed because binding site similarity analysis reveals an opportunity to design a highly selective binder.



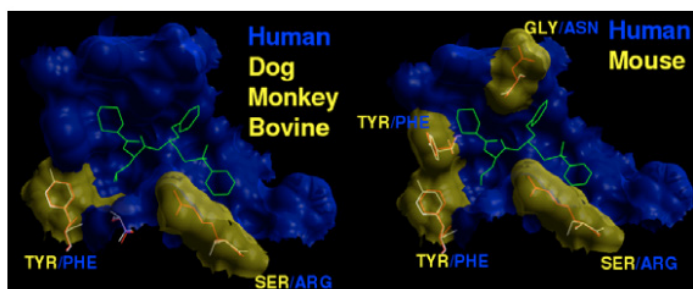
**Figure 7:** Binding site similarity analysis can reveal unwanted off-target cross-reactivities.

Figure 8 shows an example of opportunity mining in the area of anti-infectives. Once one or more projects have been mined, structural informatics can also be used to prioritize the projects by their expected feasibility.



**Figure 8:** Structural informatics can be used to mine anti-infective opportunities that cannot be discovered by comparative genomics.

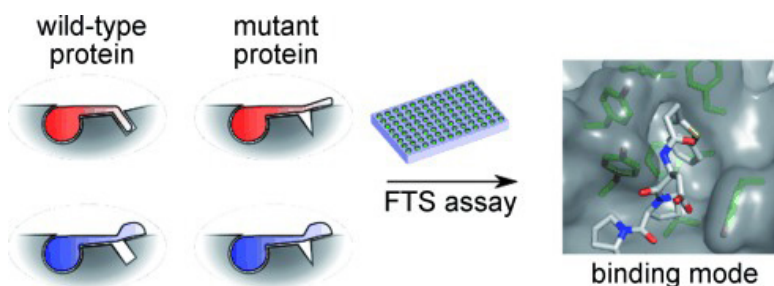
Figure 9 shows an example of a project whose feasibility has been adversely affected because the target's binding site is very different in mice, the animal model of choice.



**Figure 9:** Binding site analysis of different species can uncover potential problems with animal models for a given target.

### 4.1.4 Ligand Binding Mode Relationships

While it has long been a common practice in structure-based drug design to examine the binding modes of co-crystallized ligands to gain insight into the important principles for binding, methods for the fully automated analysis of ligand binding modes. These methods play a crucial role in structural informatics by enabling similarity based classification of the rapidly expanding database of co-crystal structural data. In the TIP database, a binding mode similarity score is determined for each of the co-crystal binding site overlays using an approach called SLiC (site-ligand contacts), which is similar to the SiFt (structural interaction fingerprint) methodology developed by Singh and co-workers at Biogen. In the SiFt and SLiC approaches, the types of contacts that a ligand makes with each of the residues of the binding pocket are coded into a bit string.



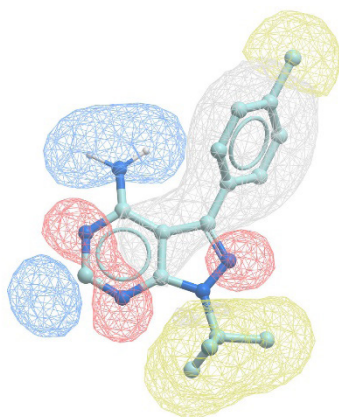
By converting the interactions important for binding into one dimensional bit strings, the SiFT and SLiC approaches can be coupled with small molecule docking approaches to find new molecules that are capable of making the same interactions. In this manner, automated binding mode analysis can be used to significantly enhance docking based approaches for inferring small molecule activity.

## 4.2 TOOLS FOR LIGAND BASED DRUG DESIGN

Ligand-based drug design or relies on knowledge of other molecules that bind to the biological target of interest. These other molecules may be used to derive a pharmacophore model



that defines the minimum necessary structural characteristics a molecule must possess in order to bind to the target. In other words, a model of the biological target may be built based on the knowledge of what binds to it, and this model in turn may be used to design new molecular entities that interact with the target. Alternatively, a quantitative structure-activity relationship (QSAR), in which a correlation between calculated properties of molecules and their experimentally determined biological activity, may be derived. These QSAR relationships in turn may be used to predict the activity of new analogs. Ligand based drug design uses ligands of the drug target— that is, molecules that bind to the drug target.



Design focuses on the structure of the ligands, for example, by the use of pharmacophore models or by QSAR models. The former model seeks to determine what ligand structures are necessary for target binding. QSAR models, on the other hand, suggest that molecular similarity, through combination molecular descriptors, predicts biological activity of the drug.

#### 4.2.1 Quantitative Structure-activity Relationship (QSAR)

Quantitative structure-activity relationship models are regression or classification models used in the chemical and biological sciences and engineering. Like other regression models, QSAR regression models relate a set of “predictor” variables to the potency of the

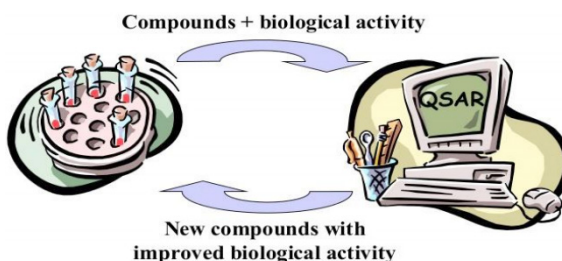


response variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable.

The number of compounds required for synthesis in order to place 10 different groups in 4 positions of benzene ring is  $10^4$ .

Solution: synthesize a small number of compounds and from their data derive rules to predict the biological activity of other compounds.

### QSAR and Drug Design



### VEGA Platform

Using the VEGA platform, you can access a series of QSAR models for regulatory purposes, or develop your own model. QSAR models can be used to predict the property of a chemical compound, using information obtained from its structure.

### DEMETRA

This project aim has been to develop predictive models and software which give a quantitative prediction of the toxicity of a molecule, in particular molecules of pesticides, candidate pesticides, and their derivatives. The input is the chemical structure of the compound, and the software algorithms use "Quantitative Structure-Activity Relationships" (QSARs). The DEMETRA software tool can be used for toxicity prediction of molecules of pesticides and related compounds. The DEMETRA models are freely available. Five models have been developed to predict toxicity against trout,

daphnia, quail and bee. The software is based on the integration of the knowledge acquired in the DEMETRA EU project in a homogeneous manner using the best algorithms obtained as the basis for hybrid combinative models to be used for predictive purposes.

### ***T.E.S.T***

Toxicity Estimation Software Tool (T.E.S.T.) will enable users to easily estimate acute toxicity using the above QSAR methodologies.

### ***OCHEM***

The OCHEM is an online database of experimental measurements integrated with the modeling environment. Submit your experimental data or use the data uploaded by other users to build predictive QSAR models for physical-chemical or biological properties.

### ***E-DRAGON***

E-DRAGON is the electronic remote version of the well-known software DRAGON, which is an application for the calculation of molecular descriptors developed by the Milano Chemometrics and QSAR. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases.

### ***SeeSAR***

SeeSAR is a software tool for interactive, visual compound prioritization as well as compound evolution. Structure-based design work ideally supports a multi-parameter optimization to maximize the likelihood of success, rather than affinity alone. Having the relevant parameters at hand in combination with real-

time visual computer assistance in 3D is one of the strengths of SeeSAR.

### *Dragon*

Dragon calculates 5,270 molecular descriptors, covering most of the various theoretical approaches. The list of descriptors includes the simplest atom types, functional groups and fragment counts, topological and geometrical descriptors, three-dimensional descriptors, but also several properties estimation (such as logP) and drug-like and lead-like alerts.

### *PaDEL-Descriptor*

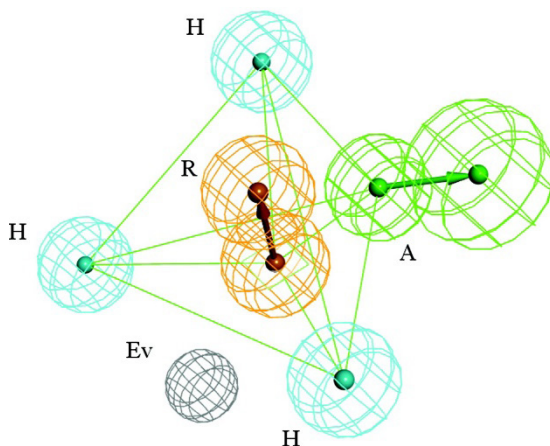
A software to calculate molecular descriptors and fingerprints. The software currently calculates 1875 descriptors (1444 1D, 2D descriptors and 431 3D descriptors) and 12 types of fingerprints (total 16092 bits). The descriptors and fingerprints are calculated using The Chemistry Development Kit with additional descriptors and fingerprints such as atom type electrotopological state descriptors, Crippen's logP and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by Laggner, and binary fingerprints and count of chemical substructures identified by Klekota and Roth.

## **4.2.2 Pharmacophore**

A Pharmacophore is an abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule. The IUPAC defines a Pharmacophore to be "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response". A Pharmacophore model explains how structurally diverse ligands can bind to a common receptor site. Furthermore, Pharmacophore models can be used to identify through de novo

design or virtual screening novel ligands that will bind to the same receptor. In modern computational chemistry, Pharmacophores are used to define the essential features of one or more molecules with the same biological activity. A database of diverse chemical compounds can then be searched for more molecules which share the same features arranged in the same relative orientation.

Pharmacophores are also used as the starting point for developing 3DQ SAR models. Such tools and a related concept of “privileged structures”, which are “defined as molecular frameworks which are able of providing useful ligands for more than one type of receptor or enzyme target by judicious structural modifications”, aid in drug discovery.



## *Pharmer*

Predicting molecular interactions is a major goal in rational drug design. Pharmacophore, which is the spatial arrangement of features that is essential for a molecule to interact with a specific target receptor, is important for achieving this goal. PharmaGist is a freely available web server for pharmacophore detection. The employed method is ligand based. It does not require the structure of the target receptor. Instead, the input is a set of structures of drug-like molecules that are known to bind to the receptor. We compute candidate pharmacophores by multiple flexible alignments of the input ligands. The main innovation of

this approach is that the flexibility of the input ligands is handled explicitly and in deterministic manner within the alignment process. The method is highly efficient, where a typical run with up to 32 drug-like molecules takes seconds to a few minutes on a standard PC. Another important characteristic of the method is the capability of detecting pharmacophores shared by different subsets of input molecules. This capability is a key advantage when the ligands belong to different binding modes or when the input contains outliers.

### *Catalyst*

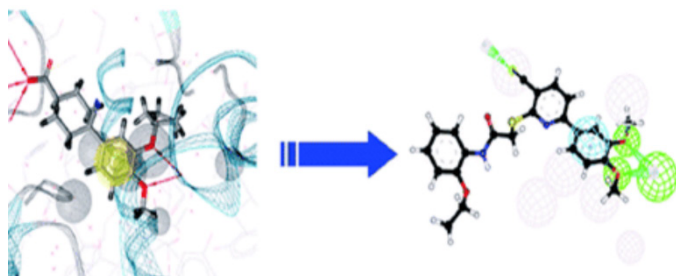
Pharmacophore Modeling and Analysis; 3D database building and searching; Ligand conformer generation and analysis tools; Geometric, descriptor-based querying; Shape-based screening.

### *LigandScout*

The LigandScout software suite comprises the most user-friendly molecular design tools available to chemists and modelers worldwide. The platform seamlessly integrates computational technology for designing, filtering, searching and prioritizing molecules for synthesis and biological assessment.

### *MOE*

MOE contains the industry-leading suite of Pharmacophore discovery applications used for fragment-, ligand and structure-based design projects. Pharmacophore modeling is a powerful means to generate and use 3D information to search for novel active compounds, particularly when no receptor geometry is available. Pharmacophore methods use a generalized molecular recognition representation and geometric constraints to bypass the structural or chemical class bias of 2D methods.



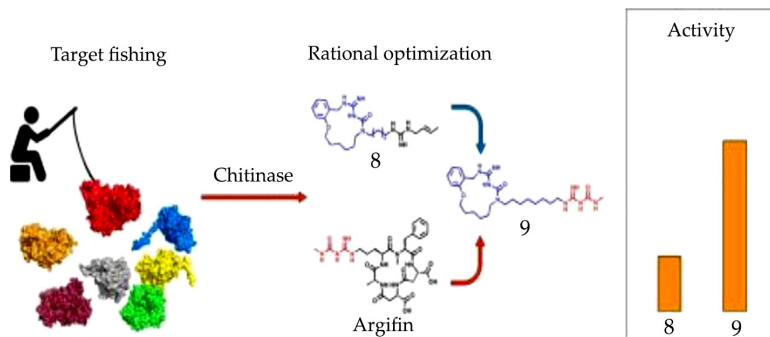
## ***Phase***

Phase is a complete, user-friendly Pharmacophore modeling solution designed to maximize performance in virtual screening and lead optimization. Fast, accurate, and easy-to-use, Phase includes a novel, scientifically validated common Pharmacophore perception algorithm.

### **4.2.3 Target Fishing**

Computational methods for Target Fishing (TF), also known as Target Prediction or Polypharmacology Prediction, can be used to discover new targets for small-molecule drugs. This may result in repositioning the drug in a new indication or improving our current understanding of its efficacy and side effects. We can set a new benchmark to validate TF methods, which is particularly suited to analyze how predictive performance varies with the query molecule.

Robust target fishing extends multitude benefits to drug research, such as avoiding unwanted side effects from poly pharmacology of small molecules at clinical stages, to reveal the mode-of-actions of a compound and also to repurpose old drugs for new targets. The rule of 'one-size-does-not-fit-all' still holds well in target fishing approaches as well. Therefore, it is important to carefully assemble the available methods and resources such that all levels of biological information, from sequences to structures to pharmacophores, are maximally utilized for fishing out the targets for the design of safer next generation drugs.



## ChemMapper

ChemMapper is a free web server for computational drug discovery based on the concept that compounds sharing high 3D similarities may have relatively similar target association profile. ChemMapper integrates nearly 300 000 chemical structures from various sources with pharmacology annotations and over 3 000 000 compounds from commercial and public chemical catalogues. Inhouse SHAFTS method which combines the strength of molecular shape superposition and chemical feature matching is used in ChemMapper to perform the 3D similarity searching, ranking, and superposition. Taking the user-provided chemical structure as the query, SHAFTS aligns each target compound in the database onto the query and calculates the 3D similarity scores and the top most similar structures are returned. Based on these top most similar structures whose pharmacology annotation is available, a chemical-protein network is constructed and a random walk algorithm is taken to compute the probabilities of the interaction between the query structure and proteins which associated with hit compounds. These potential protein targets ranked by the standard score of the probabilities. ChemMapper can be useful in a variety of polypharmacology, drug repurposing, chemical-target association, virtual screening, and scaffold hopping studies.

### ***PharmMapper Server***

The current release, i.e. version 2017, is based on the contents of PDB officially released by Jan 1st, 2016. This release applies Cavity1.1 to detect the binding sites on the surface of a given protein structure and rank them according to the corresponding druggability scores. A receptor-based pharmacophore modeling program Pocket 4.0 was then used to extract pharmacophore features within cavities. In this approach, a total of 23236 proteins covering 16159 pharmacophore models which are predicted as druggable binding sites and 52431 pharmacophore models with a pKd value higher than 6.0 are picked out, which is currently the largest collection of this kind. Compared to the last release (v.2010), target pharmacophore models included in this release have increased more than six times, from 7302 to over almost 53000.

### ***TargetHunter***

This web portal implements a novel in silico target prediction algorithm, the Targets Associated with its Most Similar Counterparts, by exploring the largest chemogenomical databases, ChEMBL. TargetHunter also features an embedded geography tool, BioassayGeoMap, developed to allow the user easily to search for potential collaborators that can experimentally validate the predicted biological targets or off targets. TargetHunter therefore provides a promising alternative to bridge the knowledge gap between biology and chemistry, and significantly boost the productivity of chemogenomics for silico drug design and discovery.

### ***ChemProt***

The ChemProt 2.0 server is a resource of annotated and predicted chemical-protein interactions. The server is a compilation of over 1 100 000 unique chemicals with biological activity for more than 15000 proteins. ChemProt can assist in the in-silico evaluation of



small molecules (drugs, environmental chemicals and natural products) with the integration of molecular, cellular and disease-associated proteins complexes.

### *SwissTargetPrediction*

This website allows you to predict the targets of a small molecule. Using a combination of 2D and 3D similarity measures, it compares the query molecule to a library of 280'000 compounds active on more than 2000 targets of 5 different organisms.

### *SuperPred*

SuperPred, which is a prediction webserver for ATC code and target prediction of compounds. Predicting ATC codes or targets of small molecules and thus gaining information about the compounds offers assistance in the drug development process. The webserver's ATC prediction as well as target prediction is based on a pipeline consisting of 2D, fragment and 3D similarity search.

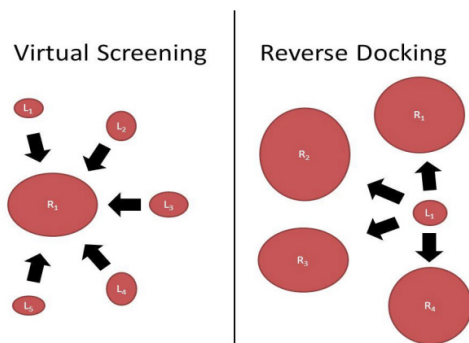
### *PASS*

PASS Online predicts over 4000 kinds of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, influence on gene expression, etc. To obtain the predicted biological activity profile for your compound, only structural formula is necessary; thus, prediction is possible even for virtual structure designed in computer but not synthesized yet.

## **4.2.4 Reverse Docking**

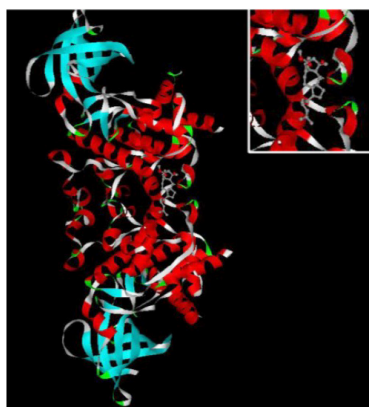
In reverse docking, one tries to find protein targets which can bind to a particular ligand. The necessary components are similar to those of forward docking methods; preparing data sets, searching

for ligand poses, and scoring and ranking the complex structures. However, several issues including high computational cost and inter-protein score bias makes reverse docking process rather complex.



### *Invdock*

A computer method, and its application software INVDOCK, have been developed for computer automated identification of potential protein and nucleic acid targets of a small molecule. The 3-D structure of the small molecule being studied is input into the programmer, the software automatically searches a protein and nucleic acid 3-D structure database to identify protein, RNA or DNA molecule that the small molecule can bind to. The identified proteins and nucleic acids are considered potential targets of the molecule.



### *idTarget*

A web server for identifying biomolecular targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach.

### *AMIDE (Automatic molecular inverse docking engine)*

Molecular docking is widely used computational techniques that allows studying structure-based interactions complexes between biological objects at the molecular scale. AMIDE was developed, a framework that allows performing inverse virtual screening to carry out a large-scale chemical ligand docking over a large dataset of proteins. Its ability to reproduce experimentally determined structures and binding affinities highlighted that AMIDE allows performing better exploration than existing blind docking methods.

### *VTS (Virtual Target Screening)*

Virtual Target Screening (VTS)", a set of small drug-like molecules are docked against each structure in the protein library to produce benchmark statistics. This calibration provides a reference for each protein so that hits can be identified for an MOI. VTS can then be used as tool for: drug repositioning, specificity and toxicity testing, identifying potential metabolites, probing protein structures for allosteric sites, and testing focused libraries for selectivity.

### *iRAISE (inverse rapid index-based screening engine)*

Integrates flexibility of hydrophilic rotatable terminal groups (such as hydroxyl groups) of the active site and the query molecule. iRAISE is an inverse screening tool based on the RApid Index-based Screening Engine (RAISE) technolog.

### ***ACTP (Autophagic Compound-Target Prediction)***

Autophagy (macroautophagy) is well known as an evolutionarily conserved lysosomal degradation process for long-lived proteins and damaged organelles. Recently, accumulating evidence has revealed a series of small-molecule compounds that may activate or inhibit autophagy for therapeutic potential on human diseases. However, targeting autophagy for drug discovery still remains in its infancy. In this study, we developed a webserver called Autophagic Compound-Target Prediction (ACTP) that could predict autophagic targets and relevant pathways for a given compound.

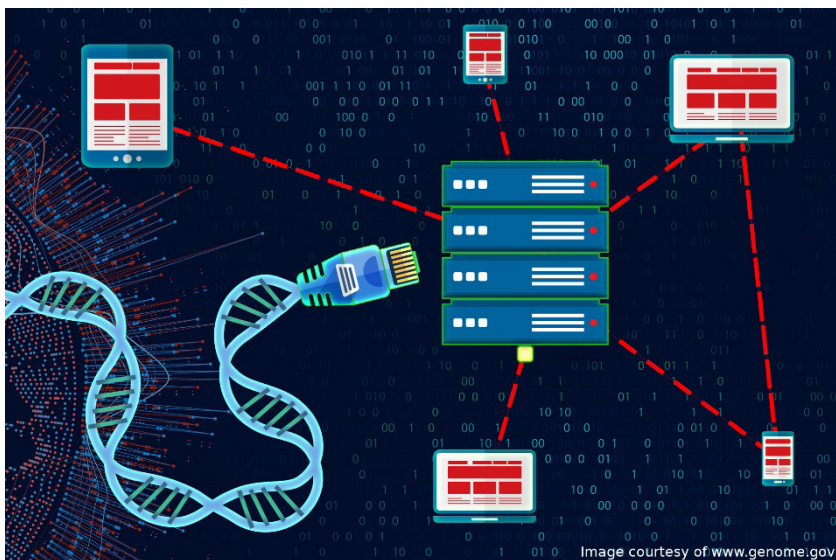
## **4.3 BIOINFORMATICS**

Bioinformatics, a hybrid science that links biological data with techniques for information storage, distribution, and analysis to support multiple areas of scientific research, including biomedicine. Bioinformatics is fed by high-throughput data-generating experiments, including genomic sequence determinations and measurements of gene expression patterns. Database projects curate and annotate the data and then distribute it via the World Wide Web. Mining these data leads to scientific discoveries and to the identification of new clinical applications. In the field of medicine in particular, a number of important applications for bioinformatics have been discovered. For example, it is used to identify correlations between gene sequences and diseases, to predict protein structures from amino acid sequences, to aid in the design of novel drugs, and to tailor treatments to individual patients based on their DNA sequences (pharmacogenomics).



### 4.3.1 Data of Bioinformatics

The classic data of bioinformatics include DNA sequences of genes or full genomes; amino acid sequences of proteins; and three-dimensional structures of proteins, nucleic acids and protein–nucleic acid complexes. Additional “-omics” data streams include: transcriptomics, the pattern of RNA synthesis from DNA; proteomics, the distribution of proteins in cells; interactomics, the patterns of protein–protein and protein–nucleic acid interactions; and metabolomics, the nature and traffic patterns of transformations of small molecules by the biochemical pathways active in cells. In each case there is interest in obtaining comprehensive, accurate data for particular cell types and in identifying patterns of variation within the data. For example, data may fluctuate depending on cell type, timing of data collection (during the cell cycle, or diurnal, seasonal, or annual variations), developmental stage, and various external conditions. Metagenomics and metaproteomics extend these measurements to a comprehensive description of the organisms in an environmental sample, such as in a bucket of ocean water or in a soil sample.



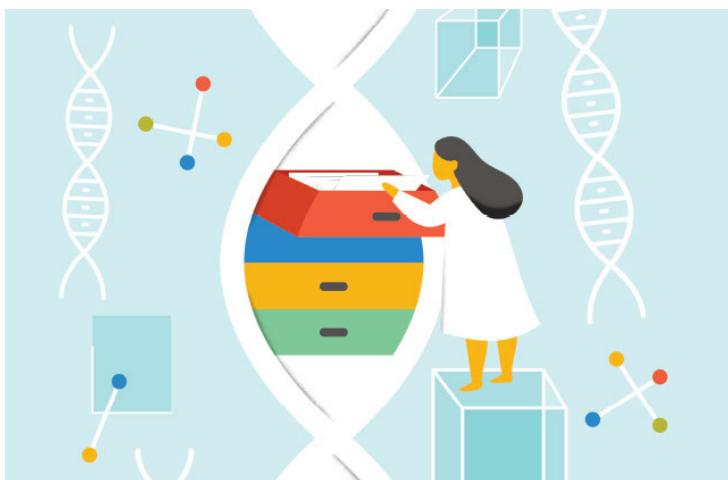
Bioinformatics has been driven by the great acceleration in data-generation processes in biology. Genome sequencing methods show perhaps the most dramatic effects. In 1999 the nucleic acid sequence archives contained a total of 3.5 billion nucleotides, slightly more than the length of a single human genome; a decade later they contained more than 283 billion nucleotides, the length of about 95 human genomes.

The U.S. National Institutes of Health has challenged researchers by setting a goal to reduce the cost of sequencing a human genome to \$1,000; this would make DNA sequencing a more affordable and practical tool for U.S. hospitals and clinics, enabling it to become a standard component of diagnosis.

#### 4.3.2 Storage and Retrieval of Data

The major database of biological macromolecular structure is the worldwide Protein Data Bank (wwPDB), a joint effort of the Research Collaboratory for Structural Bioinformatics (RCSB) in the United States, the Protein Data Bank Europe (PDBe) at the European Bioinformatics Institute in the United Kingdom, and the Protein Data Bank Japan at Ōsaka University. The homepages

of the wwPDB partners contain links to the data files themselves, to expository and tutorial material (including news items), to facilities for deposition of new entries, and to specialized search software for retrieving structures. Information retrieval from the data archives utilizes standard tools for identification of data items by keyword; for instance, one can type “aardvark myoglobin” into Google and retrieve the molecule’s amino acid sequence. Other algorithms search data banks to detect similarities between data items. For example, a standard problem is to probe a sequence database with a gene or protein sequence of interest in order to detect entities with similar sequences.



### 4.3.3 Goals

To study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:



- Development and implementation of computer programs that enable efficient access to, management and use of, various types of information.
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

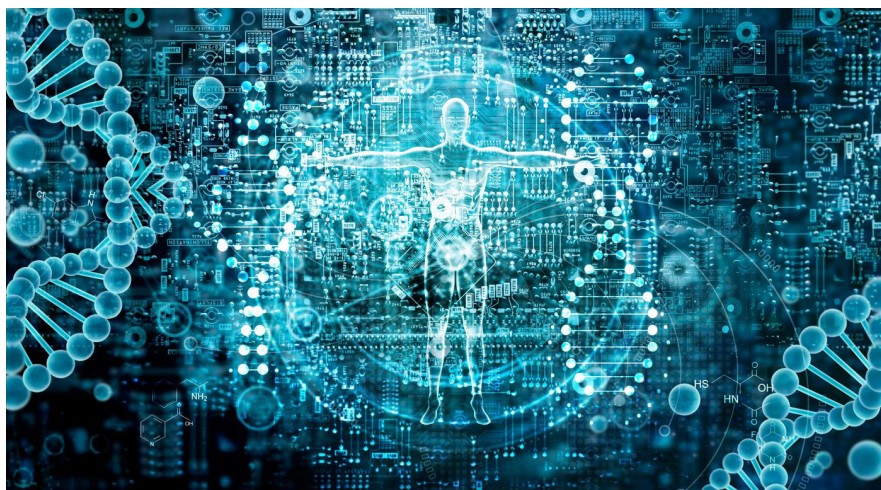
Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.



### 4.3.4 Relation to Other Fields

Bioinformatics is a science field that is similar to but distinct from biological computation, while it is often considered synonymous to computational biology. Biological computation uses bioengineering and biology to build biological computers, whereas bioinformatics uses computation to better understand biology. Bioinformatics and computational biology involve the analysis of biological data, particularly DNA, RNA, and protein sequences. The field of bioinformatics experienced explosive growth starting in the mid-1990s, driven largely by the Human Genome Project and by rapid advances in DNA sequencing technology.



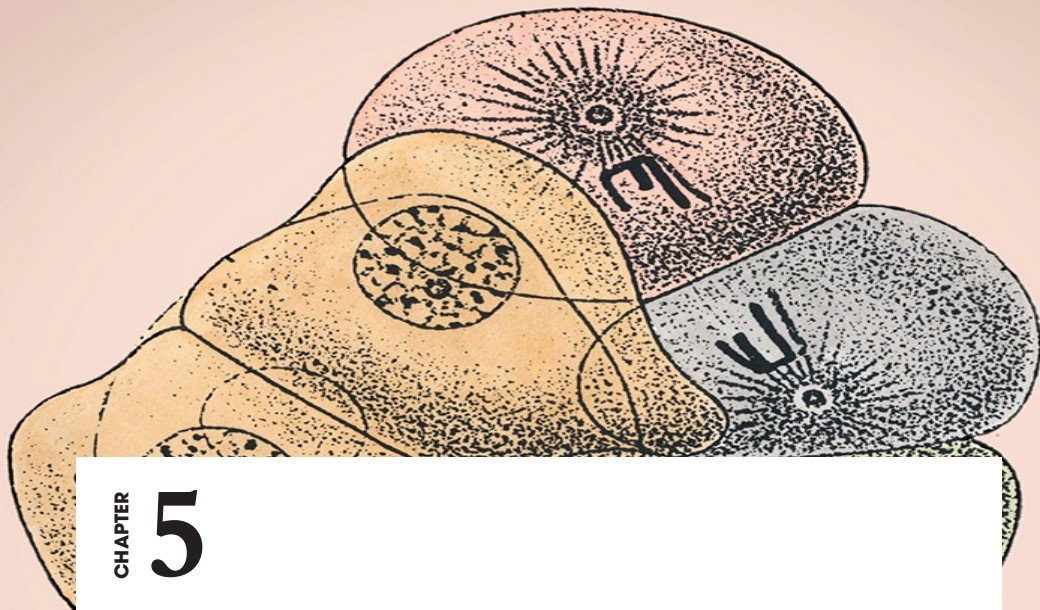
Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from graph theory, artificial intelligence, soft computing, data mining, image processing, and computer simulation. The algorithms in turn depend on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, and statistics.

## REFERENCES

1. Buratto, R. (2015) Exploring Ligand Affinities for Proteins by NMR of Long-Lived States. EPFL, Retrieved from [http://infoscience.epfl.ch/record/214545/files/EPFL\\_TH6816.pdf](http://infoscience.epfl.ch/record/214545/files/EPFL_TH6816.pdf)
2. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Pujadas, G., & Garcia-Vallve, S. (2015). Tools for in silico target fishing. *Methods*, 71, 98-103.
3. Chirag N. Patel, J. J. G., ., & et al. (2017). Molecular recognition analysis of human acetylcholinesterase enzyme by inhibitors: An in silico approach. Paper presented at the Proceedings of International Science Symposium on Recent Trends in Science and Technology (ISBN: 9788193347553).
4. Gauravi Trivedi, J. J. G. (2016). Identification of novel drug targets and its Inhibitors from essential genes of human pathogenic Gram positive bacteria. Paper presented at the Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123).
5. George, J. J. (2015). DOCKING STUDIES, ADMET PREDICTION OF PHYTOCHEMICAL INHIBITORS FOR ALZHEIMER'S DISEASE. Paper presented at the Proceedings of 8th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952116).
6. George, J. J. (2016). A Bioinformatics Approach for the Identification of Potential Drug Targets and Identification of Drug-like Molecules for Ribosomal Protein L6 of *Staphylococcus* species. Paper presented at the Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123).
7. George, J. J., & Umrana, V. (2011). In silico identification of putative drug targets in *Klebsiella pneumonia* MGH78578.
8. George, J. J., & Umrana, V. (2012). Subtractive genomics approach to identify putative drug targets and identification of drug-like molecules for beta subunit of DNA polymerase III in *Streptococcus* species. *Applied Biochemistry and*

- Biotechnology, 167(5), 1377-1395.
9. Kotadiya, R., & George, J. J. (2015). In silico approach to identify putative drugs from natural products for Human papillomavirus (HPV) which cause cervical cancer. *Life Sciences Leaflets*, 62, 1-13.
  10. Li, G.-B., Yu, Z.-J., Liu, S., Huang, L.-Y., Yang, L.-L., Lohans, C. T., & Yang, S.-Y. (2017). IFPTarget: A Customized Virtual Target Identification Method Based on Protein-Ligand Interaction Fingerprinting Analyses. *Journal of chemical information and modeling*, 57(7), 1640-1651.
  11. Patel, C. N., George, J. J., Modi, K. M., Narechania, M. B., Patel, D. P., Gonzalez, F. J., & Pandya, H. A. (2017). Pharmacophore-based virtual screening of catechol-o-methyltransferase (COMT) inhibitors to combat Alzheimer's disease. *Journal of Biomolecular Structure and Dynamics*, 1-20.
  12. Wolber, G., Seidel, T., Bendix, F., & Langer, T. (2008). Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug discovery today*, 13(1-2), 23-29.





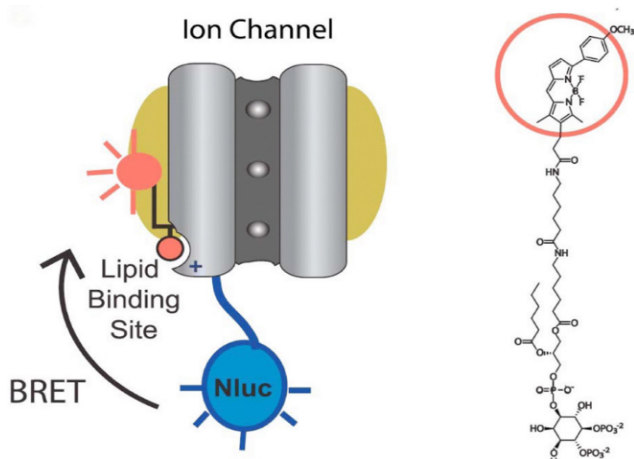
## CHAPTER

## 5

## A CHEMICAL GENOMICS APPROACH FOR ION CHANNEL MODULATORS

### INTRODUCTION

Ion channel modulators offer significant therapeutic opportunities in a number of areas, including arrhythmia, asthma, CNS disorders, coronary heart disease, hypertension, inflammation, and water retention. New ion channels are constantly being discovered and characterized in terms of their pharmacology, physiology, and structure. In addition, more and more selective ion channel modulators are emerging, upon which drug discovery programs can be initiated.



The physiological effects of ion channels are based on the regulation of ion fluxes (e.g.,  $K^+$ ,  $Na^+$ ,  $Ca^{2+}$ ,  $Cl^-$ ) across membranes, which affect, for example, osmotic pressure, nerve signal transmission, and muscle contraction. Ion permeation is extremely fast (up to  $10^7$  ions  $s^{-1}$ ) and highly selective.

Drews classified ion channels as the fourth-most important target class for drug therapies after receptors, enzymes, and hormones, and a more recent analysis considered kinases, GPCRs, and cation channels to be the most interesting target classes for pharmaceutical research. Currently, drugs targeting ion channels generate over 24 billion dollars in sales per annum.

Appropriate drug targets should meet several criteria, such as known biological functions, as well as robust assay systems for in vitro characterization and testing. Furthermore, they need to be accessible to low molecular weight compounds in vivo. Ion channels meet most of these 'druggability' criteria and can be viewed as suitable targets for small molecule drugs.

Potassium ( $K^+$ ) ion channels, for example, are recognized as critical regulators of cellular activities and are linked to several disease indications, including ventricular arrhythmias, long QT syndrome, and atrial fibrillation, as well as to insulin secretion and T-cell activation. The long QT syndrome, for instance, is associated with an inhibition of the hERG channel in the heart.



hERG inhibition represents an important safety consideration in drug discovery. Due to their hERG blocking properties and subsequent QT interval prolongation, several diverse drugs such as Terfenadine, Cisapride, and Astemizole have been withdrawn from the market. In comparison, the voltage-gated Kv1.3 channel is of interest in therapeutic immune modulation in multiple sclerosis and other T-cell mediated autoimmune diseases, and  $\text{Ca}^{2+}$ -activated potassium channels are of interest for reducing hyperactive bladder by hyperpolarization of the smooth muscle in the bladder.

Calcium-channel blockers are used for treating cardiac arrhythmia and pulmonary hypertension and for prevention of reperfusion injury. Sodium channels have been linked to epilepsy and hyperkalemic periodic paralysis.

Recently approved ion channel modulators include, for example, Nateglinide and Nimodipine. Nateglinide was approved in December 2000 as a blood glucose lowering agent. Nateglinide depolarizes pancreatic  $\beta$  cells by blocking the ATP sensitive potassium (KATP) channel, whereby calcium channels are opened, resulting in calcium influx and insulin secretion. The extent of insulin release is glucose-dependent and decreases at low glucose levels. Nateglinide is highly tissue selective with low affinity for heart and muscle.

Nimodipine was approved in August 2000 for the improvement of neurological outcome by reducing the incidence and severity of ischemic deficits in patients with subarachnoid hemorrhage.

The sodium-channel inhibitor Amiloride is used for the treatment of chronic bronchitis, and the most frequently used anesthetic drug, Lidocaine, inhibits voltage-gated sodium-channel  $\alpha$  subunits, which mediate the pathophysiology of pain.

Despite their remarkable physiological value, ion channels are still an unexploited therapeutic target class, especially in comparison to G-protein coupled receptors. Hence, lead finding and lead optimization programs for ion channel modulators are becoming more and more interesting. As ion channels are strongly related to

each other, a systematic exploration of this target family appears to be a promising way to accelerate drug discovery. Chemical genomics refers to such systematic and in-depth exploration of a target family and fosters a knowledge-driven drug design approach. This method is especially feasible for ion channel modulators, since considerable knowledge of pharmaceutically active structural classes and structure–activity relationships exists. In this chapter we summarize our current chemical genomics knowledge-based strategies for drug discovery of ion channel modulators. This includes structural information about ion channels, as well as lead-finding strategies in this field. The impact of this strategy is outlined by several successful examples.

**Table 1** Pathophysiological conditions related to ion channels

| <i>Channel</i>            | <i>Disease</i>  |
|---------------------------|---|
| Calcium channels          | Arrhythmia, diabetes, epilepsy, hypertension, migraine, stroke  |
| Chloride channels         | Cystic fibrosis, myotonia, muscoviscidos  |
| Potassium channels        | Arrhythmia, asthma, blood pressure, cardiac ischemia, cell proliferation, diabetes, epilepsy, cancer, immune suppression                      |
| Sodium channels           | Epilepsy, migraine, myotonia, pain, stroke  |
| Ligand-gated ion channels | Allergy, asthma, epilepsy, gastroesophageal reflux, inflammation, ischemia, learning and memory, migraine, neurodegenerative diseases, stroke |

We consider the highest impact of this strategy to be in lead finding, although such a target family-related approach offers further obvious advantages in the field of assay development, HTS technology, and compound optimization. In particular, the selectivity of ion channel modulators can be addressed by appropriate profiling of these compounds and by building channel-specific models applicable to lead optimization.

## 5.1 STRUCTURAL INFORMATION ON ION CHANNELS: ION CHANNEL FAMILIES

Ion channels form a large, diverse family of membrane proteins that can be grouped according to various criteria, such as the gating behavior or the ion selectivity, as shown in Table 1.



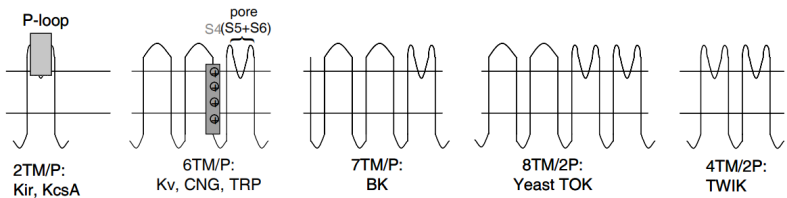
Classification according to such a scheme is not always simple, since ligandgated channels like the NMDA-activated ion channel may show voltage dependence, and, on the other hand voltage-gated channels have ligand-binding sites. Voltage gated sodium channels can be activated by drugs like Veratridine, whereas the MaxiK channel is gated by calcium ions.

The classification of ion channels by their topology is exemplified for potassium channels in Figure 1. Potassium channels can be classified into 2TM/P channels, which contain two transmembrane helices (TM) with one P loop (P) between them, 6TM/P channels, 7TM/P channels, 8TM/2P channels, and 4TM/2P channels. The 4TM/2P family is called leakage channels and is targeted by numerous anesthetics.

The 6TM/P channel family contains six transmembrane helices, labeled S1 to S6. The S4 helix contains four to seven positively charged amino acids, which are responsible for sensing the membrane potential. Therefore, the S4 helix is called the voltage sensor. The S5 and S6 helices form the ion-conducting pore by tetramerization.

**Table 2.** Classification schemes for ion channels

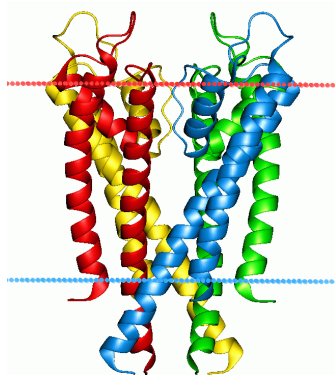
| Gating                                     | Ion selectivity  |
|--|------------------|
| Voltage                                    | Na <sup>+</sup>  |
| Ligand                                     | Ca <sup>2+</sup> |
| Mechanical                                 | K <sup>+</sup>   |
| Thermal                                    | Cl <sup>-</sup>  |
| any ion, any cation, any monovalent cation |                  |



**Figure 1.** Architectures of potassium channels: different transmembrane topologies shown together with potassium channels as examples. In the

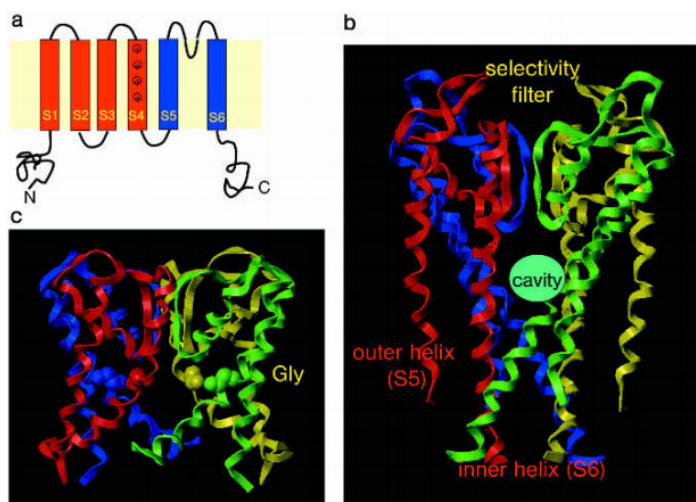
6TM family, the voltage-sensor helix S4 is highlighted, together with the pore-forming helices S5 and S6.

A number of different X-ray structures of bacterial potassium channels reveal the detailed atomic picture of the pore-forming part, helices S5 and S6. KcsA, which is crystallized in the closed conformation, has an overall structure similar to an inverted teepee. Four identical subunits surround the ion-conducting pathway. Each subunit contains two full transmembrane helices, S5 and S6, as well as the P loop. The S6 helices line the central cavity, whereas the S5 helices are involved in interactions with the lipid environment. In the closed channel conformation the transmembrane helices meet at the cytosolic side to block the ion conduction path. In the open conformation of the channel, the S6 helix kinks at a conserved glycine residue to open the ion conduction path, as shown in the structure of the bacterial channel MthK. The ion conduction path is formed by the selectivity filter and the large water-filled central cavity.



The solution of potassium channel X-ray structures has significantly contributed to the understanding of mechanistic questions like the amazing selectivity of potassium channels. Although the atomic radii of potassium (1.33 Å) and of sodium (0.95 Å) differ only slightly, potassium channels select potassium over sodium ion by a factor of 1000. This tremendous selectivity is achieved by the coordination geometry of eight amide carbonyl groups in the selectivity filter, optimized for the coordination sphere of potassium ions.

Structural data on potassium channels has also improved the understanding of the gating mechanism. Gating comprises a signaling step and the opening of the ion conduction path. The elucidation of the structure of the bacterial voltage-gated potassium channel KvAP, crystallized by using monoclonal antibody Fab fragments, yielded some unexpected insights into the design of the voltage-sensor helix S4. Mutational data demonstrated the role of the S4 helix in voltage sensing, and fluorescence labeling has shown movement of the S4 helix during gating. The prevailing model so far suggests a movement of helix S4 from one side of the membrane to the other upon changes in the membrane potential, although the findings of MacKinnon imply a different model. First, helix 3 is actually split into two different helices: the second part of helix 3 – called helix 3b – forms a helix–turn–helix motif with helix S4. This unit, called the ‘voltage-sensor paddle’, is actually oriented perpendicular to the pore unit and moves to the outer membrane side when the channel is opened.

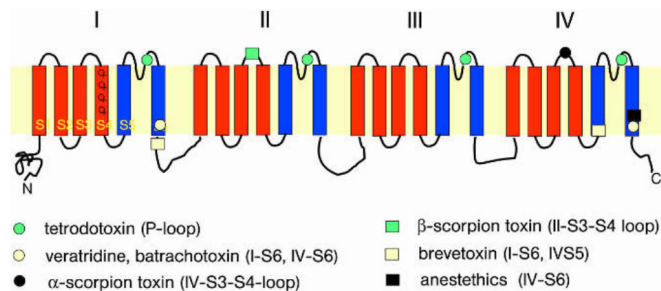


**Figure 2.** (a) Topology of 6TM/P potassium channels. (b) X-ray structure of KcsA (PDB code: 1j95). The four monomers tetramerizing to form the functional channel are shown in different colors. Important structural features such as the S5 and S6 helices, central cavity, and selectivity filter are indicated. (c) Structure of MthK pore (PDB code: 1LNQ). The structure is in the open channel conformation. The glycine residues that serve as a hinge for the bending of helix S6 are indicated.

Nevertheless, it is still unknown how the movement of the voltage-sensor paddle is linked to the opening of the ion conduction pathway, which is achieved by an outward bending of the S6 helix at the position of a conserved glycine. This helix movement opens the inner cavity to the cytosol, as shown by a comparison of the KcsA and MthK structures. For the inward-rectifying potassium channel family, a glycine residue in a different position could be the hinge position for formation of the opening pathway.

Chloride channels have a completely different structure from potassium channels. The dimeric structure has two ion pathways, one formed by each monomer. The ion pathway does not run straight through the membrane, but is U-shaped. Amino acids stabilize the ion in the pathway by forming direct interactions with the chloride atom via hydrogen bond donors, just as the carbonyl groups in the selectivity filter of potassium channels stabilize the potassium cation.

A couple of structures of ion channels have been solved, it is still a challenging task to express, purify, and crystallize these membrane proteins. Chang, for example, state that approximately 24 000 crystallization conditions were tested to solve the structure of the MscL homolog from *Mycobacterium tuberculosis*, a mechanosensitive ion channel. Therefore, the number of 3D structures of ion channels is still very small compared to the number of enzyme structures. Most importantly, no crystal structure of a ligand–ion channel complex has been obtained so far.



**Figure 3.** Topology of voltage-gated sodium channels. Known binding sites of peptides and drugs are marked. Voltage-gated sodium channels possess four 6TM/P domains.

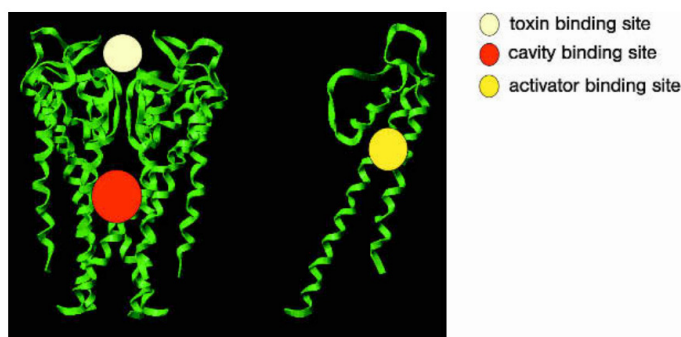
Thus, structure-based drug design in the field of ion channels still has to rely on homology models of ion channels, which can be combined with conventional methods to map the ligand binding site, such as site-directed mutagenesis or photoaffinity labeling. A number of different binding sites have thus been recognized on ion channels. For voltage-gated sodium channels, at least six different binding sites for toxins or drugs are known and are schematically depicted in Figure 3.



Voltage-gated potassium channels also have a number of different binding sites. Similar to sodium channels, there is a binding site for peptide toxins at the outer vestibule of the pore. This binding site has been identified by site-directed mutagenesis for different peptide toxins, e.g., for the toxin ShK from a sea anemone, which blocks Kv1.3, or for Charybdotoxin, which blocks various potassium channels.

Potassium channels also have binding sites within the ion conduction pore, as has been demonstrated for example for Kv1.3, Kv1.5, and hERG. Within the central cavity, there might be distinct

but possibly overlapping binding sites. Ammonium ions bind in the upper part of the cavity, close to the selectivity filter; whereas for the hERG channel or for Kv1.3, mutational data indicate drug binding sites closer to the cytosolic part of the cavity. A recent study by Milnes raises the question of whether there is a 'nonaromatic' binding site within the hERG channel, since the binding affinity of Fluvoxamine is only partially attenuated by mutations of Tyr652 and Phe656.



**Figure 4.** Structure of potassium channels with different binding sites in the pore domain marked. To show the binding location of R-L3 between helices S5 and S6, only the monomer is shown.

Recently, another binding site has been identified, the first binding site for a potassium channel activator. The benzodiazepine derivative R-L3, a partial agonist of KCNQ1, binds between the S5 and S6 helices as indicated in Figure 4. Interestingly, the structurally related compound L-7 blocks KCNQ1 binding in the central cavity.

This example illustrates the difficulty of drug design in the absence of detailed structural knowledge, since even slight modifications can have a tremendous effect on the binding site and mode of action. In the field of ion channels, rational design is even more hampered by the fact that voltage-gated ion channels cycle through at least three different states – a resting state, an open state, and an inactivated state. Electrophysiological studies give evidence that blockers can interact with open channels as well as with closed channels. Vesnarinone or MK-499 require channel



opening to bind. Other drugs, like Ketoconazole, bind to a closed state of the hERG channel, but Bertosamil binds to it in both its open and inactivated states.

## 5.2 LEAD-FINDING STRATEGIES FOR ION CHANNEL MODULATORS

Appropriate lead-finding strategies for ion channel modulators make use of as much information as possible. This includes information on modulators of closely related ion channels and presumably some 3D information about the particular target, either a homology model or available X-ray or NMR structures. The ligand information can be used for a ligand-based lead finding approach, whereas 3D structures are applicable to structure-based design.

### 5.2.1 Ligand-based Lead Finding

Ligand-based lead finding is based solely on information about putative ligands for a particular target or for a closely homologous target. This ligand information is then applied to select compounds that are closely linked to the reference molecules. This is achieved by 2D or 3D techniques. The 2D approach consists mainly of similarity and substructure searching, whereas the 3D method makes use of 3D pharmacophores built from a set of diverse compounds.

For similarity searching, all molecules are described by an appropriate binary descriptor (consisting of only zeros and ones). Such a binary fingerprint contains all structural information for a particular molecule and was applied at Aventis to identify new Kv1.5 inhibitors in the compound collection.

The Kv1.5 channel is a member of the voltage-gated K<sup>+</sup> channel family (which belongs to the 6TM/P family), whose functional form consists of four  $\alpha$  subunits each containing 6 transmembrane segments. The Kv1.5 pore domain is formed by four S5 and S6

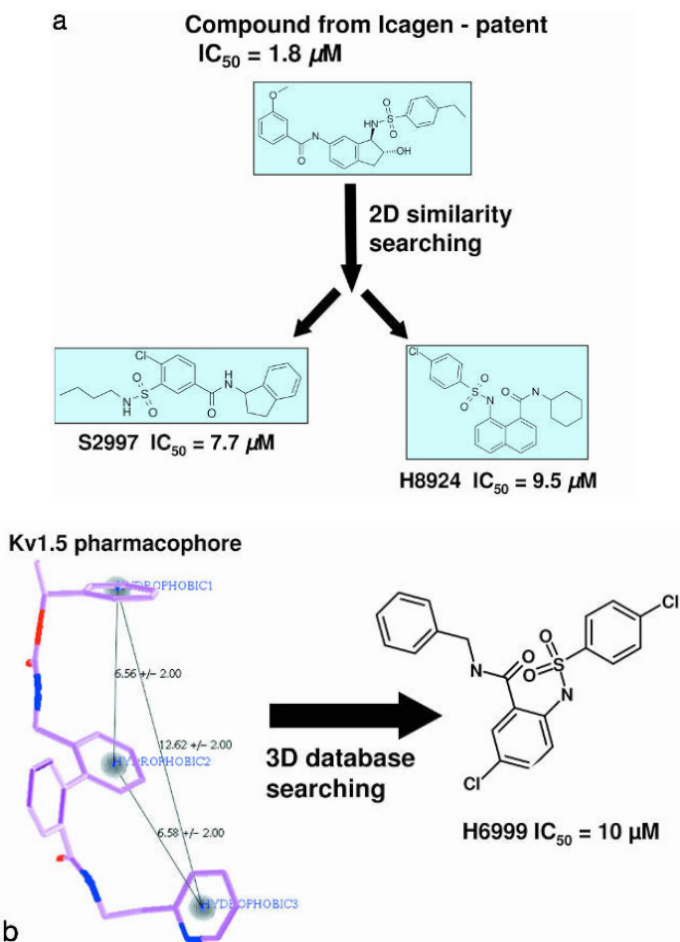
segments from four different  $\alpha$  subunits. In the human atrium, Kv1.5 is the molecular component of the repolarizing K<sup>+</sup> current *I<sub>Kr</sub>*, which contributes to the falling part of the cardiac action potential. Since *I<sub>Kr</sub>* has been found only in the human atrium, blockade of Kv1.5 has emerged as a promising approach for developing new atrial-selective antiarrhythmics devoid of undesired effects observed with the currently available antiarrhythmics. When the Kv1.5 project was started at Aventis, no high-throughput screening assay was available, and our lead-finding strategy relied on database searching.

We used a compound from an Icagen patent as a query and identified two structurally different molecules that showed almost identical Kv1.5 activity as our reference molecule. Additionally, a Kv1.5 pharmacophore was derived from a lead series of Kv1.5 inhibitors. This pharmacophore, consisting of three hydrophobic features in a specific spatial orientation, was used to identify new putative Kv1.5 inhibitors in our corporate compound collection. The 12 most promising compounds were selected based on their fit to the Kv1.5 pharmacophore. Subsequent biological profiling revealed one new lead structure.

Known side effects of a lead compound or drug can become an interesting opportunity to turn the side effect into the main pharmacological action of the compound. For the calcium antagonist Nifedipine, weak blocking of the calcium dependent potassium channels *I<sub>KCa</sub>* has been reported. *I<sub>KCa</sub>* is assumed to be involved in several diseases such as sickle cell anemia, immune disorders, and ischemic events. Blocking of this channel was also proposed to be beneficial in traumatic brain injury. Therefore, the calcium channel blocker Nifedipine was used as a starting point for developing a selective *I<sub>KCa</sub>* blocker with beneficial properties in a traumatic brain injury model. Since the NH group of the dihydropyridine ring is a prerequisite for calcium antagonistic activity, it was replaced by the isoelectronic oxygen, leading to phenylpyrans that showed significant *I<sub>KCa</sub>* blocking activity. Added electron-withdrawing substituents at the para position of the phenyl ring were able to further increase the potassium

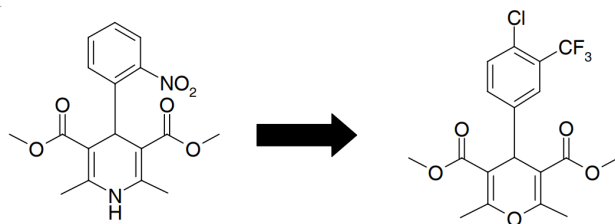


channel activity. Overall, the SAR for the phenylpyrans on the IKCa channel was found to be orthogonal to the SAR of the dihydropyridines on the L-type calcium channel, allowing for the identification of IKCa-selective compounds.



**Figure 5.** (a) Similarity-based 2D database searching for Kv1.5 inhibitors. (b) Ligand-based Kv1.5 pharmacophore and its application in 3D database searching.

Thus, this recent example nicely demonstrates that ion channel ligands can be valuable starting points for the identification of drugs acting against other members of the ion channel protein family.



**Figure 6.** Nifedipine (left) provided a good starting point to obtain a selective blocker of IKCa (right), by slight changes in the central scaffold and the substitution pattern.

### 5.2.2. Structure-based Lead Finding

Structure-based lead finding requires a target 3D structure to start with. However, experimental elucidation of ion channel structures either NMR or X-ray crystallography is extremely difficult to achieve.

Nevertheless, homology modeling of closely homologous channels to KcsA or MthK, for which 3D structures are available, makes this approach feasible

In an early structure-based design effort, a combinatorial library was designed by using LUDI for Kv1.3. Kv1.3 is involved in regulation of the membrane potential of human T cells, controlling calcium influx into the cell by voltage-dependent calcium channels. Calcium influx ultimately results in cytokine release and cell proliferation.

Therefore, Kv1.3 blockers might be interesting immunosuppressive compounds. Various peptide toxins are known to block Kv1.3. Chandy and coworkers have used these structures in combination with mutant cycle analysis to derive a model of the outer vestibule of the Kv1.3 channel.

Within this outer vestibule model, LUDI calculations were focused on three amino acids from each subunit, which are known to be important for toxin binding: His404, Gly380, and Asp386. LUDI was used to suggest fragments interacting with these key amino

acids. Fragment linking and modifications of the whole molecules, followed by molecular mechanics calculations, resulted in a phenylstilbene scaffold, which in the next step was varied in a combinatorial library comprising 400 compounds. The most active compound showed an IC<sub>50</sub> of 2.9  $\mu$ M.

This study was based on a model of the outer vestibule, which was developed using indirect evidence like the structure of known ligands and data from mutational analysis.

At that time, the KcsA crystal structure or other potassium channel X-ray structures were not available. Meanwhile, more detailed knowledge of the atomic details of potassium channels allows the development of homology models that can be successfully used in drug design, as demonstrated by the following example.

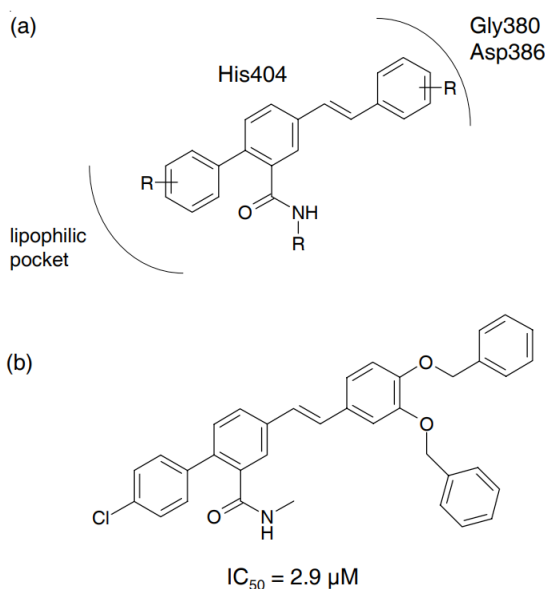
A recent structure-based lead-finding strategy was used for Kv1.5 inhibitors. The pore-forming domain of Kv1.5 exhibits 54% sequence homology with the bacterial K<sup>+</sup> channel KcsA from *Streptomyces lividans*, for which a crystal structure of the closed channel is available.

This structure was subsequently used as a template to build a homology model of the Kv1.5 pore-forming domain, using the Composer module of Sybyl 6.6.

Starting with the  $\alpha$  subunit of KcsA, a model of the S4 and S6 segments of Kv1.5 was built. Four of these segments were assembled according to the arrangement of the four  $\alpha$  subunits of KcsA, representing the pore domain of Kv1.5.

This model was refined by a two-step minimization protocol, involving minimization of the protein sidechains while keeping the backbone rigid, followed by minimization of the whole protein. The Sybyl 6.6 implementation of the AMBER forcefield was used to evaluate the energy of the system.

The minimized structure was submitted to several tests for its quality and internal consistency, which included both geometric and profile analyses.

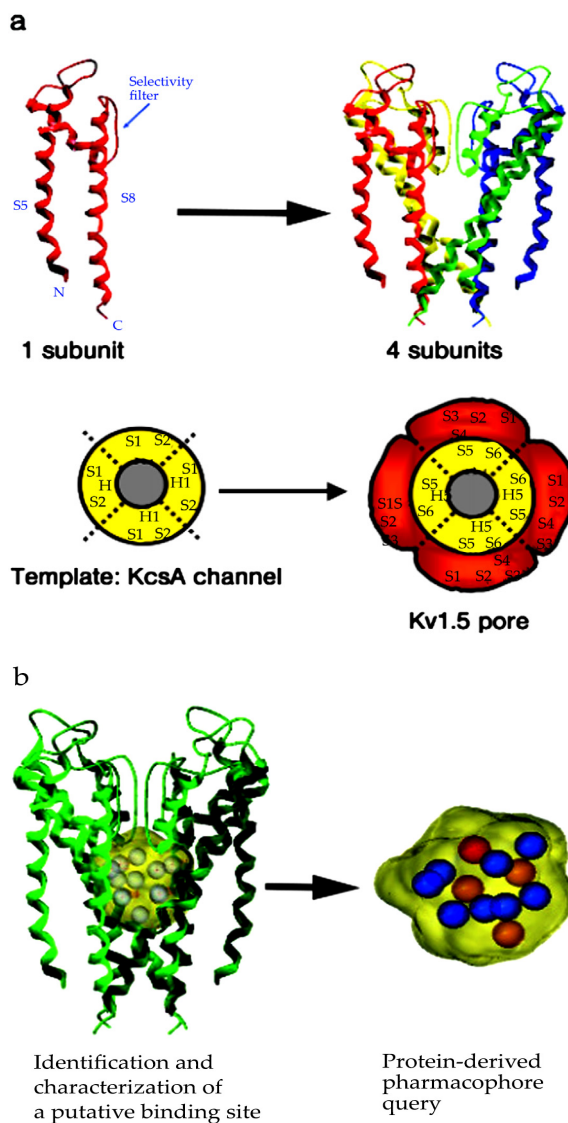


**Figure 7.** (a) Schematic representation of the proposed interactions of the phenylstilbene scaffold with residues from the outer vestibule. (b) The shown structure has the highest activity for Kv1.3 found in this library.

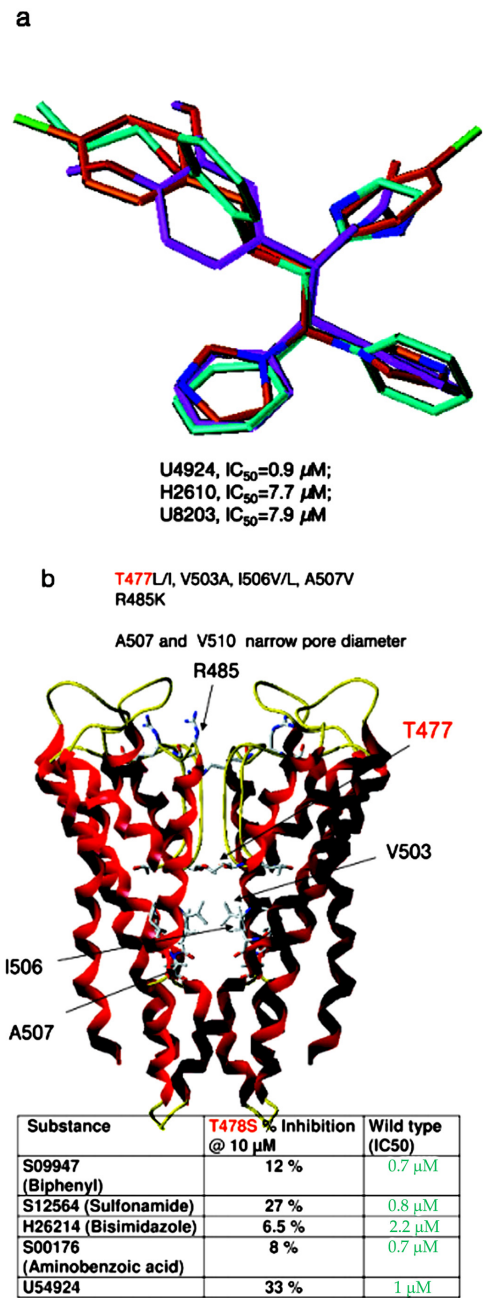
A computational elucidation of putative binding sites using the PASS algorithm revealed an internal site as most interesting for small organic molecules to interact with. Subsequent more detailed analysis resulted in the derivation of a protein based pharmacophore.

Use of this particular pharmacophore in subsequent 3D database searching of about 1 million compounds resulted in 244 interesting compounds, from which 19 compounds showed IC<sub>50</sub> values below 10  $\mu$ M.

The alignment of three of these hits is shown in Figure 9 a. Of course, these compounds exhibit high spatial similarity and fit remarkably well into the Kv1.5 binding site.



**Figure 8.** (a) Knowledge-based homology modeling of the closed Kv1.5 pore. (b) Identification of a putative Kv1.5 binding site.



**Figure 9.** (a) Alignment of the three most active Kv1.5 inhibitors. (b) Experimental validation of the Kv1.5 homology model by mutational data.

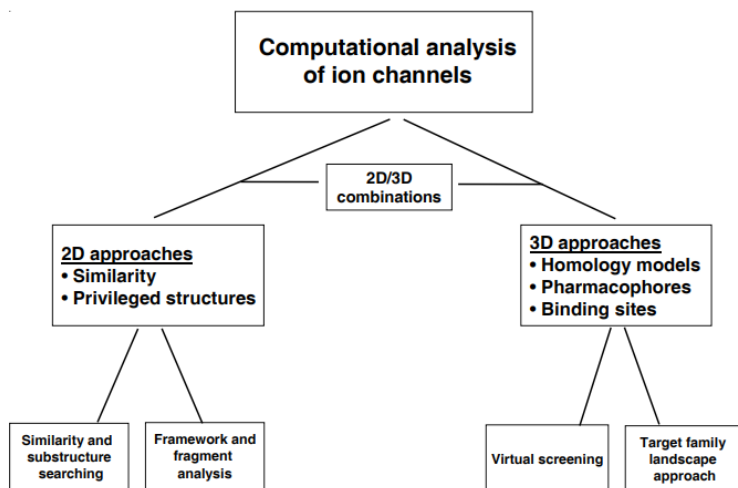
We proposed several mutations to the Kv1.5 channel to validate our model. Docking of our inhibitors in the Kv1.5 binding site revealed Thr477 as very important for binding. Indeed, mutation of Thr477 to serine left the Kv1.5 channel fully functional, but the activity of all four types of Kv1.5 inhibitors significantly decreased.

## 5.3 DESIGN OF ION CHANNEL FOCUSED LIBRARIES: CHEMICAL GENOMICS

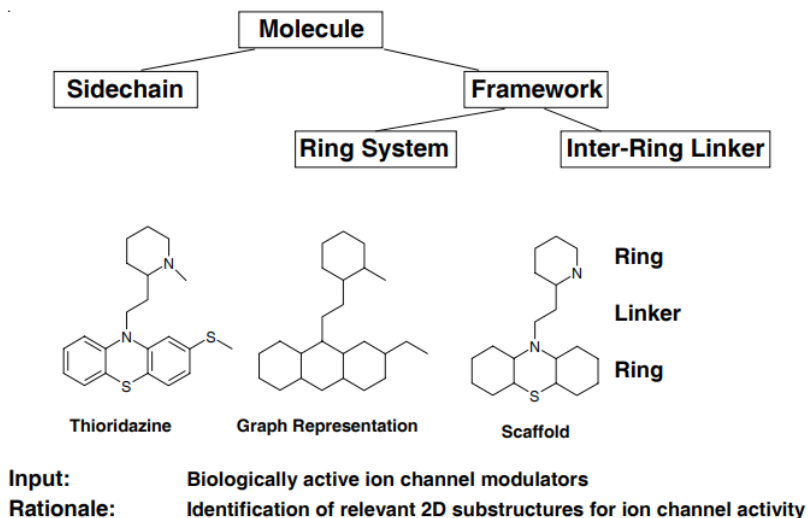
The consideration of all available ligand information concerning ion channel modulators, as well as the use of 3D structural information is required for designing appropriate ion channel focused libraries. This can be achieved by matching chemical and biological information in the target family of ion channels. The intersection of biological structures and functionalities with chemical structures and properties is derived to perform a knowledge-driven biased library design. This allows extraction of common structural features for ion channel modulators out of a practically infinite chemical space. Applying such design criteria leads to chemical libraries that are enriched in preferred features of ion channel modulators. This section covers design principles and their application.

### 5.3.1 Design Principles

Matching chemical and biological information in the field of ion channels requires combined 2D and 3D analysis. The 2D approach is based on a collection of biologically active compounds and consists mainly of similarity and substructure searching and of analysis of common frameworks and fragments to identify privileged chemotypes. Applicable 3D techniques are either ligand- or structure based. The ligand-based method requires biologically active ion channel modulators to derive 3D pharmacophores, and the structure-based technique uses a 3D structure of ion channels for subsequent virtual screening.



**Figure 10.** Computational tools for analyzing ion channels in knowledge-driven design.

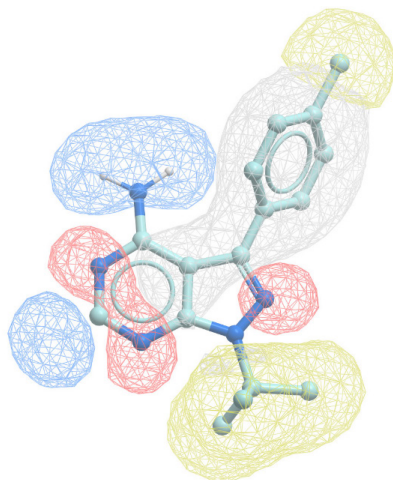


**Figure 11.** Topological framework analysis.

Substructure searching is often used in drug design and needs no further clarification. Similarity searching is also a very well-known technique described in more detail elsewhere. We usually use MACCS keys, Unity fingerprints, CATS descriptors, and



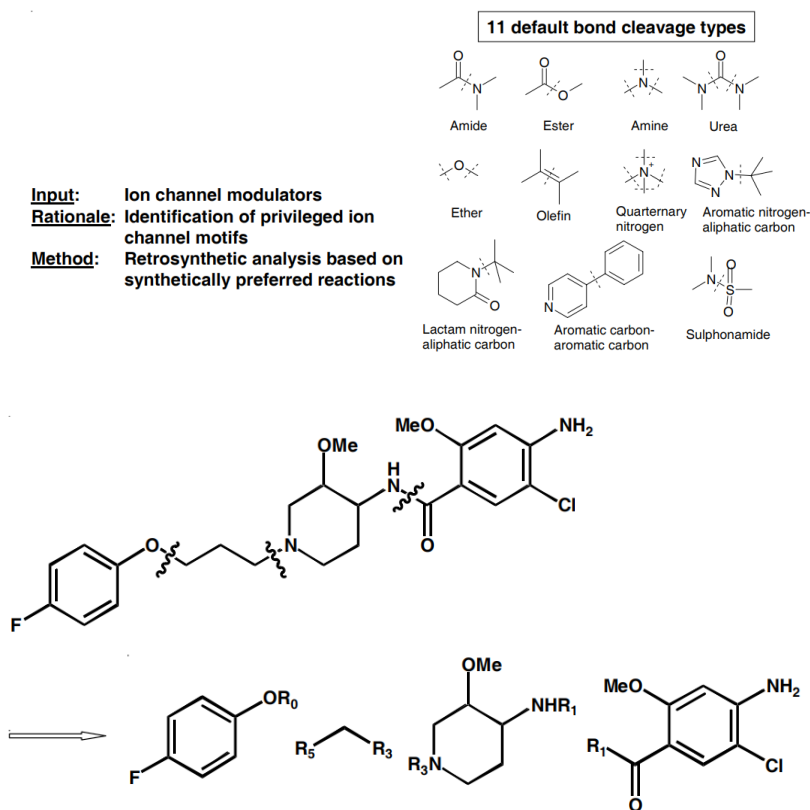
feature trees for similarity searching. Each technique has its own strengths and weaknesses, so we favor parallel application of two or three of them.



Framework analysis was published by Bemis and Murcko in 1996. They analyzed shapes of existing drugs in a commercial database to extract drug-related molecular frameworks by following a graph theoretical approach to decompose molecules into rings and noncyclic sidechains. Linkers and rings together form the framework of a molecule, whereas sidechains are omitted. For example, framework analysis of Thioridazine starts with removal of the acyclic sidechains and leaves a framework composed of two rings and one inter-ring linker.

Application of this topological framework analysis to ion channel modulators yields access to privileged ion channel chemotypes. Conversion of these frameworks into appropriate scaffolds for synthesis allows subsequent building of ion channel focused libraries.

Fragment analysis is based on the RECAP algorithm. This retrosynthetic combinatorial analysis starts with a collection of active molecules and then fragments these molecules using any set of retrosynthetic reactions. For example, Cisapride is cleaved into four fragments based on three different bond cleavage types.



**Figure 12.** Retrosynthetic combinatorial analysis procedure (RECAP).

Resulting fragments are clustered and reclassified into sets of monomers for subsequent library design. The RECAP procedure derives not only suitable chemotypes but also appropriate building blocks for scaffold decoration. Since the monomers are extracted from biologically active compounds, there is a high likelihood that new molecules derived from them will contain biologically important motifs.

The 3D approach makes use of ion channel-specific pharmacophores, ion channel X-ray structures, and homology models. Ion channel X-ray and homology models are not as precise as structures of smaller proteins. The uncertainty regarding the binding mode of ion channel modulators also adds additional complexity to structure-based virtual screening. However, valid

3D pharmacophores can be derived from these structures and subsequently used to identify privileged ion channel chemotypes by virtual screening in proprietary and public databases. Needless to say, both ligand- and structure-based pharmacophores require in-depth validation prior to their use in virtual screening.

### 5.3.2 Example: Building the Aventis Ion Channel Library

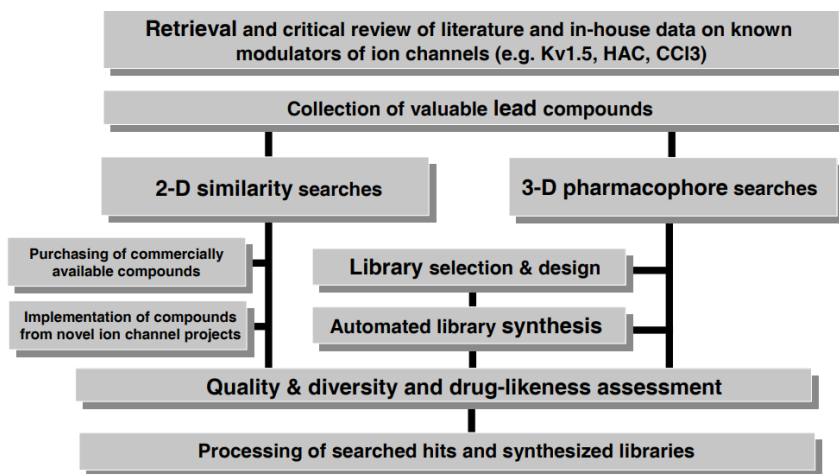
Ion channels are of potential interest in several therapeutic areas. However, appropriate high-throughput assays to test several hundred thousand compounds against a particular ion channel still lack sufficient signal-to-noise ratios. Therefore, a biased ion channel library is of high interest for lead finding.

We performed a combined 2D and 3D analysis of chemical and biological space to identify ion channel privileged chemotypes. The 2D approach was based on a collection of biologically active compounds and consisted mainly of similarity and substructure searching and of analysis of common frameworks and fragments. Our 3D approach relied on multiple ion channel pharmacophores and homology models, which were used for virtual screening.

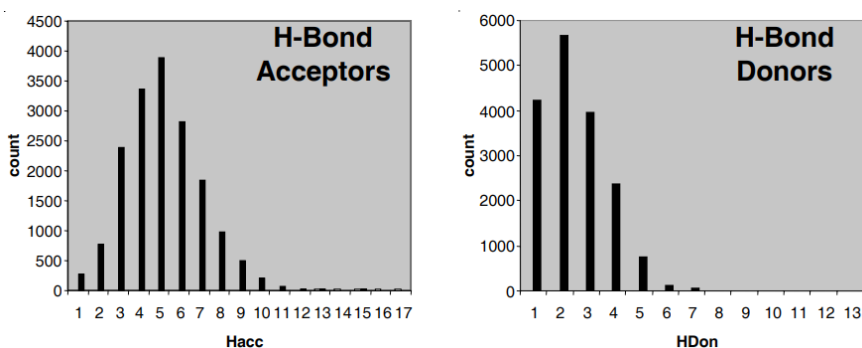
Our iterative ion channel library design process is outlined in Figure 13. Retrieval and critical review of literature and in-house data on ion channel modulators resulted in a collection of valuable lead compounds, suitable for 2D and 3D database mining in internal and external compound collections. Among others, we took into account calcium channel blockers like Clonidine, chloride channel blockers, potassium channel openers, K(ATP) channel blocker and openers (e.g., Glibenclamide), and NHE-1 inhibitors.

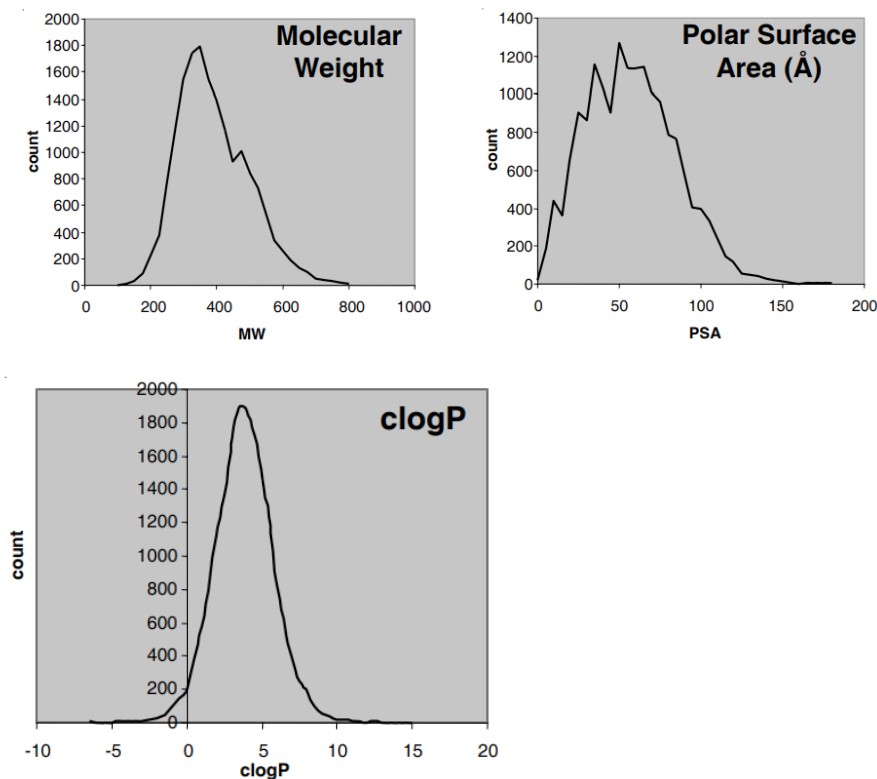
Scaffold proposals were collected and reviewed according to privileged ion channel motifs, chemical feasibility, and fit to our multiple pharmacophores. Building block selection, virtual library design, and filtering yielded small virtual libraries suitable for automated solution-phase synthesis. All synthesized compounds were finally purified and characterized prior to addition to our focused library.

Picked and purchased compounds, as well as all new designed chemotype focused libraries, were assessed in terms of quality, diversity, and drug-likeness. The remaining compounds were plated in our ion channel library, which is frequently used for screening of ion channels. New compounds and chemotypes from novel ion channel projects are continuously added to this focused library, and thus we constantly increase the value of our ion channel ligand collection.



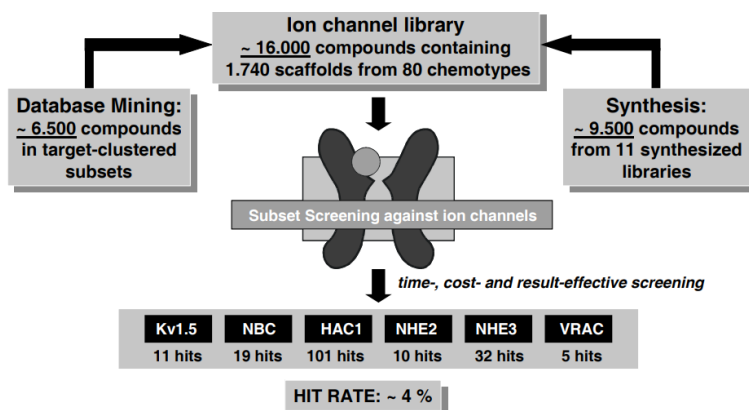
**Figure 13.** Design of the Aventis ion channel focused library.





**Figure 14.** Key properties of the Aventis ion channel library.

The key properties of our ion channel library, namely molecular weight, polar surface area, clogP, and number of hydrogen bond acceptors and donors, are within the lead-like range of compounds. A purity analysis of a small subset of our liquid ion channel collection revealed that more than 75% of the compounds had acceptable purity.



**Figure 15.** Aventis ion channel library: current status and hit rates.

Currently, the Aventis ion channel library contains approximately 16 000 compounds related to 1740 scaffolds and 80 chemotypes. Almost 6500 compounds have emerged from database mining, and 9500 compounds were synthesized.

Screening of this library or subsets of it against new ion channels revealed hit rates of approximately 4%, which is substantially higher than typical hit rates from high-throughput screening (~0.01%–0.1%). In addition, this focused screening identified highly valuable hits, since the library contains primarily drug like or lead-like compounds. Profiling this library against several ion channels not only reveals channel-subtype specific chemotypes, but also offers the opportunity to build early structure–activity relationships on scaffolds, which is very helpful especially for optimizing activity and selectivity. Hence, our chemical genomics approach has yielded improved screening hit rates and better starting points for subsequent compound optimization, thus reducing the cycle times for screening and optimization.

Although almost all our screening efforts using this focused library against new ion channel targets have resulted in good hit rates, we were recently disappointed by finding a hit rate of only 0.8% against a specific potassium channel, and we thought about opportunities for improvement. Current antiarrhythmic agents, for example Sotalol, quite often show adrenoceptor inhibition and

potassium- or multichannel blockade. Hence, a future prospect for our ion channel library is to identify common GPCR and ion channel chemotypes by profiling our GPCR library against ion channels and vice versa.

Identification of suitable lead compounds for subsequent optimization is one of the key needs in drug discovery today. Several techniques have been successfully applied, while a chemical genomics-driven approach, by building target family related compound libraries, seems to be a promising future strategy for lead finding. The high complexity of these efforts motivates a knowledge-driven design strategy, taking into account as much information as possible from targets and ligands. An in-depth scientific understanding of the intersection of biological and chemical information is crucial for enabling higher productivity in early compound identification. Hence, the relationship between certain chemotypes and biological targets within target families should drive this lead identification strategy.

The effort to bridge the chemical and biological space is called chemical genomics or chemogenomics. So far, no unique definition of chemical genomics has emerged from the literature, but the systematic exploration of target families is a common goal, based on the assumption that similar compounds bind to similar targets by similar mechanisms. This assumption is one of the foundations for the selection of compounds for our ion channel biased screening collection. Another important principle of chemical genomics is the integrated use of state-of-the-art computational tools to derive new ion channel binding motifs.

We have discussed such a chemical genomics approach for ion channel modulators. Some case studies for ion channel lead finding illustrate the opportunities of target- and ligand-related strategies. Target family-related knowledge is of course mandatory for this process. However, limited accessibility to ion channel 3D structures and uncertainties in homology models mean that structure-based approaches are feasible only after thorough validation of the underlying models.

However, identification of ion channel modulators by screening compound libraries enriched with ion channel privileged chemotypes offers rapid, efficient access to lead compounds. Flexible data-driven building and optimization of such an ion channel focused library will enable better and faster lead identification of ion channel modulators.

The increasing information in biological and chemical space and its effective transformation into a knowledge-driven ion channel focused library may foster a paradigm shift in lead identification.



## REFERENCES

1. A. Giorgetti, P. Carloni, *Curr. Opin. Chem. Biol.* 2003, 7, 150–156.
2. A. Kuo, J. M. Gulbis, J. F. Antcliff, T. Rahman, E. D. Lowe, J. Zimmer, J. Cuthbertson, F. M. Ashcroft, T. Ezaki, D. A. Doyle, *Science* 2003, 300, 1922–1926.
3. A. L. Hopkins, C. R. Groom, *Nat. Rev. Drug Discovery* 2002, 1, 727–730.
4. B. Hille, *Ionic Channels of Excitable Membranes*, 3rd ed., Sinauer Associates, Sunderland, MA 2001.
5. B. S. Jensen, L. Teuber, D. Strobaek, P. Christophersen, S. P. Olesen (Neurosearch A/S), WO 2000069794, 2000
6. D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, R. MacKinnon, *Science* 1998, 280, 69–77; (b) Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* 2003, 423, 33–41.
7. E. Jacoby, A. Schuffenhauer, P. Floersheim, *Drug News Perspect.* 2003, 16, 93–102; (b) K. Shokat, M. Velleca, *Drug Discov. Today* 2002, 7, 872–879.
8. E. Zitron, C. A. Karle, G. Wendt-Nordahl, S. Kathöfer, W. Zhang, D. Thomas, S. Weretka, J. Kiehn, *Br. J. Pharmacol.* 2002, 137, 221–228.
9. G. P. Brady, P. F. W. Stouten, *J. Comput. Aided Mol. Design* 2000, 14, 383–401.
10. G. Seeböhm, J. Chen, N. Strutz, C. Culberson, C. Lerche, M. C. Sanguinetti, *Mol. Pharmacol.* 2003, 64, 70–77.
11. G. Seeböhm, M. Pusch, J. Chen, M. C. Sanguinetti, *Circ. Res.* 2003, 93, 941–947.
12. G. Wess, M. Urmann, B. Sickenberger, *Angew. Chem.* 2001, 113, 3443–3453; (b) G. Wess, *Drug Discov. Today* 2002, 7, 533–535.
13. J. Drews, S. Ryser, *Nat. Biotechnol.* 1997, 15, 1318–1319; (b) J. Drews, *Science* 2000, 287, 1960–1964.

14. J. S. Mitcheson, *Br. J. Pharmacol.* 2003, 139, 883–884.
15. J. S. Mitcheson, J. Chen, M. Lin, C. Culberson, M. C. Sanguinetti, *Proc. Natl. Acad. Sci. USA* 2000, 97, 12329–12333.
16. J. T. Milnes, O. Crociani, A. Arcangeli, J. C. Hancox, H. J. Witchel, *Br. J. Pharmacol.* 2003, 139, 887–898.
17. J. Xu, X. Wang, B. Ensign, M. Li, L. Wu, A. Guia, J. Xu, *Drug Discov. Today* 2001, 6, 1278–1287.
18. K. Kamiya, J. S. Mitcheson, K. Yasui, I. Kodama, M. C. Sanguinetti, *Mol. Pharmacol.* 2001, 60, 244–253.
19. K. Urbahns, E. Horvath, J.-P. Stasch, F. Mauler, *Bioorg. Med. Chem. Lett.* 2003, 13, 2637–2639.
20. K.-H. Baringhaus, T. Klabunde, H. Matter, T. Naumann, B. Pirard, in *Molecular Informatics: Confronting Complexity, Proceedings of the International Beilstein Workshop May 2002* (Eds. M. G. Hicks, C. Kettner), Logos Verlag, Berlin 2003, pp. 167–178.
21. M. Hanner, B. Green, G. Ying-Duo, W. A. Schmalhofer, M. Matyskiela, D. J. Durand, J. P. Felix, L. Ana-Rosa, C. Bordallo, G. J. Kaczorowski, M. Kohler, M. L. Garcia, *Biochemistry* 2001, 40, 11687–11697.
22. N. Decher, B. Pirard, F. Bundis, S. Peukert, K.-H. Baringhaus, A. E. Busch, K. Steinmeyer, M. C. Sanguinetti, *J. Biol. Chem.* 2003, 278, 43564–43570.
23. N. Ogata, Y. Ohishi, *Jpn. J. Pharmacol.* 2002, 88, 365–377.
24. O. F. Güner, *Pharmacophore Perception, Development, and Use in Drug Design*, International University Line, La Jolla, CA 2000.
25. R. A. Pearlstein, R. Vaz, D. Rampe, *J. Med. Chem.* 2003, 46, 2017–2022 and references therein.
26. R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* 2002, 415, 287–294.
27. R. Dutzler, E. B. Campbell, R. MacKinnon, *Science* 2003, 300, 108–112.
28. S. A. Goldstein, D. Bockenbauer, I. O’Kelly, N. Zilberberg, *Nature Rev. Neuroscience* 2001, 2, 175–184.

29. S. Peukert, J. Brendel, B. Pirard, A. Brüggemann, P. Below, H.-W. Kleemann, H. Hemmerle, W. Schmidt, *J. Med. Chem.* 2003, 46, 486–498.
30. T. I. Oprea, A. M. Davis, S. J. Teague, P. D. Leeson, *J. Chem. Inf. Comput. Sci.* 2001, 41, 1308–1315.
31. Y. Jiang, A. Lee, J. Chen, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* 2002, 417, 515–522.
32. Y. Zhou, J. H. Morales-Cabral, A. Kaufman, R. MacKinnon, *Nature* 2001, 414, 43–48; (b) B. Roux, R. MacKinnon, *Science* 1999, 285, 100–102.



## PHOSPHODIESTERASE INHIBITORS: A CHEMOGENOMIC VIEW

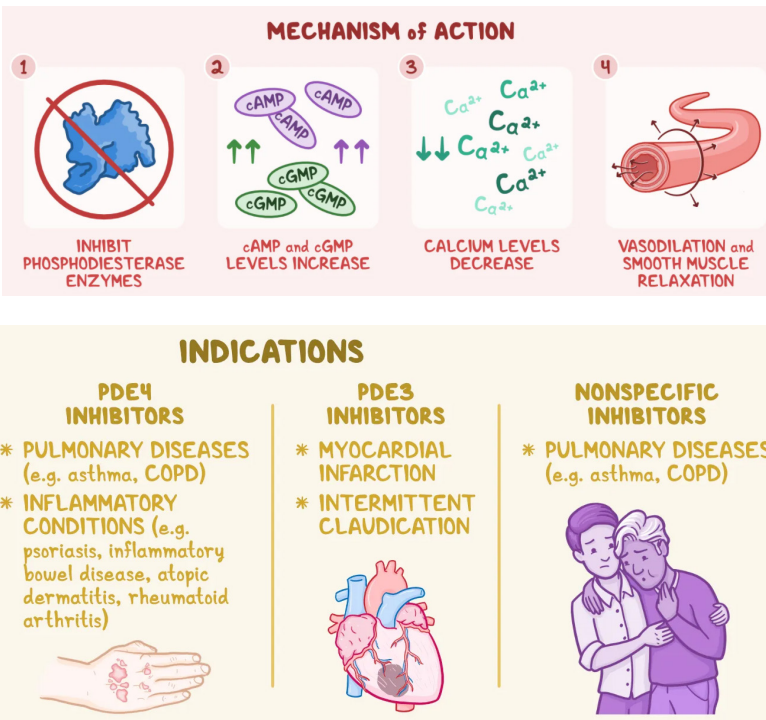
### INTRODUCTION

Phosphodiesterase inhibitors (PDE inhibitors) are a class of drugs that inhibit phosphodiesterase enzymes (PDE enzymes). PDE enzymes normally break off phosphate groups and decrease cAMP or cGMP in target cells. PDE inhibitors are classified according to which enzyme(s) they act upon as nonspecific, PDE5, PDE4, and PDE3 inhibitors. PDE5 inhibitors cause pulmonary vasodilation and penile smooth muscle relaxation, and are used for pulmonary hypertension and erectile dysfunction. PDE4 inhibitors enable bronchial dilation in severe COPD. PDE3 inhibitors have positive inotropic, vasodilator, and antiplatelet effects, which are used in acute heart failure and in peripheral vascular disease. PDE3 inhibitors are not recommended for long-term use in patients with heart failure because of their strong cardiostimulatory effects. Nitrates or alpha-blockers are strongly contraindicated in patients

taking PDE5 inhibitors because of the risk of life-threatening hypotension

# 6.1 BASICS OF PHOSPHODIESTERASE INHIBITORS

Phosphodiesterase inhibitors are a class of medications that promote blood vessel dilation (vasodilation) and smooth muscle relaxation in certain parts of the body, such as the heart, lungs, and genitals. Phosphodiesterases are a diverse family of enzymes that play a key role in regulating cell functions by indirectly increasing the intracellular levels of cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP), both of which are “second messengers” that regulate the primary effects of hormones and neurotransmitters.



| SIDE EFFECTS       |                 |                           |                        |
|--------------------|-----------------|---------------------------|------------------------|
| PDE5 INHIBITORS    | PDE4 INHIBITORS | PDE3 INHIBITORS           | NONSPECIFIC INHIBITORS |
| * HEADACHE         | * HEADACHES     | * VENTRICULAR ARRHYTHMIAS | * UPSET STOMACH        |
| * NASAL CONGESTION | * NAUSEA        | * HEADACHES               | * DIARRHEA             |
| * DYSPESIA         | * DIARRHEA      | * HYPOTENSION             | * HEADACHES            |
| * FLUSHING         |                 |                           | * RESTLESSNESS         |
| * PRIAPISM         |                 |                           |                        |

### 6.1.1 How do phosphodiesterase inhibitors work?

Phosphodiesterase inhibitors work by inhibiting the phosphodiesterase enzymes, thus preventing them from breaking down cAMP and cGMP molecules in the cell. The production of cAMP and cGMP are regulated by a molecule called nitric oxide, and their function is to help regulate physiological processes by decreasing the levels of calcium in the cell. Ultimately, cAMP and cGMP are broken down by phosphodiesterase enzymes.

When phosphodiesterase is inhibited, it is not able to break down the cAMP and cGMP. Thus, their levels inside the cell increase, which in turn leads to a decrease in the levels of calcium in the cell. Ultimately, this leads to vasodilation and smooth muscle relaxation in their target tissues.

### 6.1.2 Uses

Phosphodiesterase inhibitors are classified based on which specific phosphodiesterase enzyme they target. There are 11 families of phosphodiesterase enzymes and PDE inhibitors for each. Among these, the most widely used are four types of phosphodiesterase inhibitors: phosphodiesterase type 5 inhibitors (PDE5 inhibitor), phosphodiesterase type 4 inhibitors (PDE4 inhibitor), phosphodiesterase type 3 inhibitors (PDE3 inhibitor), and nonspecific inhibitors.

### ***PDE5 inhibitors***

PDE5 inhibitors work by increasing the levels of cGMP and specifically target the penis and the lungs. PDE5 inhibitors can be used to treat erectile dysfunction by inducing smooth muscle relaxation and increasing the blood flow to the penis, leading to an erection. Additionally, PDE5 inhibitors trigger pulmonary vasodilation, helping to regulate the pulmonary perfusion and pressure, thus they can be used to treat pulmonary hypertension when given at a lower dose compared to erectile dysfunction.

### ***PDE4 inhibitors***

PDE4 inhibitors work by increasing the levels of cAMP. They specifically target the airways, the skin and immune system, and the brain. PDE4 inhibitors work by causing smooth muscle relaxation in the airways, making them useful in the treatment of pulmonary diseases, such as asthma and chronic obstructive pulmonary disease. PDE4 inhibitors can also be used to treat inflammatory conditions that may affect the skin or other tissues, such as psoriasis, atopic dermatitis, inflammatory bowel disease, and rheumatoid arthritis. There is currently research being conducted on PDE4 inhibitors being used in the treatment of mental conditions, such as depression and anxiety.

### ***PDE3 inhibitors***

PDE3 inhibitors work by increasing the levels of cAMP. PDE3 inhibitors are typically used for cardiovascular diseases. In the heart, they help to increase cardiac contractility, or the ability of the heart to beat. They also relax vascular and airway smooth muscle, making them useful in the treatment of heart failure. In addition, PDE3 inhibitors can prevent platelet aggregation into clots, and can thus be used to prevent and treat myocardial infarction (heart attack). Finally, PDE3 inhibitors can trigger vasodilation of peripheral blood vessels and can be used to treat intermittent claudication, which is a cramping in the legs due to a decreased



blood flow.

### ***Nonspecific inhibitors***

Nonspecific inhibitors work by decreasing the destruction of the cAMP by any phosphodiesterase enzyme. They mainly induce mild dilation of the bronchioles of the lungs and help reduce airway inflammation. Nonspecific phosphodiesterase inhibitors are used in the treatment of chronic obstructive pulmonary disease, as well as for short term and long term management of asthma.

#### **6.1.3 Some Examples of Common Phosphodiesterase Inhibitors**

PDE5 inhibitors are the most common and include sildenafil, tadalafil, vardenafil, and avanafil.

The most common PDE4 inhibitors are roflumilast, apremilast, and ibudilast.

Some examples of PDE3 inhibitors are cilostazol and milrinone.

Nonspecific phosphodiesterase inhibitors include theophylline, aminophylline, and methylxanthine.

Viagra is a PDE5 inhibitor most often used for the treatment of erectile dysfunction. Viagra has also been useful in the treatment of pulmonary hypertension for both short term and long term use when given at a lower dose compared to the dosages prescribed for erectile dysfunction. Viagra has a low number of side effects and is considered relatively safe for individuals with heart disease. However, it's important to note that the use of Viagra taken along with nitrates can lead to a reduction in blood pressure and is contraindicated.

Phosphodiesterase type 5 inhibitors are competitive and reversible inhibitors. When a PDE5 inhibitor is used, it competitively binds to PDE5 to stop it from breaking down cGMP. PDE5 inhibitors are considered reversible because they bind to PDE5 for a limited

amount of time. With the decrease in the levels of the cGMP, this will reverse the smooth muscle relaxation in the penis, which will end the erection.

Caffeine is a phosphodiesterase inhibitor that has been shown to increase the levels of cAMP in the cell thus leading to smooth muscle relaxation. Caffeine is a weak inhibitor, but variations of caffeine including theophylline have been introduced as treatments for pulmonary disease.

Methylxanthines are among the first phosphodiesterase inhibitors to be discovered, and are nonspecific. The most common methylxanthine is theophylline, which is commonly used for short or long term treatment of different types of pulmonary diseases, such as asthma. Methylxanthines can help to induce bronchodilation and reduce airway inflammation.

#### **6.1.4 Side Effects of Phosphodiesterase Inhibitors**

Common side effects of PDE5 inhibitors include headache, nasal congestion, dyspepsia and flushing. A potential rare side effect of PDE5 inhibitors is priapism, or an erection lasting longer than 4 hours. Priapism is a medical emergency that requires immediate intervention. First line treatment is oral terbutaline or pseudoephedrine. If the priapism persists, needle aspiration of the blood in the penis, as well as an intracavernous injection of phenylephrine, may be needed.

Potential side effects of PDE4 inhibitors are headaches, nausea, and diarrhea.

The most common side effects of PDE3 inhibitors include ventricular arrhythmias, headaches, and hypotension.

Side effects of nonspecific inhibitors include upset stomach, diarrhea, headaches, and restlessness.

### 6.1.5 Important Facts to Know About Phosphodiesterase Inhibitors

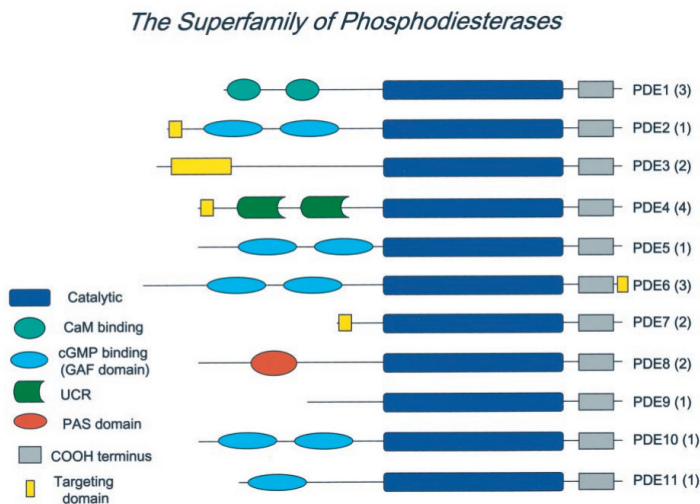
Phosphodiesterase inhibitors prevent the phosphodiesterase enzymes from breaking down cAMP and cGMP in the cell. As a result, they increase the cAMP and cGMP, leading to an increase in intracellular calcium, which causes vasodilation and smooth muscle relaxation. There are four types of phosphodiesterase inhibitors, which have an effect on different locations in the body, depending on the specific phosphodiesterase enzyme they target. PDE5 inhibitors are commonly used for the treatment of erectile dysfunction and pulmonary hypertension. PDE4 inhibitors can be used to treat asthma, chronic obstructive pulmonary disease, psoriasis, atopic dermatitis, inflammatory bowel disease, and rheumatoid arthritis. PDE3 inhibitors are indicated for the treatment of heart failure, coronary heart disease, and to prevent myocardial infarction. Finally, nonspecific phosphodiesterase inhibitors are used in the treatment of asthma and chronic obstructive pulmonary disease.

## 6.2 MODULAR STRUCTURE OF PDES

### 6.2.1 The PDE Superfamily

That more than one PDE protein degrades cyclic nucleotides in the cell was appreciated soon after PDEs were discovered. Different chromatographic PDE forms with clearly distinct kinetics, substrate specificity, and pharmacological properties were demonstrated in extracts from brain and other tissues. A more in-depth understanding of the PDE complexity in mammalian cells has come with the cloning and identification of the different PDE genes. The field is evolving rapidly, and new families and genes have been recently added to the PDE superfamily. The latest members have been identified taking advantage of the human genome project by using homologous searches of EST databases.

At last count, 21 genes coding for cyclic nucleotide PDEs have been identified in mammals (Fig. 1). Using the most widely accepted nomenclature, the PDE family is indicated by an Arabic numeral, followed by a capital letter indicating the gene within a family, and a second Arabic numeral indicating the splicing variant derived from a single gene (for example PDE1C3: family 1, gene C, splicing variant 3).



**Figure 1.** Modular Structure of the PDEs: Schematic Representation of the Domain Arrangement in Members of the 11 Families of PDEs The *number in parentheses* next to the gene family indicates the number of known genes. In the PDE6 families, only the genes coding for catalytic subunits are reported.

## 6.2.2 Structure/Function of PDEs

With the exception of a yeast and a *Dictiostelium* PDE, which may belong to a different family, all PDEs have a conserved region that corresponds to the catalytic domain (Fig. 1). This domain is structurally related to other metal-dependent phosphohydrolases with conserved HD motif, pointing to the important role of divalent cations in cyclic nucleotide hydrolysis. In addition to the signature sequence AxxHDxDHxG identified by sequence comparison, a

number of conserved residues have been identified within this catalytic domain, and their mutagenesis often impacts catalysis. Spontaneous mutations in PDE6 within the catalytic domain and their association with major defects of the retina further support the importance of this domain in the enzyme function. The recent release of the crystal structure of the catalytic domain has reconciled the many site-directed mutagenesis studies done on this region of the PDEs. The PDE4 catalytic domain is a compact structure composed of 17  $\alpha$ -helices divided into three subdomains, with the most conserved residues involved in the formation of the catalytic pocket. More importantly, this structural analysis has confirmed the presence of metal ions in the catalytic pocket, and it will certainly provide conclusive answers regarding the exact interactions of this domain with cyclic nucleotide substrates and model inhibitors.

The arrangement of domains around this catalytic core, as well as the presence of a large number of splicing variants, points to a modular structure of PDEs. Several domains with a variety of proven or putative functions have been identified at the amino terminus portion of most PDE forms (Fig. 1) and are distinctive characteristics of each family. These include protein-protein interaction domains as well as domains that bind small molecules such as cyclic nucleotides. In addition, phosphorylation domains that control the catalytic function have been mapped at the amino terminus of most PDEs. Domains present at the carboxyl terminus of PDEs may be involved in dimerization, as has been suggested for PDE4 and PDE1, or may function as a regulatory domain being a target for phosphorylation.

Although viewed by some as an oversimplification, it is probable that all these different domains regulate PDE catalysis by a common mechanism. The regulatory domains that flank the catalytic domains function as a sensor of intracellular signals. Reception of these signals produces a change in conformation of the PDE so that an inhibitory domain no longer exerts a negative constraint on the catalysis. The presence of an inhibitory domain is inferred by biochemical studies with controlled proteolysis of

PDE1, PDE2, and PDE4 and deletion mutagenesis of PDE1, PDE3, PDE4, and possibly PDE7. Moreover, the regulation of PDE6 by the inhibitory  $\gamma$ -subunit again points to the important role of inhibitory constraint in PDE catalysis. Along the same line, PDE4s have two unique modules that are conserved in the four genes that compose this family. On the basis of their conservation, they have been named upstream conserved region 1 and 2 (UCR1 and UCR2). Functionally, the UCR2 conserved domain corresponds to an autoinhibitory domain that negatively regulates the catalytic activity, while regulatory phosphorylation sites have been mapped in the UCR1. A model for the interactions between these regions has been developed on the basis of domain binding and regulation of catalysis. While this model has been recently confirmed by others, it remains to be determined to what extent it may be applied to PDEs that belong to other families.

### 6.2.3 Significance of the PDE Complexity

Domain shuffling may explain why a large number of PDE variants with divergent amino and carboxyl termini have been identified. As an example, the PDE4D gene encodes five well characterized splicing variants, a property that is inherited from *Drosophila*. These variants are generated by alternate splicing and/or different promoter usage. More importantly, these variants are subjected to different regulations (see below) or are targeted to different subcellular compartments. Targeting domains have been identified in most PDEs. PDE3s have a domain that includes six transmembrane hydrophobic helices, which target them to the endoplasmic reticulum. This domain is absent in some soluble PDE3 splicing variants. Two variants with soluble and particulate distribution have been described for PDE2 and PDE7. A domain that interacts with RACK1, a scaffold protein that binds activated protein kinase C (PKC) isoforms, was identified at the amino terminus of one of the five PDE4D variants, and putative SH3 interacting domains have been reported for one PDE4A and one

PDE4D variant. A scaffold protein that anchors PDE4D to the Golgi/centrosome structures in the vicinity of PKAs has also been reported. Although the physiological impact of this differential targeting is largely unknown, these findings lend support to the idea that PDE subcellular targeting may have a role in signal compartmentalization.

The divergent properties and the presence of distinct regulatory domains in PDEs explain, at least in part, the heterogeneity of the PDE superfamily. A cell utilizes PDEs with different properties and regulations to adapt to the large variety of signals to which it is exposed, to control cyclic nucleotide accumulation in different subcellular compartments, and to integrate different signaling pathways. More difficult to understand is why multiple genes are present within a family because the corresponding proteins have largely overlapping properties and regulations. Recently, it has been shown that inactivation of only one of the four PDE4 genes present in the mouse produces profound and unexpected phenotypes, suggesting that the functions of different PDE genes do not overlap. In addition, *in situ* studies on PDE mRNA expression in brain have uncovered some specificity in the expression of genes belonging to the same family. PDE4B, for instance, is expressed in the granular layer of the cerebellum, while PDE4D is expressed in the Purkinje cells. Thus, it is also possible that gene duplication may have occurred to increase the control of PDE expression in a tissue- and developmental-specific fashion.

### 6.3 MECHANISMS OF REGULATION OF PDES

A wide range of regulations to control cyclic nucleotide hydrolysis are operating in the cell, completely refuting the initial idea that PDEs are housekeeping enzymes with a passive role in signaling. The nature of the stimuli that modulate PDE activity is diverse, ranging from posttranslational modification and binding of small ligands to protein-protein interaction.



### 6.3.1 Protein-Protein Interaction and PDE Function

PDE1s were among the first targets for calmodulin (CaM) to be identified, and activation of cGMP hydrolysis has been used as a CaM bioassay for more than 20 yr. CaM binding modules, which consist of a basic amphipathic helix, have been identified by protein homology and by deletion mutagenesis in all proteins derived from the three PDE1 genes. Several pieces of evidence indicate that the CaM binding domain affects the catalytic domain indirectly by controlling the interaction of an autoinhibitory domain with the catalytic domain. Using deletion mutagenesis, this autoinhibitory domain has been mapped to a region between the two CaM binding sites in PDE1A1.  $\text{Ca}^{++}$  and CaM produce a major increase in PDE activity, suggesting that the enzyme may be completely inactive in the absence of  $\text{Ca}^{++}$ . Of interest is the fact that the affinity for  $\text{Ca}^{++}$ /CaM is different between the different PDE1 proteins. The PDE1A gene encodes two splicing variants, PDE1A1 and PDE1A2. CaM is 10 times more potent in activating A1 than A2, indicating that splicing is a means to regulate sensitivity to  $\text{Ca}^{++}$  and CaM. Additional data comparing isoenzymes from brain, heart, and lung have shown differences in the affinity of PDE1B and PDE1C for CaM. CaM binding is also regulated by PDE1 phosphorylation (see below).

The role of PDEs in light perception underscores the importance of PDEs in signaling. We are able to sense visual cues because light causes a dramatic decrease in cGMP in the retina via activation of a PDE. In this pathway, light-activated rhodopsin interacts with the G protein transducin that, in turn, activates PDE6, which hydrolyzes cone and rod cGMP. The decrease in cGMP results in closure of cGMP-gated channels in the membrane, thus causing hyperpolarization. The PDE6 expressed in the retina is a tetramer composed of two distinct  $\alpha$ - and  $\beta$ -subunits and two  $\gamma$ -subunits in rods, with a slightly different  $\alpha$  2-dimer expressed in cones. Two  $\gamma$ -subunits bind the function as inhibitors of the cGMP hydrolytic activity of the  $\alpha$ - and  $\beta$ -subunits. In addition, a  $\delta$ -subunit copurifies with PDE6 and may play a role in the membrane association of this PDE. The site of interaction of the  $\gamma$ -subunit on the  $\alpha\beta$ -subunits



has been identified by several laboratories and mapped to regions surrounding the catalytic domain. This interaction completely suppresses cGMP hydrolysis. Transducin controls the interaction between  $\gamma$ - and  $\alpha\beta$ -subunits, but it is unclear whether it interacts directly with the  $\gamma$ -subunits or with the  $\alpha\beta$ -catalytic subunits.

PDE5, which is widely expressed in tissues outside the retina, has considerable homology with PDE6. Hence, the idea has been put forward that the activity of PDE5 may be regulated by a homolog of the  $\gamma$ -subunit. Indeed, there are reports suggesting that proteins immunologically related to the retina  $\gamma$ -subunit are expressed outside the retina. The exact function of these novel proteins remains to be determined. Other sensory cues may use membrane signal transduction machinery involving G protein interaction with a PDE, as suggested for the taste buds.

### 6.3.2 Cyclic Nucleotide and Other Allosteric Regulations of PDEs

The regulation of PDEs through allosteric binding of cyclic nucleotides was discovered in the 1970s. PDE2 binds cGMP with an affinity of approximately 100 nM and produces an allosteric change in the catalytic domain. Because of this allosteric regulation, the enzyme hydrolyzes both cAMP and cGMP with positive cooperative kinetics. However, in the intact cell this enzyme probably functions as a cGMP-stimulated cAMP PDE. This property allows integration of the cGMP- and cAMP-regulated pathways, as suggested for atrial natriuretic factor (ANF) signaling.

Structurally related cGMP binding domains have been identified in PDE5, PDE6, PDE9, and PDE10. In PDE5, occupancy of this site may modulate the ability of the enzyme to be phosphorylated by protein kinase G. In PDE6, cGMP binding regulates the affinity of  $\alpha\beta$ -dimers for the inhibitory  $\gamma$ -subunit, while little is known about the role of cGMP binding in PDE9 and PDE10. Similar domains have been found in proteins other than PDEs and have been termed GAF domains (cGMP-specific and cGMP-stimulated

PDEs, *Anabaena* adenylyl cyclase and *Escherichia coli* Fh1A). The presence of these domains in species where cGMP is not produced has led to the recent proposal that the GAF domain in PDEs may not serve to bind cGMP but is involved in interactions with other unknown small ligands.

On the basis of sequence homology with a domain found in proteins from bacteria to eukariots, a PAS (Period, Arnt, Sim) domain was identified in PDE8. This domain functions as a signal detector and is usually associated with a heme or a chromophore cofactor. In archaea, the PAS domain of FixL is a sensor for oxygen or possibly nitric oxide. Although the function of the PAS domain in PDE8 is not known, it may be important for protein-protein interaction or for sensing concentrations of a small ligand, suggesting a novel mode of regulation for PDEs.

### 6.3.3 Posttranslational Modification

There are now reports demonstrating phosphorylation of PDE1, PDE3, PDE4, PDE5, PDE6, and possibly PDE7. With some rare exceptions, phosphorylation occurs on regulatory domains present at the amino terminus of the PDE protein. Several kinases including PKA, protein kinase B (PKB), mitogen-activated protein kinase (MAPK), and calmodulin kinase (CaMK) catalyze these regulatory phosphorylations (see below).

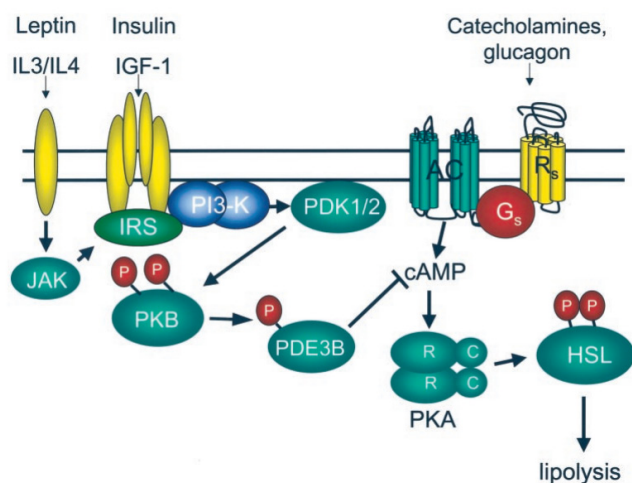
### 6.3.4 Signaling Cascades Involving PDEs in Endocrine Cells

Here we will focus on PDE regulations that have the greatest impact on endocrine systems. In several instances a PDE serves as a connection between two different pathways allowing signal integration. Some of these regulations have been extensively reviewed elsewhere in the context of cardiovascular or central

nervous system functions, PDE1 being an example of a  $\text{Ca}^{++}$  signal regulating cAMP and cGMP concentration.

### ***The PI-3 Kinase Pathway and Activation of PDE3***

Insulin and IGF-I binding activates the receptor tyrosine kinase with phosphorylation and recruitment of adapter proteins, including insulin receptor substrate 1–4 (IRS1–4). Once phosphorylated, these adapters recruit several effectors including the lipid phosphatidylinositol 3 kinase (PI-3K). The phosphatidyl triphosphate lipid formed serves as an anchor and recruits to the membrane the kinase PDK1/2 and downstream kinase PKB/AKT (Fig. 2). There is now ample evidence that PDE3s integrate this PI-3K signaling cascade with the cyclic nucleotide-regulated pathway. A large number of observations are consistent with the presence of this signaling cascade in the cell. Activation of PI-3 kinase by insulin is associated with an increase in PDE3 activity in adipocytes, and the PDE3 activation is blocked by wortmannin and LY 294002, both of which are inhibitors of PI-3 kinases. In addition, insulin treatment causes the incorporation of  $^{32}\text{P}$ -phosphate in PDE3B. That PDE3B is directly phosphorylated by PKB is demonstrated by cell-free experiments with recombinant proteins. Two possible phosphorylation sites have been identified in PDE3B. Ser302 of rat PDE3B was identified by phosphopeptide mapping of PDE3B from insulin-stimulated cells. Conversely, site-directed mutagenesis has indicated Ser273 as the major site of PKB phosphorylation. While the sequence surrounding Ser273 conforms with the consensus for PKB phosphorylation, Ser302 is an anomalous site because it is also phosphorylated by PKA. Whether both sites are used in a cell-specific fashion is unclear and requires further experimentation. Recently, Rondinone *et al.* have proposed that an additional mechanism of PDE3 activation by insulin may directly involve phosphorylation of PDE3B by the PI-3K associated with the insulin receptor.



**Figure 2.** Signaling Pathways That Control PDE3 and Lipolysis in the Cell.

The PI-3K-PKB-PDE3B signaling module has important physiological implications because it is used in several endocrine regulations and in growth factor control of the entry and exit from the cell cycle. For example, hormone-sensitive lipase (HSL) is the enzyme that catalyzes the hydrolysis of triglycerides stored in adipose tissue and is thought to be the rate-limiting enzyme for the mobilization of FFA. The activity of this enzyme is under the control of catecholamines and other lipolytic hormones that stimulate cAMP accumulation (Fig. 2). The PKA activation that follows an increase in cAMP causes activation of HSL by phosphorylation on one or more sites. In adipocytes, insulin inhibition of lipolysis is mediated by a decrease in cAMP levels and is associated with a decreased phosphorylation of HSL. Both *in vitro* and *in vivo* insulin effects are blocked by specific PDE3 inhibitors pointing to an important role of this PDE. Moreover, the phosphorylation and activation of PDE3B correlates with the inactivation of PKA and the dephosphorylation of HSL. Thus, PDE3 phosphorylation appears to be a crucial step in mediating the effect of insulin on lipid metabolism. The antiglycogenolytic effects of insulin may also be mediated, at least in part, by the same pathway involving a PDE3.

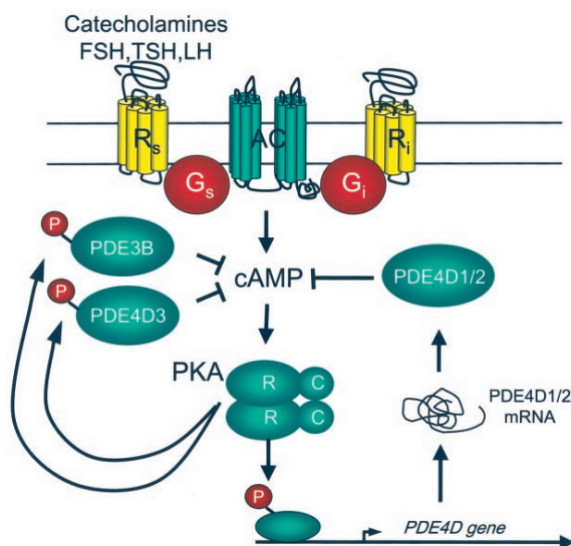
A similar PI-3K, PKB, and PDE3B cascade is activated by leptin (OB), a recently discovered hormone involved in the control of fat metabolism and food intake. The peripheral effects of leptin are mediated by the activation of receptors that are structurally related to the cytokine receptors which signal through the janus kinases. This kinase, in turn, phosphorylates IRS-1 and IRS-2 promoting the recruitment and activation of PI-3 kinase (Fig. 2). Similar to what has been shown for insulin, PI-3 kinase activation causes PKB and PDE3B activation. The resulting decrease in cAMP mediates the antiglycogenolytic effects of leptin in hepatocytes.

Because IGF-I shares the same signaling pathway with insulin, IGF-I regulation of PDE3 may be important in the regulation of cell entry and exit from the cell cycle. In *Xenopus* eggs, a PDE with the properties of PDE3 is activated by AKT and is an important step in insulin-like growth factor (IGF)-induced resumption of meiosis. A PDE3A is the predominant form that is also expressed in mammalian oocytes, and inhibition of this enzyme blocks the resumption of meiosis that follows the gonadotropin stimulation *in vitro* and *in vivo*. With the same signaling cascade, IGF-I regulates insulin secretion in islet  $\beta$ -cells by regulating a PDEB. In general, it is expected that all the growth factor pathways that use PI-3K may use the PDE3 activation to regulate cAMP levels. This could provide a means to modulate the gating effects of cAMP on the mitogen-activated protein (MAP) kinases signaling pathway, and to control exit and entry from the cell cycle.

### ***PDEs as Homeostatic Regulators***

Manipulation of hepatocytes with nonhydrolyzable cAMP analogs demonstrated that a rapid feedback controlling cAMP is operating in these cells. Accumulation of the cAMP analog in the cells activates PKA, which in turn activates a PDE. The ultimate result is a decrease in endogenous cAMP levels. With the discovery that a PDE3 is a substrate for PKA, it was proposed that this PDE is involved in these feedback mechanisms. More recently, data in thyroid cells also have shown that PDE4 is activated by hormones that increase cAMP via a PKA-dependent mechanism. PKA

phosphorylates PDE4D3, one of the variants derived from the PDE4D genes, and phosphorylation is associated with an increase in PDE activity. The residues phosphorylated by PKA have been mapped by site-directed mutagenesis to the amino terminus of PDE4D3. These observations have been confirmed and extended by demonstrating that introduction of a negatively charged amino acid in position 54 produces an activated enzyme. Thus, PDE3 and PDE4 are both rapidly activated by PKA, depending on the cell in which they are expressed (Fig. 3).



**Figure 3.** Feedback Regulation of PDE3 and PDE4 and Hormone Responsiveness.

The presence of the PKA-PDE4D feedback loop in intact cells has been recently confirmed by demonstrating that PKA inhibitors block the PDE4D3 phosphorylation/activation and cause a potentiation of TSH-dependent cAMP signaling. In a similar fashion, a stable cell line expressing PDE4D3 produces a major change in cAMP accumulation stimulated by hormones, whereas cell lines expressing a phosphorylation-deficient PDE4D3 have normal responses. Interestingly, it was observed that inactivating this feedback regulation has a major effect on the intensity of the

cAMP spike without any major change in the duration of the signal.

During the characterization of the mechanisms causing desensitization, it was observed that an increase in cyclic nucleotides produces an increase in PDE activity. This activation was thought to be a mechanism of desensitization that cooperated with receptor/G protein uncoupling. With the limited knowledge of PDE heterogeneity available at that time, little was known about the PDE involved except that the enzyme was a cAMP-specific PDE and that the regulation required protein synthesis and PKA activation. A better understanding of this second feedback mechanism has come with the cloning of the PDE4 genes. Stimulation with hormones that increase cAMP invariably produces an increase in PDE4 mRNA and *de novo* synthesis of PDE4 proteins. This is most evident for the PDE4D gene. Long-term FSH stimulation of Sertoli cells causes more than a 100-fold increase in PDE4D mRNA, the accumulation of PDE4D1/D2 variants, and more than a 10-fold increase in PDE activity. Identical induction has been observed in most cells, suggesting that this is a ubiquitous feedback regulation of cAMP, thus providing a mechanistic explanation of the early findings on long-term, cycloheximide-sensitive PDE activation (Fig. 3).

The impact of the PDE feedback on cAMP signaling is emerging from studies on PDE4D knockout mice. PDE4D-null mice display a 30–40% decrease in growth rate during puberty, and the adults are usually smaller than their littermates. The decreased growth is associated with a decrease in circulating IGF-I levels, suggesting a disruption of the GH-IGF-I axis. In addition, the homozygous PDE4D-null females display reduced fertility with litter size approximately one-third of normal. This reduction in fertility is associated with a 70–80% decrease in ovulation rate compared with wild-type littermates. Surprisingly, when the sensitivity to gonadotropin is measured in granulosa cells from the PDE4D-null mice, a decrease in responsiveness to hCG was observed. This decreased response is difficult to reconcile with the common tenet that PDE inhibition leads to an increased cAMP accumulation and



cAMP signaling. Pending additional experiments to clarify the exact cause of this decreased response, we propose that inactivation of the PDE4D-PKA feedback loop causes a desensitization of the cAMP signaling pathway at the level of receptor/G protein. Thus, PDE4 regulation allows appropriate cAMP signaling by protecting from desensitization. If confirmed, this concept may have important implications in human diseases as end-organ resistance may be associated with inherited PDE4 inactivating mutations.

Disruption of PDE4D expression also affects muscarinic cholinergic responses in the airway. Mice deficient in PDE4D do not respond to methacholine with an increase in airway resistance, in spite of a normal complement of muscarinic cholinergic receptors. This phenotype may be caused by an increase in sensitivity to noradrenaline, which causes relaxation of smooth muscle cells. Alternatively, PDEs may directly play a role in M3 muscarinic receptor signaling that mediates the contractile response of acetylcholine, again supporting a role for PDE4D as a homeostatic regulator of signaling.

The PDE4 feedback loop may have an important impact under those pathological conditions in which cAMP accumulation is deregulated. Mutations in  $G_{sa}$  produce a constitutively active protein that maintains adenylyl cyclase in a chronically activated state. These mutations are responsible for the phenotype of patients with McCune-Albright syndrome and are probable causes of a number of adenomas of the pituitary and thyroid. There is also abundant literature for constitutive activation of pituitary hormone receptors that cause chronic cAMP elevation. All these spontaneous mutations cause a marked induction of PDE4 and possibly other PDEs both *in vitro* and *in vivo*. The PDE4 activation must have an impact on growth as the proliferative effects of the  $G_{sa}$  mutations are seen *in vitro* only after inhibition of PDE4. Therefore, it is possible that the abnormal growth induced by  $G_{sa}$  or receptor mutations is modified by the presence of different PDE alleles. Ongoing experiments will determine whether polymorphisms or mutations of a PDE4 exist in humans.

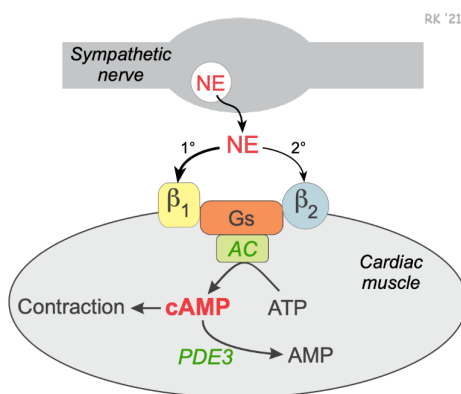


## 6.4 GENERAL PHARMACOLOGY OF CAMP-DEPENDENT PHOSPHODIESTERASE INHIBITORS

### 6.4.1 In PDE3 Inhibitors

#### *Heart*

Intracellular concentrations of cAMP play an important second messenger role in regulating cardiac muscle contraction. Activation of the sympathetic nervous system releases the neurotransmitter norepinephrine and increases circulating catecholamines (epinephrine and norepinephrine). These catecholamines bind primarily to  $\beta_1$ -adrenoceptors in the heart that are coupled to Gs-proteins. This activates adenylyl cyclase to form cAMP from ATP.



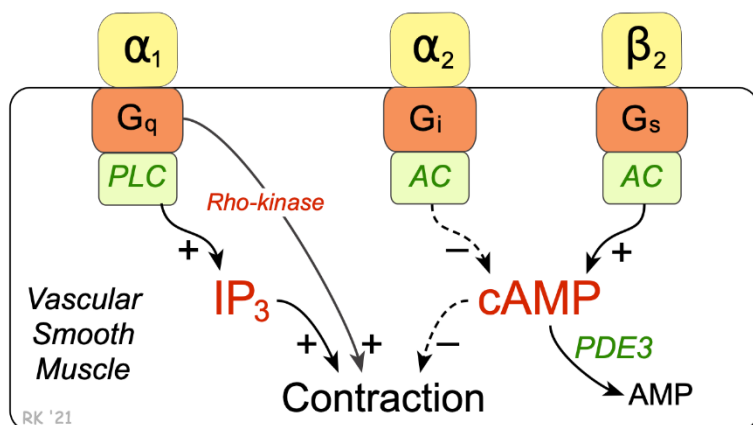
PDE3, cAMP-dependent phosphodiesterase (type 3)

Increased cAMP, through its coupling with other intracellular messengers, increases contractility (inotropy), heart rate (chronotropy) and conduction velocity (dromotropy). Cyclic-AMP is broken down by an enzyme called **cAMP-dependent phosphodiesterase (PDE)**. The isoform of this enzyme that is targeted by currently used clinical drugs is the type 3 form

(PDE3). Inhibition of this enzyme prevents cAMP breakdown and thereby increases its intracellular concentration. This increases cardiac inotropy, chronotropy and dromotropy. PDE3 inhibitors can be thought of as a backdoor approach to cardiac stimulation, whereas  $\beta$ -agonists go through the front door to produce the same cardiac effects.

## Blood vessels

Cyclic-AMP also plays an important role in regulating the contraction of vascular smooth muscle. Beta<sub>2</sub>-adrenoceptor agonists such as epinephrine stimulate the G<sub>s</sub>-protein and the formation of cAMP. Unlike cardiac muscle, increased cAMP in smooth muscle causes relaxation.



G<sub>q</sub>, G<sub>q</sub>-protein; G<sub>i</sub>, G<sub>i</sub>-protein; G<sub>s</sub>, G<sub>s</sub>-protein; PLC, phospholipase C; AC, adenylyl cyclase; IP<sub>3</sub>, inositol-triphosphate; cAMP, cyclic AMP; PDE3, cAMP-dependent phosphodiesterase (type 3)

The reason for this is that cAMP normally inhibits myosin light chain kinase, the enzyme that is responsible for phosphorylating smooth muscle myosin and causing contraction. Like the heart, the cAMP is broken down by a cAMP-dependent PDE (PDE3). Therefore, inhibition of this enzyme increases intracellular cAMP, which further inhibits myosin light chain kinase thereby producing less contractile force (i.e., promoting relaxation).

## ***Cardiovascular Actions of cAMP-dependent PDE (type3) Inhibitors***

### Systemic Circulation

- Vasodilation
- Increased organ perfusion
- Decreased systemic vascular resistance
- Decreased arterial pressure

### Cardiopulmonary

- Increased contractility and heart rate
- Increased stroke volume and ejection fraction
- Decreased ventricular preload (secondary to increased output)
- Decreased pulmonary capillary wedge pressure

## ***Overall cardiovascular effects***

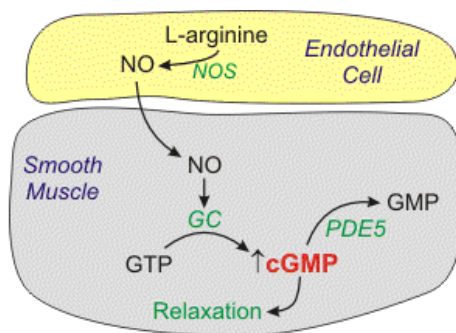
The cardiac and vascular effects of cAMP-dependent PDE inhibitors cause cardiac stimulation, which increases cardiac output, and reduced systemic vascular resistance, which tends to lower arterial pressure. Because cardiac output increases and systemic vascular resistance decreases, the change in arterial pressure depends on the relative effects of the PDE inhibitor on the heart versus the vasculature. At normal therapeutic doses, PDE3 inhibitors such as milrinone have a greater vascular than cardiac effect so that arterial pressure is lowered in the presence of augmented cardiac output. Because of the dual cardiac and vascular effects of these compounds, they are sometimes referred to as “inodilators.”

## ***Other actions***

PDE3 inhibitors also decrease platelet aggregation by increasing platelet cAMP. However, only cilostazol (see below) is used for this purpose in the treatment of intermittent claudication (ischemic leg pain associated with leg movement).

### 6.4.2 In PDE5 inhibitors

There is a second isoenzyme form of PDE in vascular smooth muscle that is a cGMP-dependent phosphodiesterase. The type 5 isoform of this enzyme (PDE5) is found in the corpus cavernosum of the penis and in vascular smooth muscle. This enzyme is responsible for breaking down cGMP that forms in response to increased nitric oxide (NO). Increased intracellular cGMP inhibits calcium entry into the cell, thereby decreasing intracellular calcium concentrations and causing smooth muscle relaxation.



Abbreviations: NO, nitric oxide; NOS, nitric oxide synthase; GC, guanylyl cyclase; PDE5, cGMP-dependent phosphodiesterase (type 5)

NO also activates  $K^+$  channels, which leads to hyperpolarization and relaxation. Finally, NO acting through cGMP can stimulate a cGMP-dependent protein kinase that activates myosin light chain phosphatase, the enzyme that dephosphorylates myosin light chains, which leads to relaxation. Therefore, inhibitors cGMP-dependent phosphodiesterase, by increasing intracellular cGMP, enhance smooth muscle relaxation and vasodilation, and cause penile erection.

### 6.4.3 Therapeutic Indications

The cardiostimulatory and vasodilatory actions of PDE3 inhibitors make them suitable for the treatment of heart failure. Arterial dilation reduces afterload on the failing ventricle and leads to an increase in stroke volume and ejection fraction, as well as increases

organ perfusion. Reducing the afterload leads to a secondary decrease in preload on the heart that helps to improve the mechanical efficiency of dilated hearts and to reduce ventricular wall stress and the oxygen demands placed on the failing heart. The cardiostimulatory effects of these drugs increase inotropy, which further enhances stroke volume and ejection fraction. Tachycardia, however, also results, and this is not beneficial; therefore, doses are used that minimize the positive chronotropic actions of the drug. A baroreceptor reflex, which occurs in response to hypotension, may contribute to the tachycardia. Clinical trials have shown that long-term therapy with PDE3 inhibitors increases mortality in heart failure patients; therefore, these drugs are not used for long-term, chronic therapy. They are very useful, however, in treating acute, decompensated heart failure or temporary bouts of decompensated chronic failure. They are not used as a monotherapy. Instead, they are used in conjunction with other treatment modalities such as diuretics, ACE inhibitors, beta-blockers or digitalis.

The somewhat selective vasodilatory actions of PDE5 inhibitors have made these compounds very useful in the treatment of male erectile dysfunction. The PDE5 inhibitor sildenafil is also approved for the treatment of pulmonary hypertension.

#### 6.4.4 Specific Drugs

Several different PDE inhibitors are available for clinical use:

- PDE3 inhibitors
  - milrinone
  - inamrinone (formerly amrinone)
  - cilostazol
- PDE5 inhibitors
  - sildenafil
  - tadalafil

The PDE3 inhibitors (except cilostazol) are used for treating acute, decompensated heart failure, whereas the PDE5 inhibitors

are used for treating male erectile dysfunction and pulmonary hypertension. *Note that the PDE3 inhibitors used in acute heart failure end in “one,” whereas the PDE5 inhibitors end in “fil”.*

Inhibition of platelet aggregation, along with vasodilation, is an important mechanism of action for cilostazol, which is used in the treatment of intermittent claudication in peripheral arterial disease. Cilostazol appears to have less cardiostimulatory effects than milrinone.

## 6.4.5 Side Effects and Contraindications

### *PDE3 inhibitors*

Milrinone and inamrinone are not used in the treatment of chronic heart failure because clinical trials have shown that long-term use of these drugs worsen outcome. The most common and severe side effect of PDE3 inhibitors is ventricular arrhythmias in about 12% of patients, some of which may be life-threatening. Headaches and hypotension occur in about 3% of patients. These side effects are not uncommon for drugs that increase cAMP in cardiac and vascular tissues, other examples being  $\beta$ -agonists.

### *PDE5 inhibitors*

The most common side effects for PDE5 inhibitors include headache and cutaneous flushing, both of which are related to vascular dilation caused by increased vascular cGMP. There is clinical evidence that nitrodilators may interact adversely with PDE5 inhibitors. The reason for this adverse reaction is that nitrodilators stimulate cGMP production while PDE5 inhibitors inhibit cGMP degradation. When combined, these two drug classes greatly potentiate cGMP levels, which can lead to hypotension and impaired coronary perfusion.

## 6.5 PDE6 INHIBITORS

Phosphodiesterase 6 (PDE6) is highly concentrated in the retina. It is most abundant in the internal membranes of retinal photoreceptors, where it reduces cytoplasmic levels of cyclic guanosine monophosphate (cGMP) in rod and cone outer segments in response to light. The rod PDE6 holoenzyme comprises  $\alpha$  and  $\beta$  catalytic subunits and two identical inhibitory  $\gamma$  subunits. Each catalytic subunit contains three distinct globular domains corresponding to the catalytic domain and two GAF domains (responsible for allosteric cGMP binding). The PDE6 catalytic subunits resemble PDE5 in amino-acid sequence as well as in three-dimensional structure of the catalytic dimer; preference for cGMP over cyclic adenosine monophosphate (cAMP) as a substrate; and the ability to bind cGMP at the regulatory GAF domains. Most PDE5 inhibitors inhibit PDE6 with similar potency, and electroretinogram studies show modest effects of PDE5 inhibitors on visual function—an observation potentially important in designing PDE5-specific therapeutic agents.

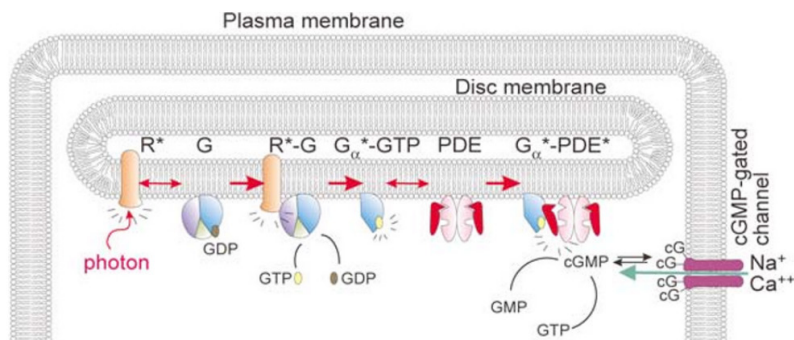
### 6.5.1 PDE6, the central effector of visual transduction in rods and cones

The retina contains cones, which operate under daylight conditions and perform color discrimination, and rods, which operate in dim light. Three types of cones exist in humans and in many primates, each one with a peak absorption at a wavelength corresponding to blue, green, or red light. In mammals, rods comprise approximately 90% of the photoreceptors, and cones the remaining 10%.

Both rods and cones contain light-sensitive pigments (opsins) that photoactivate on exposure to light. The photoexcited visual pigments stimulate the cell membranes of the rods and cones, triggering signals transmitted to the inner retina and via the optic nerve into the brain. In the dark, a circulating current flows from the inner segment of the photoreceptor cell to the outer segment. When illumination occurs, processes in the outer segment of the

photoreceptor cell interrupt this current. The plasma membrane of the outer segment encloses a stack of several thousand physically separate disc membranes, on which the initial events of the visual transduction pathway occur. Changes in cGMP levels in the outer segment transmit this signal from the disc membrane to the plasma membrane.

PDE6 is the primary regulator of cytoplasmic cGMP concentration in rod and cone photoreceptors (Figure 4). In the dark, PDE6 exists in an inactive form, and cGMP levels in the rod outer segment are relatively high (several micromolar). This permits a fraction of the cGMP-gated ion channels in the plasma membrane to remain open, allowing a current to circulate through the photoreceptor cell. Photoexcitation of the visual pigment, rhodopsin, activates the photoreceptor G-protein, transducin. The activated  $\alpha$  subunit of transducin binds to PDE6, and displaces the PDE6 inhibitory  $\gamma$  subunit from the active site of PDE6. The resulting subsecond drop in cGMP concentration causes closure of cGMP-gated ion channels, resulting in membrane hyperpolarization.



**Figure 4:** Visual excitation pathway in rod photoreceptors. The initial events of phototransduction occur on the physically separate disc membranes in the outer segment portion of the cell. Photoactivation of rhodopsin ( $R^*$ ) catalyzes the activation of hundreds of heterotrimeric G-proteins (G). Nucleotide exchange on the G  $\alpha$ -subunit is accelerated, and the dissociated  $G_{\alpha}^*$ -GTP subunit then interacts with the PDE6 holoenzyme (PDE). Displacement of the PDE6  $\gamma$  subunit from the catalytic site by  $G_{\alpha}^*$ -GTP relieves the inhibition of catalysis at one catalytic subunit, resulting in rapid hydrolysis of cytoplasmic cGMP. The drop in cGMP



levels in the outer segment causes dissociation of cGMP from the cGMP-gated ion channels in the plasma membrane, causing their closure. The reduced entry of cations into the outer segment causes membrane hyperpolarization, and ultimately, generation of the receptor potential at the photoreceptor synapse. Reactions involved in the recovery of the photoresponse and desensitization of the light response are not shown.

Precise regulation of cGMP levels is essential for normal operation of the visual transduction cascade. Indeed, a persistent imbalance in cGMP metabolism (either in its synthesis or degradation) will disrupt the visual signaling pathway and eventually lead to photoreceptor cell death and retinal degeneration (eg, retinitis pigmentosa).

### 6.5.2 Subunit composition and structure of the PDE6 holoenzyme

The rod PDE6 holoenzyme is a tetramer consisting of  $\alpha$  and  $\beta$  catalytic subunits to which two identical inhibitory  $\gamma$  subunits bind ( $\alpha\beta\gamma_2$ ). Cone PDE6 differs from rod PDE6 in that its catalytic dimer is composed of two identical  $\alpha'$  subunits. Also, the low molecular weight cone inhibitory  $\gamma'$  subunits differ slightly in size (9.4 *vs* 9.7 kDa) and amino-acid composition from the rod  $\gamma$  subunits.

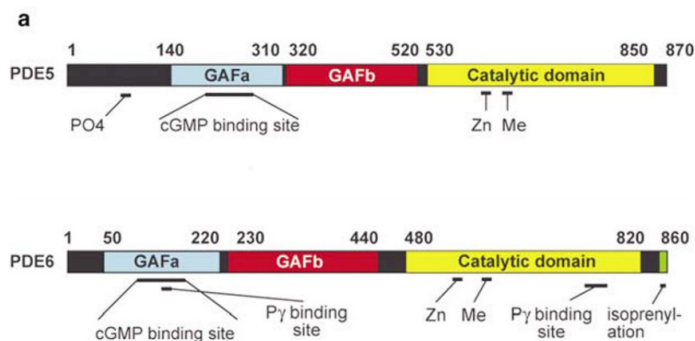
Each  $\gamma$  subunit interacts with at least two distinct sites on the catalytic subunit, and the affinity of this interaction is regulated by cGMP binding to the regulatory GAF domains of PDE6. When cGMP is bound to the GAF domain, the two  $\gamma$  molecules bind with different affinities to the catalytic dimer. When these regulatory sites are empty, both molecules of  $\gamma$  bind to  $\alpha\beta$  with the same, albeit reduced, affinity. Two major domains on  $\gamma$  interact independently with the PDE6 catalytic dimer. The extreme C-terminal residues of  $\gamma$  bind directly to the active sites of PDE6, blocking access of substrate to the catalytic core. The N-terminal half of  $\gamma$  binds to  $\alpha\beta$  with an affinity 50 times greater than its C-terminal half, and is

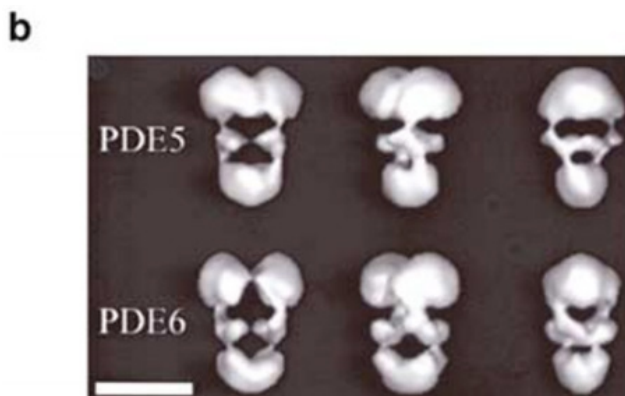
responsible for the cGMP-dependent modulation of  $\gamma$  affinity by the GAF domain.

While not yet fully elucidated, structural differences between the  $\alpha$  and  $\beta$  catalytic subunits, differences in cGMP binding to the GAF domains, and cGMP-dependent modulation of  $\gamma$  affinity for  $\alpha\beta$ , all contribute to the precise regulation of the extent and lifetime of PDE6 activation during rod visual transduction.

Electron microscopic analysis of purified rod PDE6 catalytic dimers at 2.8 nm resolution reveals a three-dimensional (3-D) structure comprised of three distinct globular domains.<sup>10</sup> The largest domain represents the catalytic domain, while the two smaller domains correspond to the two tandem GAF domains. While the resolution in the analysis was insufficient to detect low molecular weight proteins bound to the PDE6 catalytic dimer, the results unequivocally show that the primary dimerization site of the  $\alpha$  and  $\beta$  catalytic subunits resides in the N-terminal GAFa domain.

The molecular organization of PDE5 closely approximates that of PDE6, both in terms of its primary amino-acid sequence and its 3-D structure (Figure 5). For both PDE families, the two GAF domains and the catalytic domain are each likely to fold into independent globular structures, as predicted by the crystal structures of the catalytic domain of PDE4, and the tandem GAF domains of PDE2.





**Figure 5:** PDE5 and PDE6 share a similar domain organization and three-dimensional holoenzyme structure. (a) The domain organization of the catalytic subunits of PDE5 and PDE6 are compared. Both enzymes contain regulatory domains with tandem GAF domains that share significant sequence homology. For PDE5, the GAFa domain has been shown to be responsible for cGMP binding. The homologous domain for PDE6 is depicted but experimental proof is still lacking. In addition, PDE6 GAFa contains a region that interacts with the inhibitory  $\gamma$  subunit. The catalytic domains of PDE5 and PDE6 are also highly homologous, with a similar catalytic core containing metal ion binding sites. PDE6 contains a unique region at the entrance to the catalytic site responsible for inhibition of catalysis by  $\gamma$  binding. PDE5 is regulated by phosphorylation in the N-terminal region; no similar regulatory site has been identified for PDE6 to date. PDE6 contains a consensus sequence for farnesylation ( $\alpha$  subunit) or geranylgeranylation ( $\beta$  subunit) at its C-terminus. (b) The 3-D surface representation of the PDE5 and PDE6 holoenzymes shows a high degree of conservation. The largest lobe at the top represents the catalytic domain. The primary dimerization site appears to be the GAFa domain, with the smaller GAFb domains linking GAFa and the catalytic domain. The low molecular weight  $\gamma$  subunit or the prenyl binding protein of PDE6 cannot be visualized at this resolution. The three images of each enzyme are 45° rotations in the vertical axis. The bar represents 10 nm.

### 6.5.3 Similarities and differences between PDE5 and PDE6

Of the 11 mammalian PDE families, not only is PDE6 most closely related to PDE5 structurally, but it also shares several similarities in its biochemical properties. Both PDE5 and PDE6 strongly prefer cGMP over cyclic adenosine monophosphate (cAMP) as a substrate at the catalytic site. The PDE5 and PDE6 catalytic mechanisms share a requirement for divalent cations, including high-affinity binding sites for zinc ions that likely serve a structural role as well. Both PDE families bind cGMP with high affinity at the regulatory GAF domains, and most PDE5-selective pharmacological inhibitors also potently inhibit PDE6 catalysis.

However, there are several aspects of the enzymology and regulation of PDE6 that are not found in PDE5. First, while the maximum rate of cGMP hydrolysis by PDE6 achieves catalytic perfection (6000–8000 cGMP hydrolyzed per second) and operates at the diffusion-controlled limit, the catalytic constant for PDE5 is lower by almost three orders of magnitude. This extraordinary catalytic power of PDE6 may have evolved from the need of photoreceptor cells to generate a receptor potential on the millisecond time scale. Second, displacement of the inhibitory  $\gamma$  subunit from the active site by activated transducin mediates the primary mechanism of PDE6 activation. For PDE5, enzyme activation most likely proceeds via phosphorylation of the catalytic subunits and allosteric changes in cGMP binding to the GAF domains. (While PDE5 has been reported to copurify with the PDE6  $\gamma$  subunit, the physiological significance is uncertain because the  $\gamma$  subunit lacks binding or inhibitory activity toward PDE5 *in vitro*.)

Functional chimeric phosphodiesterases have been constructed using the GAF domains from cone PDE6 and the catalytic domain of PDE5 in order to identify structural and functional differences of the two PDE families. The original PDE6/PDE5 chimera retained the catalytic properties of PDE5 and the cGMP binding properties of PDE6. Site-directed mutagenesis using the PDE6/PDE5 chimera has identified two sites on PDE5 (Ala(608) and Ala(612)) that

accelerate catalysis 10-fold when substituted for glycine residues (found at positions 562 and 566 of cone PDE6). These residues are near the metal binding motif that represents the catalytic site of the enzyme. The unique sites of interaction of the  $\gamma$  subunit with the PDE6 catalytic sites have also been probed by substituting a stretch of cone PDE6 sequence (amino acids 737–784) into the catalytic domain of PDE5 and demonstrating that inhibition by the  $\gamma$  subunit occurred. Three hydrophobic residues at the entry to the catalytic core (Met(758), Phe(777), and Phe(781)) have been found to stabilize  $\gamma$  subunit binding and promote enzyme inhibition. Another  $\gamma$ -binding site has been identified in the GAFa domain of PDE6 that further stabilizes binding of this inhibitory subunit. These studies have provided insights into the unique features of transducin-activated PDE6 that distinguish PDE6 catalytic and regulatory properties from those of PDE5.

#### **6.5.4 Regulation of PDE5 and PDE6 by post-translational modifications**

PDE6 is unique among the 11 PDE families in that it undergoes a post-translational modification resulting in carboxymethylation and isoprenylation of the C-terminus of the catalytic subunits. The incorporation of a farnesyl group (rod PDE6  $\alpha$  subunit) or a geranylgeranyl group (rod PDE6  $\beta$  subunit) accounts for the high-affinity interaction of rod PDE6 with photoreceptor membranes. A 17-kDa prenyl binding protein (PrBP), originally referred to as the 'delta' ( $\delta$ ) subunit of PDE6, can bind to PDE6 and release the holoenzyme from its membrane-associated state. PrBP principally interacts with PDE6 at its prenylated C-terminus. In contrast to the catalytic and inhibitory subunits of PDE6, PrBP is widely expressed in a variety of tissues. It is also highly conserved through evolution. While PrBP has been shown to interact with many other binding partners, neither PDE5 nor other phosphodiesterases have been reported to interact with this prenyl binding protein.

In PDE5, phosphorylation at serine 92 of the bovine enzyme correlates with enhanced catalytic activity of the enzyme as well

as conformational changes in the regulatory GAF domains. The  $\gamma$  subunit of PDE6 also acts as a substrate for phosphorylation at several distinct sites within the central region of this 10 kDa protein. Phosphorylation of  $\gamma$  at Thr(22) or Thr(35) has little effect on the PDE6 holoenzyme itself, but greatly diminishes the ability of activated transducin to bind the  $\gamma$  subunit and relieve inhibition of catalysis. Phosphorylation of  $\gamma$  at Thr(62) in nonretinal tissue has been reported to regulate mitogenic signaling via interactions with proteins other than PDE6 catalytic subunits (whose expression is confined to the retina and pineal gland). Little information is available on potential regulation of PDE6 catalytic subunits by phosphorylation.

### 6.5.5 Drug selectivity for PDE5 and PDE6

One of the challenges in developing PDE5-specific inhibitors for therapeutic purposes is the similarity in the catalytic sites of PDE5 and PDE6 with respect to drug binding. For example, sildenafil is a highly selective inhibitor of PDE5 ( $K_i=4$  nM) with one exception, namely its potent inhibition of rod PDE6 ( $K_i=30$  nM). This is physiologically relevant, since one well-documented side effect of sildenafil treatment is a transient disturbance in visual function. PDE5 inhibitors are not known to cross the blood–brain barrier but they do cross the blood–retina barrier. Electroretinogram studies have shown that PDE5 inhibitors exert a modest effect on visual function. While it is likely that visual disturbances induced by sildenafil are a direct consequence of PDE6 inhibition, other modes of action must also be considered.

## 6.6 INHIBITORS OF OTHER PHOSPHODIESTERASES

### 6.6.1 PDE1

The PDE1 family is regulated in the short term by  $\text{Ca}^{2+}$ /calmodulin binding following elevation of  $\text{Ca}^{2+}$  derived primarily from

extracellular sources, and in the long term by changes in PDE protein level. PDE1 isoenzymes are widely expressed and are abundant in brain, heart, VSM, testis, macrophages, lymphocytes, and liver. PDE1 isoenzymes catalytic subunits occur as dimers, and each monomer is comprised of two calmodulin-binding sites, an autoinhibitory subdomain, and a catalytic domain.  $\text{Ca}^{2+}$ /calmodulin binding relieves autoinhibition and increases the  $V_{\text{max}}$  with no change in  $K_{\text{m}}$ . Activities of the PDE1 family are implicated in diverse processes that include regulation of smooth muscle proliferation, learning and memory, cardiac hypertrophy, and olfaction.

The PDE1 isoenzymes (PDE1A, 1B, and 1C) are derived from three genes and are represented in tissues by many splice variants that substantially differ in size. Among these isoenzymes, the relative affinities and catalytic rates for cGMP and cAMP vary considerably. Affinities for certain inhibitors and sensitivities to stimulation by  $\text{Ca}^{2+}$ /calmodulin also vary. In tissues containing primarily one isoenzyme, the relative hydrolytic contribution of the PDE1 family to breakdown of cAMP or cGMP will reflect the kinetic features of this dominant PDE1. PDE1 isoenzymes are primarily cytosolic, but they are also found in association with the plasma membrane and other particulate cellular components. Depending on location in the cell, members of the PDE1 family can differentially respond to selective changes in  $\text{Ca}^{2+}$  signaling and thereby provide for variations in the spatial and temporal changes in cGMP and/or cAMP levels. Inhibitors that show selectivity for PDE1 isoenzymes include vinpocetine, IC224, and SCH51866.

## **Discovery**

The existence of the  $\text{Ca}^{2+}$ -stimulated PDE1 was first demonstrated by Cheung (1970), Kakiuchi and Yamazaki (1970) as a result of their research on bovine brain and rat brain respectively. It has since been found to be widely distributed in various mammalian tissues as well as in other eukaryotes. It is now one of the most intensively studied member of the PDE superfamily of enzymes, which today represents 11 gene families, and the best characterized one as well.



Further researches in the field along with increased availability of monoclonal antibodies have shown that various PDE1 isozymes exist and have been identified and purified. It is now known that PDE1 exists as tissue specific isozymes.

## *Structure*

The PDE1 isozyme family belongs to a Class I enzymes, which includes all vertebrate PDEs and some yeast enzymes. Class I enzymes all have a catalytic core of at least 250 amino acids whereas Class II enzymes lack such a common feature.

Usually vertebrate PDEs are dimers of linear 50–150 kDa proteins. They consist of three functional domains; a conserved catalytic core, a regulatory N-terminus and a C-terminus [3-5]. The proteins are chimeric and each domain is associated with their particular function.

The regulatory N-terminus is substantially different in various PDE types. They are flanked by the catalytic core and include regions that auto-inhibit the catalytic domains. They also target sequences that control subcellular localization. In PDE1 this region contains a calmodulin binding domain.

The catalytic domains of PDE1 (and other types of PDEs) have three helical subdomains: an N-terminal cyclin-fold region, a linker region and a C-terminal helical bundle. A deep hydrophobic pocket is formed at the interface of these subdomains. It is composed of four subsites. They are: a metal binding site (M site), core pocket (Q pocket), hydrophobic pocket (H pocket) and lid region (L region). The M site is placed at the bottom of the hydrophobic pocket with several metal atoms. The metal atoms bind to residues that are completely conserved in all PDE family members. The identity of the metal atoms is not known with absolute certainty. However, some evidence indicate that at least one of the metals is zinc and the other is likely to be magnesium. The zinc coordination sphere is composed of three histidines, one aspartate and two water molecules. The magnesium coordination sphere involves the same aspartate along with five water molecules, one of which is



shared with the zinc molecule. The reputed role of the metal ions include structure stabilization as well as activation of hydroxide to mediate catalysis.

The domains are separated by “hinge” regions where they can be experimentally separated by limited proteolysis.

The PDE1 isozyme family (along with the PDE4 family) is the most diverse one and includes numerous splice variant PDE1 isoforms. It has three subtypes, PDE1A, PDE1B and PDE1C which divide further into various isoforms.

### Localization

The localization of PDE1 isoforms in different tissues/cells and their location within the cells is as follows:

**Table 1.** Various PDE1s location in tissues and within cells.

| Isoform                 | Tissue/cellular localization                                      | Intracellular localization |
|-------------------------|---|----------------------------|
| PDE1A ( <i>PDE1A</i> )  | Smooth muscle, heart, lung, brain, sperm                          | Predominantly cytosolic    |
| PDE1A1                  | Heart, lung   | Predominantly cytosolic    |
| PDE1A2                  | Brain   | Predominantly cytosolic    |
| PDE1B1 ( <i>PDE1B</i> ) | Neurons, lymphocytes, smooth muscle brain, heart, skeletal muscle | Cytosolic                  |
| PDE1B2                  | Macrophages, lymphocytes  | Cytosolic                  |
| PDE1C ( <i>PDE1C</i> )  | Brain, proliferating human smooth muscle, spermatids              | Cytosolic                  |
| PDE1C1                  | Brain, heart, testis  | -                          |
| PDE1C2                  | Olfactory epithelium  | Cytosolic                  |
| PDE1C4/5                | mRNA is present in the testis                                     | -                          |

Most PDE1 isoforms are reported to be cytosolic. However, there are instances of PDE1s being localized to subcellular regions but little is known about the molecular mechanisms responsible

for such localization. It is thought to be likely that the unique N-terminal or C-terminal regions of the various isoforms allow the different proteins to be targeted to specific subcellular domains.

### 6.6.2 PDE2

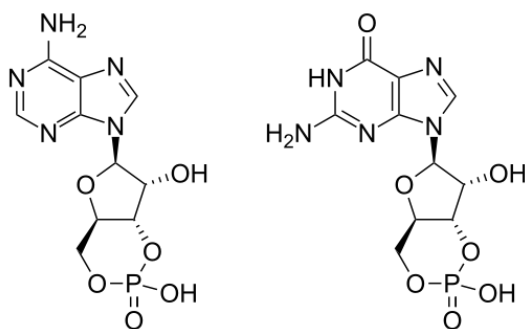
Phosphodiesterase 2 (PDE2) is a ubiquitous enzyme whose major role is to hydrolyze the important second messengers cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP). In the central nervous system, pharmacological inhibition of PDE2 results in boosted cAMP and/or cGMP signaling, which is responsible for series of changes in protein expression relevant to psychiatric and learning and memory disorders, such as depression, anxiety, and cognition deficits in Alzheimer's disease. In the periphery, inhibition of PDE2 exhibits beneficial effects in the diseased cardiovascular system, the respiratory system, skeletal muscles and *Candida albicans*-caused systemic infections. Even though blood-brain barrier penetration properties and selectivity of currently available PDE2 inhibitors have hindered them from entering clinical trials, PDE2 is still of great potential therapeutic values in different categories of diseases, and there is demand for development of new generation drugs targeting PDE2 for treatment of diseases in central nervous and peripheral systems.

#### *Crystal structure*

The crystal structure of the active site of the PDE2 enzyme has been reported. Even though amino acid sequences, for members of the PDE family show considerable difference (25-35% identity), the overall folding, functional and structural elements of the active sites are very similar. The active site is formed by residues that are highly conserved among all PDEs. The binding pocket contains metal ion (zinc and magnesium) binding sites. The two histidine and two aspartic acid residues, which bind zinc are conserved among all studied PDEs. The structure of several other PDE iso-enzymes has been elucidated and among them few

co-crystal structures, with inhibitors residing in the active site. The co-crystal structures for PDE4B, PDE4D and PDE5A have revealed two common features of inhibitor binding to PDEs. One is a planar ring structure of the inhibitors, which align in the active site of the enzymes and the other is a conserved glutamine residue (the “glutamine switch” mentioned below), which is essential for nucleotide recognition and selectivity.

### *Substrate selectivity*



**Figure 6:** cAMP (to the left) and cGMP to the right. Natural substrates for PDE2.

As mentioned above, PDE2 is able to hydrolyze both cAMP and cGMP, whereas some other members of the PDE family are selective for either of the two cyclic nucleotides. The variability in selectivity towards either cAMP or cGMP is thought to be determined by a so-called “glutamine switch”. The “glutamine switch” is an invariant glutamine found in all PDEs, for which the crystal structure has been solved. In PDE2, this residue is the Gln859. It has potential to form hydrogen bonds with the exocyclic amino group of cAMP and the exocyclic carbonyl oxygen of cGMP. In PDEs, which can hydrolyze both cAMP and cGMP, this glutamine is able to rotate freely. In PDEs that are selective for either cAMP or cGMP, this glutamine is constrained by neighboring residues to a position favoring selectivity for either cyclic nucleotide.

### 6.6.3 PDE7

High affinity cAMP-specific 3',5'-cyclic phosphodiesterase 7A is an enzyme that in humans is encoded by the *PDE7A* gene. Mammals possess 21 cyclic nucleotide phosphodiesterase (PDE) genes that are pharmacologically grouped into 11 families. PDE7A is one of two genes in the PDE7 family, the other being PDE7B. The PDE7 family, along with the PDE4 and PDE8 families, are cAMP-specific, showing little to no activity against 3', 5'-cyclic guanosine monophosphate (cGMP).

## REFERENCES

1. "Powered by Skipta technology, PharmacistSociety.com is the social network for verified Pharmacists to communicate and collaborate". [pharmacistsociety.skipta.com](http://pharmacistsociety.skipta.com). Archived from the original on 19 April 2012. Retrieved 1 May 2018.
2. Baraldi PG, Tabrizi MA, Gessi S, Borea PA (January 2008). "Adenosine receptor antagonists: translating medicinal chemistry and pharmacology into clinical utility". *Chemical Reviews*. 108 (1): 238–63. doi:10.1021/cr0682195. PMID 18181659.
3. Daly JW (August 2007). "Caffeine analogs: biomedical impact". *Cellular and Molecular Life Sciences*. 64 (16): 2153–69. doi:10.1007/s00018-007-7051-9. PMID 17514358. S2CID 9866539.
4. Daly JW (July 2000). "Alkylxanthines as research tools". *Journal of the Autonomic Nervous System*. 81 (1–3): 44–52. doi:10.1016/S0165-1838(00)00110-7. PMID 10869699.
5. de Visser YP, Walther FJ, Laghmani EH, van Wijngaarden S, Nieuwland K, Wagenaar GT (2008). "Phosphodiesterase-4 inhibition attenuates pulmonary inflammation in neonatal lung injury". *Eur Respir J*. 31 (3): 633–644. doi:10.1183/09031936.00071307. PMID 18094015.
6. Deree J, Martins JO, Melbostad H, Loomis WH, Coimbra R (2008). "Insights into the Regulation of TNF- $\alpha$  Production in Human Mononuclear Cells: The Effects of Non-Specific Phosphodiesterase Inhibition". *Clinics (Sao Paulo)*. 63 (3): 321–8. doi:10.1590/S1807-59322008000300006. PMC 2664230. PMID 18568240.
7. Essayan DM. (2001). "Cyclic nucleotide phosphodiesterases". *The Journal of Allergy and Clinical Immunology*. 108 (5): 671–80. doi:10.1067/mai.2001.119555. PMID 11692087.
8. González MP, Terán C, Teijeira M (May 2008). "Search for new antagonist ligands for adenosine receptors from QSAR point of view. How close are we?". *Medicinal Research Reviews*. 28 (3): 329–71. doi:10.1002/med.20108. PMID 17668454.

9. Lim YH, Lee YY, Kim JH, Shin J, Lee JU, Kim KS, Kim SK, Kim JH, Lim HK (2010). "Development of acute myocardial infarction in a young female patient with essential thrombocythemia treated with anagrelide: a case report". *Korean J Hematol.* 45 (2): 136–8. doi:10.5045/kjh.2010.45.2.136. PMC 2983030. PMID 21120194.
10. Peters-Golden M, Canetti C, Mancuso P, Coffey MJ (2005). "Leukotrienes: underappreciated mediators of innate immune responses". *Journal of Immunology.* 174 (2): 589–94. doi:10.4049/jimmunol.174.2.589. PMID 15634873.
11. WO patent 1985002540 Archived 2011-08-05 at the Wayback Machine, Sunshine A, Laska EM, Siegel CE, "ANALGESIC AND ANTI-INFLAMMATORY COMPOSITIONS COMPRISING XANTHINES AND METHODS OF USING SAME", granted 1989-03-22 , assigned to RICHARDSON-VICKS, INC.
12. Yu MC, Chen JH, Lai CY, Han CY, Ko WC (2009). "Luteolin, a non-selective competitive inhibitor of phosphodiesterases 1–5, displaced [(3)H]-rolipram from high-affinity rolipram binding sites and reversed xylazine/ketamine-induced anesthesia". *Eur J Pharmacol.* 627 (1–3): 269–75. doi:10.1016/j.ejphar.2009.10.031. PMID 19853596.



## CHAPTER 7

# COMPUTATIONAL CHEMOGENOMICS

### INTRODUCTION

Computational chemogenomics (CG) has demonstrated benefits of learning from entire grids of data at once, rather than building target-specific QSARs. A possible reason for this is the emergence of inductive knowledge transfer (IT) between targets, providing statistical robustness to the model, with no assumption about the structure of the targets. Computational chemogenomics models the compound–protein interaction space, typically for drug discovery, where existing methods predominantly either incorporate increasing numbers of bioactivity samples or focus on specific subfamilies of proteins and ligands. As an alternative to modeling entire large datasets at once, active learning adaptively incorporates a minimum of informative examples for modeling, yielding compact but high quality models.

## 7.1. COMPUTATIONAL CHEMOGENOMICS: IS IT MORE THAN INDUCTIVE TRANSFER

Advances in high-throughput technologies have enabled us to generate enormous volumes of cellular, functional, and target-specific bioactivity data for compounds. Despite these advances, there is still a considerable gap between our ability to link this data as a whole to phenotypical outcomes, particularly with respect to unintended drug side effects. It also frequently occurs that experimental researchers execute assays on both cellular and target specific levels using a fixed set of compounds, but are left to only speculate about the connection between the two outcomes. In these situations, development of models that explain patterns or correlation between the two levels of experimental data can be beneficial.

Computational chemogenomics (CG) recently emerged as the paradigm of learning from polypharmacological (multi-target) profiles. It characterizes each putative ligand–target complex by a composite set of both small-molecule descriptors encoding the ligand and protein descriptors encoding the target. As increasingly more emphasis is set on understanding and early prediction of drug side effects, CG naturally emerged as an attempt to address such questions, spurred by steadily accumulating multi-target activity profile data due to routine screening of pre-drug candidates over a wealth of potentially relevant biological targets.

It is important to mention that in CG (seen as the QSAR of protein–ligand complexes) both ligand and target may, formally, be considered as equivalent. CG may serve both to predict ligand affinity in virtual screening of a compound collection against a given target, as well as predict protein affinity in attempts to find novel targets that may bind a given drug (drug repositioning).

Various approaches to encode protein sequence and structure under the form of numeric descriptors have been suggested. While 3D structure-based descriptors, exploiting knowledge about the location and the geometry of the binding site, are clearly



the most information-rich, they are practically not very useful for target deorphanization (targets with well-characterized binding sites cannot qualify as “orphans”). An interesting alternative is represented by the injection of information of known or assumed key binding site residues into the protein fingerprint. The most general approach, not making any a priori assumption about the targets, include empirical amino acid sequence-derived descriptors, also used in protein sequence–activity relationship modeling.

In the CG formalism, target and ligand descriptors are perfectly interchangeable. However we will nevertheless adopt, a ligand-centric approach to CG for three main reasons. First, virtual screening for novel ligands is more often employed than the orthogonal search for new targets of a ligand. Classical ligand-centric QSAR is actively employed in the herein reported benchmarking study. Last but not least, the provided protein information does not have to be structurally relevant in order to benefit from multi-target learning, as will be emphasized in the following. Therefore, this work will describe CG as a ligand-centric approach, where the ligand structure–activity relationships are allegedly “modulated” by protein information. This point of view serves for discussion only, and has no impact on the computational strategies and their results.

The naive approach to multi-target profile prediction would simply consist in realizing, for each target  $T$ , an individual QSAR model  $\hat{A}_T(M) = f[\mathbf{D}(M)]$  (circumflex cap meaning in silico calculated value throughout this text), where a molecule  $M$  is described by  $\mathbf{D}(M)$ . In their simplest form, such relationships are linear:

$$\hat{A}(M) = \alpha_0 + \alpha_1 D_1(M) + \alpha_2 D_2(M) + \cdots + \alpha_n D_n(M) \quad (1)$$

where coefficients (weights)  $\alpha_i$  represent the relative impacts of the ligand feature  $i$  encoded by  $D_i$  of the activity  $A$  on the current target. They will be termed “feature weights” in the following.

Although this work exclusively deals with non-linear models, the linear approach will be used to illustrate the concepts that

are central to this work. These concepts are independent of the actual functional form of the models, and hence easiest explained on the basis of maximum simplicity approaches. The reader is encouraged to visit the data mining for a more formal treatment.

Predicting activity  $\hat{A}$  of ligand  $M$  by individual models trained for each  $T$  will mechanically allow the completion of the ligand-by-target matrix of predicted activities. This amounts to determining the matrix of feature weights for targets (symbolically,  $\alpha_i^t$ ), by successively fitting each vector  $\alpha^t$  for every  $t$ . Here, each  $\alpha_i^t$  is implicitly associated to a target  $t$ , the unique data source serving to fit its value. No explicit knowledge on the relationships between these weights and the nature of the target is generated here. Fitting ( $\alpha_i^t$ ) assumes that the initial training data supports—in terms of the per-protein ligand set size and diversity (chemical space coverage)—the building of all of these individual models. Unfortunately, experimental activity profiles  $A_T(M)$  are often sparse; few  $M, T$  pairs were subjected to experimental scrutiny, and fewer still provide examples of high-affinity complexes. Moreover, the only way to update the  $\alpha_i^t$  matrix in order to cover a novel target  $T$  is by fitting  $\alpha^T$  values. This is conditioned by the existence of sufficient and diverse training examples of binders and non-binders, thus obviously unfeasible for orphan targets.

However, the paradigm of CG, that is the simultaneous machine learning from the entire available activity matrix, may significantly outperform the -mentioned naive strategy. There are two main reasons for this:

### ***Inductive transfer (IT)***

Under its most simple form, the principle of IT can be outlined as follows. Suppose that the activity of ligands with respect to a target  $t$  is conditioned by  $n$  ligand features, as highlighted in Equation 1. Now consider a related target  $T$ , depending on exactly the same  $n$  features, where selectivity stems from a single feature that is weighted differently:  $\hat{A}_T(M) = \beta_1 D_1(M) + \alpha_2 D_2(M) + \dots$ . In the naive approach, fitting either equation would require, by rule-of-thumb,

20n or more training examples in order to grant some statistical robustness to the models. The strength of IT is the transfer of knowledge (e.g.,  $\alpha_i$  values) obtained from analysis of a related problem to enhance solving of a new one. Suppose that enough data is available to train  $\hat{A}_t$ . Then, predicted values  $\hat{A}_t$  could be employed as a new molecular descriptor for the related model  $\hat{A}_T = \hat{A}_t + (\beta_1 - \alpha_1)D_1$ . A few ligands, including several actives, tested on T would suffice to robustly determine the two coefficients ( $1.0$  and  $\beta_1 - \alpha_1$ ) of the latter regression.

In practice, IT is not bound to such a sequential training scheme as outlined (output of primary models serving as input for the IT-enhanced approaches, often referred to as “feature nets”), but may also be achieved by simultaneous (multi-task) learning for related sets of tasks (endpoints). Related tasks share a common latent subspace of descriptors, enhanced with elements responsible for the specificity of each target.

### ***Explicit learning (EL)***

This requires target structural information to be injected into the learning process, by means of a protein descriptor vector  $\Delta(T)$ . Intuitively, one may think about an EL model as a QSAR equation in which the weights at  $\alpha_i = f_i[\Delta(t)]$  are now functions of the protein descriptors. The reasoning for this is that the relation between ligand structures and their affinities to observable endpoints should be related to and explained by the explicit provision of the protein information. As each protein responds differently to the presence of a ligand feature  $i$  (a substructure, for example), the intensity of this response is dependent on protein structure. If the relevant protein structure features are captured by the descriptor  $\Delta$ , then the conditioning of  $\alpha_i$  with respect to  $\Delta$  should be learned during CG model building, leading to true E Lenhanced models. For example, if affinity is represented as a linear combination of ligand and protein descriptor cross-terms, as shown in Equation 2:

$$\hat{A}(m, t) = \alpha_0 + \sum_{i,j} \beta_{ij} \times D_i(m) \Delta_j(t) \quad (2)$$

$$\text{then } \alpha_i^t = \sum_j \beta_{ij} \Delta_j(t).$$

Both the IT and EL concepts are independent of machine learning approaches, and therefore should be insertable into any given algorithm, including, but not limited to, the well-known support vector machines (SVMs) or neural networks. Hence, it is possible to employ the same algorithm to generate both IT enhanced and potentially EL-enabled models, where the difference is toggled by supplied protein information. The key difference between the IT-enhanced and EL-enabled approaches is that the former, like naive single-endpoint QSARs, are completely ignorant of the nature of the targets. The IT approaches are injected with indicator variables which replace actual protein information. In other words,  $\Delta(t)$  should stand for actual physicochemical and structural target properties in EL, while proteins are represented by mere labels in IT. For example, in previous work led by Vert, EL-enabled models rely on calculated or biology-inspired protein kernel values, while IT-enhanced models are based on the so-called “multi-task” kernel. These studies are, to our knowledge, the most extensive analysis of IT versus EL, done in terms of classification, based on the “kernel trick”. This “trick” operates under the tensor-product working hypothesis: if ligand–target complexes denoted by  $m:t$  are described by the tensor  $\mathbf{D}(m) \otimes \Delta(t)$ , then the expensive cross-product calculation can be avoided by alternatively computing the product of ligand and target kernels:

$$\mathcal{K}(m : t, M : T) = \mathcal{K}^{lig}(m, M) \times \mathcal{K}^{prot}(t, T) \quad (3)$$

Despite reported success in previous studies, it is yet to be made clear whether injection of actual protein information in intended EL models actually leads to the desired EL model, or whether machine learning merely exploits those protein descriptors in the same way as it would handle target labels in IT processes. Moreover, the few realizations of explicit EL-enhanced approaches

based on Equation 2 only marginally outperformed the simpler linear combination of stand-alone ligand and protein descriptors. This is further evidence that granting the technical feasibility for the fitting of an EL-enhanced approach is not a guarantee that the resulting model will actually attain such status. The present work aims to shed some more light on this issue.

To this purpose, a rigorous benchmarking protocol was designed, in order to compare (a) single-endpoint QSAR models, (b) IT-enhanced single-endpoint QSARs, (c) IT enhanced (multi-endpoint) CG models, and eventually (d) EL-enabled (protein information-supplied) CG models. This should enable us to weight multiple perspectives for discovering hidden knowledge in phenotypical and other endpoint assay data.

We used this opportunity to focus on quantitative support vector regression (SVR). This is more challenging than SVM classification, and such studies were so far, only performed as proof-of-concept. To our knowledge, SVR was never explicitly investigated in the context of IT versus EL benchmarks. Like in the abovementioned SVR-driven studies, we opted for the maximum simplicity option for ligand–target descriptor pairs: concatenation of their respective descriptors. This means that a putative ligand–protein complex  $M:T$  will be considered as one object represented by a vector  $\mathbf{D}(M), \Delta(T)$  resulting from concatenation of ligand and protein terms respectively.

Based on 9,642 accurate GPCR-ligand complexes of measured  $pK_i$  values (approximately 4,500 ligands for 31 rhodopsin-like GPCRs), the herein used training set is one of the largest coherent multi-target sets seen in CG studies so far. Featuring actual high and low-affinity ligand-protein pairs, it has no need to rely on artificially generated, experimentally untested entries as decoys.

ISIDA property-labeled fragment counts and fuzzy pharmacophore triplets were used to describe ligands. In terms of genuine protein descriptors, we employed a two-pronged approach. On one hand, we utilize sequence-based terms that can be easily calculated for poorly-studied proteins, and are thus potentially usable in a non-

simulated, true deorphanization attempt. On the other hand, a fingerprint of protein–protein affinity-focused similarity scores based on directly measured experimental affinities was exploited. EL-enabled models in our approach use the genuine protein descriptors, whereas IT-enhanced models introduce indicator variables (identity fingerprints); both are concatenated to ligand descriptors.

The analyses executed were as follows. First, the various approaches have been benchmarked in terms of model building and cross-validation propensities. A challenging SVR cross-validation protocol was based on a 10-trial randomized leave-1/3-out scheme. A genetic algorithm (GA)-driven optimization of the SVR operational parameters has been employed to build optimal models within each of the CG modeling approaches tested. Cross-validated prediction propensities of models were monitored in terms of residual errors, allowing us to locate benchmarked machine learning strategies in a “strategy space” and to report mutual closeness relationships.

Second, a target deorphanization study was carried out. Whilst there is not a priori expectation to see EL-enabled models outperform IT-enhanced approaches in terms of cross-validation propensities, target deorphanization challenges are the actual stumbling block for genuine EL-enabled CG. Indeed, EL should display a decisive advantage, for it is expected to explicitly adapt the  $\alpha'_i = f_i[\Delta(t)]$  weights to the orphan target. By contrast, as already mentioned, IT methods are unaware of the nature of orphans. However, while a rigorous deorphanization protocol cannot proceed without providing relevant target information, some less rigorous alternatives do exist. A baseline deorphanization strategy, herein termed “deorphanization by substitution”, advocates using a predictive model for some training set proteins as a predictor of the affinity of the presumed orphan protein. While there is no fundamental reason for the success of such a strategy, in practice this may well be the case, if the presumed orphan has at least some close analogs among training set proteins. Unfortunately, published deorphanization success stories rarely explicitly report

the closeness to training set proteins, or how well the single-endpoint models of those proteins would have fared instead of the advocated CG approach. Therefore, the target deorphanization protocol here assigns paramount importance to this aspect.

Our study has interestingly revealed that, while CG methods once more confirmed their advantages over classical QSAR, most of this advantage seems to be due to IT. No significant boost of EL approaches over IT strategies was evidenced. The direct consequence of this is that successful deorphanization was restricted to ‘trivial’ cases of targets being quite close to one or several training set proteins. It further shows that future CG studies should benchmark EL against some baseline IT experiment to provide sufficient proof of EL, and that the field of protein descriptor needs further improvements to truly realize the expected benefit of EL. A future implication of this is that once one has established proof of concept for an EL-enabled model, they can return to the critical overarching task of applying their CG model to molecule design.

### **7.1.1. New Insights in Protein Kinase Conformational Dynamics**

Protein kinases are a large family of signaling proteins that are involved in the regulation of a wide range of cellular processes. Their mis-regulation is associated with a plethora of diseases, ranging from infections to diabetes and cancer. The development of novel kinase inhibitors has experienced a significant acceleration following the successful approval of imatinib (Gleevec<sup>®</sup>), a drug that inhibits the BcrAbl fusion protein, which is the main causative agent in chronic myeloid leukemia. In the last few years, approximately a 30% of the scientific documenting drug development efforts has been focused on kinase inhibitors; more than a hundred different compounds are currently under clinical trials.

The activation of a kinase generally involves one or more conformational transitions from “inactive” to “active” (catalytically



competent) states. Since inactive conformations in different kinases are more structurally diverse than their respective active states, some inhibitors, like imatinib, attain a restricted selectivity by targeting an inactive state. Thus, understanding the structure and the dynamics of the different conformations of kinases is of great importance. However, the lifetime of inactive conformations is generally short, while the energetic barriers that separate them from the active state can be significant, making their direct observation by structural and spectroscopic methods difficult. Accordingly, computational predictions of the structures and energy landscapes associated with conformational changes are becoming increasingly widespread. Many of the new kinase inhibitors in development have benefited from computational models, and, as the case of HIV integrase has clearly shown, the transient cavities only visible in molecular dynamics simulations are expected to help the design of more potent and selective kinase inhibitors.

### *A Complex Conformational Landscape*

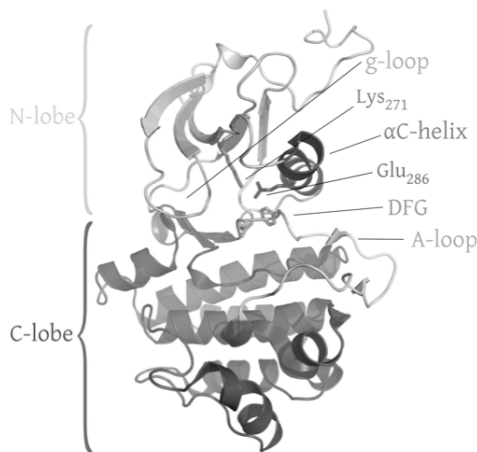
The crystal structures have shown that most protein kinases can assume at least two different states, an active state, able to phosphorylate a substrate upon ATP binding and an inactive one. Besides those large scale changes, the local configuration of several structural elements can vary during the transition from active to inactive, leading to a wide ensemble of structures. Taking the structure of the Abl kinase 1 as a reference (Figure 1), the C-helix in the N-lobe can adopt a “in” conformation, in which a glutamate (Glu<sub>286</sub> in Abl, UniProt sequence P00519) points towards the ATP site and interacts with a lysine on the  $\beta 3$  (Lys<sub>271</sub> in Abl), and an “out” conformation in which the same glutamate points towards the solvent and the salt-bridge is broken. The conserved DFG motif preceding the activation loop can analogously adopt an “in” and an “out” conformation depending on which residue between the glutamate (DFG-in) and the phenylalanine (DFG-out) points towards the ATP cavity. Finally, the activation loop (A-loop) can undergo a large structural rearrangement from an “open”



structure to a “close” one. While the structural hallmarks of the active state are well known and common to most members of the kinase family, namely  $\alpha$ C-helix in, DFG in and activation loop open (Figure 1), many inactive states exists, whose features are less prone to be generalized across different kinases. What is more, the full sequence of events leading from the on-state to the off-state is known only for a very limited set of kinases, as the insulin receptor kinase or the cyclindependent kinase 2. Just considering the local arrangement of the three structural elements described so far, several possible off-states, not fulfilling the requirements to be active conformations, arise (Table 1), whose existence, so far, has been mainly assessed by crystallographic means.

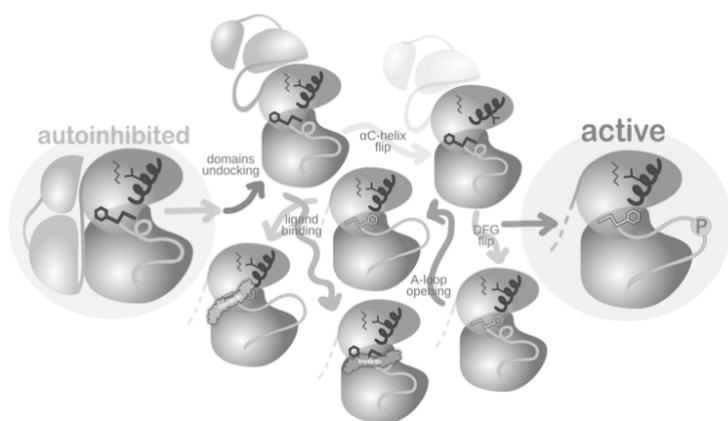
**Table 1.** Different Structures of the Abl Kinase in the Protein Data Bank

| DFG | A-loop | $\alpha$ C-in                     | $\alpha$ C-out            |
|-----|--------|-----------------------------------|---------------------------|
| In  | open   | active<br>2F4J, 2GQG, 2G2I, ...   |                           |
| In  | close  |                                   | Src-like inactive<br>2G1T |
| Out | open   | 1OPK, 1OPL, 2G2F(B), 2G2H(B)      | 2G2F(A), 2G2H(A)          |
| Out | close  | Imatinib-bound<br>1OPJ, 1IEP, ... |                           |



**Figure 1.** Structure of the human Abl kinase. The N and C lobes are represented in white, black, respectively. Important regulatory regions as the  $\alpha$ C-helix, the g-loop, the DFG motif and the activation loop are labeled.

The picture is further complicated by the existence of several “ancillary” domains and ligands that interact with the kinase domain in different steps throughout the on-to-off transition and lead to an even wider ensemble of conformations. The Abl kinase, for example, is auto-inhibited by forming a complex with the two Src Homology (SH) domains SH2 and SH3 and can be regulated by several endogenous compounds as imatinib or dasatinib (Figure 2). Cyclin-dependent kinases bind to cyclins to carry out their functions and receptor tyrosine kinases, as the epidermal growth factor receptor (EGFR), dimerize to form an active complex.



**Figure 2.** Abl kinase exists in several different conformations, varying for the conformations of the DFG motif, the A-loop and the  $\alpha$  C helix. Some of these conformations have been targeted by several drugs, such as imatinib, dasatinib, etc.

The need for an atomic level description of how these phenomena take place, able to overcome the unsatisfactory “static picture” representation, has led to an increase of expectations towards molecular modeling and, in recent years, several features of kinases conformational dynamics have been elucidated by simulations. In the remainder of this review, we summarize some of the most recent and promising advancements achieved by computation.

## *Long Molecular Dynamics Simulations*

Molecular dynamics (MD) is the most established approach to study proteins at an atomic level. However, the time scales accessible to MD are limited by the small integration step (typically  $10^{-15}$ ) used to evolve Newton's equations, that needs to be of the order of magnitude of the fastest molecular motion (i.e. bonds vibration). Large structural rearrangements, as the transitions between different kinase conformations, take place in time scales of the order of hundreds of nanoseconds, at best, and are generally considered "rare events", difficult to observe during MD simulations. Only recently, thanks to considerable software and hardware advancements, especially in special-purpose machines, massively-distributed computing and GPU infrastructures, the time-scales needed to sample these phenomena have become accessible. Thanks to these advancements, Shaw and co-workers were able to perform 2.2 s of total simulation time to study the "flip" from the DFG-in to the DFG-out state of the Abl kinase. The study represented the first all-atoms MD simulation of a kinase in the s time frame and the first unbiased observation of the DFG flip. In their study, the authors identified a new intermediate in which the C-helix is in "out" conformation and the DFG assumes a characteristic configuration with the aspartate pointing towards the C-lobe and the phenylalanine occupying a cavity in the N-lobe. The role of DFG protonation in favoring different conformations was also highlighted, proving a pH-dependency in the binding of ligands, as imatinib. Besides deepening the understanding of the conformational transition itself, these finding open new possible scenarios for drug discovery. Yet, the authors were forced to introduce an  $\alpha$ C-helix destabilizing mutation to observe the conformational change. A considerable step forward in the study of kinase conformational dynamics is represented by a recent research on the EGFR kinase. In their study, the authors perform long MD simulations (a total of 47 s) to study the binding of the inhibitor lapatinib (Tycerb<sup>R</sup>) to EGFR and postulated the existence of a third unknown state of the C-helix from discrepancies in the calculated rate of association. According to their interpretation, the slower binding rate was due to the existence of a new

predominant conformation assumed by the  $\alpha$ C-helix, different from the “out” inactive conformation that binds lapatinib, and to which the protein must evolve in order for the drug to bind. Deploying a massive body of simulations and experiments, the author demonstrated the existence of a partially disordered state of the  $\alpha$ C-helix, more stable than the “in” active conformation. What is more, they were able to explain the role of the important and widespread mutant L845R, affecting the dimerization dynamics of EGFR despite being distant from the dimerization interface. Indeed, they demonstrated that L845R stabilize the  $\alpha$ C-helix “in” conformation by suppressing the disorder of the helix through the formation of new interactions between the mutated arginine and a cluster of negatively-charged residues. The same authors successfully reproduced also the binding of cancer drug dasatinib (Sprycel<sup>®</sup>) to Src kinase using long “unguided” MD simulations (35 s) starting from the unbound ligand, recovering the correct crystallographic bound structure and discovering the crucial role of cavity desolvation and a potential allosteric site between helices F and G. These two examples demonstrate the usefulness of very long all-atom molecular dynamics simulations and the remarkable correctness of the recent all-atom force fields that are able to correctly predict the main conformations of protein kinases. Based on the continuous increase of available computational power, it is easy to forecast that soon such simulations will be feasible on standard high-performance computing platforms.

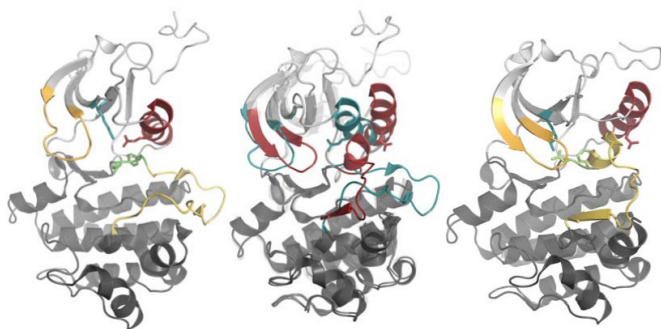
### *Coarse Grain*

A different approach to overcome the time and length scale limitations of all-atoms MDs is simplifying the description of the system, giving up the all-atom representation in favor of a coarser picture for the sake of speed and feasibility. Several CG approaches exist, and describing them all is well beyond the scope of the present review. They can be loosely classified into Go-like models or “beads” based models. The first class draws on the original ideas of Go describing the Hamiltonian of the system as a network of native contacts, thus obtaining a structure-based potential, i.e.

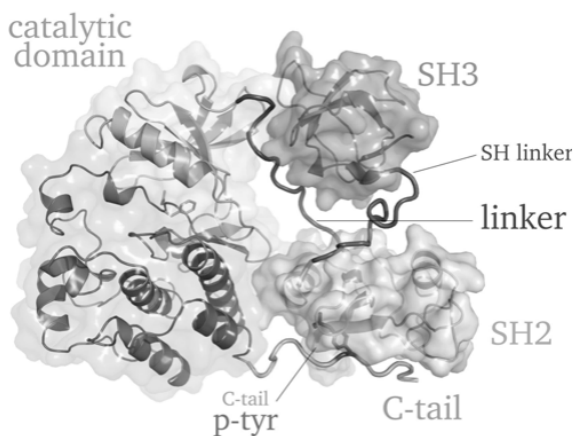
a potential in which the native structure is the minimum of the potential energy by construction. The second class, instead, maps the all-atoms system onto a chain of “beads”, each describing a group of atoms from which they inherit their properties (generally hydrophobic, hydrophilic or neutral). Depending on the process of interest, the Go-like representation might, or might not, be preferred to the more general beads-based one. Using a multiple state Go model, Roux and Yang described the transition from inactive to active of Src kinase, identifying two different routes and explaining in details the sequence of events leading to activation along the lowest energy pathways. In their study, they started from a close conformation of the A-loop and an Chelix “out” and observed that the activation loop is able to fluctuate between active and inactive conformations while the helix has a slower transition; they also observe that the transition to C-helix “in” take place only after the activation loop has reached an active-like structure. The first event along the transition was identified as a loss of contacts between the A-loop and the C-helix, leading to a subsequent detachment of the helix from the N-lobe strands. An alternative path, involving partial unfolding of the N-lobe itself, was also described. Beside the importance of understanding the process underlying Src activation, the study highlight significant details on how the information might flow across the highly coupled structural elements of kinases, unleashing information important for the understanding, and drugrelated harvesting, of allosteric regulation. A hybrid potential, in which the bonded terms were treated with a classical force field description, while the non-bonded term were treated as a structure-based potential. In their study, the authors identified a new intermediate state along the close-to-open transition, characterized by a predominantly close structure, but with few open-like contacts in the C-terminal region. No correlation between the A-loop motion and the  $\alpha$  C-helix was observed as both close and open conformations has an even population of helix “in” and “out” structures. A similar study, investigating the activation mechanism of Lyn kinase using a two-state Go-like potential, was recently published by Post and co-workers. The importance of the C-helix transition, constituting the highest barrier along the activation path, emerges once again

from the CG simulations. However, the order of the events is not the same as the one obtained in ref. 19. The  $\alpha$ C-helix “in” to “out” movement was instead found to be the last event (going from active to inactive), preceded by the breaking of the contacts between the C-helix and the activation loop, and by the formation of a helical region in the N-terminal segment of the A-loop itself. As suggested by the authors, the discrepancies might arise from differences in the Go models and in the description of the activation loop. The transition from active to inactive was also studied for another kinase, the Protein Kinase A (PKA) by Onuchic and co-workers. In their study, the authors applied structure-based Hamiltonians to study the closing mechanism upon ATP binding and confirmed the existence of motions with different time scales in the catalytic domain. The ATP binding site was found to be considerably more flexible and moving faster than the overall kinase motion. Upon ATP binding, a suppression of the fluctuations in this region is observed, that led to the formation of additional contacts that shift the structure towards the close conformation. The partial unfolding of some elements of the N-terminal lobe was also observed as in the study on Src kinase. The use of simplified CG models also allowed to perform simulations of Hck kinase catalytic domain together with the regulatory domains SH2 and SH3 (Figure 4). In their work, the authors successfully generated a wide ensemble of putative topological arrangements for the three domains, ranging from fully docked auto-inhibited structures to completely detached ones, and clustered them into 9 different representative states. Using this 9-states representation, the authors were able to fit experimental smallangle X-ray scattering (SAXS) data and determine the different population of these states under different conditions. Interestingly, the analysis pointed out that, in the wild-type complex, the auto-inhibited state is predominant and populated for the 82%, regardless of the phosphorylation state of the C-terminal tail tyrosine, which was believed to be essential for deactivation, thus disproving the theory that rapid disassembling is triggered by the C-tail dephosphorylation. On the other hand, they observed that the shift of population towards the disassembled state in the presence of potent SH2- and SH3-binding peptides is less prominent for the high affinity C-tail mutant Hck-

YEEL, thus confirming the crucial role of the C-terminal region for the formation of the complex.



**Figure 3.** Abl kinase conformations; (left) active conformation with the A-loop open, the  $\alpha$  C-helix “in” and the DFG “in”, (right) inactive conformation with A-loop close and C-helix “out”; (center) super-imposition of the active and inactive structures.



**Figure 4.** Src kinase auto-inhibited state.

### **Enhanced Sampling**

If the results of the simulations are to be used in rational drug design, CG simulations may not be sufficiently accurate and a fully atomistic representation is required. Several methods, that can be loosely classified as “enhanced sampling” approaches have



been developed to use an atomically accurate description while achieving a statistically significant sampling of the free energy landscape. Many of these methods are based on the addition of some sort of position or time-dependent bias to the system, in order to (partially) compensate the free energy of the starting minimum and allow the system to escape, thus enhancing the exploration of the conformational space and allowing to observe the process of interest in a shorter simulation time. The well-known approach is perhaps the Umbrella Sampling (US) method. It is based on the dimensional reduction paradigm, according to which the dynamics of the system is studied as a function of a collective variable (CV), function of the atomic coordinates, which should properly approximate the real reaction coordinate. In US, a fixed harmonic potential is added corresponding to a particular value of the CV, so that the system gradually moves from its equilibrium conformation towards the minimum of the added potential. After a sufficiently long sampling time, the real probability distribution of the different states can be recovered re-weighting the populations for the additional bias. In practical situations, it is often convenient to divide the process in different “windows” along the reaction coordinate and to perform different US runs for each window, centering the umbrella potential in the window itself. Using US, the opening of Src activation loop was characterized, highlighting the importance of the exchange of Glu<sub>310</sub> salt-bridge between Arg385 (in the Clobe) and Arg<sub>409</sub> (on the A-loop). The authors also identified a possible intermediate state that, in their setup, was found to be the most stable conformation of the A-loop. Targeted MD (TMD) is conceptually similar to US, but the harmonic potential is not kept fixed throughout the simulation. Instead, the potential is gradually moved along the CV, contributing a force that pushes the system towards a final (desired) state. Mendieta and Gago investigate the inter-lobe motion in Src kinase and the effect of inter-lobe rearrangements on the conformation of the A-loop. In their study, the authors constrained the N-lobe and the C-lobe and used TMD to stretch the inter-lobe hinge peptide towards a open-like configuration. Interestingly, they found that the activation loop didn't open as a consequence of the lobes being pulled apart; instead, the

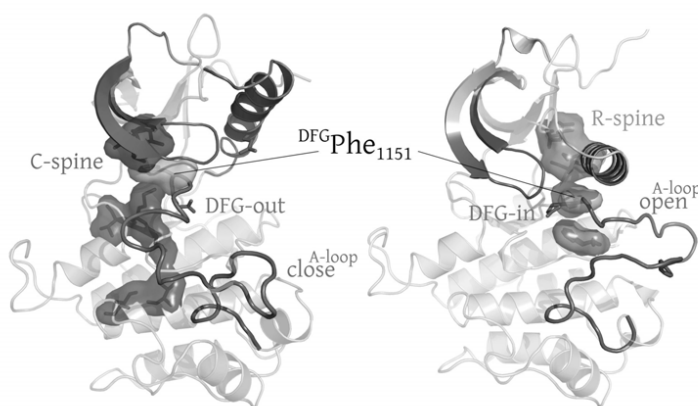


A-loop packed more effectively in the inter-lobe cleft, pushing the tyrosine of the phosphorylation site towards the ATP pocket. Drawing on this, an autophosphorylation mechanism, triggering the A-loop opening, was postulated. Similar “in silico” pulling experiments on Src A-loop opening were also performed to obtain starting point for subsequent unbiased multiple-trajectories MD simulations and to study the release of ADP from kinase A. The use of a fixed potential, as in US simulations, require a certain knowledge of the system and of the different conformations that it might explore. To overcome this limitation, especially for complex systems, several “adaptive” techniques have been developed, such as local elevation, conformational flooding, self-healing US or metadynamics. The latter, in particular, is rapidly becoming the method of choice for the study of complex conformational changes, especially when the use of 2 or more CVs is needed. Metadynamics is based on an extended Lagrangian formalism and on the addition of a bias potential in a history dependent fashion. Every  $\tau$  steps, a small repulsive Gaussian is added in the corresponding position of the CV space, discouraging the system to visit already explored basins and favoring the sampling of new regions of the conformational space. When the dynamics in the CV space becomes diffusive, the bias potential, sum of the deposited Gaussian, has compensated the underlying free energy that can hence be recovered from the bias potential itself. The closure mechanism of cyclin dependent kinase 5 (CDK5) was studied by Berteotti using metadynamics with path collective variables (PCV). A complex two step mechanism, resembling the one obtained with CG for Src kinase, was described, involving first a 45° rotation of the C-helix and a subsequent concerted movement of the helix and the activation loop that leads to the complete rotation of the first to the “out” conformation and the closure of the latter. An accurate free energy surface was obtained, allowing to quantify the stability of the close state over the open one by 4-6 kcal·mol<sup>-1</sup>. Most importantly, the study identified an intermediate along the path, in which the helix has already reached its final position, but the activation loop is still in an open conformation, that might unleash important information for the development of selective inhibitors. Metadynamics was also used to rationalize the

different activity of imatinib towards the two highly homologous kinase  $\alpha$ c-Src and  $\alpha$ c-Abl. In the study, the author applied parallel tempering metadynamics (PTMetaD) variant using 28 replicas of the system in the temperature range 308-399 K, summing up to 22 s of total simulation time, to investigate the DFG motif in-to-out flip transition. It was found that, while imatinib binds with the same pose to both kinase, the different stability of the imatinib-binding DFG-out conformation largely accounts for the difference in activity (0.08 M for Abl, vs. 7.2 M for Src). The authors also identified the structural determinants behind the different DFG-out stability, underlining the higher intrinsic degree of flexibility of  $\alpha$ c-Abl structure, allowing for a larger opening of the active site and for a larger number of water molecules, able to better solvate the DFG aspartate. Metadynamics, in the Bias Exchange MD (BEMD) scheme, was used together with unbiased MD to rationalize the differences between the dynamics of the apo and holo structures of PI3K. The analysis suggested a long-range conformational selection mechanism for the binding of inhibitors, coupled with a short-range ligand-induced fit in the binding cavity. It is to note that metadynamics was also used to predict the binding affinity of several ligands to protein kinases as CDK2.

A similar method, based on the same extended representation used in metadynamics, is Temperature Accelerated MD (TAMD). TAMD is based on the same extended representation used in metadynamics. However, in this case, the slower degrees of freedom described by the CVs are selectively subjected to a different temperature  $T$ , higher than the target temperature  $T$ , and can hence lead the system to explore a wider portion of the conformational space, without distorting the free energy itself. Using TAMD, Abrams studied the inactive-to-active transition in the insulin receptor kinase, refining the lowest energy path by means of the string method. Interestingly, the authors paid much attention to the dynamics of the so-called C-spine and R-spine (Figure 5), two networks of stacked hydrophobic residues that have been observed in several protein kinases. The formation of a small helical segment in the N-terminal region of the Aloop was found to be the first event in the activation pathway. Along this

path, the subsequent unfolding of this helix is responsible for a rearrangement in the dihedrals of the DFG motif preceding the A-loop, leading to the counterclockwise flip of the DFG itself to the “in” conformation. The rotation of the  $\alpha$ C-helix to its final position was observed only in the last part of the path, when the A-loop moves towards a completely open conformation. It is to note that, even after the DFG flip, the R-spine remains broken, due to the incorrect placement of the  $\alpha$ C-helix, suggesting an alternative mechanism for the regulatory spine disassembly.



**Figure 5.** Structure of the insulin receptor kinase; (left) inactive conformation with the phenylalanine of the DFG motif inside the ATP cavity and forming the so-called C-spine, (right) active conformation with the DFG motif flipped in the “in” position and the phenylalanine joining a different hydrophobic network known as the regulatory spine (R-spine).

## 7.2. CHEMOGENOMICS APPROACHES FOR THE QUANTITATIVE COMPARISON OF BIOLOGICAL TARGETS

Chemogenomics is a term coined about 10 years ago. Traditional approaches for the identification of bioactive compounds use a chemical library, a single target protein, and an assay, which allows us to measure the activity of these compounds against the selected

target. In contrast chemogenomics aims at the identification of the bioactivity of all these compounds against multiple targets and even beyond: in a very general sense the goal of chemogenomics is the exploration of all possible ligand–target interactions, or in other words the identification of bioactive compounds from the chemical space for all targets of the biological space.

The compound–target matrix plays a central role in chemogenomics. Its columns are formed by the set of all possible targets encoded in the genes of organisms (not necessarily only human genes), and the rows represent all the compounds that span the huge chemical space of fragments and lead- or drug-like compounds. The matrix elements describe the biological interaction, for example, a classification as active/inactive or a quantitative description by  $IC_{50}/EC_{50}$  or raw % CTRL values. Each row of this matrix displays the activity profile (the bioprint) of a compound, and each column displays the compound-binding profile of a target (the chemoprint).

Regarding experimental data the compound–target matrix is and will remain extremely sparse. Given the huge size of the relevant chemical space and ten thousands of potential targets, it is obviously impossible to fill the matrix with assay data. Hence, in silico approaches are the alternative to complement the bioprints of the compounds and the chemoprints of the targets.

Calculating the interaction strength of a wide diversity of compounds and targets represents a challenging goal, and computational chemogenomics is by far not yet mature enough to always provide reliable predictions. Despite this, it is a very attractive goal for pharmaceutical research. The prediction of the biological profile of compounds would allow the identification of potential off-targets, which may cause unwanted side effects of a drug. This information would help to prioritize the targets for the safety profiling and could be used to optimize compounds toward reduced side effects. Knowledge of the similarity between proteins can pave the way to chemical starting points or tools for innovative targets. There is also increasing evidence that most if not all drugs bind to a variety of targets (called polypharmacology)

with relevance for the therapeutic action of the drugs and/or for the side effects. The knowledge of the target spectrum of a drug is crucial information for the so-called drug repurposing where a known drug is applied to a new disease. It can also help to get better insight into the disease relevant targets and pathways and to identify new and better approaches to treat a disease, for example, by multitarget drugs. Moreover in phenotypic screening the target is mostly unknown and the activity profile of an active compound may be the key to the identification of the relevant target(s).

The basic assumption that guides all computational approaches in chemogenomics is that similar compounds bind to similar targets and therefore show a similar binding profile. Conversely targets that bind similar ligands have similar binding sites. The fundamental question in chemogenomics is how to measure and compare the similarity of compounds and targets, respectively. It is worth to mention that compounds with a similar bio profile may nevertheless have dissimilar structures which is the reason why biological fingerprints have a potential for scaffold hopping. The successful use of these descriptors in virtual screening goes back to the 1990s and was one of the earliest applications of the concept of chemogenomics.

An increasing wealth of experimental data about target–ligand interactions is available in the public domain, which is at least partly compiled in annotated chemical libraries. Therefore most of the chemogenomics studies rely on this data even if its quality, comparability, and completeness may be difficult to assess.

The term “chemogenomics” has been defined as the discovery and description of all possible drugs to all possible drug targets. The interaction of chemical and biological matter may be tackled starting from either end. Target-based approaches are grounded on a similarity measure derived from the quantitative comparison of sequence and/or three-dimensional (3D) structural information on targets. The subsequent binding profile prediction is based on the assumption that known ligands of a similar target may also bind to the target of interest, which allows us to make predictions for orphans without known ligands, too. Docking of compounds

to experimental or calculated protein structures is an alternative approach to use the 3D structure of targets. Whereas sequence data is available for all interesting targets due to the mapping of the human genome, 3D information is still incomplete but rapidly growing.

Ligand-based approaches, which are limited to targets with at least one known active compound, start from the other end and can include available information about target–ligand interactions in different ways. They all build upon existing knowledge about bioactive compounds. Some approaches just categorize ligands as active or inactive and compare targets based on the similarity of their ligand sets. Others consider explicitly the strength of the target–ligand interaction in the model-building process. Last but not least both protein and compound information can be used in target–ligand-based approaches. Some methods combine target similarity with information about known ligands; others take the details of the interaction on the atomic level into account to derive predictive models with machine-learning techniques.

### 7.2.1. Target-Based Similarity Methods

Similarity measures, which are grounded on target properties alone, provide the most direct access to target comparison. Such an approach does not need bioactive compounds to derive a similarity relationship between targets. It therefore allows searching for off targets for a drug without limitation to targets with known ligands. In the case of orphans ligands of similar targets can serve as a chemical starting point in deorphanization.

On the other hand ligand-based methods have a direct relationship to the pharmacological action of chemical matter, while any kind of similarity derived from sequence or structural data alone needs to be translated into a pharmacologically relevant scale, which is a critical and error-prone step.

Methods and tools to compare and to hierarchically cluster proteins based on their sequence similarity are well established,

and the so-called phylogenetic trees not only are the basis for the reconstruction of evolutionary history of life but also are routinely employed for the identification of potential targets for selectivity testing or safety profiling in drug research. It is, however, well known that targets from different families with low sequence similarity may nevertheless bind similar ligands. One example is the family of serotonin (5-HT) receptors. They are all activated by the neurotransmitter serotonin and belong to the superfamily of G protein-coupled receptors (GPCRs), with the exception of 5-HT<sub>3</sub>, which is an ion channel with very low overall similarity to the other members. The same phenomenon can be observed with the nicotinic and muscarinic acetylcholine receptors. The inducible cyclooxygenase-2 (COX-2) and the totally unrelated carbonic anhydrase (CA) show affinity to the same inhibitors celecoxib and valdecoxib. Glycogen synthase kinase 3 beta (GSK3 $\beta$ ) and cyclin-dependent kinase 4 (CDK4) show very similar structure-activity relationships despite a sequence identity of only 28%. Another example is the frequently observed binding of compounds to the human Ether-à-go-go-Related Gene (hERG) ion channel. The primary target of the vast majority of these compounds is phylogenetically unrelated to the hERG channel, but yet persistent binding to this antitarget is one of the major reasons for the early termination of drug research projects.

It may therefore be misleading to compare the overall sequence of targets. Instead, the focus should be on those parts that are relevant for ligand binding. These binding sites can be described by their (discontinuous) amino acid sequences or by the 3D properties of the binding pockets. The latter approach requires either experimental 3D structures of proteins or reliable homology models. The accurate identification of the binding sites is a prerequisite for meaningful results.

An example for the use of sequence data combined with limited structural information regarding the location of the binding site can be found in the work The authors aligned the transmembrane domain sequences of 369 human GPCRs and used the X-ray structure of the bovine rhodopsin-retinal complex to identify 30



discontinuous amino acids that are most likely to form the ligand-binding site. Clustering of the receptors based on these 30 amino acids yielded a phylogenetic tree that displays the relationship between the receptors. Gloriam revisited this approach by considering the additional GPCR X-ray structures that had been published in the meantime. They expanded the set of relevant amino acids to 44, including all previously identified 30 residues. Even if there are differences in the details both the clustering published by Surgand and by Gloriam reflects very well the results of the phylogenetic analysis by Fredriksson which was based on the full sequence of 342 GPCRs. Interestingly a clustering that is based only on those residues out of the 44 that most likely interact with the group of bioaminergic ligands yields different results that better reflect the pharmacologically relevant receptor similarity of the respective GPCRs. Milletti came to a similar conclusion when they compared binding sites of kinases. They found that the prediction of the target-binding profile of a compound requires the selection of the appropriate sub pockets that are actually occupied by the compound.

Even a close neighborhood of receptors measured in terms of the binding relevant amino acid sequence does not always result in a similar ligand-binding profile. The bradykinine receptors B1 and B2 are closely related in all sequence-based analyses. The endogenous peptidic ligand bradykinine is bound by B2 but not by B1, which is caused by a single residue exchange between B1 and B2. A Ser residue in the transmembrane helix 3 of B2 is replaced by Lys in B1. The positive charge in the binding site of B1 repels the C-terminal

Arg in bradykinine. In contrast the C-terminally truncated desArg<sup>9</sup>-bradykinine is bound by B1 since the negative charge of the C-terminus is attracted by Lys. This example demonstrates that seemingly minor changes in either binding sites or ligands may drastically influence the binding profile. These “activity cliffs,” that is, discontinuities in structure–activity relationships, are found when looking both at targets and at ligands. Activity cliffs have been in the focus of interest since a number of years,



and the interested reader is referred to a recent review.

The comparison of targets using the detailed 3D structure analysis of binding sites typically involves the following four steps:

- (i) Identification of binding pockets,
- (ii) Conversion of the residues lining the pockets into a simplified representation,
- (iii) Alignment of the patterns, and
- (iv) Quantitative assessment of the similarity between the patterns by a scoring function.

There are several programs available for the first step, the identification of binding pockets, which were recently reviewed. The authors conclude that encouragingly the programs perform well and can tolerate deviations up to 2 Å (heavy atoms) between ligand-unbound and ligand-bound protein structures. Limitations are encountered if binding pockets are very narrow in the unbound state. The authors also tested the ability of the programs to cope with homology models and found that the quality of predictions was comparable to the one found with native proteins in the tested cases. They observed again that too narrow binding sites in the homology model are a hurdle for the prediction and noticed that the overall quality of the structure in terms of root mean square (RMS) deviation does not correlate with the modeling quality of the binding site. It is suggested to use molecular dynamics calculations to generate a more realistic picture of the plasticity of the binding site. An alternative could be to incorporate the ligand into the homology modeling process.

A number of recent reviews focuses on steps 2–4 in the abovementioned binding site comparison process. A multitude of approaches were developed to tackle the problem, and examples of the successful identification of similar binding sites not closely related by sequence are presented for all these methods. In addition to alignment-dependent approaches alignment-free methods were also published in recent years, which avoid the time-consuming and critical alignment step.

Despite all advances in this field there are still challenges. Most methods are able to detect highly similar binding sites but may vary in their performance if the binding sites are of medium similarity. The definition that residues of a protein are involved in binding to a particular ligand is not clear without referring to the X-ray structure of the protein–ligand complex. It was already mentioned that the degree of the sequence-based similarity between two targets changes if only the residues that are relevant for binding are taken into consideration. The same is true if the comparison is grounded on the 3D properties of the binding site. Ligands do not need to fill a binding site completely to achieve sufficient affinity but instead may interact only with sub pockets. The ability of an algorithm to identify similarity of targets on the sub cavity level is therefore a necessary and important feature.

Even similar binding sites may exhibit variations in shape upon ligand binding due to the plasticity of the protein structure. Hence, algorithms for binding site comparison need to show some degree of fuzziness, while on the other hand a sufficient accuracy is required.

There is no generic and unambiguous similarity score threshold that separates similar binding sites in terms of ligand-binding properties from dissimilar ones. Similarity searches in the binding pocket space produce only an enrichment of true-positive results among false positives and will miss false negatives. Any cutoff is a compromise between recall and precision and may be case dependent, quite similar to searching for bioactive compounds by virtual screening in the chemical space. Moreover one should also be aware of cases where a target is predicted to bind a specific ligand. A target–ligand complex may actually be formed in agreement with the prediction, but the affinity of the ligand may be too weak (e.g.,  $>10\ \mu\text{M}$ ) to be recognized in the particular assay, and the target will erroneously be considered as a false-positive hit.

No binding site comparison approach can currently successfully deal with a situation where a ligand binds to two different targets but in different orientations that fit into binding sites that hardly

show any substantial similarity. In principle, docking could be an alternative but requires a substantial computational effort and a careful assessment of the docking poses due to the limitations of current scoring functions. The interested reader is referred to recent publications.

### 7.2.2. Ligand-Based Target Comparison

At first glance it might appear surprising that ligands, that is, small molecules, are suited to deduce quantitative similarity information on targets. The reason for this is the nature of the ligand–target interaction. Emil Fischer was the first to postulate in 1894 that the interaction of a protein with a small molecule can be described in a simplified analogy by the lock-and-key principle. Although it is known today that the binding process between ligand and target is much more complex than Fischer supposed, a complementarity between both partners is required. Therefore, a single ligand can be considered as a negative imprint of its own specific interaction site and the sum of all ligands of a target provides a negative imprint of the complete binding site. The latter statement, however, is true only if the known ligands of a target describe the ligand–target interaction in its entirety.

Regarding the available knowledge that is applied to gain information from the ligands, three types of target comparison procedures can be distinguished. Chemocentric approaches use the ligand sets of targets themselves to deduce knowledge on the respective target/binding site similarities. Chemoprint comparisons make use of biological activity data or comparable parameters that indicate the strength of the relationship between ligand and target, and proteochemometric approaches describe the interaction of each ligand with its target on a very detailed level. All approaches require known ligands for each target to be described and thus cannot be used for orphans. Despite this limitation the development and application of ligand-based methods is in the focus of many publications, and substantial progress has been gained during recent years.

## *Chemocentric Approaches*

The similarity ensemble approach (SEA). SEA is an application of the “similarity principle,” which states that structurally similar compounds have similar biological activity, reflecting the experience medicinal chemists have made since a long time. The first systematic studies regarding the validity of the similarity principle were published in the mid-nineties. In the following years, different research groups have reached different conclusions. In a careful analysis of biological data collected at Abbott, Martin found that there is only a 30% chance that a compound with a Tanimoto similarity  $\geq 0.85$  to an active compound is active itself. Even if it is much better than random, it is a surprisingly low value. The similarity principle with respect to the neighborhood behavior within a combinatorial library was recently revisited. The authors basically confirmed the conclusions of the former study and in particular found a strong and unpredictable dependence of search results on the employed query, in spite of a variety of descriptor spaces that were used. These results shed some light on the limitations of the similarity principle. As pointed out by Maggiora rugged activity landscapes with activity cliffs are much more frequent than assumed in the past. For this reason the similarity principle should not be taken as a fundamental principle but more as a valuable guideline with exceptions.

SEA goes beyond a simple similarity search. Inspired by the bioinformatics method BLAST SEA uses a statistical model to derive an expectation E-value for the comparison of compound sets rather than using the Tanimoto similarity itself. This E-value describes the significance of the pairwise similarity of compounds or compound sets. Thus, if compound sets for two distinct targets are compared, the E-value can also be considered as the strength of the relationship between these two targets. To derive the underlying statistical model, the similarity of random compound sets of different sizes from the MDL Drug Data Report (MDDR, Accelrys, San Diego, CA, USA) using Daylight fingerprints and the Tanimoto coefficient  $T_c$ . For each combination of compound sets a raw score was calculated as the sum of all pairwise comparisons

of the compounds from set 1 with all compounds from set 2. Raw scores were calculated for 300,000 pairs of random sets in a size interval between 10 and 1,000. The mean raw score, which was linearly dependent on the product of the set sizes as well as the standard deviation of the raw scores, was fit against the product of the set sizes, resulting in two functions for set-size-dependent expected mean raw scores and mean raw score standard deviations. Based on these expectation values and the individual raw scores, Z-scores were calculated by reproducing the above-described procedure using  $T_c$  thresholds for the calculation of raw scores in the range between 0.00 and 0.99 and by a fit to an extreme value distribution. The distribution derived from a threshold of 0.57 resulted in the best fit. For this reason only  $T_c$  values  $\geq 0.57$  are used in SEA calculations.

Thus, if two ligand sets do not have a single pair of compounds with a similarity of at least 0.57 the raw score is 0. It is important to note that this value needs to be recalibrated if other than Daylight descriptors are used.

With the functions for the expectation values of the mean raw score, raw score standard deviation, and the Z-score distribution, an E-value can be calculated for every comparison of two ligand sets. This E-value quantitatively describes the probability to obtain the same or a better raw score just by chance. Keiser used this method to analyze a 246-receptor subset of the MDDR. The pairwise comparison of all ligand sets as described above yields the result that the majority of the compound sets had a similarity not better than random. Only 5% of the calculated E-values could be interpreted as a statistically significant similarity between the targets based on their ligand sets and displayed as a cross-target similarity network. The authors found that on average any given receptor was similar to 5.8 other receptors with an E-value  $< 10^{-10}$ .

Moreover, the ligand-derived E-value from this statistical model can directly be compared with a sequence-derived E-value from a BLAST search. The MDDR database with known sequences of the relevant targets. The pairwise comparison of the ligand sets and targets using SEA and BLAST, respectively, yielded a rank order

of similarities, which were analyzed by the Spearman rank-order correlation coefficient. The authors found few examples (e.g., serine proteases) where ligand and target sequence similarities agreed well, but such correspondences were more the exception than the rule. In general there was no correlation between the sequence- and the ligand-based similarity.

To further support the validity of the statistical model and of the SEA approach in a prospective manner, Keiser predicted novel targets for the three known drugs methadone, emetine, and loperamide. According to SEA methadone should be an M3 receptor antagonist, emetine should antagonize the  $\alpha 2$  receptor, and loperamide was predicted to be an NK2 antagonist. All three predictions could be confirmed by experiment (methadone: 1  $\mu$ M antagonist at M3; emetine and loperamide: micromolar antagonists of the  $\alpha 2$  and the NK2 receptor, respectively).

While SEA describes the similarity of ligand sets and thereby the similarity of their targets by comparing the ligands themselves, Bender presented a different approach, introducing the "Bayes Affinity Fingerprint" (BAF). Rather than comparing two compounds by their binary fingerprint, which indicates the presence or absence of substructures, BAFs describe compounds by the scores calculated by multiple-activity class-specific Bayesian models. Bayesian models are predicated on Bayes's theorem, named after the English mathematician and Presbyterian minister Thomas Bayes (~1701–1761). In the implementation used by the authors, the Bayesian model calculates the probability that any compound, containing a feature F from the ECFP\_4 feature space, belongs to an activity class A, given the total number of compounds containing the feature F and the number of compounds with feature F that belong to activity class A. Bender used the WOMBAT 2005.01 database containing more than 100,000 bioactivity data points to train 1,003 activity class-specific Bayesian models. The similarity of two compounds is then expressed by the Pearson correlation coefficient of their activity class scores.

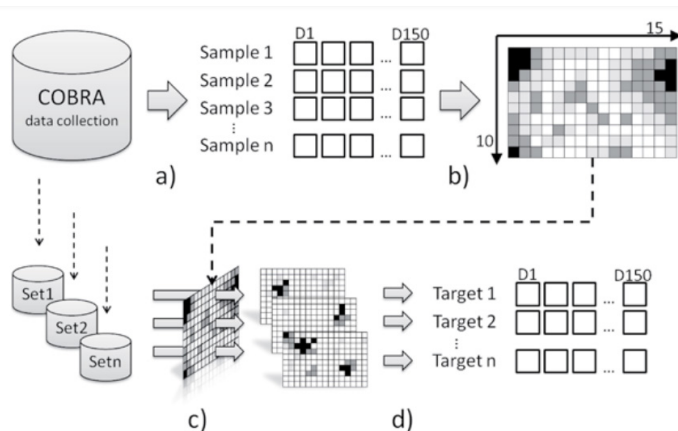
The approach used by Bender is very similar to the Prediction of Activity Spectra for Substances (PASS). However, while PASS

calculates the probability of a compound being active at a given target, Bender use the combined information from 1,003 targets as a compound descriptor to calculate intercompound similarities, comparable to the approach of Kauvar who used experimental assay data or the *in silico*-generated fingerprints of Briem who employed the program DOCK. Applying the BAFs as a similarity descriptor to a benchmark dataset, Bender improved the retrieval rates by about 24% in the top 5% of the hit list compared to the ECFP\_4 as a descriptor set. These improved retrieval rates indicate that the transformation from the graph based compound descriptor into a bioactivity space descriptor incorporates some knowledge about the chemical space of the 1,003 targets.

Moreover, Bender used this approach to compare the targets themselves. The generated Bayesian models are composed of a set of ECFP\_4 features with positive and negative coefficients, depending on the frequency of the feature in the active and inactive compounds, respectively, of the training set. Model comparison and the subsequent target comparison could be achieved by comparing the coefficients of identical features for different activity classes. Due to the large number of substructural features this is a computationally very expensive task. Bender therefore evaluated a different approach: for the 100 largest activity classes, Bayesian scores for 102,500 compounds from the MDDR database were calculated. Afterwards, principal components for this matrix were calculated, yielding only 9 of 100 eigenvalues larger than 1, an observation that reflects a low dimensionality of the BAF space. In principle, the variable loadings of the 100 Bayesian models could have been used to determine model similarities and derived target similarities. While this was not in the focus of the work of Bender they could nevertheless show that the nine models with the highest correlation with the selected nine principal components are sufficient to achieve retrieval rates similar to those of the ECFP\_4 themselves. This indicates that the information content described by these nine Bayesian models is similar to the information content given by the ECFP\_4 descriptor, notwithstanding the dramatically reduced dimensionality.



Obviously, the BAF-derived target similarity depends on the ligand set that is used to generate the matrix for the principal component analysis. A more direct insight into the bioactivity feature space would be provided by the comparison of the learned features themselves. Such an analysis has been introduced. The authors trained a self-organizing map (SOM) with a set of 10,840 known drugs from the COBRA data collection, encoded by the 150-dimensional two-dimensional (2D) topological descriptor CATS. After training of the SOM, the dataset was split into 174 target specific subsets and each subset was presented to the SOM in such a way that each compound was assigned to 1 of the 150 neurons of the trained SOM, applying the “winner-take-all” function. Subsequently, the distribution of the target specific ligand sets was scaled to the interval and transformed into a 150-dimensional target descriptor by assigning each neuron of the SOM to one position of the descriptor. The different steps of this process are illustrated in Figure 6. Thus, by using a nonlinear, robust, and noise-tolerant projection method like SOM, Schneider transformed a target-specific ligand set into a target fingerprint descriptor of the same dimensionality as the ligand descriptor. This target fingerprint was then used to describe target similarities by the pairwise Pearson correlation between targets.



**Figure 6.** Workflow for the generation of target fingerprints.



In a subsequent analysis of the pairwise target fingerprint correlations, Schneider generated a target network, using a cutoff value for the intertarget connections of  $r > 0.2$ . Comparing this network with the equivalent network, derived directly from similarity comparisons of the CATS descriptor-encoded compound sets, reveals the advantage of the SOM-based approach: targets that belong to related protease subfamilies (like serine proteases or metalloproteases) are clustered. Moreover for particular targets like the  $\gamma$ -secretase or the hepatitis C virus protease NS3, the target fingerprint representation suggests that a relationship to other targets that is based solely on the catalytic mechanism has to be reconsidered with regard to the nature of their ligands.

### *Chemoprint Approaches*

Solely make use of ligand sets that are annotated as active or inactive at their specific target. Quantitative information on the activity is at best used to distinguish between actives and nonactives. Other approaches go beyond this binary classification and use a quantitative description of the interaction between ligands and targets. The approaches discussed have the common feature that they build on a compound–target matrix, which is almost completely filled by experimental or calculated interaction data. While the bio prints of compounds can be used to establish a similarity relationship between ligands the chemoprints allow us to calculate the similarity of the targets. We want to shed some light on selected aspects of the history of these approaches.

The first who introduced the systematic analysis of activities of compounds at multiple targets as a bioactivity compound descriptor, a method referred to as “target-related affinity profiling,” TRAP. Their “affinity fingerprints” were assembled from the dose response values ( $IC_{50}$ ) of 122 compounds at eight targets (reference panel). The compounds and targets were selected out of a larger matrix based on correlation tests of the chemo prints of the targets and structural diversity of the compounds. A subset of 12 compounds, representing the most diverse binding profiles, was tested at two new targets, glutathione reductase (GRd) and

aldehyde dehydrogenase (AdDH). Based on the dose response values of the 12 compounds at the two new targets and their known affinity fingerprints from the reference panel, multivariate linear regression models were fitted for GRd and AdDH. In other words, the activity of a compound at a new target is described by a linear combination of its activities at each of the proteins of the reference panel. Subsequently, the regression models were applied to the entire set of 122 affinity fingerprints to select a new set of 10 compounds predicted to represent the range of relevant potencies more evenly. Final multivariate linear regression models for the 22 compounds were trained and applied to all 122 compounds. Fitting coefficients of 0.85 and 0.86 for GRd and AdDH, respectively, between the linear regression models and the experimentally derived values for all compounds after the second iteration demonstrated the potential of the TRAP approach for iterative screening, which was further highlighted by subsequent applications. It is worth to emphasize that the linear relationship of affinities between different targets was found to be valid also for sequentially unrelated proteins.

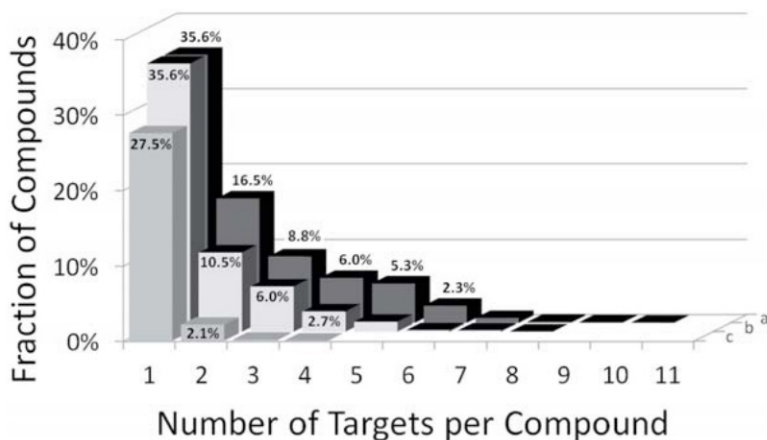
Encouraged by these results, Briem presented a similar approach by using DOCK scores instead of measured affinity data. The application of docking scores replaces the *in vitro* experiment by an *in silico* experiment and therefore broadens the applicability not only by reducing the experimental overhead but also by making it possible to derive fingerprints for not yet synthesized compounds. In contrast to Kauvar Briem did not aim for the prediction of quantitative data but for the classification of compounds into actives and in actives and the enrichment of known actives in retrospective similarity searches. In fact, using docking scores for the binding sites of eight known 3D protein structures from the Brookhaven Protein Databank, they found enrichment factors for the known ligands of different targets in the range of two to five within the first 5% of the scored hit sets. Lessel and Briem improved the method by using the program FlexX, which allows flexible docking. Even if the so-called FlexSimX fingerprints are somewhat inferior to the Daylight descriptors regarding the enrichment in similarity searches, the strength of the method is

its demonstrated scaffold hopping potential. The approach taken by Bender who described the BAFs, shows some relationship to the docking-based fingerprints. Both fingerprints encode the interaction of compounds with a set of targets. The BAFs, however, require a target-specific training of Bayesian models, which is not required for the docking-based fingerprints.

Recent publications on the usability of the compound–target matrix are the “biological spectra analysis,” which was introduced by Fliri and the application of GPCR affinity profiling at Hoffmann-La Roche with the objective of finding a nonpeptidic somatostatin receptor subtype 5 (SSTR5) antagonist. Based on a dataset of % inhibition values at 10  $\mu$ M of 1,567 compounds at 92 ligand-binding assays from the BioPrint database of the bio print is a completely sufficient descriptor to cluster compounds in accordance to biological responses to an additional target, which was not part of the 92 assays. Guba could corroborate this result. Following the finding that astemizole, an antagonist at the histamine H1 receptor, also antagonizes the SSTR5 receptor in the micromolar range, a set of 5,000 ligands from an in-house GPCR-directed library was tested against a panel of 15 GPCRs at Cerep and the resulting affinity profile was compared with that of astemizole, resulting in a new lead structure.

All mentioned publications demonstrate applications of the chemogenomics matrix for the purpose of compound screening or, more generally, for the quantification of compound similarity by means of affinity fingerprints. There have been fewer attempts to quantify the similarity of the targets by employing chemogenomics data. Metz list some metrics that have been applied to the question of chemoprint-derived target similarities. Besides the usage of the Pearson correlation coefficient, a common approach to analyze chemoprints with the objective of target comparison is to first transform the chemoprint into a binary vector of actives and inactives by introducing some threshold that separates the two classes. Subsequently Tanimoto similarities between the binary representations are calculated.

Pharmacology interaction strength which is calculated as the fraction of compounds that do not exceed a certain selectivity threshold between two targets, normalized by the total number of compounds that are active at both targets. The selectivity cutoff is set arbitrarily, which is an obvious disadvantage of this and similar procedures. The polypharmacological potential clearly depends on the hit threshold (Figure 7).



**Figure 7.** Distribution of multiple target hitters with respect to different hit thresholds.

Probably the first to use the chemogenomics matrix without a threshold were Vieth who introduced the “SAR similarity,” which is basically the mean of the selectivity’s of all compounds measured in two assays, normalized by the total range of the measurements in log units and subtracted from 1.

In our opinion the chemogenomics similarity of two targets depends not only on the selectivity of the compounds but also on the potency. Nonselective compounds with high potencies should be more relevant for target similarities than nonselective compounds with borderline potency. This reflects the experience of medicinal chemists who typically observe a correlation of selectivity’s to off-targets and the potency of compounds during the phase of lead optimization. Therefore we suggested an “assay-related target similarity” (ARTS) metric as the sum of all compounds, measured

in two assays, weighted by an affinity-dependent score, as given in Equation 4:

$$ARTS = \frac{\sum_{i=1}^n \left( k + \frac{pIC_{50,a1_i} + pIC_{50,a2_i}}{2} \right)^2 e^{-(pIC_{50,a1_i} - pIC_{50,a2_i})^2}}{\sqrt{\sum_{i=1}^n (k + pIC_{50,a1_i})^2 \sum_{i=1}^n (k + pIC_{50,a2_i})^2}} \quad (4)$$

where ARTS is the similarity of two assays  $a1$  and  $a2$ ,  $k$  is constant ( $k = -4$ ), and  $IC_{50,a1i}$  and  $IC_{50,a2i}$  are the dose response values of compound  $i$  with  $i = 1, 2, \dots, n$  in assays  $a1$  and  $a2$ , respectively

The score rewards potent compounds and considers the selectivity of a compound by introducing a smooth Gaussian function that adds a nonlinear penalty term to the score, depending on the selectivity of the compounds. The advantage of this function is the fact that the penalty term can be adjusted in accordance to the observed accuracy of the assays. Fluctuations of dose response values that may exist due to the assays can be treated less restrictively than by using a simple linear penalty function. By implementing these two functions into the score, we are able to deprioritize compounds with low potency or high selectivity and to prioritize highly potent and nonselective compounds without introducing any arbitrary cutoff value.

We applied ARTS to a dataset of 3,500 compounds measured in dose response at 11 pharmaceutically relevant GPCRs, calculated all pairwise ARTS similarities between the targets, and compared them to sequence-based similarities. Surprisingly we found some degree of similarity between the cannabinoid receptor type 2 (CB2) and the somatostatin receptor type 4 (SST4). To analyze to what extent our results depend on the dataset, we compared hierarchical clustering results of a randomly assigned subset of the dataset with the results obtained from the entire dataset. We found that using ARTS as a similarity metric, we could decrease the dataset by 60% and still obtain stable and reproducible target similarities, while other metrics like the Pearson correlation coefficient showed a massive decrease in reproducible target similarities.

Typically, datasets are gathered from the collecting published data for activities. If compounds are not explicitly described as actives for a certain target, they are defined to be inactive. Obviously, this is not necessarily true. We analyzed the influence of simulated sparsity on the stability of the pairwise similarities of the targets in our chemogenomics dataset by randomly deleting single activities. Surprisingly we could show that while the deletion of entire compounds does not influence the similarity of the targets significantly, the insertion of holes into the matrix does have a severe influence. However, the target similarities calculated by ARTS are still more stable than those derived by Pearson correlation. Nevertheless this shows that careful collection and curation of the dataset are extremely important, and in the case of holes in the matrix, well-defined procedures for the usage of imputation techniques like inverse distance weighting are more suited than defining unknown affinities as inactives.

### **7.2.3. The Impact of Data Quality on Chemogenomics**

In general, models are only as good as the data they are based on. Whenever models are used to predict the properties of compounds or targets this inherent limitation needs to be taken into account.

In that respect there is no difference between, for example, in silico models of absorption, distribution, metabolism, and excretion (ADME) and chemogenomics models. However, in the field of in silico ADME the data quality issue and the reliability of models have been systematically investigated and discussed for many years. There is still a lack of such systematic studies in chemogenomics, in spite of an increasing number of publications that address this topic.

#### ***Experimental Variations and Errors***

Issues in data quality may arise from experimental values measured under different conditions that are assembled to larger datasets. Variations of assay conditions, for example, regarding

the used cell line, the chemical nature of the displaced ligand, or the type of readout to name only a few important parameters, may lead to inconsistent data. Drug-like molecules frequently exhibit low solubility, and undetected precipitation during the assay may seriously distort the results and lead to false negatives. It is difficult to collect homogeneous data of high quality even within a single institution, for example, a pharmaceutical company. It is even more difficult to assess the quality of published data. A careful manual curation is needed, as highlighted frequently. Even doing so some experimental errors like the mentioned solubility problem may remain hidden. In recent years databases with detailed binding or functional assay results have become publicly available. Still, there is a careful analysis of the data regarding experimental variations required using them for model building.

Vidal and Mestres, in a systematic analysis of the affinity profiles of 13 antipsychotic drugs for 34 protein targets, recently studied the prediction of these profiles using more than 21,000 reference compounds from public databases with measured affinities to these 34 targets. They noticed that the variations in the affinity data for compounds with more than one data point per target showed an average standard deviation of 0.5 log units irrespective of the affinity range, which is within usual error limits of such experimental data. Other authors noticed variations up to 1 log unit. Vidal and Mestres characterized the molecules with a combination of descriptors that were based on pharmacophoric fragments, featurepair distributions, and Shannon entropy. The affinity of each of the 13 drugs for each target was derived from the affinity landscape of the reference compounds surrounding the drugs using an inverse distance-weighting interpolation. They were able to predict 65% of all affinities within a 1 log unit error with a precision of more than 90%. This is an encouraging result that could be achieved despite the fact that the authors had to rely on experimental data from diverse sources. Nevertheless date that explicitly employ quantitative affinity data, and the publication of Vidal and Mestres is the first attempt to predict the affinity data of a complete compound–target matrix. It is therefore still too early to assess the influence of experimental errors and neglected assay



variations of published compound–target interaction data on the results of chemogenomics studies.

### ***Thresholds and Cutoff Values***

Most publications in the field of computational chemogenomics on approaches where compounds are classified into active and inactive ones. Thus, the elements of the compound–target matrix are reduced to binary data. Classification models built upon such data do not depend on experimental errors and may even tolerate variations in assay conditions as long as there is no misclassification of training compounds. The risk of misclassification is low in case of highly active or completely inactive compounds and increases if the activity of a compound is close to the cutoff that separates active and inactive compounds. Often a cutoff value of 10  $\mu\text{M}$  is used similar to the cutoff that is very common in high-throughput screening to distinguish “hits” from “nonhits.” There is no sharp boundary between active and inactive compounds, and potentially interesting compounds and even whole compound classes may be excluded from the analysis due to either experimental variations and errors or minor structural modifications that reduce the affinity of the compounds just beyond the cutoff. In a recent publication Briansó systematically investigated the influence of both biological and similarity thresholds on cross-pharmacology profiles of GPCR ligands within and outside of the GPCR target family. The authors observed a pronounced effect of both thresholds on the resulting profiles. The inclusion of less similar and less active compounds in the analysis clearly increased the crosspharmacology beyond the GPCR target family of the compounds.

### ***The Dataset Composition***

The composition of the datasets used for model building in terms of structural and biological diversity has a strong impact on the results of any chemogenomics analysis. The situation is similar to conventional quantitative structure–activity relationship (QSAR) models where ligands and their affinities to a single target



provide the input to machine-learning techniques. The situation in chemogenomics, however, is more complex since the goal is to fill the compound–target matrix with predicted values for many targets.

The currently available annotated chemical libraries have only a limited diversity with respect to the represented chemo types and often contain series of similar compounds, which is critically pointed out by some authors. These databases reflect to a large extent the history of pharmaceutical research and the targets and compound classes that were in the focus of interest years ago. It would be desirable that research groups in the field of chemogenomics perform a detailed analysis of the diversity of the chemical libraries that are used in their studies and moreover characterize the compound sets by the distribution of physicochemical properties (e.g., MW and clog P) and the degree of drug-likeness. It was shown that the topology of drug–target networks derived from annotated chemical libraries implicitly depends on these parameters.

Another factor that has a general influence on the quality of QSAR and chemogenomics models is the composition of the decoy set. This is an intensively discussed matter also in virtual screening. Comprehensive assay data on inactive compounds for a given target is hardly ever published. Therefore decoy sets need to be artificially composed. It is important that a decoy set be structurally sufficiently different from the active compounds in order to assemble almost certainly nonbinders but still have a similar distribution of physicochemical parameters. A frequently used approach is to assume that compounds not reported to be ligands at a given target are inactive at this target even if this bears some risk of false negatives. Emphasize the importance of a completely random sampling of these presumed inactives. Weill and Rognan studied the influence of the decoy selection on the prediction of ligand–GPCR complexes in detail. If the decoys are selected from ligand sets of GPCRs that have a distantly related binding site to the receptors where the active compounds are taken from, the prediction of targets for a given compound is improved, while the

prediction of ligands for a given receptor is less accurate. So far, to our best knowledge this is the only detailed report regarding the impact of the decoy selection process. It would be worthwhile to investigate the consequences of the decoy selection in further studies.

### *The Lack of Data Completeness*

The lack of data completeness in chemogenomics has gained increasing attention in recent years. It is clear that sparse compound–target matrices provide an incomplete picture of multitarget activity landscapes and conclusions derived from such incomplete data must be interpreted with caution. In many studies drug–target interaction networks are constructed based on the accumulated information in protein–ligand interaction databases. In these networks targets are connected if they share common ligands and ligands are connected if they share common targets. There is no consensus in the scientific community how to best decide on the thresholds that should be set in such networks. Paolini connect two proteins if both bind one or more compounds, if the affinity of the compounds is  $< 10 \mu\text{M}$ , if the affinity difference between the compounds is not larger than a factor of 1,000, and if at least 10 compounds were commonly tested at both targets. Keiser connect two targets by an edge in the network if they share ligands sets with at least five members each, with a pairwise Tanimoto coefficient of at least 0.57 between two compounds of each set, and an E-value of  $\leq 1$ . In many studies two targets are connected if they share at least one active compound that is listed in one of the public databases. Nisius and Bajorath draw an edge between two targets if they share at least five active compounds, Gregori-Puigjané and Mestres construct a network only from targets that share at least 10 drugs. It should be emphasized that our method ARTS in contrast does not use any arbitrary cutoff.

The influence of the data space used for the network construction. Depending on the kind and number of databases and the number of unique drug–target interactions, the number of targets per drug increased from 1.8 to 5.9, while at the same time the networks

changed from a clearly organized modular structure to a highly connected and complex topology. The value of 5.9 connections per drug increases even further up to 13 if in silico predictions of additional targets are added. There is a broad consensus in the scientific community that polypharmacology is a widespread property of most or even all drugs. However, networks derived from incomplete data should be considered with caution. The strength of the relationship between targets can increase or decrease with the publication of new or revised data, and relationships that are based on only a single shared bioactive compound are built on shaky ground.

### *The Applicability Domain*

Last but not least all models have a limited applicability domain. If a QSAR model was trained with a given set of compounds and associated data and the properties of new compounds never seen by the model are to be predicted, it is a key question how reliable the predictions will be. To provide an estimate of the expected accuracy is a topic that has been intensively discussed in many publications in recent years. Closely related to this question is the use of local versus global models. Local models use a limited set of objects, be it compounds or targets, which share some common properties, for example, structural similarity in the case of compounds and sequence or 3D similarity in the case of targets. Global models, in contrast, use the full range of available data. Local models may reach a higher accuracy but with the price of a more limited applicability domain. Global models may show the opposite profile.

In the context of chemogenomics the situation is even more complex than in other QSAR applications, for example, in silico ADME, since not only the properties of new compounds for a given target are to be predicted but also the question if a new target belongs to the applicability domain of a model describing a different target needs to be answered. In other words, how different can a compound or a target be to still belong to the applicability domain of the model, and how can this difference be measured?

There are not too many publications in chemogenomics that explicitly address that topic. Vidal and Mestres used a set of different descriptors to predict the affinity of compounds for given targets based on the affinity landscape of similar compounds. They calibrated the similarity threshold for each descriptor by its ability to discriminate active from random compounds in public databases and used this threshold to define an applicability domain for their chemogenomics studies. Weill and Rognan investigated the performance of local versus global models for GPCRs and their ligands with proteochemometric approaches. They concluded that local models clearly outperform the global ones. They also made the interesting observation that the most predictive models in terms of cross-validation of the training sets were not the best ones regarding the prediction of test sets. A recent review suggest the use of Gaussian processes to measure the reliability of predictions. Gaussian processes can take the statistical uncertainty of the training data, for example, by experimental fluctuation, into account. They are nevertheless bound to the diversity of the training set. If compounds or targets are structurally outside of the applicability domain, Gaussian processes, too, cannot make a reliable prediction anymore.

To the best knowledge of the authors of this review there is only one publication that addresses explicitly the applicability domain issue of the target space. Investigated a set of compounds against 14 human immunodeficiency virus (HIV) reverse transcriptase (RT) mutants. They employed a proteochemometric approach using 6,314 known compound–target combinations, including 14 different HIV RT sequences. They predicted prospectively the activity of 130 compounds and 317 untested compound–mutant pairs. The predictions were measured subsequently and a rootmean-square error (RMSE) of the predictions comparable to the assay reproducibility was achieved (0.6 log units for the calculated model vs. 0.5 log units for the experimental data).

## REFERENCES

1. Abernethy J, Bach F, Evgeniou T, Vert JP (2009) A new approach to collaborative filtering: operator estimation with spectral regularization. *J Mach Learn Res* 10:803–826
2. Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. *Mach Learn* 73(3):243–272
3. Bock JR, Gough DA (2002) A new method to estimate ligandreceptor energetics. *Mol Cell Proteomics* 1(11):904–910
4. Bock JR, Gough DA (2005) Virtual screen for ligands of orphan G protein-coupled receptors. *J Chem Inf Model* 45(5):1402–1414
5. Frimurer T, Ulven T, Elling C, Gerlach LO, Kostenis E, Hogberg T (2005) A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg Med Chem Lett* 15:3707–3712
6. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl Acids Res* 40(D1):D1100–D1107
7. Gozalbes R, Rolland C, Nicola E, Paugam MF, Coussy L, Horvath D, Barbosa F, Mao B, Revah F, Froloff N (2005) QSAR strategy and experimental validation for the development of a GPCR focused library. *QSAR Comb Sci* 24(4):508–516
8. Harrell F (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Graduate texts in mathematics. Springer, Berlin
9. Horvath D, Marcou G, Varnek A (2013) Do not hesitate to use tversky—and other hints for successful active analogue searches with feature count descriptors. *J Chem Inf Model* 53(7):1543–1562
10. Ivanciuc O (2007) Applications of support vector machines in chemistry. Wiley, New York, pp 291–400
11. Jacob L, Hoffmann B, Stoven V, Vert JP (2008) Virtual

- screening of GPCRS: an in silico chemogenomics approach. *BMC Bioinform* 9(1):363
12. Jacob L, Vert JP (2008) Protein–ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 24(19):2149–2156
  13. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg J (2008) Proteochemometric modeling of hiv protease susceptibility. *BMC Bioinform* 9(1):181
  14. Lapins M, Prusis P, Gutcaits A, Lundstedt T, Wikberg J (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug–receptor interactions. *Biochim Biophys Acta* 1525:180–190
  15. Li S, Xi L, Wang C, Li J, Lei B, Liu H, Yao X (2009) A novel method for protein–ligand binding affinity prediction and the related descriptors exploration. *J Comput Chem* 30(6):900–909
  16. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucl Acids Res* 34(Suppl. 2):W32–W37
  17. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA (2013) Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov Today* 18(9–10):495–501
  18. Rosenbaum L, Dorr A, Bauer MR, Boeckler FM, Zell A (2013) Inferring multi-target QSAR models with taxonomy-based multitask learning. *J Cheminform* 5:1–20
  19. van Westen G, Swier R, Cortes-Ciriano I, Wegner J, Overington J, IJzerman A, Van Vlijmen H, Bender A (2013) Benchmarking of protein descriptors in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptors. *J Cheminform* 5:42
  20. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A (2010) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* 2(1):16–30

21. Weill N, Rognan D (2009) Development and validation of a novel protein–ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J Chem Inf Model* 49(4):1049–1062
22. Yabuuchi H, Niiijima S, Takematsu H, Ida T, Hirokawa T, Hara T, Ogawa T, Minowa Y, Tsujimoto G, Okuno Y (2011) Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Mol Syst Biol* 7(472)





# INDEX

## A

Absorption, distribution, metabolism, and excretion (ADME) 260  
academic environment 66  
acid secretion 81, 88  
Affinity fingerprints 255, 256, 257  
Aldehyde dehydrogenase (AddH) 256  
Algorithm Comparison Procedure 57  
aminophylline 183  
Angiotensin-converting enzyme (ACE) 10  
apremilast 183  
Arylpiperazine groups 5  
ATP citrate lyase (ACL) 28  
atrial natriuretic factor (ANF) 191  
Autophagic Compound-Target Prediction (ACTP) 138

## B

Bacterial potassium channels 152  
Bioinformatics 138, 139, 140, 142, 143, 144  
Bioinformatics-based inference 117  
biological system 82, 86  
bipartite local models (BLMs) 55  
blood pressure 81

## C

Calcium-channel blockers 149  
calmodulin 190, 192, 212, 213, 214  
cAMP-dependent phosphodiesterase (PDE) 199  
Carbonyl groups 152  
catecholamines 194, 199  
chemical genetics 79, 85, 86,

87, 89, 92, 109, 110  
 Chemical genetics 3, 4, 35  
 ChemMapper 133  
 chemogenomic features 46, 49, 50, 51  
 Chemogenomic informatics 115, 117  
 Chemogenomics 41, 42, 45, 63  
 Chemogenomics application 120  
 Chloride channels 154  
 Classical genetics 4  
 collective variable (CV) 238  
 Computational chemogenomics (CG) 222  
 computational technique for predicting 42  
 Convergent evolution 120  
 cyclic adenosine monophosphate (cAMP) 180, 205, 210, 216  
 cyclic guanosine monophosphate (cGMP) 180, 205, 216, 218  
 Cyclin dependent kinase 5 (CDK5) 239  
 cytotoxicity 81

## D

Data mining 142, 143  
 Dictiostelium 186  
 DNA sequences 138, 139  
 DNILMF algorithm 60  
 Drosophila 188  
 drug discovery 41, 43, 44, 52, 62, 63, 74, 78, 79, 81, 83, 84, 86, 89, 102, 110

Drug discovery programs 147  
 drug–drug interactions (DDIs) 53  
 Druggability 148  
 drug-induced biology 83  
 Drug phenotypic effects 45  
 Drug Target Interaction projection enactment 43  
 Drug–target interactions 43, 46, 47  
 DTI predictions 53, 61, 62

## E

Electrophysiological studies 156  
 Embryonic stem (ES) 4  
 endoplasmic reticulum 188  
 Escherichia coli 192  
 Extended topochemical atom (ETA) 129  
 Extraction of Chemogenomic Features 49

## G

Gene3D 120  
 Gene product 115  
 genetic interactions 41  
 Glycogen synthase kinase-3 $\beta$  (GSK-3 $\beta$ ) 4  
 G protein-coupled receptors (GPCRs) 12

## H

Histone deacetylase (HDAC) 28  
 Homology 120, 121  
 Human druggable genome 119  
 human genome 80, 81, 82

Human immunodeficiency  
virus (HIV) 266  
Hyperkalemic periodic paraly-  
sis 149

## I

ibudilast 183  
Informatics 115, 117, 118, 119,  
120  
insulin 193, 194, 195  
insulin-like growth factor (IGF)  
195  
Ion channel modulators 147  
Ion channels 148

## L

Ligand Binding Mode Rela-  
tionships 125  
Ligand information 157  
LigBase 121  
Lipid environment 152

## M

Mechanosensitive ion channel  
154  
Membrane proteins 150  
metabolism 81, 87, 98  
methylxanthine 183, 184  
Mycobacterium tuberculosis  
154

## N

National Institutes of Health  
(NIH) 4  
Nimodipine 149  
Nonspecific inhibitors 183

Nuclear magnetic resonance  
(NMR) 23

## P

PDE3 inhibitors 179, 182, 183,  
184, 185, 194, 200, 201,  
202, 203, 204  
PDE4 inhibitors 179, 182, 183,  
184, 185  
PDE5 inhibitors 179, 182, 183,  
184, 185, 202, 203, 204,  
205, 212  
PFAM database 47  
pharmaceutical industry 79,  
80, 82, 83, 93  
pharmacokinetics 81, 100  
Pharmacophore 129, 130, 131,  
132, 145  
phenotypic effects 45, 46  
phosphodiesterase enzymes  
(PDE enzymes) 179  
Phosphodiesterase inhibitors  
(PDE inhibitors) 179  
Potassium channels 151, 155  
Prediction of Activity Spectra  
for Substances (PASS)  
252  
probable targets 64  
protein 82, 83, 85, 86, 87, 89, 90,  
91, 94, 95, 101, 102, 107  
protein complexes 82  
protein sequence, position-spe-  
cific score matrix (PSSM)  
53  
Protein structure information  
119

Proteomics 139  
PubChem database 44

## Q

Quantitative structure-activity  
relationship (QSAR) 126

## R

Reverse Docking 135  
Reverse transcriptase (RT) 266  
roflumilast 183

## S

second messengers 180, 216  
Selective optimization of side  
activities\'\' (SOSA) 7  
Self-organizing map (SOM)  
254  
signaling pathway 195, 198,  
207  
Similarity ensemble approach  
(SEA) 250  
Site alignments 118  
small-molecule action 41  
Sodium channels 149  
sparse canonical correspond-  
ence analysis (SCCA) 46  
Structure-activity relationship  
(SAR) 18, 33  
Structure-based drug design  
116, 125

Suberoylanilide hydroxamic  
acid (SAHA) 28  
support vector machine (SVM)  
48, 53

Support vector regression  
(SVR) 227

## T

Target Fishing 132  
Target-identification and mech-  
anism-of-action 41  
target identification (TI) 86  
Target validation (TV) 84  
theophylline 183, 184  
Toxicity Estimation Software  
Tool (T.E.S.T.) 128  
toxicology 81, 87  
Traumatic brain injury model  
158

## V

vasodilation 179, 180, 181, 182,  
185, 202, 204  
Virtual Target Screening (VTS)  
137  
Voltage-gated channels 151  
Voltage gated sodium channels  
151

## W

Water-filled central cavity 152

## X

X-ray structures 152, 161, 168



## Chemogenomics

Until the recent sequencing of the human genome, drug discovery has long been a multi-disciplinary effort to optimize ligands properties (potency, selectivity, and pharmacokinetics) towards a single macromolecular target. A robust knowledge of the interactions between small molecules and specific proteins aids the development of new biotechnological tools and the identification of new drug targets, and can lead to specific biological insights. Chemogenomics is a complementary strategy for the investigation of chemically related compounds and libraries against various members of a target family. It is largely based on the intelligent application of automated parallel synthesis. Chemogenomics is a new strategy in drug discovery which, in principle, searches for all molecules that are capable of interacting with any biological target. Because of the almost infinite number of drug-like organic molecules, this is an impossible task. Therefore Chemogenomics has been defined as the investigation of classes of compounds (libraries) against families of functionally related proteins.

This thorough book provides a collection of techniques used in the emerging field of Chemogenomics. Chemogenomic modeling is concerned with the application of techniques to extract patterns in ligand-target binding, aiming to exploit similarity of bioactivity between similar molecules which is then expected to contribute to the identification of bioactive pairs more efficiently than random molecule selection and activity measurement. In this book, the concepts and core elements to build a computational Chemogenomic platform are presented, with special emphasis on adaptive instance selection by the active learning technique. Despite its adaptation to drug discovery almost two decades prior, it is not until recently that active learning has been investigated in Chemogenomic contexts; nonetheless, the technique is demonstrating the ability to build models of ligand-target binding that are also predictive on external prediction challenges.

**Sandro Castro** holds MS in Genetic Engineering and PhD in Genetics and Genomics. He has published papers in peer reviewed journals on topics genetics and biotechnology. His research interest areas are bioinformatics, Chemogenomics, data mining and web application.