

Chemogenomics and Chemical Genetics

Grenoble Sciences

The aims of Grenoble Sciences are double:

- ▶ to produce works corresponding to a clearly defined project, without the constraints of trends or programme,
- ▶ to ensure the utmost scientific and pedagogic quality of the selected works: each project is selected by Grenoble Sciences with the help of anonymous referees. Next, the authors work for a year (on average) with the members of an interactive reading committee, whose names figure in the front pages of the work, which is then co-published with the most suitable publishing partner.

Contact: Tel.: (33) 4 76 51 46 95 - E-mail: Grenoble.Sciences@ujf-grenoble.fr
website: <http://grenoble-sciences.ujf-grenoble.fr>

Scientific Director of Grenoble Sciences: **Jean BORNAREL**,
Emeritus Professor at Joseph Fourier University, Grenoble, France

Grenoble Sciences is a department of Joseph Fourier University, supported by
the **French National Ministry for Higher Education and Research**
and the **Rhône-Alpes Region**.

Chemogenomics and Chemical Genetics is an improved version of the original
book ***Chemogénomique - Des petites molécules pour explorer le vivant***
sous la direction de Eric MARÉCHAL, Sylvaine ROY et Laurence LAFANECHÈRE,
EDP Sciences - Collection Grenoble Sciences, 2007, ISBN 978 2 7598 0005 6.

The Reading Committee of the French version included the following members:

- ▶ **Jean DUCHAINE**, Principal Advisor of the Screening Platform, Institute for Research in Immunology and Cancer, University of Montreal, Canada
- ▶ **Yann GAUDUEL**, Director of Research at INSERM, Laboratory of Applied Optics (CNRS), *Ecole Polytechnique*, Palaiseau, France
- ▶ **Nicole MOREAU**, Professor at the *Ecole Nationale Supérieure de Chimie*, Pierre and Marie Curie University, Paris, France
- ▶ **Christophe RIBUOT**, Professor of Pharmacology at the Faculty of Pharmacy, Joseph Fourier University, Grenoble, France

*Translation performed by **Philip SIMISTER***

*Typesetted by **Centre technique Grenoble Sciences***

*Cover illustration: **Alice GIRAUD***

(with extracts from a DNA microarray image - *Biochip Laboratory/Life Sciences Division/CEA* - and a photograph of actin filaments array and adhesion plates in a mouse embryonic cell - *Yasmina SAOUDI, INSERM U836 Grenoble, France*)

Eric Maréchal • Sylvaine Roy • Laurence Lafanechère
Editors

Chemogenomics and Chemical Genetics

A User's Introduction for Biologists,
Chemists and Informaticians

 Springer

Editors

Dr. Eric Maréchal
Laboratory of Plant Cell Physiology
UMR 5168, CNRS-CEA-INRA-
Joseph Fourier University
Rue des Martyrs 17
38054 Grenoble Cedex 9
France
eric.marechal@cea.fr

Sylvaine Roy
Laboratory of Plant Cell Physiology
UMR 5168, CNRS-CEA-INRA-
Joseph Fourier University
Rue des Martyrs 17
38054 Grenoble Cedex 9
France
sylvaine.roy@cea.fr

Laurence Lafanechère
Albert Bonniot Institute
Department of Cellular Differentiation and Transformation
Rond-point de la Chantourne
38706 La Tronche Cedex
France
laurence.lafanechere@ujf-grenoble.fr

Translator:

Philip Simister
Weatherall Institute of Molecular Medicine
University of Oxford
Oxford OX3 9DS, UK

Originally published in French: *Chemogénomique - Des petites molécules pour explorer le vivant* sous la direction de Eric MARÉCHAL, Sylvaine ROY et Laurence LAFANECHÈRE, EDP Sciences - Collection Grenoble Sciences, 2007, ISBN 978 27598 0005 6.

ISBN 978-3-642-19614-0 e-ISBN 978-3-642-19615-7
DOI 10.1007/978-3-642-19615-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930786

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover illustration: Alice Giraud

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

CONTENTS

Preface	1
----------------------	----------

Introduction	3
---------------------------	----------

FIRST PART

AUTOMATED PHARMACOLOGICAL SCREENING

Chapter 1 - The pharmacological screening process: the small molecule, the biological screen, the robot, the signal and the information	7
--	----------

Eric MARÉCHAL - Sylvaine ROY - Laurence LAFANECHÈRE

1.1. Introduction	7
1.2. The screening process: technological outline.....	8
1.2.1. Multi-well plates, robots and detectors	8
1.2.2. Consumables, copies of chemical libraries and storage	10
1.2.3. Test design, primary screening, hit-picking, secondary screening	10
1.3. The small molecule: overview of the different types of chemical library	12
1.3.1. The small molecule	12
1.3.2. DMSO, the solvent for chemical libraries.....	12
1.3.3. Collections of natural substances	12
1.3.4. Commercial and academic chemical libraries.....	14
1.4. The target, an ontology to be constructed	14
1.4.1. The definition of a target depends on that of a bioactivity.....	14
1.4.2. Duality of the target: molecular entity and biological function	15
1.4.3. An ontology to be constructed	16
1.5. Controls	17
1.6. A new discipline at the interface of biology, chemistry and informatics: chemogenomics	17
1.7. Conclusion.....	18
1.8. References	19

Chapter 2 - Collections of molecules for screening: example of the french national chemical library	23
--	-----------

Marcel HIBERT

2.1. Introduction	23
2.2. Where are the molecules to be found?	25
2.3. State of progress with the European Chemical Library	27
2.4. Perspectives	28
2.5. References	28

Chapter 3 - The miniaturised biological assay: constraints and limitations 29*Martine KNIBIEHLER*

3.1. Introduction	29
3.2. General procedure for the design and validation of an assay	30
3.2.1. Choice of assay	31
3.2.2. Setting up the assay	33
3.2.3. Validation of the assay and automation	35
3.3. The classic detection methods	36
3.4. The results	36
3.4.1. The signal measured: increase or decrease?	36
3.4.2. The information from screening is managed on three levels	37
3.4.3. Pharmacological validation	40
3.5. Discussion and conclusion	40
3.6. References	41

Chapter 4 - The signal: statistical aspects, normalisation, elementary analysis 43*Samuel WIECZOREK*

4.1. Introduction	43
4.2. Normalisation of the signals based on controls	44
4.2.1. Normalisation by the percentage inhibition	44
4.2.2. Normalisation resolution	44
4.2.3. Aberrant values	46
4.3. Detection and correction of measurement errors	48
4.4. Automatic identification of potential artefacts	49
4.4.1. Singularities	49
4.4.2. Automatic detection of potential artefacts	50
4.5. Conclusion	52
4.6. References	52

Chapter 5 - Measuring bioactivity: K_i , IC_{50} and EC_{50} 55*Eric MARÉCHAL*

5.1. Introduction	55
5.2. Prerequisite for assaying the possible bioactivity of a molecule: the target must be a limiting factor	55
5.3. Assaying the action of an inhibitor on an enzyme under Michaelian conditions: K_i	56
5.3.1. An enzyme is a biological catalyst	57
5.3.2. Enzymatic catalysis is reversible	57
5.3.3. The initial rate, a means to characterise a reaction	59
5.3.4. Michaelian conditions	59
5.3.5. The significance of K_m and V_{max} in qualifying the function of an enzyme	60
5.3.6. The inhibited enzyme: K_i	60
5.4. Assaying the action of a competitive inhibitor upon a receptor: IC_{50}	62
5.5. Relationship between K_i and IC_{50} : the CHENG-PRUSOFF equation	63
5.6. EC_{50} : a generalisation for all molecules generating a biological effect (bioactivity)	64
5.7. Conclusion	64
5.8. References	65

Chapter 6 - Modelling the pharmacological screening: controlling the processes and the chemical, biological and experimental information.....	67
<i>Sylvaine ROY</i>	
6.1. Introduction	67
6.2. Needs analysis by modelling.....	68
6.3. Capture of the needs	69
6.4. Definition of the needs and necessity of a vocabulary common to biologists, chemists and informaticians	69
6.5. Specification of the needs	69
6.5.1. Use cases and their diagrams	70
6.5.2. Activity diagrams	72
6.5.3. Class diagrams and the domain model.....	73
6.6. Conclusion.....	78
6.7. References	78
Chapter 7 - Quality procedures in automated screening.....	79
<i>Caroline BARETTE</i>	
7.1. Introduction	79
7.2. The challenges of quality procedures.....	79
7.3. A reference guide: the ISO 9001 Standard.....	80
7.4. Quality procedures in five steps	82
7.4.1. Assessment	82
7.4.2. Action plan - planning.....	83
7.4.3. Preparation	83
7.4.4. Implementation.....	83
7.4.5. Monitoring.....	83
7.5. Conclusion.....	84
7.6. References	84

SECOND PART

HIGH-CONTENT SCREENING AND STRATEGIES IN CHEMICAL GENETICS

Chapter 8 - Phenotypic screening with cells and forward chemical genetics strategies.....	87
<i>Laurence LAFANECHÈRE</i>	
8.1. Introduction	87
8.2. The traditional genetics approach: from phenotype to gene and from gene to phenotype.....	88
8.2.1. Phenotype	88
8.2.2. Forward and reverse genetics	89
8.3. Chemical genetics	89
8.4. Chemical libraries for chemical genetics	90
8.4.1. Chemical library size.....	91
8.4.2. Concentration of molecules.....	91
8.4.3. Chemical structure diversity.....	91
8.4.4. Complexity of molecules	93

8.4.5. Accessibility of molecules to cellular compartments.....	93
8.4.6. The abundance of molecules.....	94
8.4.7. The possibility of functionalizing the molecules.....	94
8.5. Phenotypic tests with cells.....	94
8.6. Methods to identify the target.....	96
8.7. Conclusions.....	99
8.8. References.....	99

Chapter 9 - High-content screening in forward (phenotypic screening with organisms) and reverse (structural screening by NMR) chemical genetics..... 103

Benoît DÉPREZ

9.1. Introduction.....	103
9.2. Benefits of high-content screening.....	104
9.2.1. Summarised comparison of high-throughput screening and high-content screening.....	104
9.2.2. Advantages of high-content screening for the discovery of novel therapeutic targets.....	104
9.2.3. The nematode <i>Caenorhabditis elegans</i> : a model organism for high-content screening.....	105
9.2.4. Advantages of high-content screening for reverse chemical genetics and the discovery of novel bioactive molecules.....	108
9.3. Constraints linked to throughput and to the large numbers.....	110
9.3.1. Know-how.....	110
9.3.2. Miniaturisation, rate and robustness of the Assays.....	110
9.3.3. Number, concentration and physicochemical properties of small molecules.....	111
9.4. Types of measurement for high-content screening.....	111
9.4.1. The critical information needed for screening.....	111
9.4.2. Raw, numerical results.....	111
9.4.3. Results arising from expert analyses.....	112
9.5. Conclusion.....	112
9.6. References.....	112

Chapter 10 - Some principles of Diversity-Oriented Synthesis..... 113

Yung-Sing WONG

10.1. Introduction.....	113
10.2. Portrait of the small molecule in DOS.....	114
10.3. Definition of the degree of diversity (DD).....	116
10.3.1. Degree of diversity of the building block.....	116
10.3.2. Degree of stereochemical diversity.....	118
10.3.3. Degree of regiochemical diversity.....	119
10.3.4. Degree of skeletal diversity.....	121
10.4. Divergent multi-step DOS by combining elements of diversity.....	124
10.5. Convergent DOS: condensation between distinct small molecules.....	127
10.6. Conclusion.....	130
10.7. References.....	130

THIRD PART

TOWARDS AN IN SILICO EXPLORATION OF CHEMICAL AND BIOLOGICAL SPACES

Chapter 11 - Molecular descriptors and similarity indices	135
<i>Samia ACI</i>	
11.1. Introduction.....	135
11.2. Chemical formulae and computational representation.....	136
11.2.1. The chemical formula: a representation in several dimensions.....	137
11.2.2. Molecular information content.....	137
11.2.3. Molecular graph and connectivity matrix.....	139
11.3. Molecular descriptors.....	140
11.3.1. 1D descriptors.....	141
11.3.2. 2D descriptors.....	141
11.3.3. 3D descriptors.....	144
11.3.4. 3D versus 2D descriptors?.....	146
11.4. Molecular similarity.....	147
11.4.1. A brief history.....	147
11.4.2. Properties of similarity coefficients and distance indices.....	147
11.4.3. A few similarity coefficients.....	148
11.5. Conclusion.....	148
11.6. References.....	150
Chapter 12 - Molecular lipophilicity: a predominant descriptor for QSAR	153
<i>Gérard GRASSY - Alain CHAVANIEU</i>	
12.1. Introduction.....	153
12.2. History.....	153
12.3. Theoretical foundations and principles of the relationship between the structure of a small molecule and its bioactivity.....	154
12.3.1. QSAR, QPAR and QSPR.....	154
12.3.2. Basic equation of a QSAR study.....	155
12.4. Generalities about lipophilicity descriptors.....	155
12.4.1. Solubility in water and in lipid phases: conditions for bioavailability.....	155
12.4.2. Partition coefficients.....	156
12.4.3. The partition coefficient is linked to the chemical potential.....	156
12.4.4. Thermodynamic aspects of lipophilicity.....	157
12.5. Measurement and estimation of the octanol/water partition coefficient.....	158
12.5.1. Measurement methods.....	158
12.5.2. Prediction methods.....	160
12.5.3. Relationship between lipophilicity and solvation energy: LSER.....	163
12.5.4. Indirect estimation of partition coefficients from values correlated with molecular lipophilicity.....	163
12.5.5. Three-dimensional approach to lipophilicity.....	165
12.6. Solvent systems other than octanol/water.....	166
12.7. Electronic parameters.....	167
12.7.1. The HAMMETT parameter, σ	167
12.7.2. SWAIN and LUPTON parameters.....	168

12.8. Steric descriptors	169
12.9. Conclusion	169
12.10. References	169

Chapter 13 - Annotation and classification of chemical space in chemogenomics 171

Dragos HORVATH

13.1. Introduction	171
13.2. From the medicinal chemist's intuition to a formal treatment of structural information	171
13.3. Mapping structural space: predictive models.....	174
13.3.1. Mapping structural space	174
13.3.2. Neighbourhood (similarity) models	175
13.3.3. Linear and non-linear empirical models	178
13.4. Empirical filtering of drug candidates.....	180
13.5. Conclusion.....	181
13.6. References	181

Chapter 14 - Annotation and classification of biological space in chemogenomics 185

Jordi MESTRES

14.1. Introduction	185
14.2. Receptors	186
14.2.1. Definitions	186
14.2.2. Establishing the 'RC' nomenclature	187
14.2.3. Ion-channel receptors	188
14.2.4. G protein-coupled receptors	189
14.2.5. Enzyme receptors	190
14.2.6. Nuclear receptors	190
14.3. Enzymes	191
14.3.1. Definitions.....	191
14.3.2. The 'EC' nomenclature	192
14.3.3. Specialised nomenclature.....	193
14.4. Conclusion.....	193
14.5. References	194

Chapter 15 - Machine learning and screening data..... 197

Gilles BISSON

15.1. Introduction	197
15.2. Machine learning and screening.....	199
15.3. Steps in the machine-learning process	202
15.3.1. Representation languages.....	203
15.3.2. Developing a training set	205
15.3.3. Model building.....	206
15.3.4. Validation and revision	207
15.4. Conclusion.....	209
15.5. References and internet sites	209

Chapter 16 - Virtual screening by molecular docking.....	213
<i>Didier ROGNAN</i>	
16.1. Introduction.....	213
16.2. The 3 steps in virtual screening.....	213
16.2.1. Preparation of a chemical library.....	213
16.2.2. Screening by high-throughput docking.....	216
16.2.3. Post-processing of the data.....	218
16.3. Some successes with virtual screening by docking.....	220
16.4. Conclusion.....	221
16.5. References.....	222

APPENDIX

BRIDGING PAST AND FUTURE?

Chapter 17 - Biodiversity as a source of small molecules for pharmacological screening: libraries of plant extracts.....	227
<i>Françoise GUERITTE, Thierry SEVENET, Marc LITAUDON, Vincent DUMONTET</i>	
17.1. Introduction.....	227
17.2. Plant biodiversity and North-South co-development.....	229
17.3. Plant collection: guidelines.....	230
17.4. Development of a natural-extract library.....	231
17.4.1. From the plant to the plate.....	231
17.4.2. Management of the extract library.....	231
17.5. Strategy for fractionation, evaluation and dereplication.....	233
17.5.1. Fractionation and dereplication process.....	233
17.5.2. Screening for bioactivities.....	235
17.5.3. Some results obtained with specific targets.....	236
17.5.4. Potential and limitations.....	239
17.6. Conclusion.....	240
17.7. References.....	240
Glossary.....	241
The authors.....	253

PREFACE

Jean CROS

Having completed the reading of this work, one can only feel satisfied for having encouraged Laurence LAFANECHÈRE, Sylvaine ROY and Eric MARÉCHAL, who attempted and succeeded in achieving the impossible: writing, along with colleagues from the public sector, a book that will endure, concerning a technology the mastery of which had remained until this point the domain of the pharmaceutical industry. Indeed, this work has arisen from the competence and practical knowledge of fifteen or so academic scientists who, often against the tide of strategies defined by their host organisations, have established automated pharmacological screening for fundamental research ends. It is important to recall that the first book *High throughput screening*, edited in 1997 by John P. DEVLIN, which enabled all scientists to discover the importance of robotics in the discovery of new medicines, was written by about a hundred contributors, all of whom were industrial scientists involved in drug discovery.

Over the last ten years, we have seen appear in the scientific literature much more about ‘small’ molecules coming from robotic screens that have been used with success in revealing new biological mechanisms. From drug candidate, the molecule has thus become a research tool.

The successful experience at Harvard is a fertile example which should serve as a model for some of our research centres: basic research in chemical genetics, discovery of new drug candidates and training of young researchers. May this book, which has developed out of training workshops organised by the CNRS, CEA and INSERM, be the stimulus for future careers in a field which is eminently multidisciplinary and which brings together biologists, chemists, informaticians and robotics specialists. The great merit of this book is to have simply, from everyday experiences, united researchers and competencies that until now had not associated with one another.

Beyond the new terms that we discover or rediscover throughout the chapters: chemical genetics, cheminformatics, chemogenomics etc., there are the techniques, certainly, but also and above all there are the scientific questions to which these technologies will henceforth help to find answers. In addition, there are the economic issues that from now on become the duty of every researcher to take into account.

Congratulations to all of the authors and editors.

INTRODUCTION

André TARTAR

Over the last two decades, biological research has experienced an unprecedented transformation, which often resulted in the adoption of highly parallel techniques, be it the sequencing of whole genomes, the use of DNA chips or combinatorial chemistry. These approaches, which have in common the repeated use of trial and error in order to extract a few significant events, have only been made possible thanks to the progress in miniaturisation and robotics informatics.

One of the first sectors to put into practice this approach was within pharmaceutical research with the systematic usage of high-throughput screening for the discovery of new therapeutic targets and new drug candidates. Academic research has for a long time remained distanced from this process, as much for financial as for cultural reasons. For several years, however, the trivialisation of these techniques has led to a considerable reduction in the cost of accessing them and has thus permitted academic groups to employ such methods in projects having generally more cognitive objectives.

Nevertheless, it is no less vital, as with all involved methods, to take into account the cost factor as a fundamental parameter in the development of an experimental protocol relative to the expected benefit. The value of a chemical library is in effect an evolving notion resulting from the sum of two values that evolve in opposite directions:

- » On the one hand, the set of physical samples whose value will fatally decrease due both to its consumption in tests, but above all to the degradation of the components. The experience of the last few years also shows that it will be subjected to the effects of fashion, which will contribute rapidly to its obsolescence: no-one today would assemble a chemical library as would have been done only five years ago. Since the great numbers that dominated the first combinatorial chemical libraries, a more realistic series of criteria has progressively been introduced, bearing witness to the difficulties encountered. ‘Drugability’ has thus become a keyword, with LIPINSKI’s rule of 5 and the ‘frequent hitters’ becoming the *bête noire* of screeners having given them too often cause for hope, albeit unfounded.
- » On the other hand, the mass of information accumulated over the different screening tests is ever increasing and will progressively replace the physical chemical library. With a more or less distant expiry date, the physical chemical

library will have disappeared and the information that it has allowed to accumulate will be all that remains. This information can then be used either directly, constituting the ‘specification sheet’ of a given compound, or as a reference source in virtual screening exercises or *in silico* prediction of the properties of new compounds.

A very simple strategic analysis shows that with the limited means available to academic teams, it is easier to be competitive with respect to the second point (quantity and quality of information) than to the first (number of compounds and high thoughtput). This also shows that the value of an isolated body of information is much less than that of an array organised in a logical manner based on two main dimensions: the diversity of compounds and the consistency of the biological tests.

It is in this vein that high-content screening should become established, permitting the collection and storage of the maximum amount of data for each experiment. This high-content screening will be the guarantee for the optimal evaluation of physical collections. It is interesting to note that the problem of information loss during a measurement was at the centre of spectroscopists’ preoccupations a few decades ago. In the place of dispersive systems (e.g. prisms, networks) that sequentially selected each observation wavelength but let all others escape, they have substituted non-dispersive analysis techniques entrusting deconvolution algorithms and multi-channel analysers with the task of processing the global information. Biology is undergoing a complete transformation in this respect. Whereas about a decade ago one was satisfied by following the expression of a gene under the effect of a particular stimulus, today, thanks to pan-genomic chips, the expression profile of the whole genome has become accessible. It is imperative that screening follows the same path of evolution: no longer losing any information will become the rule. In the longer term, it will be necessary for this information to be formatted and stored in a lasting and reusable manner.

With this perspective, this book appears at just the right moment since it constitutes a reference tool enabling different specialists to speak the same language, which is essential to ensure the durability of the information accrued.