CHEMOGENOMICS AND CHEMICAL GENETICS

A user's introduction for biologists, chemists and informaticians

Edited by
Eric MARÉCHAL - Sylvaine ROY - Laurence LAFANECHÈRE

Translation by Philip SIMISTER

Springer
2011

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

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*)

EXTRACTS

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 thoughput). 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.

Chapter 1

THE PHARMACOLOGICAL SCREENING PROCESS: THE SMALL MOLECULE, THE BIOLOGICAL SCREEN, THE ROBOT, THE SIGNAL AND THE INFORMATION

Eric MARÉCHAL - Sylvaine ROY - Laurence LAFANECHÈRE

1.1. INTRODUCTION

Pharmacological screening implements various technical and technological means to select from a collection of molecules those which are active towards a biological entity. The ancient or medieval pharmacopeia, in which the therapeutic effects of mineral substances and plant extracts are described, arose from pharmacological screening but whose operative methods are either unknown or very imprecise (see chapter 17). Due to a lack of documentation, one cannot know if this ancient medicinal knowledge resulted from systematic studies carried out with proper methods or from the accumulation of a collective body of knowledge having greatly benefitted from individual experiences. Over the centuries, along with the classification and archiving of traditional know-how, the research into new active compounds has been oriented towards rational exploratory strategies, or screens, in particular using plants and their extracts. The approaches based on systematic sorting have proved their worth, for example, through the research into antibiotics.

The recent progress in chemistry, biology, robotics and informatics have, since the 1990s, enabled an increase in the rate of testing, giving rise to the term **high-throughput screening**, as well as the measurement of multiparametric signals, known as **high-content screening**. Beyond the medical applications, which have motivated the growth of screening technologies in pharmaceutical firms, the **small molecule** has become a formidable and accessible tool in fundamental research. The know-how and original concepts stemming from robotic screening have given rise to a new discipline, **chemogenomics** (BREDEL and JACOBY, 2004), a practical component of which is **chemical genetics**, which we shall more specifically address in the second part of this book.

Pharmacological screening involves very diverse professions, which have their own culture and jargon, making it difficult not only for biologists, chemists and informaticians to understand each other, but so, too, for those within a given discipline. What is an 'activity' to a chemist or to a biologist, a 'test' to a biologist or to an informatician, or even a 'control'? **Common terminology** must remove these ambiguities. This introductory chapter briefly describes the steps of an automated screening **process**, gives a preview of the different types of collections of molecules, or **chemical libraries**, and finally tackles the difficult question of what are the definitions of a **screen** and of **bioactivity**.

1.2. THE SCREENING PROCESS: TECHNOLOGICAL OUTLINE

1.2.1. MULTI-WELL PLATES, ROBOTS AND DETECTORS

Automated pharmacological screening permits the parallel testing of a huge number of molecules against a biological target (extracts, cells, organisms). For each molecule in the collection, a test enabling measurement of an effect on its biological target is implemented and the corresponding signal is measured. Based on this signal a choice is made as to which of the molecules are interesting to keep (fig. 1.1).

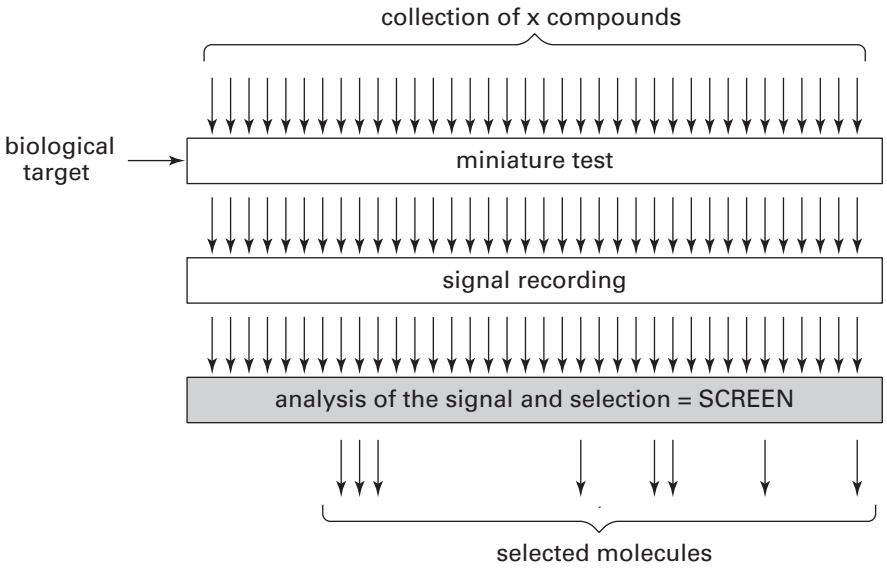


Fig. 1.1 - Scheme of a pharmacological screening process

The mixture of molecules and target as well as the necessary processes for the test are carried out in plates composed of multiple wells (termed **multi-well plates,** or **microplates**, fig. 1.2). These plates have standardised dimensions with 12, 24, 48, 96, 192, 384 or 1536 wells.

1.3. THE SMALL MOLECULE: OVERVIEW OF THE DIFFERENT TYPES OF CHEMICAL LIBRARY

1.3.1. THE SMALL MOLECULE

Small molecule is a term often employed to describe the compounds in a chemical library. It sums up one of the required properties, *i.e.* a molecular mass (which is obviously correlated with its size) less than 500 daltons. A small, active molecule is sought in collections of pure or mixed compounds, arising from natural substances or chemical syntheses.

1.3.2. DMSO, THE SOLVENT FOR CHEMICAL LIBRARIES

Dimethylsulphoxide (DMSO, fig. 1.5) is the solvent frequently used for dissolving compounds in a chemical library that were created by chemical synthesis. DMSO improves the solubility of hydrophobic compounds; it is miscible with water.

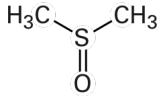


Fig. 1.5 - Dimethylsulphoxide (DMSO), the solvent of choice for chemical libraries

One of the properties of DMSO is also to destabilise the biological membranes and to render them porous, allowing access to deep regions of the cell and may provoke, depending on the dose, toxic effects. Although DMSO is accepted to be inert towards the majority of biological targets, it is important to determine its effect with appropriate controls before any screening. In case the DMSO is found harmful for the target, it is critical to establish at what concentration of DMSO there is no effect on the target and consequently to adapt the dilution of the molecules in the library. Sometimes, it may be necessary to seek a solvent better suited to the experiment.

1.3.3. COLLECTIONS OF NATURAL SUBSTANCES

Natural substances are known for their diversity (HENKEL *et al.*, 1999) and for their structural complexity (TAN *et al.*, 1999; HARVEY, 2000; CLARDY and WALSH, 2004). Thus, 40% of the structural archetypes described in the data banks of natural products are absent from synthetic chemistry. From a historical point of view, the success of natural substances as a source of medicines and bioactive molecules is evident (NEWMAN *et al.*, 2000).

Current methods for isolating a natural bioactive product, called bio-guided purifications, are iterative processes consisting of the extraction of the samples using solvents and then testing their biological activity (see chapter 17). The cycle of purification and testing is repeated until a pure, active compound is obtained. While allowing the identification of original compounds arising from biodiversity, this type of approach does present several limitations (LOCKEY, 2003). First of all,

the extracts containing the most powerful and/or most abundant bioactive molecules tend to be selected, whereas interesting but less abundant compounds, or those with a moderate activity,would not be not retained. Cytotoxic compounds can mask more subtle effects resulting from the action of other components present in the crude extract. Synergistic or cooperative effects between different compounds from the same mix may also produce a bioactivity that disappears later upon fractionation. Pre-fractionation of the crude extracts may, in part, resolve these problems (ELDRIGGE *et al.*, 2002). Mindful of these pitfalls, some pharmaceutical firms choose to develop their collections of pure, natural substances from crude extracts. This strategy, despite requiring significant means, can prove to be beneficial in the long term (BINDSEIL *et al.*, 2001). Lastly, with chemical genetics approaches (second part), the strategies adopted for identifying the protein target may necessitate the synthesis of chemical derivatives of the active compounds, which can present a major obstacle for those natural substances coming from a source in short supply and/or that have a complex structure (ex. 1.1).

Depending on the synthesis strategy used (see chapter 10), **two types of collection can be generated**: **target-oriented collections**, synthesised from a given molecular scaffold, and **diversity-oriented collections** (SCHREIBER, 2000; VALLER and GREEN, 2000). Each of these types of collection has its advantages and disadvantages. Compounds coming from a **target-oriented** collection have more chance of being active than those selected at random, however, they may only display activity towards a particular class of proteins. A diversity-oriented collection (chapter 10), on the other hand, offers the possibility of targetting entirely new classes of protein., Each individual molecule has, however, a lower probability of being active.

Example 1.1 - An anti-cancer compound from a sponge

Obtaining 60 g of discodermolide, an anti-cancer compound found in *Discodermia dissoluta* (GUNASEKERA *et al.*, 1990), a rare species of Caribbean sponge, would require 3,000 kg of dry sponges, *i.e.* more sponges than in global existence. Therefore, chemists have attempted to synthesise the discodermolide molecule. Only in 2004 did a pharmaceutical group announce that, after two years of work, they managed to produce 60 g of synthetic discodermolide, by a process consisting of over thirty steps (MICKEL, 2004). Discodermolide is now under evaluation in clinical studies for its therapeutic effect on pancreatic cancer.

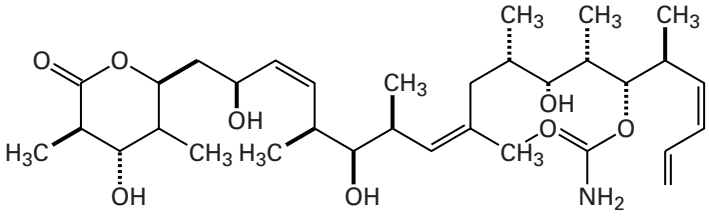


Fig. 1.6 - Discodermolide

Several groups have attempted to reproduce, with the help of combinatorial organic synthetic methods, the diversity and complexity of natural substances. The current developments in combinatorial synthesis are moving towards the simultaneous and

bioactive molecule seems sufficiently central to help remove the ambiguities over the terms target, screen, test, signal and control. In case of doubt, the reader is encouraged to consult the general glossary at the end of this book.

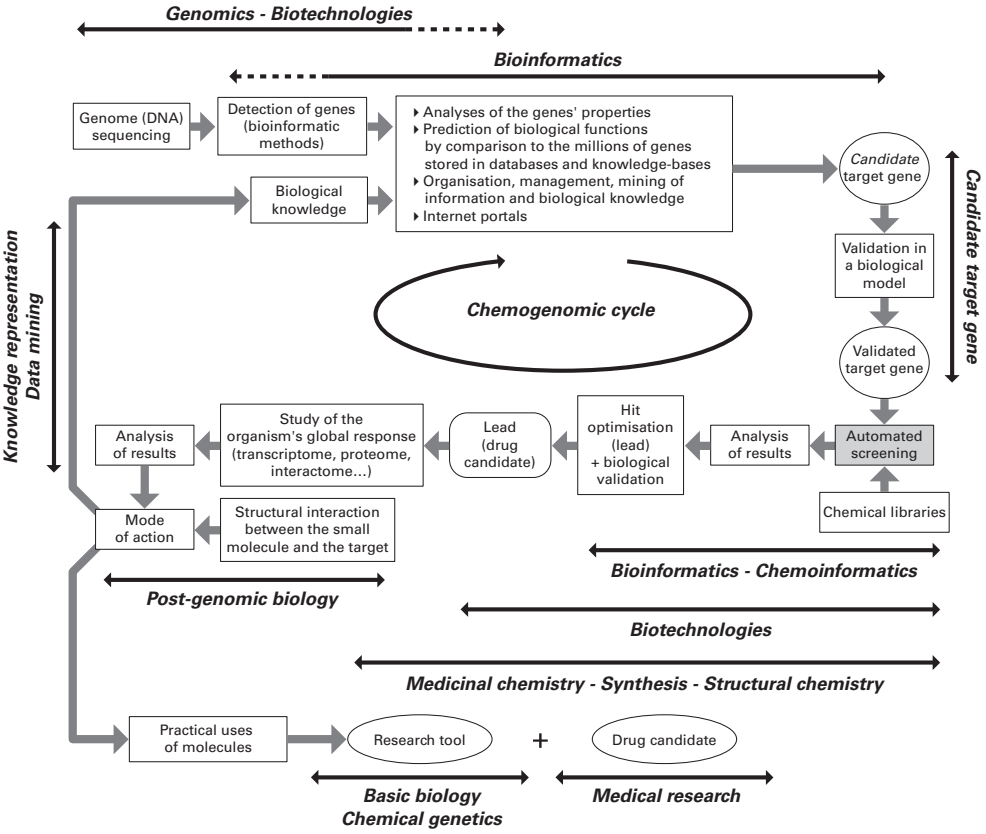


Fig. 1.9 - Chemogenomics, at the interface of genomics and post-genomic biology, chemistry and informatics

Chemogenomics aims to understand the relationship between the biological space of targets and the chemical space of bioactive molecules. This discipline has been made possible by the assembly of collections of molecules, the access to automated screening technologies and significant research in bioinformatics and chemoinformatics.

1.8. REFERENCES

[Harvard]
http://www.broad.harvard.edu/chembio/lab_schreiber/anims/animations/smdbSplitPool.php

ASHBURNER M., BALL C.A., BLAKE J.A., BOTSTEIN D., BUTLER H., CHERRY J.M., DAVIS A.P., DOLINSKI K., DWIGHT S.S., EPPIG J.T., HARRIS M.A., HILL D.P., ISSEL-TARVER L., KASARSKIS A., LEWIS S., MATESE J.C., RICHARDSON J.E., RINGWALD M., RUBIN G.M., SHERLOCK G. (2000) Gene ontology: tool for the unification of biology.
The Gene Ontology Consortium. *Nat. Genet.* **25**: 25-29

Chapter 8

PHENOTYPIC SCREENING WITH CELLS AND FORWARD CHEMICAL GENETICS STRATEGIES

Laurence LAFANECHÈRE

8.1. INTRODUCTION

A commonly used method to understand the role of complex biological systems and how they function is **to disrupt** them and then **to observe** the result of this disruption. A classic way to create such disruptions is to generate **genetic mutations** and then to observe the effect of these mutations on the cell or the organism. **Small organic molecules** can also cause disruption to the functioning of biological systems and can be employed to understand the role of the protein with which they interact. The history of biology is full of examples of complex systems whose **molecular functioning can be understood thanks to the use of drugs or ligands as well as to the characterisation of the protein targets of these ligands**. One such example is the role of colchicine in the discovery of tubulin, a component protein of microtubules (SHELANSKI and TAYLOR, 1967; PETERSON and MITCHISON, 2002).

The development of automated screening methods in an academic setting and the access to large collections of organic molecules has permitted systematising the exploration of the chemical world's diversity for isolating molecules active on biological systems. In parallel, the concept of 'chemical genetics' or 'pharmacological genetics' was born. This term may seem to be a corruption of language because in chemical genetics approaches we are not dealing with the gene but, most of the time, with the gene product. In fact, this concept designates a group of approaches that aims to use small molecules to interfere with proteins systematically and therefore to determine their function, in the same way as mutations are utilised in actual genetics. Conceptually therefore, chemical genetics and genetics are rather analogous. More recently the term **chemogenomics** was proposed to designate the multidisciplinary approaches aiming to dissect biological functions with the help of small molecules (BREDEL and JACOBY, 2004). Genetics has been a formidable engine for discovery in the biomedical sciences. Small molecules offer advantages compared to genetic technologies: they are versatile research tools, which can be quickly adopted by different laboratories and used for the precise control of the

function of certain proteins in a cellular context, particularly in the case of difficult, indeed impossible, genetic manipulation. Furthermore, these small molecules can be a first step towards the development of therapeutic agents. Indeed, on the one hand, they offer a means to test the possible involvement of a given protein in a pathology; on the other hand, their chemical structure can be the starting point for drug development.

This chapter will be devoted more specifically to the approaches for **phenotypic screening with cells**, in the context of **forward chemical genetics**. This approach, which allows the identification and characterisation of new proteins, has taken an increasingly important place in basic research and has proved to be complementary of biochemistry and genetics. For applied research, forward chemical genetics also offers the advantage of selecting drug candidates capable of penetrating cells straightaway and being active in a cellular context.

8.2. THE TRADITIONAL GENETICS APPROACH: FROM PHENOTYPE TO GENE AND FROM GENE TO PHENOTYPE

8.2.1. PHENOTYPE

Phenotype refers to the set of physical and physiological characteristics of an individual resulting from its genome and its environment. A phenotype can be defined on different levels, from the cell (fig. 8.1) to the whole organism (chapter 9).

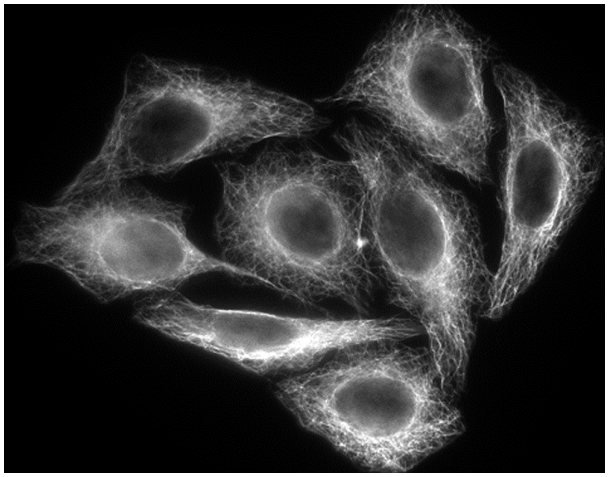


Fig. 8.1 - The cellular phenotype is complex. It can be in part analysed with the help of molecular markers enabling certain structures to be observed. In this image fluorescent probes allow visualisation of the microtubular cytoskeleton.

By using mutagenic agents or radiation, it is possible to cause random mutations in the whole genome of model organisms. With the techniques of molecular biology it is also possible to mutate specifically a given gene. **Genetic mutations** can induce in this way a modification of quite a specific nature at the level of the cell or the whole organism, such as, for example, retarded growth, or an altered appearance or behaviour. We refer to this as a **mutant phenotype**.

Chapter 11

MOLECULAR DESCRIPTORS AND SIMILARITY INDICES

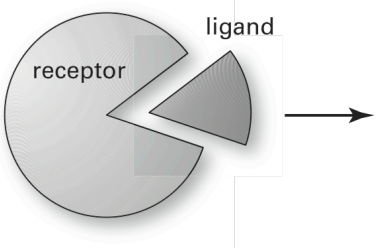
Samia ACI

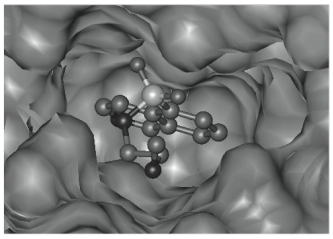
11.1. INTRODUCTION

High-throughput and high-content pharmacological screening methods generate such a flood of chemical and biological data that their analysis would barely be feasible without the use of computational tools (chapters 6 and 15). As is the case with forward chemical genetics (chapter 8), for which the target itself is unknown and the biological information incomplete, the small molecule represents the best defined piece of data. Chemists, biologists and informaticians have to pull together as best they can in order to create from their joint observations ingenious hypotheses about the reasons for a molecule's bioactivity, or lack thereof. An apparently simple question to ask is: what is a molecule? And what sensible representation to make of it? Example 11.1 illustrates the range of possible answers.

Example 11.1 - different ways to 'look' at a molecule

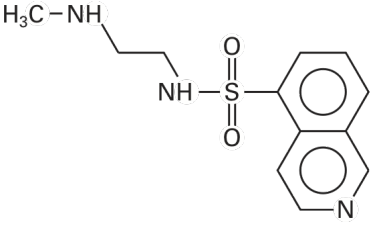
For the biologist





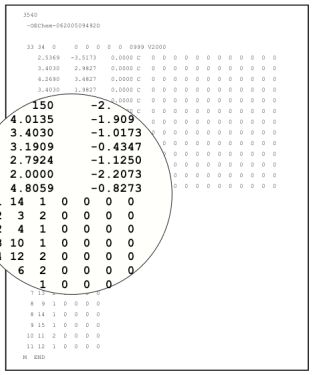
Global properties:
- IC50
- log *P*
- molecular mass
- etc.

For the chemist



Functional properties:
- pKa (acid/base character)
- H-bond donor/acceptor character
- nucleophilic/electrophilic character
- etc.

For the informatician



Graph properties:
- number of nodes
- number of edges
- number of node neighbours
- etc.

How do we go about bringing together these different views, each of which represents only one facet of the chemical entity, to enable the extraction of innovative concepts as to the reasons for bioactivity? More particularly, with regards to **virtual screening** (chapter 16), which **representation language** should be adopted to code for the chemical and structural properties of molecules in a format exploitable by the informatician and thus enabling him or her to extract chemically and biologically relevant information? This chapter presents a certain number of descriptors utilised in informatics for the characterisation of molecules. The aim of such a description is to introduce the concept of **similarity** between objects evaluating their degree of coverage in the space of the descriptors used to characterise them. How should this evaluation be carried out? Which measurement index should be adopted depending on the properties of the molecules to be evaluated for similarity or dissimilarity? Several of these indices will be detailed here. This chapter is not an exhaustive inventory of the whole set of molecular descriptors – which number in their hundreds or indeed thousands – nor of similarity measures (barely less numerous) but seeks to give to the reader an overview of the general categories that are available (see also chapter 12 for the particular point about hydrophobicity and chapter 13 for the recent developments in the annotation and classification of chemical space). The reader desiring more complete information is invited to consult the works dedicated to these subjects by TODESCHINI and CONSONNI (2000), and WILLET *et al.* (1998).

11.2. CHEMICAL FORMULAE AND COMPUTATIONAL REPRESENTATION

The calculation of, as well as the type of information given by, molecular descriptors depend on the chemical formula. They also depend on the computational representation of this chemical formula. Here, therefore, follows firstly a brief reminder of the concept of chemical formula. We shall see thereafter a way in which this information may be represented in calculations.