ELSEVIER

The principle of complementarity: chemical *versus* biological space

Stephen J Haggarty

Chemical genomics aims to systematically explore the interactions between small molecules and biological systems. These efforts aim to annotate genomes using the language of chemistry, and to provide information-rich profiles of chemical and biological systems. Here, I describe recent conceptual and experimental advances toward the goal of mapping multidimensional chemical and biological descriptor spaces. In doing so, I will focus on the complementary nature of these efforts, the importance of recognizing the distinction between computed versus observed descriptors, and highlight recent 'landmark' examples of small molecules discovered using phenotypic screens. Future computation and experimental advances will be needed to fully realize the goals of chemical genomics. For those willing to consider both local and global properties of chemical and biological space, and to venture into uncharted territory, there promises to be new vistas and principles to be discovered.

Addresses
Eli & Edythe L Broad Institute, 320 Charles Street, Cambridge, MA 02141, USA

Corresponding author: Haggarty, Stephen J (haggarty@broad.mit.edu)

Current Opinion in Chemical Biology 2005, **9**:296–303

This review comes from a themed issue on
Combinatorial chemistry
Edited by Hans-Jörg Roth and Prabhat Arya

Available online 25th April 2005

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2005.04.006

Introduction

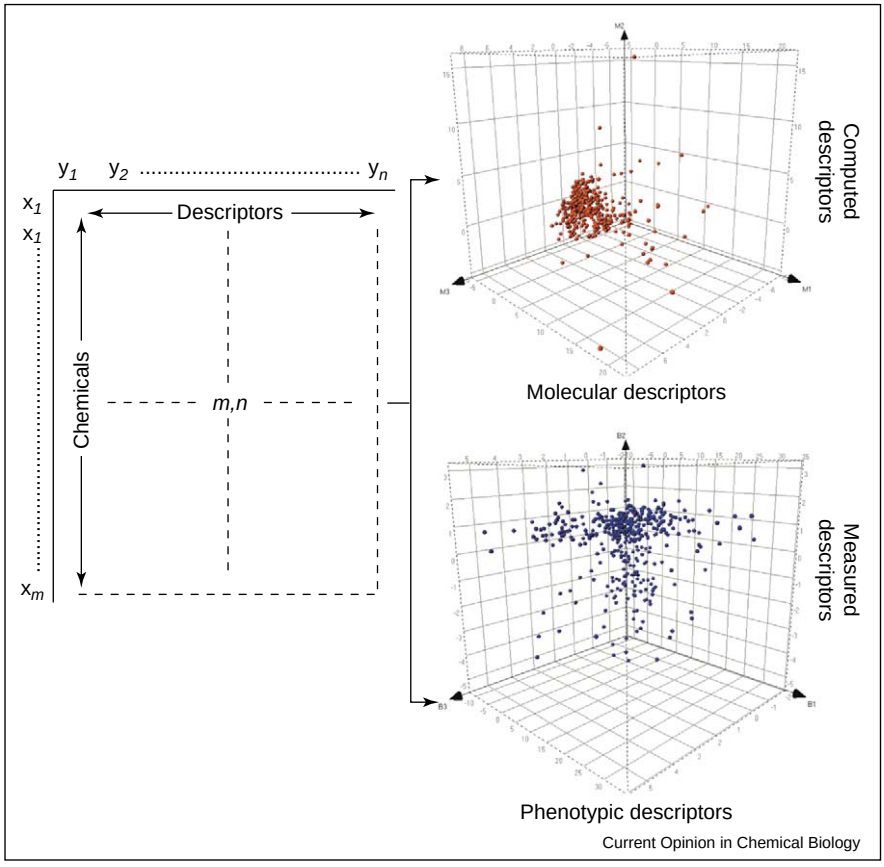
In addition to examining known bioactive molecules, and therapeutically useful drugs, chemical genomics requires the efficient synthesis and screening of novel collections of small molecules having rich skeletal and stereochemical diversity, to increase the likelihood of discovering new probes of biological and disease mechanisms [1,2,3•,4,5•]. With growing interest in the use of chemicals as probes in basic and clinical research, the field of chemical biology in general, and chemical genomics in particular, is facing the challenge of transitioning from the *ad hoc* discovery of small molecules to the systematic discovery and elucidation of novel targets and mechanisms of action. An important part of this transition is the development of an experimentally driven, computational framework that can inform both chemists and biologists. The success of this framework holds the potential to revolutionize the discovery of small-molecule probes for basic research and, potentially, the discovery of novel therapeutic targets and agents. As evidenced by the efforts toward the development of ChemBank (http://chembank.med.harvard.edu/), the Biomolecular Interaction Network Database (BIND; http://bind.ca/), Blueprint's Small-Molecule Interaction Database (SMID; http://smid.blueprint.org), PubChem (http://pubchem.ncbi.nlm.nih.gov/), the National Institutes of Health's Roadmap [6], and a flurry of recent reviews and meetings discussing the topic of 'chemical and biological space' [7••,8•,9••], the appeal of venturing into this uncharted territory is rapidly growing.

Chemical space

The key organizing factor for analyzing chemical space is derived from the multidimensional data structure generated when multiple descriptors are used to annotate a small molecule's structure and observed activities and properties. This data structure is most often that of an array, or matrix, denoted by **S**, consisting of an ordered array of n columns and m rows (Figure 1). Each column ($\mathbf{y}_j$) in **S**, corresponds to a descriptor, and is denoted by a bold-face, lower-case letter subscripted j (where $j = 1$ to n). Each row ($\mathbf{x}_i$) in **S** corresponds to a chemical, and is denoted by a bold-face, lower case letter subscripted i (where $i = 1$ to m). Accordingly, an element (e_n) of **S** encodes information (m, n) about chemical m for descriptor n. This structure allows the elements of **S** to be considered as coordinates in a multidimensional space spanned by the descriptor axes, which, in turn, allows each chemical to be represented as a vector, whose magnitude and direction is given by the corresponding values in **S**, $\mathbf{x}_i = [e_1, e_2, \ldots, e_n]$. In the corresponding chemical space, the relative distance between chemicals $\mathbf{x}_i$ becomes a measure of their similarity with respect to the particular descriptors considered, and thus a tool for navigation and further analysis.

As depicted in Figure 1, when considering chemical space there are two fundamentally different classes of descriptors that are used: computed and measured (this is also true for biological space; see below). These classes differ insofar as the former are generally calculated using a computer and various algorithms designed to determine the value of a specified mathematical function [9••,10••,11•,12,13••,14••], whereas the latter involve the observation of the effect of a chemical on, for example, the function of a gene product (nucleic acids, proteins) or metabolite (carbohydrate, lipid, other organic

Figure 1



Computed *versus* observed descriptors used to create maps of chemical space. Principle component models of chemical space for 480 small molecules analyzed using 24 computed molecular descriptors and 60 measured phenotypic descriptors derived from a cell-based assay of cell proliferation data from [49].

molecules). Recognizing the distinction between chemical spaces derived from computed descriptors as compared to measured descriptors is of fundamental importance. Whereas the former can be explicitly defined using specific algorithms, the latter involves the process of observation, and as such involves error inherent to the process of measurement. Furthermore, as discussed below, phenotypic descriptors are also subject to the influence of a variety of other variables, including the dose of the chemical, length of treatment, and the genotype of the biological system.

Much has been written lately about the use of molecular descriptors to assess the diversity of small molecules and improve the coverage of unpopulated regions of chemical space, which will not be reiterated here [10••,11•,13••]. One challenge in the use of molecular descriptors to create maps of chemical space that can both locally and globally predict biological activity, is that a given chemical can exist as a variety of structures corresponding to various protonation, tautomeric and stereochemical states depending on the molecule's environment. Another major challenge is the ability of enzymes to metabolize small molecules into what might be either an active or inactive component. Together, these, and other, factors contribute to the difficulty of predicting the function of a small molecule, particularly in the context of an intact living system as complex as the human body. Despite these limitations, since chemical space can be explicitly defined using specific algorithms to compute molecular descriptors, it seems possible that a universally agreed upon set of molecular descriptors could be used to create maps, much like those of early geneticists or the existing assemblies of the human genome, that investigators can annotate systematically with various observational data [11•,12,13••,14••] (Figure 2b).

In contrast to computed molecular descriptors, observed, or phenotypic, descriptors involve the measurement of the effects of a small molecule on a biological system [8•,14••,15••,16,17]. Although the term phenotype is mostly widely used in genetics to refer to any part of the observable structure, function or behavior of a living organism, it also includes the observable physical parts: