COMPUTER-AIDED COMBINATORIAL CHEMISTRY AND CHEMINFORMATICS

ALEXANDER TROPSHA

School of Pharmacy, University of North Carolina, Chapel Hill, NC
(alex_tropsha@unc.edu)

ROBERT S. PEARLMAN
College of Pharmacy, University of Texas, Austin, TX
(pearlman@vax.phr.utexas.edu)

We live in the time of "–ics" sciences. Genomics, proteomics, genetics, bioinformatics, cheminformatics are a few examples of recent terms that refer to relatively new and rapidly growing areas of both academic and industrial research. This growth implies an unprecedented accumulation of biomolecular information stored in ever growing databases. These include both gene databases and various databases of organic molecules, which contain millions of individual molecular entries. All these molecular information has to be stored, manipulated, understood and used, in a rational way, in designing new drugs. Computational analysis of molecular diversity and similarity, database mining, combinatorial library design has become one of the most vital areas of biocomputing.

It is well known that many genes and their products have an interesting yet unidentified function, and the analysis of sequence-structure-function relationships has been a traditional area of bioinformatics. Perhaps it is less obvious but experimental medicinal chemists face, in many respects, the same type of problems as experimental biologists. Practically every chemical has certain, frequently unknown, biological effect (cf. gene sequences with unknown function). There is a huge array of chemical molecules (cf. millions of genes), many of which could potentially be drugs, but their specificity against a particular biological target is yet to be determined (as well, in many cases, as the target itself). The computational aspects of macromolecular vs. chemical database analysis appear strikingly common yet complementary. Indeed, the exact challenge is that of matchmaking: how to find the right drug for the right target? Both computational chemists and computational biologists address this challenge by developing fast and accurate methods of biomolecular database analysis to enhance our ability to discover or design lead molecules of pharmaceutical significance. These approaches rely on our understanding of three-dimensional structure of both organic and biological macromolecules and the development of rigorous quantitative models that explain experimental structure-activity (for organic molecules) and sequence-structure-

function (for macromolecules) relationships. This similarity of computational aspects of macromolecular vs. chemical computing makes it especially interesting to have this session on the analysis of chemical diversity and similarity.

In a complete agreement with the laws of dialectics, the generation of new quantities of data required qualitatively novel methodologies for data analysis. New approaches are being developed to establish Quantitative Structure-Activity Relationships (QSAR) not for dozens but for hundreds, or even thousands, of molecules. Database mining and pharmacophore searching techniques should now afford the analysis of databases or virtual libraries containing millions compounds. These new methods should be robust yet extremely computationally efficient to compete with newly developing experimental techniques of data generation and analysis such as combinatorial chemistry and high-throughput screening.

The following papers address novel methods and developing ideas in the areas of chemical database analysis and bioactive structure prediction. They discuss novel metrics for the description and comparison of organic molecules (Godden et al.), current approaches to diversity and similarity sampling of molecular databases (Dunbar), and combinatorial library design applications (Mason and Cheney, and Zheng et al.). The Godden et al. paper describes a novel mini-fingerprint (MFP) representation for small organic molecules and application of these descriptors for effective molecular similarity searches using newly developed fingerprint profiling method. J. Dunbar presents an overview of methodologies used in compound acquisition, which is one of the most common tasks performed by the majority of pharmaceutical and agrochemical companies. These methodologies involve using adequate molecular descriptors and similarity and diversity searching algorithms that afford effective sampling of external molecular databases for compounds either similar or different from "in-house" compounds in terms of their biological function. The Mason and Cheney paper presents the continuing work of this group in the area of four-point pharmacophore development of biologically active molecules (now using the shape of the target site as an additional constraint). The authors further discuss an exciting application of the paharmacophores for virtual screening of combinatorial libraries. Finally, the Zheng et al. paper presents a novel computational tool for combinatorial library design, which optimizes reagent selection on the basis of simultaneous optimization of several important criteria of virtual libraries such as synthetic feasibility, developability, druglikeness, cost, etc. Thus, papers presented in this session reflect many important developments and applications in the field of chemical informatics and should be of great interest to all scientists working in the area of biocomputing.