Receptor-ligand molecular associations

Responsable du projet :

François Major (CERCA)
Département d'Informatique et de R.O.
Université de Montréal
Tél. : (514)343-7091
Fax : (514)343-5834
Courriel : major@iro.umontreal.ca

Yoshua Bengio (CRM)
Département d'Informatique et de R.O.
Université de Montréal
Tél. : (514)343-6804
Fax : (514)343-5834
Courriel : bengioy@iro.umontreal.ca

Chercheurs principaux :

Dr. Laurent David	Informatique et R.O., Université de Montréal
Dr. Michael Gilson	Center for Advanced Research in Biotechnology
Dr. Enrico O. Purisima	Institut de recherche en biotechnologie
Dr. Suzanne Sirois	CERCA

Description du projet

Objectives

The primary goal of this research is to improve our understanding and, thus, the reliability and speed of molecular association predictive methods so that they can be used in practice in high-throughput screening studies. An indirect goal is the formation and training of qualified personnel, technicians, students and post-doctoral trainees in the field of computational chemistry and cheminformatics. The long-term objectives of this research are to: 1) Identify and understand the most important factors in receptor-ligand associations, and 2) formalize the physicochemical principles of receptor-ligand associations in computer data structures and algorithms. More specifically, this research will allow us to develop: 3) Models of evaluation functions of association constants (approximation of standard bonding free-energy) for receptor-ligand associations (scoring functions), 4) Simplified and precise representations of receptor-ligand association information, 5) Formal tests to evaluate the models and representations implemented in 3) and 4), and 6) Develop novel approaches for predicting receptor-ligand association constants in a pharmaceutical context, i.e. to filter thousands of different ligands at very high speed (high-throughput screening).

Introduction

Molecular docking and scoring functions

Molecules, and in particular a receptor with its ligand, interact in a highly specific manner. A commonly used analogy is to see the receptor as a *lock*, and the ligand as a *key*. Receptor-ligand [*RL*] interactions involve specific physical and chemical contacts during a period of time

resulting in attractive forces. Different approaches for estimating the free energy of binding have been developed, based on the importance of specific contributions to binding. Most approaches use a model function composed of the sum of contribution terms, usually referred to as the *master equation*. This equation accounts for the contributions due to the *solvent*, conformational changes (*conf*) in the receptor and the ligands, receptor-ligand interactions (*int*), and *motion*. Molecular *docking* consists in sampling, usually on a discrete grid of points, the conformational space of the molecules involved in the association, here for instance the receptor and its ligand, in the search of the most stable association conformations. A *scoring function* is used to evaluate and compare association conformations. In general, docking can be thought of as a task that brings complementary chemical groups in contact. Thus, the scoring function of docking programs comes from pre-calculated force fields, not estimations of the free energy of binding (cf. Flexx[1], GOLD[2], AutoDock[3], DOCK[4], etc). Docking methods can predict binding affinities in the nanomolar precision, but they usually perform poorly in lower resolution. In an attempt to identify low affinity ligands to specific receptors, Charifson et al. [7] have evaluated two docking methods and thirteen scoring functions. They concluded that combinations of various scoring functions result in an enhancement of the ability to discriminate between active and inactive enzyme inhibitors. They divided the thirteen scoring functions into three categories: 1) empirical functions, 2) molecular-mechanics- (MM) based (MM) functions, and 3) others. The empirical subgroup includes *Böhm*, *ChemScore*, *SCORE* and *Piecewise Linear Potential* (PLP). The MM-based subgroup comprises the *Merck Molecular Force Field* (non-bond energy), *DOCK* (energy score), *DOCK* (chemical score) and *Flexible* (ligands on a grid (FLOG)). Finally the others include *Poisson-Boltzmanm*, *Buried Lipophilic Surface Area*, *DOCK* (contact score) and *Volume Overlap*.

The research challenge

It is important to distinguish between the *precision, reliability, and granularity* of molecular association computer programs and scoring functions. The *precision* corresponds to the range of values returned by the program. For instance, force fields approaches return energy values in kcal/mol, whereas other approaches could return a score on a scale of one to ten. The *reliability* of a computer method is its molecular association soundness, that is, the computer program returns values that correlate well with experimental results; a reliable computer program makes good predictions. Finally, the *granularity* is a degree of details in which a program is defined. On a grid search small distances between points mean a low granularity, whereas long distances mean high granularity. In a scoring function, lots of physicochemical details in the terms mean a low granularity (or high resolution), whereas few rough terms mean high granularity (low resolution). For instance, in a low granularity model that introduces perturbation in the free energy [5], lots of details must be considered. The main advantage of such approaches is precision, whereas their principal drawback is a high computational cost. On the other side of the granularity scale, some empirical approaches combine several factors and complex mathematical equations into few symbols. The main advantage of high granularity approaches is a low computational cost, but they are rarely reliable. In fact, there exist no empirical approaches of satisfying reliability at this moment, and one of the goals of this research is to develop a new approach of low computational cost (high-throughput screening), high reliability (good prediction), and good precision (selection).

Experimental training data set

The approach described above highlights the importance of using refined structural and kinetic data from pharmaceutical important molecular associations. Now, consider a different

approach, the one of Wang et al. who developed a scoring function by regression analysis upon a training set of 170 molecular complexes [6]. Taken as a typical empirical approach, it exemplifies again the needs for a good experimental training data set of pharmaceutical importance. In fact, building the training data set is an important and critical step that must be taken a priori of the development of any method using any approach. The data set will be composed of a spectrum of experimental receptors with altered ligands associations with associated kinetics. This database will be composed of publicly available examples, as well as new ones that will be provided by the CERCA pharmaceutical partners. The examples will include molecular association systems such as protein-drugs, and MHC-TCR. The training data set will allow us to evaluate and study currently available scoring functions, and therefore to describe new ones.

Research Plan

1. Identify and understand the most important factors in receptor-ligand associations

This will be accomplished performing a comparative study of various scoring functions and their precision to identify the appropriate molecular association constants. A database, which will include various scoring functions from the literature and three-dimensional structures of molecular associations, will be built. A computer system that applies scoring functions to each association will be developed. Both databases and the map operator will be made available to all partners. Particular efforts will be devoted to accessibility and ease of use, in order to encourage the development of new scoring functions (steps 3 to 5). Scoring functions in object (compiled) code will be accepted from industrial and university partners who do not want their function to be exposed to the others (see collaborative agreement). For the final report, each function, in source or object code, will be assigned an identifier. The database and the map operator will be developed, in preference, using the *Linux* operating system to simplify distribution and future development. No file or database formats are determined at this time for this project. However, standard *mmCIF* formatted files, as used by the *Protein Data Bank* (PDB), could be used for molecular associations, for instance. Molecular association structures will be pre-processed to contain hydrogen and all other necessary atoms. An object (as in object-oriented programming language) will be defined in a high-level computer programming language such as C++ or Java, the language in which the map operator will be developed. At this moment, let us call the molecular association object a *molecularAssociation*, which will be parsed in the high-level programming language, such as C++, in order to get a memory internal representation. Scoring functions will be defined in the high-level programming language. The results of mapping a scoring function to a molecular association will be made available to molecular docking (see above) programs. The scores will be saved in structured files allowing us to provide reports to the research partners, as well as a pipe mechanism for input to molecular docking and molecular dynamics programs. The results will be saved in a fast cache system (currently developed in the laboratory of F. Major). A docking program, for instance, would evaluate the scores returned by the mapping for refining a given association.

2. Formalize the physicochemical principles of receptor-ligand associations in computer data structures and algorithms

The data structures and algorithms used in existing scoring functions will be studied. All ligand properties from all scoring functions will be collected, and implemented in a generalized data structure model suitable to the development of novel scoring schemes.

3. Models of evaluation functions of association constants (approximation of standard bonding free-energy) for receptor-ligand associations (scoring functions)

New models of scoring functions will be developed, based on the study in 1 using a combined AI and energy-based approach. The geometric method is first applied to rapidly identify potential candidates of good affinity. Then, high-precision energy-based scoring functions are applied to confirm the predictions. Pareto optimality is the theory of optimizing a function composed of parameters that contribute in contradictory manners to an optimization criterion. For instance, in evaluating the largest common three-dimensional substructure of a set of molecules, the superposition of two different structures involves contradictory factors: the RMSD vs. the number of atoms in the substructures. This is an important step in the identification of a pharmacophoric pattern for molecules that bind to the same receptor. For instance, the use of a method as described in [8] can help determine the affinity of binding of molecules to a common receptor. The ligand of a molecular association is used for comparison, and is superimposed over a set of molecules. Then, the molecules that superimpose well on the ligand, and the ligand, define a set of potential pharmacophores. The superposition method described in [8] combines a genetic algorithm with a numerical optimization method. The goal is to adequately address the conformational flexibility of ligand molecules. The genetic algorithm is used to optimize the number of atoms, whereas the geometric fit of the substructures is further improved by a numerical optimization that samples torsion angles. In this geometrical approach, no energy evaluation is performed. However, once as few pharmacophores have been proposed, high-precision energy-based scoring functions can be applied to support the predictions. We propose to investigate this combination of approaches, which we previously applied to the development of MC-SYM where the conformational space of a RNA is explored empirically without the use of energy evaluation. In a second step, the proposed conformations are refined using energy minimisation. The approximations made in the first step allow for rapid exploration of alternative solutions, and the high-precision process is only applied to potential candidates.

4. Simplified and precise representations of receptor-ligand association information

From the studies in 3, simplified representations of receptor-ligand association will be developed.

5. Tests to evaluate the models and representations

Specific tests built from the three-dimensional structures that will be inserted in the database will be developed. The tests will be systematically applied and analyzed for each new representation and scoring function that will be developed.

6. High-throughput screening

The database and scoring function “map” operator constitutes a high-throughput system. When detailed structural information about the target receptor is unknown, it is possible to derive an abstract model called a *pharmacophore*. This term refers to a set of active molecules that can be found in the active region of the receptor. Positively and negatively charged groups, hydrogen-donors, hydrogen-acceptors, and hydrophobic groups are the common types of chemical groups that are involved in the active site. A three-dimensional pharmacophore specifies the spatial representation of those chemical groups. The characteristics of pharmacophores are expressed in distances, distance ranges, angles, and angles between the planes defined by the chemical groups. Once a “good” pharmacophore is determined, the procedure of pharmacophore mapping is used to screen a very large database of possible ligands. This method, called *very large scale virtual screening*, is often applied to

eliminate most of the ligands that would not fit in the active site of the receptor. Our research will also be dedicated to the development of computer algorithms to determine "good" pharmacophores that fit into the active site of the receptor. These algorithms will include methods for generating pharmacophores, screening the database, and analyzing the dynamic conformations of ligands prior to the molecular docking (see references [9] to [13]).
The code will be optimized according to the needs of the partners. A parallel implementation for the Beowolf cluster in my laboratory will be developed. The Beowolf is composed of dual pentium processor units, and a front-end that manages the jobs. This is a cheap way of building a cluster of processors, where every dual processor unit costs approximately $2,000.

Reference

[1] B. Kramer, M. Rarey, T. Lengauer, *Proteins*, **Suppl**, 221-5 (1997)
[2] G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol*., **267**, 727-48 (1997)
[3] D, S. Goodsell, A. J. Olson, *Proteins*, **8**, 195-202 (1990)
[4] E. C. Meng, B. K. Schoichet, I. D. Kuntz, *J. Comput. Chem*., **13**, 505-524 (1992).
[5] P. Kollman, *Chem. Rev*., **93**, 2395-2417 (1993)
[6] R. Wang, L. Liu, L. Lai and Y. Tang, *J. Mol. Model*., **4**, 379-394 (1998).
[7] P. S. Charifson, J. J. Corkery, M. A. Murcko, W. P. Walters, *J. Med. Chem*., **42**, 5100-5109 (1999).
[8] S. Handschuh, M. Wagener, J. Gasteiger, *J Chem Inf Comput Sci*, **38**, 220-232 (1998).
[9] Martin, YC, Bures, MG, Willett, P. *In Rev. in Computational Chemistry*; Lipkowitz, K., Boyd, D, Eds; VCH Publishers, Inc., New York, 1990; pp 213-263.
[10] Y.C. Martin. *J. Med Chem.* **35(12)**:2145-2154(1992).
[11] J.S. Mason. *In Molecular Similarity in Drug Design*; Dean, PM, Ed.; Blackie Academic & Professional, New York, 1995; pp 138-162.
[12] R.P. Sheridan, A. Rusinko III, N. Ramaswany and R. Venkataraghavan. *Proc. Natl. Acad. Sci. USA* **86**:8165-8169(1989).
[13] G.R. Strobl, S. von Kruedener, J. Stockigt, F.P. Guengerich and T. Wolff. *J. Med Chem.* **36**:1136-1145(1993).

Interaction avec les partenaires :

The principal objective of the Canadian pharmaceutical industry is the discovery of novel drugs for improving the overall Canadian quality of life. The costs of new drug discoveries are astronomical, and at the same time the international competition is very aggressive. It follows that the pharmaceutical industries are in constant need of novel and more effective approaches in order to be more cost effective and also to bring to the market new drugs that are much more potent than the current ones. Fortunately, significant progress in how to investigate and model complex biomolecular systems has been made during the last decade. This progress, combined into an increasing understanding of the relationships between chemical structures and biological activities, allows us to study, using computers, biological systems of importance for the pharmaceutical industry, and in particular for developing new drugs. For instance, theoretical predictions of molecular association constants represent an alternative to classical experiments used in the pharmaceutical industry. Computer predictive tools can reduce considerably the costs of drug discoveries. In order to develop a specific expertise to the Canadian industry, it is important to formalize state-of-the-art knowledge about receptor-ligand associations, which will be transferred into novel approaches. The goal of these methods is to rapidly and theoretically select ligands, from a database, that have a great potential of becoming drug candidates, the *high-throughput screening* of drugs and their rational design.