A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP

Yogendra Patel1,†, Valerie J. Gillet1,*, Gianpaolo Bravi2 & Andrew R. Leach2
1Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom
2GlaxoSmithKline, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom

Received 29 July 2002; accepted 18 October 2002

Key words: pharmacophore identification, protein ligand interactions, pharmacophoric features, structural alignment, conformational flexibility

Summary

Three commercially available pharmacophore generation programs, Catalyst/HipHop, DISCO and GASP, were compared on their ability to generate known pharmacophores deduced from protein-ligand complexes extracted from the Protein Data Bank. Five different protein families were included Thrombin, Cyclin Dependent Kinase 2, Dihydrofolate Reductase, HIV Reverse Transcriptase and Thermolysin. Target pharmacophores were defined through visual analysis of the data sets. The pharmacophore models produced were evaluated qualitatively through visual inspection and according to their ability to generate the target pharmacophores. Our results show that GASP and Catalyst outperformed DISCO at reproducing the five target pharmacophores.

Introduction

A pharmacophore is the spatial arrangement of key chemical features that are recognised by a receptor and are thus responsible for ligand-receptor binding [1]. Pharmacophore models are typically used when some active compounds have been identified but the three-dimensional (3D) structure of the target protein or receptor is unknown. The active compounds are superimposed to determine their common features and hence to provide a pharmacophore model that explains ligand-receptor binding. Once such a model has been derived it can be used to: search for other molecules that contain the same pharmacophore and that may also be active; to explain structure activity relationships within a series of molecules; and to form a basis for the design of new potentially active molecules.

Given a set of active molecules, the identification of a pharmacophore involves two steps: analysing the molecules to identify pharmacophoric features, that is, atoms that can interact with a receptor, and aligning the active conformations of the molecules to find the best overlay of the corresponding features. The main difficulty in pharmacophore generation is in the handling conformational flexibility since the active conformations of the molecules are usually unknown.

Several programs have been developed for the automatic identification of pharmacophore models [2]. The main differences between the programs lie in the algorithms used for the alignment and in the way in which conformational flexibility is handled. Here, three commercially available programs DISCO, Catalyst/HipHop and GASP are compared with regard to their ability to reproduce known pharmacophores which have been determined by analysing crystallographic data for a series of ligands bound to the same protein.

The main features of the programs are given in the next section which is then followed by a description of the methodology and the data sets used in the study. Finally the results are presented and conclusions drawn.
The programs

In DISCO [3, 4], each molecule is characterised by ligand points and site points. Ligand points include atoms with positive charge, negative charge, hydrogen bond donor, hydrogen bond acceptor and hydrophobic character. Site points represent the hypothetical position of complementary atoms in a receptor and are determined from the position of heavy atoms in the ligand structure. Conformational flexibility is handled by precomputing a series of low energy conformers for each molecule with each conformer being treated as a rigid body during the alignment step. A conformer is represented by the interpoint distances calculated for the ligand and site points and a clique detection algorithm is used to align structures based on these distances. The Bron-Kerbosh clique-detection algorithm has been modified to allow multiple alternative conformations of molecules to be considered and to preserve chirality of molecules.

The molecule with the fewest conformations is used as a reference molecule. DISCO takes each conformation of the reference molecule in turn and compares it to all conformations of the other molecules. The cliques identified are examined in an attempt to find one that is common to at least one conformation of every molecule. This process is repeated for every conformation of the reference molecule. If no solution is found the tolerances on the clique detection process are increased until either a solution is found or the maximum tolerance is reached. The output from a DISCO run is a ranked list of all possible pharmacophore mappings where each feature of a pharmacophore must be present in all molecules. This requirement can result in good pharmacophores being missed, hence, DISCO has the option of finding solutions where some molecules are excluded from the model.

Catalyst comprises two modules: HipHop and HypoGen [5]. This study is based on HipHop [6] which attempts to derive a pharmacophore based on features that are common to active molecules. HypoGen, on the other hand, makes use of quantitative activity data to derive the pharmacophore. The pharmacophoric features identified in HipHop are hydrogen bond donors and acceptors, negative and positive charge centres, and surface accessible hydrophobic regions [7]. As in DISCO, both ligand atoms and projected positions of complementary site atoms are considered as hydrogen bonding features. The handling of conformational flexibility is also similar to that used in DISCO, with each molecule being represented by a set of low energy conformations that are subsequently treated as being rigid. Conformations can be generated using the Poling technique [8] that ensures broad coverage of conformational space, or by using any external structure generation program.

Each molecule is treated as reference molecule in turn. Different configurations of feature points are identified in the reference molecule using a pruned exhaustive search which starts with small sets of features and extends them until no larger configuration is found. Each configuration is then compared with the remaining molecules in an attempt to identify configurations that are common to all molecules. A molecule matches a configuration if it possesses a set of features that can be superimposed on the configuration. The requirement that all molecules match all features in the configuration can be relaxed so that not all molecules are required to possess all the features identified in the pharmacophore. Thus, the assumption is that a molecule can be active despite lacking a feature relevant in the binding of other molecules. The resulting hypotheses are ranked using a combination of how well the molecules map onto the proposed model and the rarity of the model.

GASP [9,10] is based on a genetic algorithm and differs from both DISCO and Catalyst in its handling of conformational flexibility which is performed on-the-fly. Thus each molecule is input as a single conformation and random rotations and a random translation are applied before any superimposition is made. The pharmacophoric features in the molecules are determined (hydrogen bond donor protons, acceptor lone-pairs, and ring centres including projected site points). The molecule with the least number of pharmacophoric features is chosen as the base molecule to which the other molecules are fitted (cf the reference molecule in DISCO).

A chromosome in GASP encodes the angles of rotation of the rotatable bonds in all of the molecules and the mapping of the pharmacophoric features in the base molecule to corresponding features in each of the other molecules. For a data set of \( N \) compounds the chromosome consists of \( 2N - 1 \) strings: \( N \) binary strings to represent the conformations of the molecules and \( N - 1 \) integer strings that represent the mapping from the base molecule to each of the other molecules. Thus, the length of the integer strings is determined by the number of pharmacophoric features in the base molecule. The fitness function first generates conformations for each molecule and then uses a