Ligand-Info, searching for similar small compounds using index profiles

Marcin von Grotthuss, Jakub Pas and Leszek Rychlewski*

BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland

Received on July 19, 2002; revised on October 11, 2002; December 12, 2002; accepted on December 20, 2002

ABSTRACT

Motivation: The Ligand-Info system is based on the assumption that small molecules with similar structure have similar functional (binding) properties. The developed system enables a fast and sensitive index based search for similar compounds in large databases. Index profiles, constructed by averaging indexes of related molecules are used to increase the specificity of the search. The utilization of index profiles helps to focus on frequent, common features of a family of compounds.

Results: A Java-based tool for clustering and scanning of small molecules has been created. The tool can interactively cluster sets of molecules and create index profiles on the user side and automatically download similar molecules from databases of 250,000 compounds. The results of the application of index profiles demonstrate that the profile based search strategy can increase the quality of the selection process.

Availability: The system is available at http://Ligand.Info. The application requires the Java Runtime Environment 1.4, which can be automatically installed during the first use on desktop systems, which support it. A standalone version of the program is available from the authors upon request.

Contact: leszek@bioinfo.pl

Small molecule search engines, which can scan (Csizmadia, 2000; Ihlenfeldt et al., 2002) and cluster (Wild and Blankley, 1999) compounds are available on the Internet. The main objective for the development of the new server was to provide the possibility to create novel index profiles for compound families. The application of profiles of protein sequences for families of homologs as used in PSI-BLAST has dramatically increased the sensitivity of sequence similarity search methods in the past (Altschul and Koonin, 1998). The new small compound scanning server follows the same rationale and increases the accuracy of the search process.

The clustering and search algorithms are based on two-dimensional structure similarity, where molecular structures are represented by modified hashed fingerprints (Daylight Chemical Information Systems, 2000). The Modified Tanimoto Coefficient (MTC) is used as similarity metric between molecules. The client program of the server uses a complete linkage algorithm for clustering of molecules. The user can specify an MTC cutoff value (between 0 and 1) for the minimum similarity between every pair of molecules in the cluster. The clustering algorithm generates groups (clusters), which can be manually adjusted by the user, by selecting or deselecting molecules. All loaded molecules are automatically displaying using the Marvin Applet ChemAxon Ltd. http://www.chemaxon.com/marvin in a separate window.

Small compound scanning is conducted on the server side and can be based on either one of two search strategies: multiple search or profile search. The multiple search algorithm iteratively downloads molecules found for each of the molecules of the cluster. If a database molecule exceeds the query similarity threshold to multiple comparison molecules in a cluster, its final ranking is determined by its maximum similarity value to any cluster member. In contrast one single virtual molecule is used in the profile search strategy. A virtual molecule has all the indices that occurred in molecules of the cluster with frequencies averaged over the whole cluster. The profile search algorithm is obviously much faster than multiple searching. In both methods the downloaded molecules are displayed below the molecules of the selected group (cluster) and they are sorted by the similarity value in descending order.

Both similarity search strategies can be applied on three data sets: the Anti-HIV Database, the Drug-likeness Database and the total NCI Database. The Anti-HIV Database contains 42,689 molecules, where 423 were experimentally determined to protect the CEM cells from HIV-1 infection, 1081 are moderately active (protective) and 41,185 are inactive (National Cancer Institute http://dtp.nci.nih.gov/docs/aids/aids_data.html). The Drug-likeness Database contains 192,323 molecules, where 69,045 are described as drug-like (TORVS Research Team http://zabib.chemie.uni-erlangen.de/services/...
A search accuracy test was conducted on 42,689 anti-HIV-1 inhibitors. 1081 moderately active inhibitors were used as starting set. The other inhibitors (423 active and 41,185 inactive) remained in the database on the server. The goal of this scan was to find the maximum number of active inhibitors in the database using the moderately active inhibitor molecules as queries. The results (Fig. 1) show the superiority of the profile search strategy compared with standard bit-wise fingerprints.

The Ligand.Info system can be used to download and visualize compounds, which are similar to the molecules provided by the user. The system offers the possibility to apply the index profile based search strategy. This strategy has shown superior performance in our Anti-HIV drug search benchmark. The Ligand.Info database will be continuously updated with new compounds if they become publicly available and can be populated with compounds provided by the user.

REFERENCES

Fig. 1. A comparison of the two structure-based searching methods using different kinds of indices. 1081 moderately active inhibitors were clustered using the complete linkage algorithm with an MTC cutoff value from 0.7 to 0.9. Only molecules which belonged to a cluster of a size of at least three molecules were used. The multiple search strategy was performed for classic bits fingerprints with bond paths from 2 to 8 (Daylight Chemical Information Systems, http://daylight.com/dayhtml/doc/theory.figer.html) and the profile strategy was applied for Ligand-Info indices. The results show that the profile index based search strategies find the most active inhibitors and perform best with a cluster similarity cutoff between 0.75 and 0.85 MTC.

The NCI Database (Voigt et al., 2001) contains 250,249 molecules (235,012 molecules contained in previous subsets and additional 15,237, not functionally annotated molecules).