*Bioinformatics*, 31(24), 2015, 4020–4022
doi: 10.1093/bioinformatics/btv489
Advance Access Publication Date: 26 August 2015
Applications Note

OXFORD

Structural bioinformatics

# LIBRA: LIgand Binding site Recognition Application

**Le Viet Hung[1,2], Silvia Caprari[1], Massimiliano Bizai[1], Daniele Toti[1] and Fabio Polticelli[1,3,*]**

[1]Department of Sciences, University of Roma Tre, 00146 Rome, Italy, [2]Department of Science and Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam and [3]National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy

[*]To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on March 5, 2015; revised on July 7, 2015; accepted on August 2, 2015

## Abstract

**Motivation:** In recent years, structural genomics and *ab initio* molecular modeling activities are leading to the availability of a large number of structural models of proteins whose biochemical function is not known. The aim of this study was the development of a novel software tool that, given a protein's structural model, predicts the presence and identity of active sites and/or ligand binding sites.

**Results:** The algorithm implemented by ligand binding site recognition application (LIBRA) is based on a graph theory approach to find the largest subset of similar residues between an input protein and a collection of known functional sites. The algorithm makes use of two predefined databases for active sites and ligand binding sites, respectively, derived from the Catalytic Site Atlas and the Protein Data Bank. Tests indicate that LIBRA is able to identify the correct binding/active site in 90% of the cases analyzed, 90% of which feature the identified site as ranking first. As far as ligand binding site recognition is concerned, LIBRA outperforms other structure-based ligand binding sites detection tools with which it has been compared.

**Availability and implementation:** The application, developed in Java SE 7 with a Swing GUI embedding a JMol applet, can be run on any OS equipped with a suitable Java Virtual Machine (JVM), and is available at the following URL: http://www.computationalbiology.it/software/LIBRAv1.zip.

**Contact:** polticel@uniroma3.it

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

An important area of investigation in structural biology related to protein function recognition is the identification of ligand binding sites on the surface of proteins. This has a number of different applications ranging from function identification to drug discovery (Petrey *et al.*, 2015). Current tools, implemented in software applications like eF-Seek (Murakami *et al.*, 2013), G-LoSA (Lee and Im, 2013) and APoc (Gao and Skolnick, 2013), predict putative binding sites of a given input protein by comparing the properties of its surface patches with those of known binding sites. In addition, web applications such as ProFunc provide similar predictions based on a combination of sequence and structural information (Laskowski *et al.*, 2005a, b). In this study, LIBRA, a novel application employing a graph theory approach for function recognition and ligand binding sites detection, is described. LIBRA is based on the comparison of an input protein structure with a database of ligand binding sites extracted from PDB (Berman *et al.*, 2000). Besides, the application has been also designed to overcome some

of the weak points evidenced in the function prediction software ASSIST (Caprari *et al.*, 2014) previously developed in our lab. In fact, LIBRA can be used for the prediction of the catalytic activity as well in conjunction with a database derived from the CSA (Furnham *et al.*, 2014). Extensive tests against the LigaSite (Dessailly *et al.*, 2008) non-redundant database (381 apo-proteins) indicate that LIBRA is able to identify the correct binding/active site in 90% of the cases analyzed. Furthermore, comparative tests on a set of 30 apo-proteins extracted from LigASite demonstrate that LIBRA is able to recognize the correct function/ligand in all of the cases analyzed, outperforming other applications such as eF-Seek and SiteSeer, the functional site identification service implemented in ProFunc (Laskowski *et al.*, 2005a,b).

## 2 Methodology

### 2.1 Generation of the databases

Given an input protein P, LIBRA detects the largest subsets of similar residues between P and not only known functional sites, but also their local environment. As far as active site prediction is concerned, this approach required the construction of a database of known protein active site templates, starting from the CSA entries (Furnham *et al.*, 2014). For the prediction of ligand binding sites, a database has also been generated by extracting residues surrounding the ligand from 75 460 structures of protein–ligand complexes deposited in PDB (as of June 2015). The final database contains 173 240 ligand binding sites in PDB format. A detailed description of the procedure used to build the two databases is given in the Supplementary Materials.

### 2.2 Implementation of LIBRA

The similarity between a subset of residues belonging to an input protein and a binding site with its neighboring residues is defined in LIBRA by a graph (Biggs *et al.*, 1986) formed by nodes and edges whose definitions are given as follows: a residue $R_1$, belonging to the input protein, and a residue $R_2$, belonging to a known protein, form a node if they are identical or equivalent, according to a residues equivalence table (see Supplementary Table S1). If Node $A$ is formed by the residues pair ($A_1$, $A_2$), Node $B$ by the residues pair ($B_1$, $B_2$), $d_1$ is distance$(A_1, B_1)$, $d_2$ is distance$(A_2, B_2)$ and $\delta =$ error distance threshold, there is an edge which connects $A$ and $B$ if $|d_1 - d_2| < \delta$. Once the relationships between the input protein and the known binding site (and neighboring residues) are defined as a graph, a subset of equivalent residues is then defined as a clique (a subset of nodes such that every two nodes in the subset are connected by an edge), and the largest subset of equivalent residues is defined as the maximum clique. The advantage of this approach is that the similarity between the ligand environment of the template and the input protein is well described as a graph, and as such it is possible to implement an exact algorithm to find the maximum clique within it. As a matter of fact, LIBRA employs an exact algorithm for solving the problem of maximum clique detection (Carraghan and Pardalos, 1990) and incorporates novel pruning techniques based on the biological context. A more detailed technical implementation of the algorithm is presented in the Supplementary Materials.

### 2.3 LIBRA's input/output

In LIBRA's input window, the user can select the database to be used and set a series of parameters for the execution of the program, such as: the minimum percentage of similar residues between the input protein site and the known site, the minimum size of the similar site, the maximum rmsd value of the alignments, etc. For the detection of ligand binding sites, the user can also choose to eliminate from the final output all the alignments in which the residues of the input protein clash with the rototranslated ligand. LIBRA's input procedure is described in detail in the Supplementary Materials. LIBRA's results are displayed to the user in the output window, in the form of an interactive table. For each of the alignments the user is given the possibility of displaying three-dimensional representations of the alignment via the embedded JMol program (Hanson, 2010). Lastly, an info button provides information regarding the known protein (and ligand in the case of binding sites) and links to the relevant PubMed and Protein Data Bank web pages. The alignments displayed in the output window can be sorted by different criteria, including holo protein name, ligand ID, known site size, percentage of conservation, clique size and rmsd. Further details are given in the Supplementary Materials.

## 3 Results

### 3.1 Tests on catalytic sites and binding sites recognition

LIBRA's effectiveness in the prediction of ligand binding sites and of ligand identity has been extensively tested via two sets of apo-proteins from the LigASite database (Dessailly *et al.*, 2008). This database is a collection of proteins for which three-dimensional structures are available both in the apo, ligand-free form and in the holo form. The first test set, whose results can be found in Supplementary Table S2, included 381 apo-proteins and showed that LIBRA identified the correct binding site in 349 cases (91.6%), 332 (95%) of which featured the identified site as ranking first. The second test set, used for comparative purposes and whose results are reported in Supplementary Table S4, included 30 randomly-chosen apo-proteins and demonstrated that LIBRA was able to recognize the correct function/ligand in all of the cases analyzed. Interestingly enough, the correct ligand binding site was detected also when the corresponding hit was represented by a known protein displaying an amino acid sequence identity with the input protein as low as 18%. Two other applications for protein function/ligand binding sites prediction were tested against this same test set, i.e. eF-Seek and SiteSeer. Results indicate that LIBRA performed better than SiteSeer, the latter failing to find a solution in three cases out of 30, while it significantly outperformed eF-Seek, which did not return any significant results for ~30% of the cases. Besides, LIBRA has also been tested on the same test set used for the validation of ASSIST (Supplementary Table S5). Here, the correct active site was detected in 45 cases (83%), 39 (86.7%) of which ranking first, showing a significantly higher accuracy in comparison with ASSIST and a higher number of cases where the correct result ranked first.

### 3.2 System performance

LIBRA is a stand-alone software application, whose execution time may depend on the capabilities of the machine running it. On an I7 2860QM with 32-GB RAM, an Intel SSD 520 and a 64-bit OS and JVM, average running times of 25 min have been detected. This appears more than acceptable given the number of entries that make up the ligand binding sites database. When using LIBRA for the recognition of active sites with the much smaller catalytic sites database (1000 entries), the running time is always lower than 1 min. It is important to stress out that, in order to achieve these levels of performance, LIBRA takes advantage of a multi-threading algorithm.

## 4 Discussion

In this article, LIBRA, a software tool for ligand binding/active sites recognition, has been described. Regarding ligand binding sites detection, tests on the entire LigaSite non-redundant database (381 apo-proteins) demonstrate that LIBRA correctly predicts the binding site/ligand identity in 90% of the cases. Besides, on a smaller, randomly chosen, subset of LigaSite, LIBRA outperforms both SiteSeer and Ef-Seek. As far as active sites detection is concerned, LIBRA showed a significant improvement over ASSIST (Caprari *et al.*, 2014), which in turn had shown to yield results comparable to SiteSeer's. Most importantly, in LIBRA the correct prediction ranks first in more than 70% of the cases, as opposed to 60% in ASSIST. Finally, a web application fully interoperable with the stand-alone program and packed with additional features is currently under development and will be announced soon.

## Funding

This work was supported by the COST Action TD1102.

*Conflict of Interest:* none declared.

## References

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–242.

Biggs,N. *et al.* (1986) *Graph Theory*. Oxford University, Oxford, MA, pp. 1736–1936.

Caprari,S. *et al.* (2014) ASSIST: a fast versatile local structural comparison tool. *Bioinformatics*, **30**, 1022–1024.

Carraghan,R. and Pardalos,P.M. (1990) An exact algorithm for the maximum clique problem. *Oper. Res. Lett.*, **9**, 375–382.

Dessailly,B.H. *et al.* (2008) LigASite: a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**(Database issue), D667–D673.

Furnham,N. *et al.* (2014) The Catalytic Site Atlas 2. *Nucleic Acids Res.*, **42**, D485–D489.

Gao,M. and Skolnick,J. (2013) A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.*, **9**, e1003302.

Hanson,R.M. (2010) Jmol—a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.

Laskowski,R.A. *et al.* (2005a) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **3**, W89–W93.

Laskowski,R.A. *et al.* (2005b) Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**, 614–626.

Lee,H.S. and Im,W. (2013) Ligand binding site detection by local structure alignment and its performance complementarity. *J. Chem. Inf. Model.*, **53**, 2462–2470.

Murakami,Y. *et al.* (2013) Exhaustive comparison and classification of ligand-binding surfaces in proteins. *Protein Sci.*, **22**, 1379–1391.

Petrey,D. *et al.* (2015) Template-based prediction of protein function. *Curr. Opin. Struct. Biol.*, **32C**, 33–38.