BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 7 2014, pages 1022–1024
doi:10.1093/bioinformatics/btt664

Structural bioinformatics

Advance Access publication November 15, 2013

ASSIST: a fast versatile local structural comparison tool

Silvia Caprari[1,†], Daniele Toti[2,†], Le Viet Hung[1], Maurizio Di Stefano[3] and Fabio Polticelli[1,4,*]

[1]Department of Sciences, University of Roma Tre, 00146 Rome, [2]Department of Information and Electric Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy, [3]Artificial Solutions, 08026 Barcelona, Spain and [4]National Institute of Nuclear Physics, Roma Tre Section, 00146 Rome, Italy

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Structural genomics initiatives are increasingly leading to the determination of the 3D structure of target proteins whose catalytic function is not known. The aim of this work was that of developing a novel versatile tool for searching structural similarity, which allows to predict the catalytic function, if any, of these proteins.

Results: The algorithm implemented by the tool is based on local structural comparison to find the largest subset of similar residues between an input protein and known functional sites. The method uses a geometric hashing approach where information related to residue pairs from the input structures is stored in a hash table and then is quickly retrieved during the comparison step. Tests on proteins belonging to different functional classes, done using the Catalytic Site Atlas entries as targets, indicate that the algorithm is able to identify the correct functional class of the input protein in the vast majority of the cases.

Availability and implementation: The application was developed in Java SE 6, with a Java Swing Graphic User Interface (GUI). The system can be run locally on any operating system (OS) equipped with a suitable Java Virtual Machine, and is available at the following URL: http://www.computationalbiology.it/software/ASSISTv1.zip.

Contact: polticel@uniroma3.it

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

Received on July 7, 2013; revised on October 23, 2013; accepted on November 12, 2013

1 INTRODUCTION

In recent years the dogma 'one gene—one protein—one function' had to be corrected based on the observation that a wealth of proteins thought in the past to carry out a unique specific function show a multifunctional or 'moonlighting' character (Jeffery, 1999). In addition, the flourishing of structural genomics initiatives (Burley *et al.*, 1999; Chance *et al.*, 2002) is leading to the determination of the 3D structure of a number of target proteins whose precise catalytic function is not known. On the other hand, the availability of a huge number of enzymes' 3D structures has allowed to code structure–function relationships in databases, among which the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) is the most prominent example.

The possibility to predict the function of a protein based on its sequence or structure would be of great utility in the aforementioned cases, and prediction of proteins' functional sites by using sequence information is a long sought goal with early attempts dating back to several years ago such as the PROSITE (Bairoch and Bucher, 1994) and PRINTS (Attwood *et al.*, 1994) databases.

More recently, several algorithms and tools have been developed based on structural motifs similarity. Probably the first tool developed with this aim is TEmplate Search and Superposition (TESS) (Wallace *et al.*, 1997), which allows to search 3D templates in new structures to identify functional sites. Along the same line, SPatial Arrangements of Side chains and Main chains (SPASM) and RIGOR are two programs for the analysis and identification of functionally relevant spatial motifs in protein structures (Kleywegt, 1999).

The JESS algorithm is an evolution of TESS that allows to search protein structures using templates consisting of constraints on physical, sequential and geometric properties (Barker and Thornton, 2003). Among the web resources for protein function recognition, one of the most widely used is ProFunc, which analyzes the input protein sequence, fold and structure using several methods among which JESS is used as functional motifs identification algorithm (Laskowski *et al.*, 2005).

Recently, a similar approach has been implemented in the server Mark-Us that, in addition to sequence and structure, uses information such as protein cavities and electrostatic potential profile (Petrey *et al.*, 2009).

In this framework, the purpose of the present work was that of developing a fast and exclusively structure-based tool for functional sites and, more generally, structural similarity identification in proteins. The program, named ASSIST from Active Site Similarity Search Tool, is based on a local geometric/chemical comparison algorithm designed to find the largest subset of similar residues between input proteins and query sites/motifs. ASSIST is not meant to outperform current function prediction methods but rather to provide the user with a standalone program, which can be readily adapted to various problems that can be solved through local structural similarity searches. In the following sections the approach used in developing the tool is described and performance tests are presented.

2 METHODOLOGY

2.1 System overview

ASSIST is a system whose core functionality lies in trying to identify, within an input protein whose function is unknown, substructures

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

potentially similar to known catalytic sites. Basically, given an input protein P, the program tries to find sets of similar structures, or alignments, between P and a list of known catalytic sites, which can be either functional sites or structural motifs. These alignments are detected by means of geometric and chemical similarity among groups of residues coming from the input protein and from the known sites, at different and configurable structural levels (side chain, backbone or both). The execution flow of the system is made up of the following phases.

(1) A preliminary setup phase, where catalytic sites are retrieved and stored, by analyzing the information available online in the CSA and the corresponding information from the Protein Data Bank.

(2) A preprocessing phase, where the input protein is scanned and its pairs of residues are stored into a hash table according to their amino acid types and the distance between their geometric centers. This is done via a technique called *geometric hashing*, where information related to the protein's *residues*, taken as pairs, is indexed and mapped into a 3D hash index in terms of a so-called nine-tuple, which includes the name of the input protein and the name, number, chain id and geometric center of the residues from the pair; therefore, each 3D point of the hash structure will potentially contain a list of nine-tuples from the input protein. The 3D hash index thus built is exemplified in Figure 1.

(3) An alignment recognition phase, where the data structure created during the previous phase is compared with the known catalytic sites earlier acquired, so that a list of suitable alignments between those sites and the input protein are returned (if any) and filtered by several criteria (minimum score, maximum rmsd, etc).

(4) A result display phase, where the alignments found are presented to the user, who is given a number of options for sorting them, analyzing them and graphically displaying them as 3D models: this includes visualization via the JMol program (Hanson, 2010).

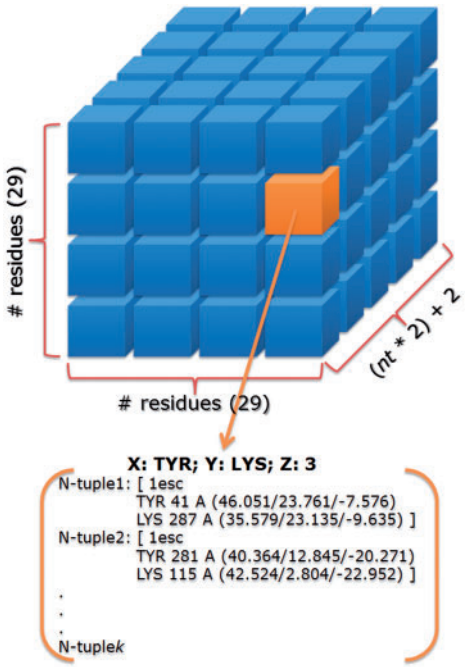

Fig. 1. Schematizationof the 3D hash index built during the preprocessing phase. Each 3D element of the index may contain a list of residue pairs R-N whose residue codes (respectively, from R and N) and mutual distance correspond to its X, Y and Z coordinates. The element highlighted in the figure (clearly not in scale) features, for instance, a set of 3D coordinates corresponding to X = 8 (the code for Tyr), Y = 5 (the code for Lys) and Z = 3 (the indexed value of their mutual distance), and thus includes a list of R-N pairs, expressed as nine-tuples, with these values

A more detailed description of the aforementioned phases can be found in the Supplementary Materials.

2.2 Technology, deployment and availability

ASSIST is a software application developed in Java SE 6, with an underlying HSQLDB (http://hsqldb.org) and a Java Swing Graphic User Interface (GUI). It also embeds the JMol visualization program as earlier described. The system can be run locally on any OS equipped with a suitable Java Virtual Machine and is available at the following URL: http://www.computationalbiology.it/software/ASSISTv1.zip.

3 RESULTS

3.1 Effectiveness on protein function recognition

To assess the effectiveness of ASSIST in recognizing the catalytic activity of an enzyme, tests were performed on 54 randomly chosen enzymes belonging to the six classes of the Enzyme Commission classification, i.e. oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5) and ligases (EC 6). Detailed results of the tests are reported in the Supplementary Table S2. As can be seen from the table, the program is able to correctly predict the catalytic activity of the enzymes analyzed in the vast majority of the cases. In >60% of the cases (34 of 54) ASSIST predicts the exact catalytic activity of the input protein. Furthermore, in 26 of 34 cases the exact solution is the best match, either by rmsd or by score. In 10 other cases, the program identifies a close functional homolog of the input protein even though only in 6 of these cases the functional homolog is the best match. Interestingly, in several cases ASSIST is able to predict the correct catalytic function of an input protein even when the sequence identity with the known functional homolog is well below the twilight zone (1–10% sequence identity; see Supplementary Table S2), cases in which sequence-based function recognition methods would likely fail. Finally, in nine cases the program fails to identify the correct enzyme function. However, this is due to the fact that either no entry is present in the CSA for the specific catalytic activity of the input protein or the corresponding entry in the CSA is defined by just one residue and thus ASSIST's distance-based approach is not applicable to these cases. Comparative tests show that, aside from these latter cases, ASSIST's performance is comparable with that of other function recognition methods such as ProFunc (see Supplementary Table S2). Some additional sample usage cases meant to illustrate the usefulness of the program are also illustrated in the Supplementary Materials. Among these we also provide an example of ASSIST's versatility showing the usage of the program to identify the binding site of the drug imatinib to human serum albumin by simply plugging in a database of known binding sites for that molecule in the place of the CSA.

3.2 System performance

ASSIST, in its current form, is a standalone software application, whose performance in terms of execution time may depend on the capabilities of the machine running it. On a I7 2860QM with 32GB RAM, an Intel SSD 520 and a 64-bit OS, with a light load from other tasks, an average of 9.5 s have been detected for testing a single protein against the known catalytic sites of all the 967 entries from the CSA (the whole execution flow of the system including all its three core phases); an average of 23 s for the initial, one-time-only setup of the application. On a 32-bit OS with lower specifications these values tend to double, keeping, however, run times at an acceptable level for a reasonably satisfactory user experience.

4 DISCUSSION

In this article, the development of ASSIST, a software tool to search for local structural similarities between a protein structure and a set of structural motifs, has been reported. Its application to the problem of function recognition has been presented, demonstrating that, notwithstanding the simplicity of its approach, ASSIST correctly predicts the specific catalytic function of a given input protein in the vast majority of the cases analyzed. Particularly interesting from this point of view is the case of the protein BfR192 (presented in the Supplementary Materials), which does not share any significant sequence or structure homology with proteins whose function is known. In this case, the exclusively structure-based approach of ASSIST is able to recognize the local structural similarity between BfR192 and the proline-specific aminopeptidase active site even in a different sequence/structure context. The tests also underlined the limits of ASSIST when coupled to the current CSA in which, for each enzyme, only residues directly involved in the catalytic reaction are present. To improve the performance of ASSIST, the tool should be used in conjunction with a collection of active sites in which also substrate binding residues and cofactors are present, a sort of 'Active Site Atlas'. The development of an Active Site Atlas according to the aforementioned criteria requires careful analysis of the literature related to all the CSA entries and, as such, is a long term effort. Currently this work has been undertaken and completed for ~10% of all the CSA entries. Finally, as reported in the 'Sample usage cases' section of the Supplementary Materials, ASSIST can be readily adapted to other biological problems involving local structure similarity searches by simply plugging in different structural motif databases in the place of the CSA.

ACKNOWLEDGEMENT

The authors thank Dr Giovanni Minervini for helpful discussion.

Funding: The authors wish to acknowledge financial support from the University of Roma Tre (CAL to F.P.).

Conflict of Interest: none declared.

REFERENCES

Attwood,T.K. *et al.* (1994) PRINTS-a database of protein motif fingerprints. *Nucleic Acids Res.*, **22**, 3590–3596.

Bairoch,A. and Bucher,P. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, **22**, 3583–3589.

Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.

Burley,S.K. *et al.* (1999) Structural genomics: beyond the human genome project. *Nat. Genet.*, **23**, 151–157.

Chance,M.R. *et al.* (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.*, **11**, 723–738.

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

Jeffery,C.J. (1999) Moonlighting proteins. *Trends Biochem. Sci.*, **24**, 8–11.

Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.

Laskowski,R.A. *et al.* (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.

Petrey,D. *et al.* (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc. Natl Acad. Sci. USA*, **106**, 17377–17382.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Wallace,A.C. *et al.* (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.