Conference Review

In silico identification of functional protein interfaces

Rachel E. Bell and Nir Ben-Tal*
Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

*Correspondence to: Nir Ben-Tal, Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.
E-mail: bental@ashtoret.tau.ac.il; Web: http://ashtoret.tau.ac.il/

Abstract

Proteins perform many of their biological roles through protein–protein, protein–DNA or protein–ligand interfaces. The identification of the amino acids comprising these interfaces often enhances our understanding of the biological function of the proteins. Many methods for the detection of functional interfaces have been developed, and large-scale analyses have provided assessments of their accuracy. Among them are those that consider the size of the protein interface, its amino acid composition and its physicochemical and geometrical properties. Other methods to this effect use statistical potential functions of pairwise interactions, and evolutionary information. The rationale of the evolutionary approach is that functional and structural constraints impose selective pressure; hence, biologically important interfaces often evolve at a slower pace than do other external regions of the protein. Recently, an algorithm, Rate4Site, and a web-server, ConSurf (http://consurf.tau.ac.il/), for the identification of functional interfaces based on the evolutionary relations among homologous proteins as reflected in phylogenetic trees, were developed in our laboratory. The explicit use of the tree topology and branch lengths makes the method remarkably accurate and sensitive. Here we demonstrate its potency in the identification of the functional interfaces of a hypothetical protein, the structure of which was determined as part of the international structural genomics effort. Finally, we propose to combine complementary procedures, in order to enhance the overall performance of methods for the identification of functional interfaces in proteins. Copyright © 2003 John Wiley & Sons, Ltd.

Keywords: functional regions; inter-protein interfaces; evolutionary relations; evolutionary rate; evolutionary conservation; homologous proteins

Introduction

The discrimination of functional oligomeric protein–protein interfaces observed in three-dimensional (3D) structures from contacts that are artifacts of crystallization is a challenging question. Many criteria have previously been used to identify and characterize protein–protein interfaces. These rely on considerations of the solvent accessible surface area buried upon association [14,13]; free energy changes upon alanine-scanning mutations [31]; experimental [35] and in silico [25] two-hybrid systems; scoring functions based on statistical potentials [21,26]; and physicochemical and geometric properties of the surface, such as electrostatics [28], hydrophobicity [32], amino acid composition [22] and shape complementarity or planarity [18,15,12]. Other approaches are based on scoring evolutionary conservation. The simplest of these are based on estimating the local amino acid’s information content or ‘relative entropy’ [7,34] and the more advanced involve phylogenetic reconstruction [17]. Other forms of utilizing phylogenetic information to this effect include the tracing of correlated mutations [23] and similarities between phylogenetic trees [24].

The most rudimentary approach for the identification of biological inter-protein interfaces uses
the solvent accessibility of the protein surface area (ASA). This approach is based on two fundamental assumptions: (a) biologically significant interfaces must be thermodynamically stable; and (b) relative stability is, by and large, proportional to the number of interacting atoms. Thus, the larger the surface area buried upon association, the more likely the interface is to be biologically relevant. Pongstingl et al. [26] showed that ASA correctly classified homodimers and monomers in 85% of cases.

In cases where a sufficient number of homologous proteins are available, evolutionary-based methods may aid in the discrimination between biological and non-biological contacts. Elcock and McCammon [7] showed that a simple relative entropy criterion can be used to correctly discriminate between X-ray crystal interfaces observed for homodimeric and monomeric proteins in 86% of cases. Valdar and Thornton [34] combined the ASA and conservation measures using statistical tests and accurately identified 92% of the examined inter-protein interfaces. Such methods are also commonly used for the identification of other functional interfaces involved, e.g. in catalytic activity and ligand-, peptide- or DNA-binding. Numerous applications using different evolutionary-based methods have been developed to this effect [1,3,4,6–9,11,16,19,29]. For example, Madabushi et al. [20] and Yao et al. [36] applied the Evolutionary Trace (ET) method for the identification of various known functional interfaces, and reported a success rate of 90% or more. Valdar [33] presented a critical review of 19 different methods for the scoring of evolutionary conservation, including the ET method. The conclusion was that none of the methods achieved statistical or biological rigour, presumably since they all suffer from inadequate treatment of the evolutionary process. The Rate4Site algorithm [27], which was recently developed in our laboratory and not reviewed by Valdar, provides a more accurate treatment of the process. Rate4Site accepts as input a phylogenetic tree reconstructed from a multiple-alignment of the homologous sequences, and provides a maximum likelihood estimate of the evolutionary rates of the amino acid sites. The topology and branch lengths of the tree, as well as the underlying stochastic process of the evolution of the homologues, are explicitly taken into account in the calculations, which makes Rate4Site very accurate and sensitive; we demonstrated that the algorithm is superior to other methods in the identification of various functional regions [27]. We have also developed a web-server, ConSurf, which implements this algorithm for proteins with a known 3D-structure (http://consurf.tau.ac.il/) [10].

Recently, a large-scale analysis was carried out using the ET method in combination with several statistical-significance tests, in an attempt to identify known functional interfaces in a set of 86 proteins [36]. With a few exceptions, the method correctly identified these interfaces, further demonstrating the power of evolutionary-based methods. Here we demonstrate that ConSurf was sensitive enough to identify the ATP binding-pocket of putative protein Mj0577, which was overlooked in the ET analysis.

Methods

For comparison, we attempted to use an input that is as similar as possible to the one that was generated by Yao et al. [36]. The reference structure (PDB code: 1mjh [37]) was entered into the ConSurf web-server; 19 homologues were retrieved from the SWISSPROT database [5], using PSI–BLAST ([2]; E-score threshold of 0.05), and aligned using CLUSTAL W [30]. The alignment is available at: http://ashtoret.tau.ac.il/~rebell.

Results and discussion

One of the major problems with hypothetical proteins is how to relate them to known protein families. In the case of the putative protein Mj0577 the homologues are also classified as hypothetical; furthermore, the sequence identity among the homologous proteins can be as low as ~15%. We demonstrate here that ConSurf successfully detected the ATP binding-groove of the protein (Figure 1). The residues comprising the ATP binding-site are not strictly conserved among the homologues, which may be the reason why the ET analysis of Yao et al. [36] failed to significantly detect it. Nevertheless, ConSurf assigns these residues with reasonably high conservation scores. Interestingly, the ConSurf analysis also shows that the homodimer–protein interface, which was observed in the X-ray crystal structure and was not examined by
Figure 1. The conserved ATP-binding groove in the ATP-binding domain of hypothetical protein Mj0577 (PDB code: 1mjh [37]). The domain is represented as a space-filled model, and the ATP and Mn ion are shown as ball-and-stick models. The evolutionary rates are colour-coded onto the domain as follows: slowly evolving residues are maroon, residues that evolve at average rates are white, and rapidly evolving residues are turquoise. Residues aligned with less than 10 sequences out of a total of 19 and whose conservation grades are of low confidence are marked in yellow. Several amino acids that are directly involved in ATP-binding (Asp13, Val41, Gly127, Gly130, Gly140, Ser141, Val142 and Thr143) receive the highest conservation grades; yet none of them is exclusively invariant among the homologous proteins. They interact with the ATP molecule mainly through their backbone.

Yao et al., is highly conserved, suggesting that this interface is biologically important (Figure 2).

Recent progress in structure determination techniques has resulted in a major increase in the number of novel high-resolution 3D structures with unknown function. Evolutionary analysis of these proteins can provide means for the identification of their functional regions. We have analysed dozens of proteins with unknown function (data not shown). The bottleneck in the identification of the functional interfaces in such proteins appears to be the number of putatively related homologues and their degree of similarity. Methods that make use of evolutionary information require a minimum number of homologous proteins to provide a sufficient span of their evolutionary history. Our experience has been that even ConSurf, in spite of its high accuracy and sensitivity, is somewhat limited when dealing with less than 10 homologous proteins. It is also noteworthy to recall that some rather important functional interfaces are not evolutionarily conserved; the hyper-variable peptide recognition groove in MHC molecules is an excellent example of this. In cases where there is an insufficient number of homologues, some of the complementary methods mentioned above may be used.

Acknowledgements

We are grateful to the Bioinformatics Unit at the George S. Wise Faculty of Life Sciences at Tel Aviv University for providing technical assistance and computational facilities, and to Karen B. Avraham for her assistance. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund.
References


Submit your manuscripts at http://www.hindawi.com