BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 8 2004, pages 1322–1324
DOI: 10.1093/bioinformatics/bth070

ConSeq: the identification of functionally and structurally important residues in protein sequences

Carine Berezin[1], Fabian Glaser[1], Josef Rosenberg[1], Inbal Paz[1], Tal Pupko[2], Piero Fariselli[3], Rita Casadio[3] and Nir Ben-Tal[1,*]

[1]Department of Biochemistry and [2]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, 69978 Israel and [3]Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Via Irnerio 48, 40126 Bologna, Italy

Received on June 10, 2003; revised on August 22, 2003; accepted on September 9, 2003
Advance Access publication February 10, 2004

ABSTRACT

Motivation: ConSeq is a web server for the identification of biologically important residues in protein sequences. Functionally important residues that take part, e.g. in ligand binding and protein–protein interactions, are often evolutionarily conserved and are most likely to be solvent-accessible, whereas conserved residues within the protein core most probably have an important structural role in maintaining the protein's fold. Thus, estimated evolutionary rates, as well as relative solvent accessibility predictions, are assigned to each amino acid in the sequence; both are subsequently used to indicate residues that have potential structural or functional importance.

Availability: The ConSeq web server is available at http://conseq.bioinfo.tau.ac.il/

Contact: bental@ashtoret.tau.ac.il

Supplementary information: The ConSeq methodology, a description of its performance in a set of five well-documented proteins, a comparison to other methods, and the outcome of its application to a set of 111 proteins of unknown function, are presented at http://conseq.bioinfo.tau.ac.il/ under 'OVERVIEW', 'VALIDATION', 'COMPARISON' and 'PREDICTIONS', respectively.

INTRODUCTION

We have developed Rate4Site, an algorithmic tool for the identification of functionally important regions in proteins by estimating the rate of amino acid substitutions at each position in a Multiple Sequence Alignment (MSA) of homologous proteins (Pupko *et al*., 2002). The underlying assumption of this approach is that, in general, structurally and functionally important residues are slowly evolving. (The hypervariable peptide binding sites in MHC molecules are prominent counter examples). We have recently presented the ConSurf server (http://consurf.tau.ac.il; Glaser *et al*., 2003), which automates these tools for proteins of known three-dimensional (3D) structure. A sequence-based tool and web server, ConSeq, which is suitable for proteins of unknown 3D structure, is presented here.

METHODOLOGY

A brief description of the methodology follows; a more detailed description is provided at http://conseq.bioinfo.tau.ac.il/ under 'OVERVIEW'.

Computations: Given a query protein sequence, the server automatically collects its homologous sequences and multiply aligns them. A phylogenetic tree for the homologues is derived. The server then calculates the substitution rate at each position in the MSA, taking into account the evolutionary relations between the homologues, using the maximum likelihood paradigm (Pupko *et al*., 2002). Based on the MSA, the server also implements a neural network prediction scheme to discriminate between buried and exposed residues in globular proteins (Fariselli and Casadio, 2001).

Visualization: Conservation scores are grouped into a nine-colour grade scale. The query protein, with the conservation grades colour-coded onto its amino acid sequence, can be visualized on-line (Fig. 1). The buried or exposed predicted status of each residue is marked in the first row below the sequence. Slowly evolving (colour grades 8 and 9) and 'exposed' residues are predicted to be functional, whereas slowly evolving (colour grade 9) and 'buried' residues are predicted to be structurally important. Both are indicated in the second row below the sequence, as 'f' and 's', respectively.

RESULTS

Validation: A description of the performance of the ConSeq server is available at http://conseq.bioinfo.tau.ac.il/ under

*To whom correspondence should be addressed.

Bioinformatics 20(8) © Oxford University Press 2004; all rights reserved.

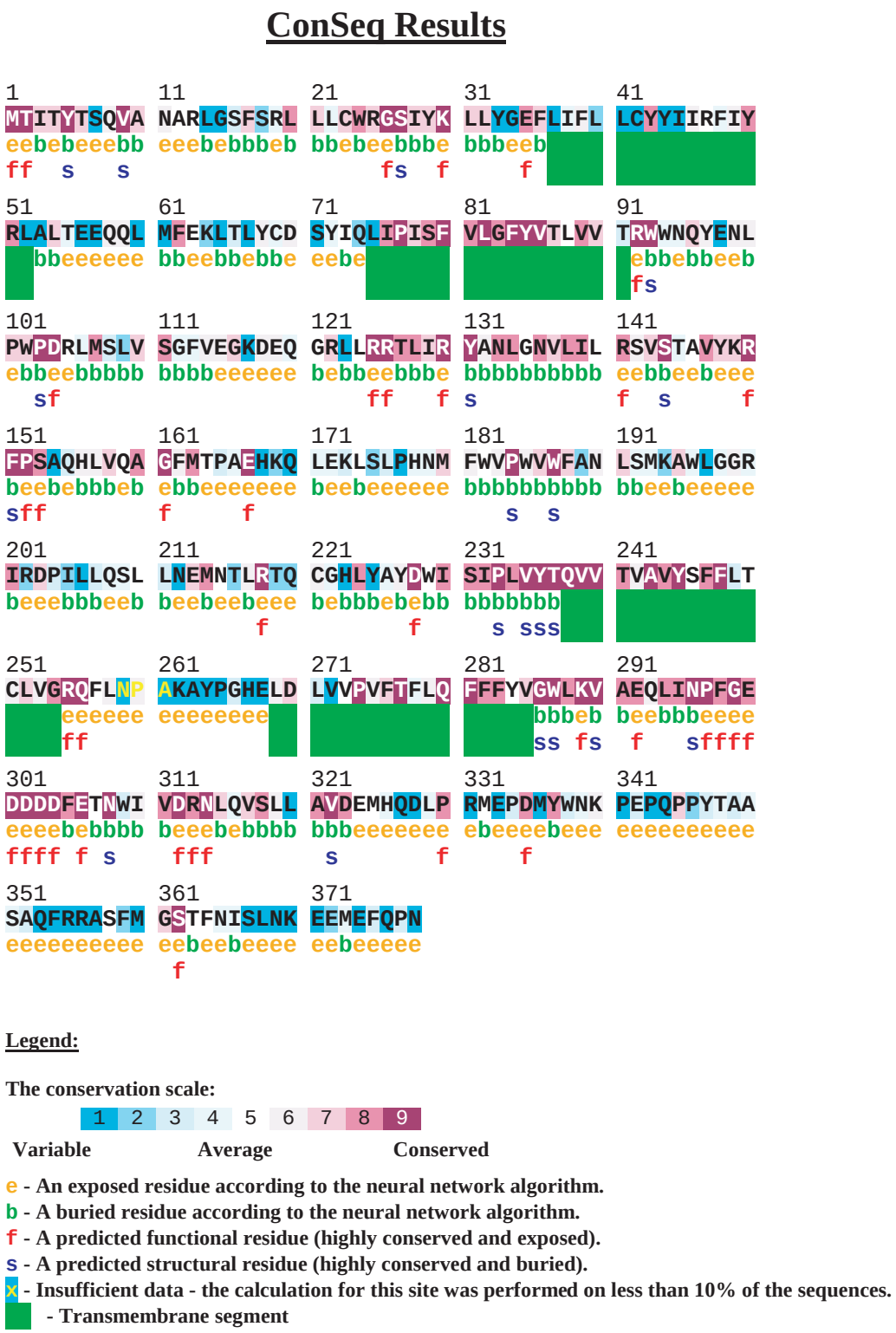

Fig. 1. ConSeq predictions demonstrated on human bestrophin (SWISS-PROT: VMD2_HUMAN), using 43 homologues obtained from the Pfam database (family code: DUF289). The sequence of the query protein is displayed with the evolutionary rates at each site colour-coded onto it (see legend). The residues of the query sequence are numbered starting from 1. The first row below the sequence lists the predicted burial status of the site (i.e. 'b'—buried versus 'e'—exposed). The second row indicates residues predicted to be structurally and functionally important: 's' and 'f', respectively. The four transmembrane segments according to Bakall *et al*. (1999) are marked as green boxes.

'VALIDATION'. In summary, five proteins with known 3D structure and precise annotation of their functional sites were studied using ConSeq; these proteins have eight functional sites of a different nature, i.e. ligand, protein and DNA/RNA binding sites. The average success rate in identifying the functional residues, i.e. those that are highly conserved and exposed to solvent is 56%. As anticipated, the error arises primarily from a wrong assignment of the buried/exposed status of the residues. We are investigating various options to minimize this error. In any event, the assignment of sequence conservation scores may be valuable for biochemists, molecular biologists and geneticists.

Predictions: We used ConSeq to make predictions of the functional and structural residues in 111 domains and proteins of unknown function and structure. One example is given below; the remainders are presented at http://conseq.bioinfo.tau.ac.il/ under 'PREDICTIONS'.

Pfam DUF289

The DUF289 domain, from the Pfam database (Bateman *et al*., 2002), is found in a family of 41 eukaryotic proteins. We carried out a ConSeq analysis using the 378-residue long human bestrophin domain as a query sequence. The full-length bestrophin protein contains 585 amino acids, and was predicted from the open reading frame of the VMD2 gene; thus far, its function is still unknown. Several mutations in the VMD2 gene have been associated with Best macular dystrophy, a progressive autosomal dominant disease that causes vision loss (Bakall *et al*., 1999). Bestrophin is predicted to include four transmembrane segments and the analysis is performed only on the soluble regions of the domain.

The ConSeq results (Fig. 1) reveal that 37 residues are predicted to be functionally important and 20 residues to have a structural role. ConSeq assigned conservation grades above the average (6–9 colour grades) in this domain to 40 out of 49 missense mutations, which have been reported to be associated with Best disease (http://www.uni-wuerzburg.de/humangenetics/vmd2.html). About 10 of these 49 residues are predicted to be functional residues, while five are predicted to have an important structural role.

Thus, ConSeq is a fast and useful tool for the design and analysis of mutagenesis studies.

ACKNOWLEDGEMENTS

We thank Amit Kessel and Rachel E. Bell for helpful discussions. 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. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund to N.B.-T.

REFERENCES

Bakall,B., Marknell,T., Ingvast,S., Koisti,M.J., Sandgren,O., Li,W., Bergen,A.A.B., Andreasson,S., Rosenberg,T., Petrukhin,K. and Wadelius,C. (1999) The mutation spectrum of the bestrophin protein—functional implications. *Hum. Genet*., **104**, 383–389.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.

Fariselli,P. and Casadio,R. (2001) RCNPRED: prediction of the residue co-ordination numbers in proteins. *Bioinformatics*, **17**, 202–204.

Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.