InterPreTS: protein Interaction Prediction through Tertiary Structure

Patrick Aloy* and Robert B. Russell

EMBL, Meyerhofstrasse, 1, D-69117, Heidelberg, Germany

Received on May 23, 2002; revised on July 15, 2002; accepted on July 24, 2002

ABSTRACT
Summary: InterPreTS (Interaction Prediction through Tertiary Structure) is a web-based version of our method for predicting protein–protein interactions (Aloy and Russell, 2002, Proc. Natl Acad. Sci. USA, 99, 5896–5901). Given a pair of query sequences, we first search for homologues in a database of interacting domains (DBID) of known three-dimensional complex structures. Pairs of sequences homologous to a known interacting pair are scored for how well they preserve the atomic contacts at the interaction interface. InterPreTS includes a useful interface for visualising molecular details of any predicted interaction.

Availability: http://www.russell.embl.de/interprets.

Contact: aloy@embl.de

Protein interactions and their associated networks are key to understanding complex cellular processes. Much recent effort has been put into experimental and computational methods to identify protein interaction and complexes (reviewed by von Mering et al., 2002). However, despite these efforts in interaction discovery, only limited attention has been paid to one of the most informative sources of protein–protein interactions data: complexes of known three-dimensional (3D) structure.

We have recently described a method to model putative interactions on known 3D complexes (Aloy and Russell, 2002). The method assesses the fit of two potential interacting partners on a complex of known 3D structure and infers molecular details of how the interaction is likely to occur (i.e. which residues are in contact). Here we describe the web-based version of this method: InterPreTS—Interaction Prediction through Tertiary Structure.

The Figure provides an overview of the method. Given sequences of two query proteins we first assign the sequences to one or more Pfam (Bateman et al., 1997) domains by means of Blast2 (Altschul et al., 1997). Domains are assigned to regions that have $E$-values $\leq 10^{-10}$ for any sequence in the ‘full’ set for a Pfam family. We then check whether any combination of domain pairs is present in our database of interacting domains (DBID). This database correctly contains a total of 1131 complexes of known structure, which correspond to 429 unique interacting pairs of Pfam domains. If one or more complexes are found, we align the query sequences to the closest homologues of known structure with HMMer (Eddy, 1998) and the HMM profile for the Pfam domains. The user can also provide different alignments of the query proteins to a particular complex of known structure.

Given an alignment of the query protein sequences to a complex of known structure, we apply our method of protein interaction prediction (Aloy and Russell, 2002). In brief, after identifying the residues that make atomic contacts in the complex of known structure, we check whether the query protein sequences preserve these interactions by means of empirical potentials. We then estimate a statistical significance for the potential interaction based on a background of random sequences.

The output page includes links to information relating to the interacting Pfam domains as well as to the complex structure in the Protein Data Bank (Berman et al., 2000). It also includes the score and Z-score for the interacting protein pair. A Z-score $\geq 2.3$ indicates a significance of the prediction of 99%, Z-scores $\geq 1.3$ indicate a significance of 90% and when the Z-score is $<1.3$ the two proteins are predicted not to interact in the same way as the known complex structure. This can arise owing to either a genuine non-interacting pair of proteins or to conformational changes (i.e. different loop conformations, indels, etc.) modifying the nature of the interaction. We make no attempt to model such changes here.

InterPreTS uses the program ModView (Ilyin and Sali, 2001) to combine a display of the two interacting domains of known structure and sequence alignments of these to the query protein sequences. A backbone representation of the two interacting chains is, by default, shown in red and cyan respectively. Residues interacting via side-chain are shown in wireframe and ‘cpk’ coloured. Protein sequences are coloured according to the structure (i.e. domain-1 red and domain-2 cyan). All the residues involved in the interaction are shown in capital letters and those interacting

*To whom correspondence should be addressed.
via side-chain contacts are coloured in grey. As ModView only runs under Linux/Unix, the two interacting chains can also be displayed with RasMol (Sayle and Milner-White, 1995). However, the combination of sequence and structure information is only possible via ModView.

Future implementations will include a browsing system to easily access the results obtained from running InterPreTS on entire genomes. The databases used by InterPreTS will be regularly updated to keep pace with updates to Pfam and the Protein Data Bank.

ACKNOWLEDGEMENTS
We thank C. Hadley, A. Stark and T. Sardon (EMBL) for testing the server.

REFERENCES


Ilyin, V. and Sali, A. (2001) Protein superfamily analysis by ModView. CSHL meeting.
