BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 15 2009, pages 1963–1965
doi:10.1093/bioinformatics/btp335

Sequence analysis

InterMap3D: predicting and visualizing co-evolving protein residues

Rodrigo Gouveia-Oliveira†, Francisco S. Roque†, Rasmus Wernersson, Thomas Sicheritz-Ponten, Peter W. Sackett, Anne Mølgaard and Anders G. Pedersen*

Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received on January 31, 2009; revised on April 29, 2009; accepted on May 26, 2009

Advance Access publication June 15, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: InterMap3D predicts co-evolving protein residues and plots them on the 3D protein structure. Starting with a single protein sequence, InterMap3D automatically finds a set of homologous sequences, generates an alignment and fetches the most similar 3D structure from the Protein Data Bank (PDB). It can also accept a user-generated alignment. Based on the alignment, co-evolving residues are then predicted using three different methods: Row and Column Weighing of Mutual Information, Mutual Information/Entropy and Dependency. Finally, InterMap3D generates high-quality images of the protein with the predicted co-evolving residues highlighted.

Availability: http://www.cbs.dtu.dk/services/InterMap3D/

Contact: gorm@cbs.dtu.dk

1 INTRODUCTION

Co-evolution of amino acid residues occurs when two or more residues in a protein exert selective pressure on each other, so each residue has an influence on the evolution of the rest. Co-evolution has been conceived of as occurring in a variety of settings, but mostly between amino acids that are close to each other in a protein's 3D structure. For this reason, visualization of the location of co-evolving sites in the 3D structure is of interest.

InterMap3D is a tool for detection and visualization of co-evolving residues useful also to non-expert users. InterMap3D is able to take a single sequence as input, from which it automatically finds a set of homologous sequences, constructs a multiple alignment, discovers co-evolving sites and produces as output an image of the protein 3D structure with co-evolving residues highlighted. The tool can also accept a user-generated alignment. A manually curated dataset will most often be of better quality than the automatically generated one, thus improving the quality of the predictions. An additional goal with InterMap3D is to make already existing methods for detection of co-evolution available to the protein research community. While there has been considerable interest in detecting co-evolving protein residues in the past years, the dissemination of these methods has not been as active, and only a few, such as Dependency (Tillier and Lui, 2003), CAPS (Fares and McNally, 2006), CoMap (Dutheil and Galtier, 2007) and PCOAT (Qi and Grishin, 2004) have been made available to the general community. InterMap3D tries to remedy this situation by providing several methods in a freely available web server, built in a modular fashion that enables easy expansion to new methods.

2 IMPLEMENTATION

InterMap3D is a synergy of several tools in the fields of biological data representation, phylogeny and detection of co-evolution. InterMap3D can take the user's alignment or create one from a single sequence given by the user (in FASTA format). If the user cannot provide an alignment, but only a sequence, InterMap3D compares it with UniProt (Apweiler *et al*., 2004) via BLASTP. All significant database hits covering a minimum of 50% of the protein length are retrieved (this value can be set by the user). All compatible homologs are then aligned using either MAFFT (Katoh *et al.*, 2005), MUSCLE (Edgar, 2004) or ClustalW (Thompson *et al*., 1994). That alignment is then processed by MaxAlign (Gouveia-Oliveira *et al*., 2007), diminishing the number of gapped columns in the alignment, and passed to the tools predicting co-evolving residues. The prediction of co-evolving residues is done by one or more of the three methods currently implemented in InterMap3D: Row and Column Weighing of Mutual Information (RCW-MI) (Gouveia-Oliveira and Pedersen, 2007), Mutual Information/Entropy (MI/E) (Martin *et al*., 2005) and Dependency. Finally the results are mapped onto a 3D structure if possible, using the FeatureMap3D program (Wernersson *et al*., 2006). Briefly, FeatureMap3D searches for the most similar homologous protein with an experimentally determined 3D structure, and then uses PyMOL (Delano, 2002) to plot the predicted pairs of co-evolving sites onto that structure. The final result is a 3D image of the protein structure in several formats. Highlighted in this image are the co-evolving pairs (or networks) and also completely conserved sites, as co-evolution analysis cannot rule out interactions between these. The reliability of inferring the location of the co-evolving residues, for proteins that have not themselves been structurally characterized obviously depends on sequence similarity. To help the user judge how representative a structure is for a given protein sequence, it is possible to create a figure of the structure, color-coded by sequence conservation.

From the output page, the user has several options for getting additional information related to the analysis. This includes a PNG-format plot of the labeled 3D structure, the corresponding PyMol script and PDB file, the alignment, etc. For each predicted pair of co-evolving sites, the user can also plot a phylogenetic tree showing

*To whom correspondence should be addressed.

†These authors contributed equally.

how the amino acids present at these two sites change over the tree (Zmasek and Eddy, 2001). If the same pair of amino acids arise independently at several places in the tree, then co-evolution is more likely to be true.

3 PREDICTION OF CO-EVOLUTION

There are currently three methods in InterMap3D for predicting co-evolving pairs of residues: RCW-MI, MI/E and Dependency. The user can also choose to use the intersection between predictions produced by these methods.

MI/E extracts the entropy dependency from the signal by dividing MI by the joint site's entropy.

$$\mathrm{MI}/H(X,Y)=\frac{\sum_i\sum_j P(x_iy_j)\log\left(P(x_iy_j)/P(x_i)\cdot P(y_j)\right)}{-\sum_i\sum_j P(x_iy_j)\log\left(P(x_iy_j)\right)}$$

RCW-MI and Dependency aim at extracting both the phylogenetic signal and the entropy dependency. RCW-MI does that by considering that the signal in most pairwise comparisons results from both the phylogenetic signal and the entropy-driven signal. Thus, it weighs down pairwise MI score by the average MI of each site as:

$$\mathrm{RCW}(A;B)=\frac{\mathrm{MI}_{ij}}{\left(\mathrm{MI}_{.j}+\mathrm{MI}_{i.}-2\mathrm{MI}_{ij}\right)/\left(n-1\right)}$$

Where MI_{ij} represents the MI between sites i and j, and a dot stands for the summation over all sites. RCW-MI also allows discarding the top hits in each summation in order to accommodate for more than two-way co-evolution.

The Dependency method, on the other hand, considers that the above weighing, used by RCW-MI, extracts only the phylogenetic signal, and uses a second weighing to account for the entropy-driven signal, producing a set of best hits entitled S1. This set is then filtered, yielding S2, which also contains information about *P*-values. In Intermap3D, the output provided is the S1 set, while the S2 set can be downloaded in text format.

For both the MI/E and RCW-MI methods, we also estimate *P*-values for the predicted co-evolving sites. This is done using a three-step, heuristic approach: first, we estimate a maximum likelihood phylogenetic tree and substitution model parameters from the processed alignment using the PhyML program (Guindon and Gascuel, 2003). The tree and other fitted model parameters are then used to generate a number of simulated alignments using the program Seq-Gen (Rambaut and Grassly, 1997). (The default number of simulated alignments is two, but this can be controlled by the user). By design, these alignments do not contain any co-evolving site pairs. Finally, we compute the score distribution from the simulated alignments, and use this as a null distribution based on which *P*-values for real, biological scores can be estimated. Specifically, this is done by fitting a generalized Pareto distribution (GPD) to the right tail of the null distribution (the top 2%) and then using the fitted GPD for computing tail-probabilities (i.e. *P*-values) for each of the predicted co-evolving pairs. The GPD is well suited to model tails of a wide variety of distributions (Coles, 2001). We here use it as a heuristic shortcut for providing tail probabilities with much finer resolution than what is supported by the empirical cumulative distribution function based on the relatively few simulations we perform. This way we partially avoid the computationally expensive construction and analysis of a large number of simulated alignments, while still having a sound basis for *P*-value estimation. GPD fitting and computation of GPD tail probabilities is done using the R-packages ismev and evd (R development core team, 2008; Stephenson, 2002).

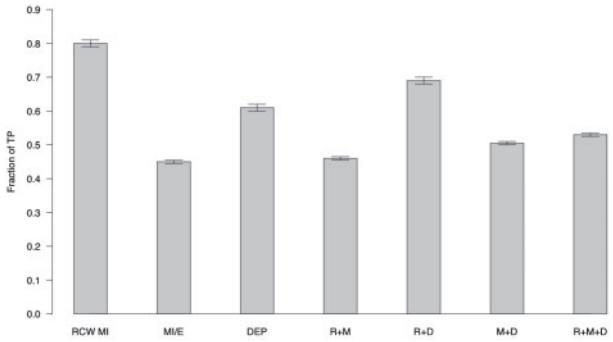

Fig. 1. Comparison of the performance for all the methods available in Intermap3D (DEP: Dependency; combined methods are indicated by the initials of the used methods). Performance was measured on a synthetic dataset with mostly independently evolving sites, at four different rates. The fraction of positives was calculated at the threshold of 20 best hits.

We compared the performance of the different methods and their intersections using 100 simulated datasets of 64 taxa and 300 residues each, evolved along balanced trees. In each alignment, there were 20 pairs of co-evolving residues. Residues were divided into four classes evolving at different rates, both when evolving independently and when co-evolving. The results, shown in Figure 1, suggest RCW-MI to be the best method in these conditions. All methods were very good at spotting pairs of residues co-evolving slowly, but RCW-MI performed better at residues evolving at intermediate rates.

ACKNOWLEDGEMENTS

The authors thank the Danish Center for Scientific Computing.

Funding: Foundation for Science and Technology, Portugal (grant SFRH/BD/12448/2003 to R.G.-O.).

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Coles,S.G. (2001) *An Introduction to Statistical Modelling of Extreme Values*. Springer, London.

Delano,W. (2002) *The PyMOL Graphics System*. DeLano Scientific, San Carlos, CA.

Dutheil,J. and Galtier,N. (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol. Biol.*, **7**, 242

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Fares,M.A. and McNally,D. (2006) CAPS: coevolution analysis using protein sequences. *Bioinformatics*, **22**, 2821–2822.

Gouveia-Oliveira,R. and Pedersen,A.G. (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Alg. Mol. Biol.*, **2**, 12.

Gouveia-Oliveira,R. *et al.* (2007) MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics*, **8**, 312.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518

Martin,L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Qi,Y. and Grishin,N.V. (2004) PCOAT: positional correlation analysis using multiple methods. *Bioinformatics*, **20**, 3697–3699.

R Development Core Team (2008) R: a language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria.

Rambaut,A. and Grassly NC. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci*., **13**, 235–238.

Stephenson,A.G. (2002) evd: Extreme Value Distributions. *R News*, **2**, 31–32.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22,** 4673–4680.

Tillier,E.R. and Lui,T.W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**, 750–755.

Wernersson,R. *et al.* (2006) FeatureMap3D–a tool to map protein features and sequence conservation onto homologous structures in the PDB. *Nucleic Acids Res.*, **34**, W84–W88.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.