Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W69–W72***
DOI: 10.1093/nar/gkh471

ADVICE: Automated Detection and Validation of Interaction by Co-Evolution

Soon-Heng Tan*, Zhuo Zhang and See-Kiong Ng

Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore

Received February 15, 2004; Revised and Accepted April 30, 2004

ABSTRACT

ADVICE (Automated Detection and Validation of Interaction by Co-Evolution) is a web tool for predicting and validating protein-protein interactions using the observed co-evolution between interacting proteins. Interacting proteins are known to share similar evolutionary histories since they undergo coordinated evolutionary changes to preserve interactions and functionalities. The web tool automates a commonly adopted methodology to quantify the similarities in proteins' evolutionary histories for postulating potential protein–protein interactions. ADVICE can also be used to validate experimental data against spurious protein interactions by identifying those that have few similarities in their evolutionary histories. The web tool accepts a list of protein sequences or sequence pairs as input and retrieves orthologous sequences to compute the similarities in the proteins' evolutionary histories. To facilitate hypothesis generation, detected co-evolved proteins can be visualized as a network at the website. ADVICE is available at http://advice.i2r.a-star.edu.sg.

INTRODUCTION

Co-evolution is a process whereby two or more species interact and influence genetic changes in one another. The process is also evident at the molecular level, where interacting proteins exhibit coordinated mutations to evolve at a similar rate (1). Mutation—a mechanism of evolution—disrupts protein interactions when residue changes occur within inter-protein contact sites or at regions implicated in the structural integrity of proteins. When a disrupted interaction leads to reduced fitness, the mutated sequence will be selected against and removed by natural selection. However, the mutated sequence will be retained if compensatory mutations that preserve the interaction occur in its interacting partners. As a result, interacting proteins will seem to evolve at the same rate and have similar evolutionary histories. This is a phenomenon that has been well characterized in various receptor–ligand systems (2–4) such as two-component signal transduction (5).

Observed co-evolution between interacting proteins has been used previously to predict protein interaction sites (6) and to improve docking algorithms (7,8). Recently, Goh *et al.* (9) adopted a statistical method to quantify the similarities in the evolutionary histories of proteins to predict the interactions of chemokines with their receptors based on the high correlation in the distance matrices constructed from multiple sequence alignments. Pazos and Valencia (10) extended the idea to genome-wide prediction of protein–protein interactions in *Escherichia coli*. The co-evolution approach was later further exploited to successfully pinpoint a family of ligands to its specific receptors (11). In these works, the methodology adopted to detect co-evolved interacting proteins consisted of the following sequential steps: (i) searching and retrieving pairs of orthologous sequences from databases, (ii) constructing distance matrices from the multiple sequence alignments of the retrieved orthologous sequences and (iii) measuring similarities in evolutionary histories of proteins by comparing the distance matrices constructed.

We have implemented ADVICE (Automated Detection and Validation of Interaction by Co-Evolution)—a web-based tool—that automates the steps needed to compute the similarities between proteins' evolutionary histories. The web tool can aid biologists in postulating potential protein–protein interactions using co-evolution. We also propose to use co-evolution between interacting proteins to rapidly validate experimentally derived protein–protein interactions against artificial interactions. It is possible that non-biological interactions that do not occur in nature may be detected under experimental conditions. However, these artificial interactions will not be subject to natural selection to exhibit co-evolution. As a consequence, ADVICE can be used to identify such spurious experimental interactions by finding interacting pairs that have little or no similarities in their evolutionary

*To whom correspondence should be addressed. Tel: +65 6874 6929; Fax: +65 6774 8056; Email: soonheng@i2r.a-star.edu.sg

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors histories. ADVICE can be useful for rapidly assessing the quality of large volumes of interaction data from high-throughput detection methods such as yeast-two hybrid (12,13), affinity purification (14,15) and protein chip experiments (16).

INPUTS

ADVICE allows both interactive and batch modes for processing. In the interactive mode, a user submits a pair of protein sequences in raw or FASTA format, or a list of protein sequences where all possible pairwise combinations of sequences will be permuted automatically by ADVICE for processing. When more than one pair of protein sequences is provided as input, ADVICE allows the detected co-evolved protein pairs to be visualized as a network. In the batch mode, the web tool accepts a list of sequence pairs for processing. The computed results will be sent to an email address provided by the user.

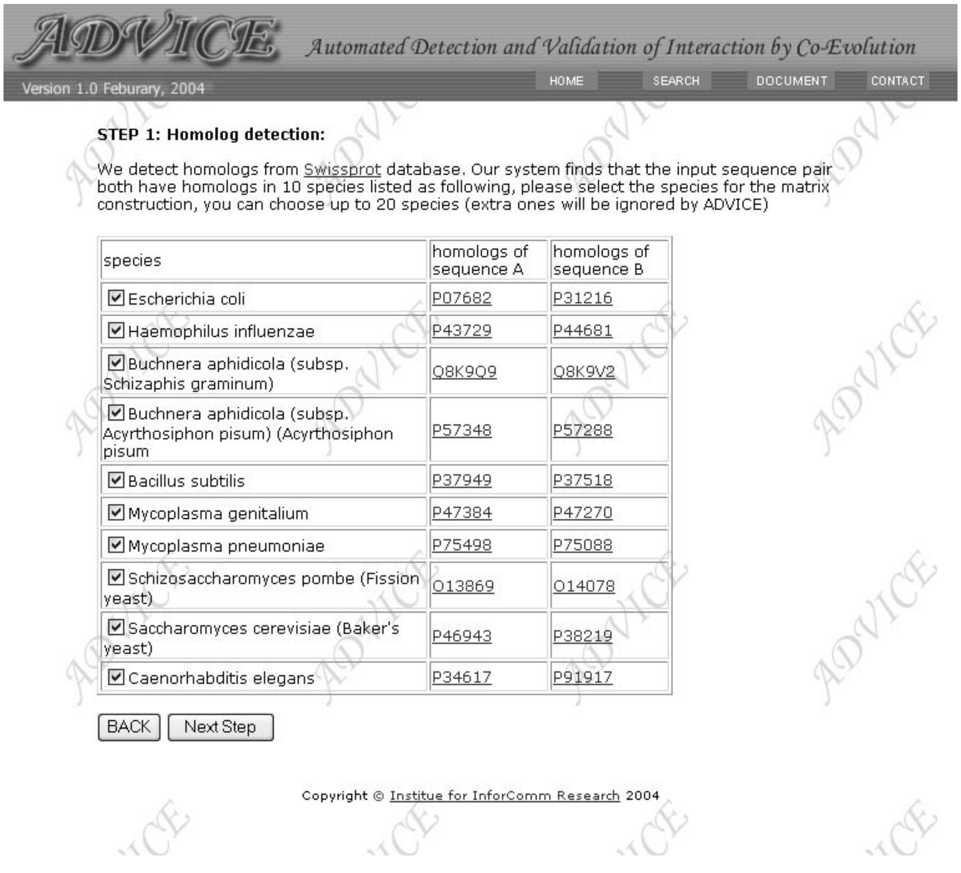
METHODOLOGY

Identifying orthologous sequences

The pair of sequences submitted by the user is used to search sequence databases for orthologous sequences based on sequence similarities. Identified orthologous sequences will be used to compute each input protein's evolutionary history. ADVICE allows users the option to search for orthologous sequences either from one of the four kingdoms of life (Eukaryota, Prokaryota, Archaebacteria and Viridae) or from the Swiss-Prot (release 42.9) and/or TrEMBL (release 25.9) databases (17). BLAST v2.2.4 (18) is used to search these databases and the user can control the sensitivity of the search by setting an *E*-value threshold for the BLAST hits.

Constructing distance matrices

To detect co-evolved proteins from their evolutionary histories, we use only pairs of orthologous sequences occurring together in the same species for constructing the distance matrices. By default, ADVICE uses sequence pairs from the top 10 species (based on highest average *E*-value of the BLAST hits) to construct the respective distance matrices from multiple sequence alignments, excluding those species where more than one orthologous sequence of the input sequences is found (since it would be difficult to determine which is the actual ortholog). In the interactive mode, the user can manually inspect annotations of the sequences and remove/add orthologous sequence pairs (Figure 1). ClustalW v1.84 (19) is used to construct the two distance matrices from respective multiple sequence alignments of the pairs of orthologous sequences.

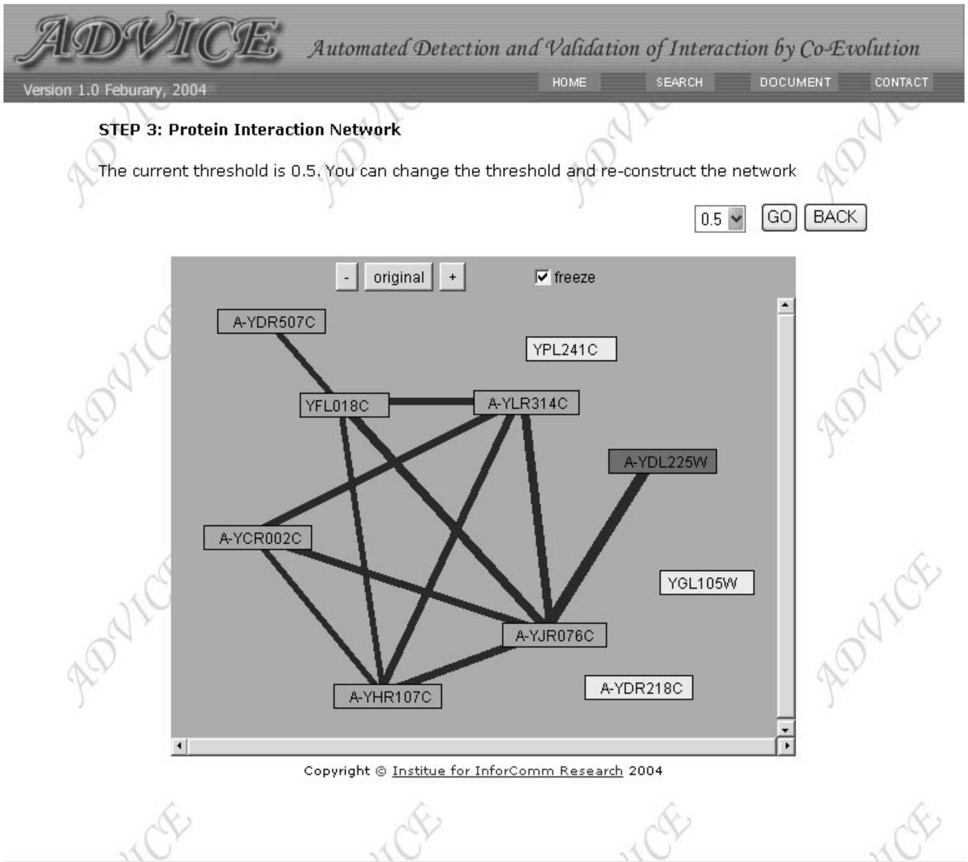

ADVICE
Automated Detection and Validation of Interaction by Co-Evolution
Version 1.0 Feburary, 2004
HOME
SEARCH
DOCUMENT
CONTACT
BACK
Next Step
Copyright © Institue for InforComm Research 2004
ADVICE

STEP 1: Homolog detection:

We detect homologs from Swissprot database. Our system finds that the input sequence pair both have homologs in 10 species listed as following, please select the species for the matrix construction, you can choose up to 20 species (extra ones will be ignored by ADVICE)

species	homologs of sequence A	homologs of sequence B
☑ Escherichia coli	P07682	P31216
☑ Haemophilus influenzae	P43729	P44681
☑ Buchnera aphidicola (subsp. Schizaphis graminum)	Q8K9Q9	Q8K9V2
☑ Buchnera aphidicola (subsp. Acyrthosiphon pisum) (Acyrthosiphon pisum	P57348	P57288
☑ Bacillus subtilis	P37949	P37518
☑ Mycoplasma genitalium	P47384	P47270
☑ Mycoplasma pneumoniae	P75498	P75088
☑ Schizosaccharomyces pombe (Fission yeast)	O13869	O14078
☑ Saccharomyces cerevisiae (Baker's yeast)	P46943	P38219
☑ Caenorhabditis elegans	P34617	P91917

Figure 1. Pairs of orthologous sequences identified in different species using protein sequences input by users (sequence A and sequence B). Users can select the desired set of orthologous pairs to compute the similarity in the proteins' evolutionary histories.



STEP 3: Protein Interaction Network

The current threshold is 0.5. You can change the threshold and re-construct the network

Copyright © Institue for InforComm Research 2004

Figure 2. Detected co-evolved proteins visualized as a protein network. The edge thickness increases linearly with the computed correlation coefficient. Users can specify the coefficient cut-off value for the construction of the network.

Measuring similarities in evolutionary distances

The correlation coefficient (r) between two distance matrices is computed using Pearson's correlation coefficient equation:

$$r = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\bar{X}\right)\left(Y_{ij}-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(Y_{ij}-\bar{Y}\right)^2}},$$

where X and Y are two $N \times N$ distance matrices and N is equal to the number of orthologous sequence pairs retrieved (here, N is equal to the number of species as we allow only one sequence pair per species). X_{ij} refer to the pairwise distance between sequences x_i and x_j from species S_i and S_j, respectively. Similarly, Y_{ij} refers to the pairwise distance between sequences y_i and y_j from species S_i and S_j respectively. This statistical approach is the same method used by Goh *et al*. (9) to quantify the correlation between two distance matrices for measuring the similarities in proteins' evolutionary histories.

OUTPUT

ADVICE outputs the computed correlation coefficient (r), ranging from -1 to 1, on the web page for each pair of input sequences. The distance matrices used to compute the correlation coefficient are also presented on the web page. In batch processing, the output data will be sent to an email address provided by the user.

When more than one pair of proteins is provided as input, in addition to computing the correlation coefficient score between proteins' evolutionary histories, ADVICE also provides the facility to visualize the computed co-evolved associations between proteins as a non-directional weighted graphical network (Figure 2). Each node on the network corresponds to an input protein. The edge thickness between proteins corresponds to the computed correlation coefficient. The thickness of the edges increases linearly with coefficient score. In this way, users can identify highly co-evolved protein pairs easily. Users can also filter out edges by specifying a correlation coefficient threshold. All these facilities provide users with a global view of the detected associations between proteins.

INTERPRETATION

The computed correlation coefficient ranges from -1 to 1. A correlation coefficient of 1 corresponds to 100% correlation or similarities in the input proteins' evolutionary histories, while a score of -1 implies 100% anti-correlation. A coefficient of 0 will mean that there is no correlation. Goh *et al*. and Pazos *et al*. in their separate works have determined a lower coefficient limit of 0.8 to be a good indicator of interacting

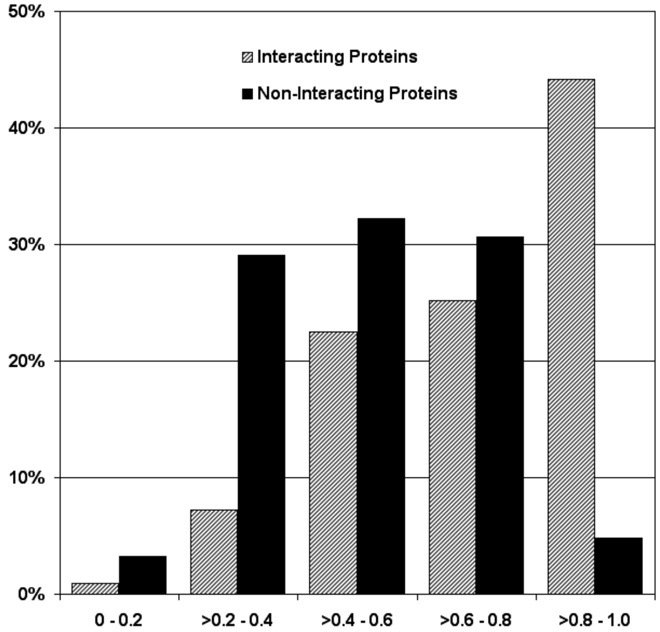


Figure 3. Distribution of computed correlation coefficients between high-confidence interacting proteins and putative non-interacting protein pairs in yeast.

proteins; users can therefore use this value to identify potential interacting proteins. To assess the sensitivity of this particular threshold, we have also computed the correlation coefficient for 111 yeast protein–protein interactions (20) (supplementary data) which represent a confident set of true interactions as they have been detected by multiple methods. Figure 3 shows the distribution of computed coefficients. The result indicates that the user can detect ~45% of these high-confident interactions using a cut-off value of 0.8. In addition, we also tested ADVICE on a set of 63 putative non-interacting yeast protein pairs where one protein is localized in the nuclear membrane while the other is localized in the mitochondrial inner membrane. Of these protein pairs, <5% were found to have correlation coefficients >0.8. For a suitable upper bound for detecting spurious interactions, we have observed that ~23% of these false interactions have coefficients <0.3. For the high-confidence interactions, only 2.7% of them have correlation coefficients <0.3. Thus, for the purpose of validating experimental interactions, users can adopt a cut-off value of ~0.3 to detect potential spurious interactions. The use of a higher cut-off will need to be treated prudently or done in conjunction with other validation methods such as gene expressions for best result.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Suisheng Tang and Han Hao for proofreading the manuscript.

REFERENCES

1. Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
2. Moyle,W.R., Campbell,R.K., Myers,R.V., Bernard,M.P., Han,Y. and Wang,X. (1994) Co-evolution of ligand–receptor pairs. *Nature*, **368**, 251–255.
3. van Kesteren,R.E., Tensen,C.P., Smit,A.B., van Minnen,J., Kolakowski,L.F., Meyerhof,W., Richter,D., van Heerikhuizen,H., Vreugdenhil,E. and Geraerts,W.P. (1996) Co-evolution of ligand–receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. *J. Biol. Chem.*, **271**, 3619–3626.
4. Hughes,A.L. and Yeager,M. (1999) Coevolution of the mammalian chemokines and their receptors. *Immunogenetics*, **49**, 115–124.
5. Koretke,K.K., Lupas,A.N., Warren,P.V., Rosenberg,M. and Brown,J.R. (2000) Evolution of two-component signal transduction. *Mol. Biol. Evol.*, **17**, 1956–1970.
6. Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **271**, 511–523.
7. Jespers,L., Lijnen,H.R., Vanwetswinkel,S., Van Hoef,B., Brepoels,K., Collen,D. and De Maeyer,M. (1999) Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *J. Mol. Biol.*, **290**, 471–479.
8. Jucovic,M. and Hartley,R.W. (1996) Protein–protein interaction: a genetic selection for compensating mutations at the barnase–barstar interface. *Proc. Natl Acad. Sci. USA*, **93**, 2343–2347.
9. Goh,C.S., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.
10. Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.
11. Ramani,A.K. and Marcotte,E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.
12. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
13. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan *C.elegans*. *Science*, **303**, 540–543.
14. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
15. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
16. Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.
17. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
18. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.
20. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.