BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S71–S77

Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues

Tal Pupko[1,†], Rachel E. Bell[2,†], Itay Mayrose[2], Fabian Glaser[2] and Nir Ben-Tal[2,]*

[1]The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan and [2]Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Motivation: A number of proteins of known three-dimensional (3D) structure exist, with yet unknown function. In light of the recent progress in structure determination methodology, this number is likely to increase rapidly. A novel method is presented here: 'Rate4Site', which maps the rate of evolution among homologous proteins onto the molecular surface of one of the homologues whose 3D-structure is known. Functionally important regions often correspond to surface patches of slowly evolving residues.

Results: Rate4Site estimates the rate of evolution of amino acid sites using the maximum likelihood (ML) principle. The ML estimate of the rates considers the topology and branch lengths of the phylogenetic tree, as well as the underlying stochastic process. To demonstrate its potency, we study the Src SH2 domain. Like previously established methods, Rate4Site detected the SH2 peptide-binding groove. Interestingly, it also detected inter-domain interactions between the SH2 domain and the rest of the Src protein that other methods failed to detect.

Availability: Rate4Site can be downloaded at: http://ashtoret.tau.ac.il/. It is implemented as a web server at: bioinfo.tau.ac.il/ConSurf

Contact: tal@ism.ac.jp; rebell@ashtoret.tau.ac.il; fabian@ashtoret.tau.ac.il; bental@ashtoret.tau.ac.il

Supplementary Information: Multiple sequence alignment of homologous SH2 domains, the corresponding phylogenetic tree and additional examples are available at http://ashtoret.tau.ac.il/~rebell

Keywords: rate variation among sites; evolutionary conservation; protein evolution; maximum likelihood; SH2 domains.

*To whom correspondence should be addressed.
† These authors contributed equally.

INTRODUCTION

The rate of evolution is not constant among amino-acid sites; some positions are highly conserved while others vary substantially (Uzzell and Corbin, 1971; Yang, 1993). These rate variations correspond to different levels of the purifying selection acting on these sites. This purifying selection can be either the result of geometrical constraints on the folding of the protein into its three-dimensional (3D) structure, constraints on amino-acids involved in enzymatic activity or in ligand binding, or alternatively at sites that take part in protein-protein interactions (Branden and Tooze, 1999).

In this paper we are concerned with methods of identifying functionally important regions in proteins with known 3D-structures and with many close sequence homologues. We developed a rigorous statistical method for infering the level of amino acid conservation at each amino acid site, taking into account the phylogenetic relations between the sequences as well as the stochastic process underlying their evolution.

The underlying assumption in this and related approaches, such as the 'Evolutionary Trace' (ET) method (Lichtarge *et al.*, 1996) and 'ConSurf' (Armon *et al.*, 2001), which was based on the maximum parsimony approach (MP-ConSurf), is that the slow evolution rate of surface residues is the result of constraints due to binding, e.g., to other proteins, ligands or DNA molecules. The ET method is based on the following steps: (1) building a phylogenetic tree from a multiple sequence alignment using the UPGMA method; (2) using this tree to cluster closely related sequences; (3) finding a 'consensus' sequence for each cluster, and each position; (4) comparing

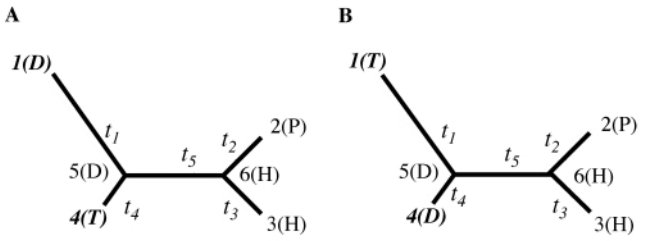

Fig. 1. A shortcoming of the maximum parsimony (MP) method. While the conservation score in case A should be higher than in case B, MP assigns the same conservation score to both cases. The differences between the two cases are marked in bold italics characters; see text for details. The tree is unrooted with 4 taxa. All nodes are labeled: Leaves (1–4) and internal nodes (5–6). t_i are the branch lengths. Capital letters in parentheses are one-letter abbreviations for amino acids.

the consensus sequences and assigning each position with a status of either 'variable' or 'conserved'; (5) mapping the status of each site onto the 3D structure of the protein. Using this method, a considerable amount of the original data is lost during the computation of the consensus sequences for each group. Furthermore, only the variation among clusters is taken into account. In addition, the conservation score is binary in nature.

A significant improvement to this method was introduced in later versions of the ET method (e.g., Landgraf *et al.*, 1999, 2001) and in MP-ConSurf (Armon *et al.*, 2001), when weighting schemes were devised to better quantify amino-acid replacements. In the MP-ConSurf algorithm, the first step is to construct a maximum parsimony tree from the multiple sequence alignment. The MP approach strives to reconstruct a tree that minimizes the number of replacements. The MP method, also reconstructs the ancestral sequences (the characters in the internal nodes of the phylogenetic tree; e.g., Figure 1). Thus, amino-acid replacements are mapped onto the tree. The conservation score is defined as the total number of replacements, weighted by the physiochemical distance between each pair of amino acids.

This method is better than ET in several respects: The MP trees are more reliable than the UPGMA trees (Graur and Li, 2000). Furthermore, by weighting the replacements according to the physicochemical distances, and averaging over all equally parsimonious reconstructions, the algorithm takes into account the fact that amino acids differ in frequency as well as the uncertainty of the reconstructed ancestral sequences.

Nevertheless, this approach has several shortcomings that are a direct outcome of the use of the MP criterion. One such fault is demonstrated in Figure 1. MP-ConSurf assigns conservation scores without taking branch lengths into account. Thus, identical scores will be given in the two similar but not identical cases described in Figures 1A and 1B. In reality, conservation scores for the two cases should differ because in the first example (Figure 1A) the replacement from D to T is mapped to a short branch (t_4), while in the second (Figure 1B), it is mapped to a long branch (t_1). Long branches correspond to either a high rate of evolution, or a long evolutionary time. Hence, replacements are more likely to occur in long branches. The MP method does not take this factor into account. An additional problem with the MP approach is the use of equally parsimonious reconstructions, for example, assigning T, D or H in node 5 of Figure 1, requires three replacements. In MP-ConSurf, the conservation score is obtained by averaging over these three possibilities. However, it is expected that T, D or H, should not necessarily be attributed with the same probability. A more robust approach would be to average over all 20 possible reconstructions in this node, weighted by their probabilities, or preferably, weighted by the average of all 20^2 possible reconstructions of the entire tree. (The tree includes two internal nodes, hence 20 is raised to the power of 2.) We present here 'Rate4Site': a novel algorithm that overcomes these alleged problems using the Maximum Likelihood (ML) method for phylogeny (Felsenstein, 1981). The ML approach assumes an underlying stochastic process in describing sequence evolution. Based on this process, amino-acid replacement probabilities are computed for each branch in the phylogenetic tree.

SYSTEM AND METHOD

Like the ET and the MP algorithm of ConSurf, the input of our algorithm (Rate4Site) is a multiple sequence alignment (MSA); if the input sequences are not aligned, an MSA is automatically generated using CLUSTAL W with default parameters (Thompson *et al.*, 1994). Another input is an amino-acid replacement model. For nuclear genes, the most widely used model is the JTT model (Jones *et al.*, 1992). Based on this model, one can easily compute the probability that an amino acid i will be replaced by amino-acid j along a branch of length t. We denote this probability by $P_{ij}(t)$. The next step in our algorithm is to reconstruct a phylogenetic tree. When the number of sequences is less than 20, the ML method can be used to find the most likely tree (e.g., Friedman *et al.*, 2002). However, when the number of sequences is larger, one must resort to a heuristic approach, such as the neighbor-joining (NJ) algorithm (Saitou and Nei, 1987). Both methods are regarded superior to the MP approach for constructing phylogenetic trees (Felsenstein, 1996). We note that in many cases the phylogenetic tree is known in advance. In such cases, the program can obtain an input tree. Once the phylogenetic tree is constructed, a

maximum likelihood rate is calculated for each position, treating gaps as missing data. Computation of the roles is the heart of our new algorithm and is described in the Section Algorithm below. Once a score is assigned to each of the positions, the conservation grades are mapped onto the 3D structure, and can be visualized using GRASP (Nicholls *et al.*, 1991) or equivalent molecular graphics programs.

To demonstrate the algorithm, we studied SH2 and SH3 domains. These domains are common mediators of inter-protein interactions and are involved in intracellular signaling. (The SH2 results are presented below and the SH3 results can be found at http://ashtoret.tau.ac.il/~rebell.)

Input sequences and trees

An excessive MSA of 233 homologous SH2 domains generated by a method that combines sequence and structure alignment was obtained from the Honig group (Al-Lazikani *et al.*, 2001); it includes remote homologues with sequence identity as low as ~15%. A limited MSA of 34 Src-like SH2 domains (sequence identity of 60% or more) was obtained using CLUSTAL W. Phylogenetic trees consistent with each of the MSAs were computed using the NJ algorithm. The MSAs and trees are available at http://ashtoret.tau.ac.il/~rebell.

The assumed stochastic process

In this study, probabilistic models based on amino acid sequences were used. The replacement probabilities among amino acids were calculated using the JTT matrix (Jones *et al.*, 1992). We assume that different sites evolve independently. Thus, we compute rates one site at a time. Hereafter we address the rate inference of a single site.

ALGORITHM

Among site rate variation

Consider two sequences of M positions. Suppose that the average distance between the two sequences is l. This means that we expect $l \times M$ replacements altogether. How many replacements should we expect at each site? We assume that the number of replacements is $l \times r[j]$, where $r = r[j]$ is the rate parameter for this position. Our method finds the maximum likelihood estimate of this rate parameter. The higher the variability of the site, the bigger r is. Since the mean rate over all sites is l, the mean r must be equal to one. A similar approach was also used by Yang (1993) for modeling sequence evolution when the rate varies among sites. However, Yang averaged rates, whilst here we are interested in estimating a rate parameter for each site.

Estimating the rate for a specific site

Assume that we are given the phylogenetic tree and the stochastic process. Suppose also that the internal nodes are labeled as in Figure 1A. Let the character assignments in the internal nodes be $\{D, H\}$. The probability of this assignment given a rate parameter r is:

$$P(\{\mathrm{D,H}\}, data|r) = \pi_D \times P_{D,D}(r \cdot t_1) \times P_{H,P}(r \cdot t_2) \times P_{H,H}(r \cdot t_3) \times P_{D,T}(r \cdot t_4) \times P_{D,H}(r \cdot t_5) \quad (1)$$

where π_D is the frequency of aspartic acid (D), and $P_{X,Y}(r \cdot t)$ is the probability that amino acid X will be replaced by amino acid Y along a branch of length t given that the rate of the site is r. Because of the reversibility of our Markov process, the tree could have its root anywhere (Felsenstein, 1981); in Eq (1) we arbitrarily chose internal node 5.

In practice, we do not have the character state assignment in the internal nodes. Hence, we sum over all possible assignments X, Y of the internal nodes, and obtain:

$$P(data|r) = \sum_{\mathrm{X,Y}\in\{\mathrm{Amino-acids}\}} \pi_x \times P_{X,D}(r \cdot t_1) \times P_{Y,P}(r \cdot t_2) \times P_{Y,H}(r \cdot t_3) \times P_{X,T}(r \cdot t_4) \times P_{X,Y}(r \cdot t_5) \quad (2)$$

The only unknown variable here is r. Given the rate r, the expression in (2) is calculated by a standard dynamic programming algorithm (e.g., Felsenstein, 1981). Our estimate of r is the maximum of (2) over all possible rates. Thus, we have a method to evaluate the rate at each position. We repeat this calculation for all positions in the multiple sequence alignment. We then normalize these rates so that the average is zero and the standard deviation is one.

Computer program

A C++ implementation of the maximum likelihood method described above with a JAVA-based Graphic User Interface is available at: http://ashtoret.tau.ac.il/~rebell

IMPLEMENTATION

The Rate4Site method for calculating amino acid conservation grades described above is demonstrated and compared to the MP algorithm of ConSurf (Armon *et al.*, 2001) in studies of the SH2 domain. This domain mediates protein-protein interactions in cellular signaling cascades, and is found in many proteins, including the Src family. SH2 domains essentially bind polypeptide segments containing a phosphotyrosine (Gonfloni *et al.*, 1997; Superti-Furga *et al.*, 1993).

The SH2 domain contains two main functional regions, shown in Figures 2 and 3. The first is the phosphopeptide

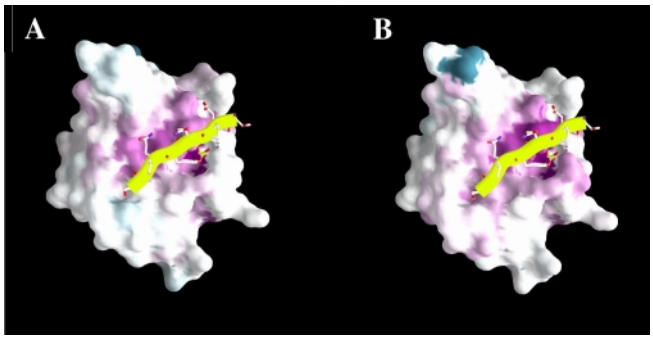

Fig. 2. The peptide-binding groove of the SH2 domain. The structure of the SH2 domain in complex with the C-tail of the tyrosine kinase domain (PDB entry 1fmk, Xu *et al.*, 1997) and MSA of 233 homologues were used. The conservation pattern obtained using MP-ConSurf (**A**) and Rate4Site (**B**) is color-coded onto the molecular surface of the domain: dark violet corresponds to maximal conservation, white corresponds to average conservation level and dark turquoise to maximal variability. The peptide is shown as a balls-and-sticks model with a yellow tube tracing its backbone. The picture was drawn using GRASP (Nicholls *et al.*, 1991). The domain boundaries were set to Trp148 to Pro246 (Al-Lazikani *et al.*, 2001).

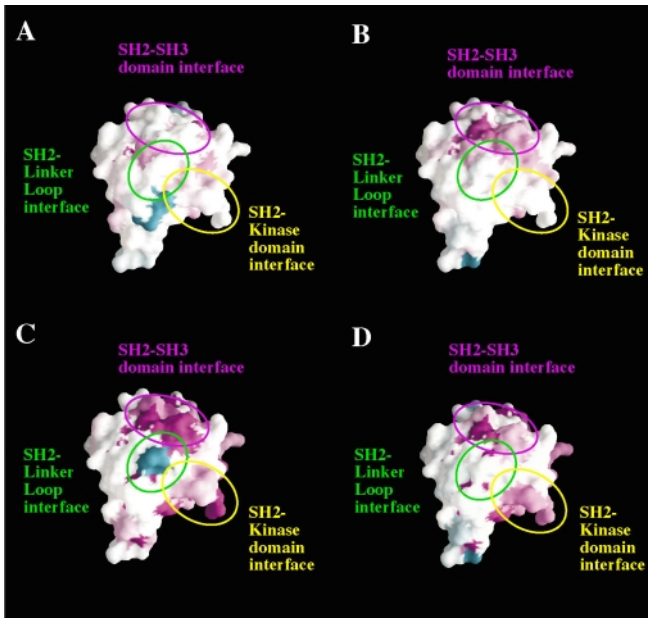

Fig. 3. SH2 domain within the context of intact Src. The interfaces with the SH3 and kinase domains are circled in pink and yellow, respectively, and the linker loop interaction site is circled in green. (**A-B**) The conservation pattern obtained using the 233 homologous sequences and the MP-ConSurf and Rate4Site methods, respectively. (**C-D**) The conservation pattern obtained using only 34 close SH2 homologues from the Src family and the MP-ConSurf and Rate4Site methods, respectively. The picture was drawn using GRASP and conservation is color-coded as in Figure 2. The domain boundaries were set to Trp148 to Pro246.

binding-groove, which is detected as highly conserved using both methods (Figures 2A and 2B). This region represents the major biological role of SH2. It was also detected using the Evolutionary Trace method, mentioned above, and can easily be traced by visual inspection of virtually any MSA of SH2 domains. Encouragingly, when using Rate4Site, the peptide position makes a better match with the conservation pattern obtained than when using the MP-ConSurf or Evolutionary Trace methods. (Supplementary information.)

The second significant SH2 surface is that of the interfaces between the SH2 domain and the rest of Src. The SH2 domain makes contacts with the SH3 domain, the kinase domain and the SH2-kinase linker loop (Figure 3). Here the two methods show different results, where Rate4Site proves more powerful than MP-ConSurf in detecting the biologically important patches. In analysis of all 233 SH2 homologues, Rate4Site displays a patch of conserved residues, which overlaps nicely with the SH2-SH3 interface (Figure 3B, circled in pink). In contrast, this interface appears to have an average conservation grade using the MP-ConSurf (Figure 3A). Within the context of human Src, this conserved patch contains three hydrophobic residues, Trp148, Tyr149 and Phe150, at the N-terminal of the SH2 domain. It has been suggested that the residues coupling the SH2 and SH3 domain, play a significant role in regulation of the catalytic domain (Young *et al.*, 2001). The SH2 and SH3 domains are connected by the SH2-SH3 connector, which adapts their orientations to one another and couples their dynamic fluctuations. The connector forms crucial hydrogen bonds with the N-terminus of the SH2 domain, thus enabling formation of a rigid linkage between the SH2 and SH3 domain. This rigid coupling of the SH2 and SH3 domains enables regulation of Src by locking the protein into a closed and inactive state. Simulations have shown that flexibility in the SH2-SH3 connector mimics the effect of activation of the protein. Furthermore, mutation studies have shown that disrupting the water mediated interactions formed with the highly conserved Trp148 disrupts the regulatory interactions between the SH2 and SH3 domains (Young *et al.*, 2001).

The conservation pattern of 34 close SH2 homologues from the Src family, including Src, Yes, Hck, Lck, Lyn, Blk, Fgr, Fyn and Yrk, displays wide patches of highly conserved residues, due to limited divergence between the sequences. In the Rate4Site analysis (Figure 3D) the interfaces with the kinase and SH3 domains are highly conserved. These results are anticipated, as the Src family is a highly conserved signaling family of proteins,

in which coupling of the SH2 and SH3 domains and formation of intramolecular interactions between these and the kinase domain enables repression of Src's catalytic activity. This is in contrast with the results obtained in the analysis of all 233 SH2 sequences (Figure 2B), in which these interfaces were detected as averagely and moderately conserved, respectively, suggesting that SH2 domains may function differently depending on whether they comprise a Src-like or another protein.

Analysis of the same 34 Src homologues using MP-ConSurf (Figure 3C) also detects these interfaces well, although as opposed to Rate4Site, the high conservation appears to expand beyond the interface boundaries. Furthermore, the linker loop interface appears to be variable (Figure 3C; circled in green) while Rate4Site assigns it conservation scores slightly above the average (Figure 3D). This is probably due to the method of handling gapped positions (see Discussion).

DISCUSSION

We recently developed MP-ConSurf for calculating amino acid conservation scores, based on phylogenetic reconstruction of the evolutionary relations between sequence homologues using the parsimonious principle (Armon *et al.*, 2001). Comparison has shown that MP-ConSurf is more sensitive than previously suggested methods, e.g., the Evolutionary Trace method. We proposed here a novel method, Rate4Site, for calculating conservation grades by estimating the rate of evolution at each site using the maximum likelihood paradigm. The comparison displayed sensitivity even higher than that of MP-ConSurf. For example, it detected a patch of conserved residues at the SH2-SH3 inter-domain interface, even when using an MSA containing distant homologues with a sequence identity as low as ∼15%; this region appears to be averagely conserved using MP-ConSurf and the same MSA (Figures 3B versus 3A; pink circles).

The main conceptual advantage of Rate4Site over MP-ConSurf is that it takes branch lengths explicitly into account. Branch lengths correspond to the expected number of substitutions per site. If two close sequences have a small branch length, Rate4Site 'expects' fewer exchanges between them, i.e., higher conservation. The longer the branch length, the more divergence is 'expected' between the sequences. MP-ConSurf on the other hand, weighs the exchanges between different sequences equally, without taking into account their branch length. Phe150, an SH2 residue from the linker loop patch, provides an example. Examining its position in the MSA of the 233 sequences reveals that it is conserved within the closely related sequences. However, in more distant sequences it diverges mainly to His, and rarely to other residues; a few gaps appear in this position as well. In the Rate4Site analysis, it appears conserved, yet in MP-ConSurf it is variable. Such differences between the methods are not major, as generally residues involved in the most important biological functions such as binding sites or active sites residues will receive very high conservation scores in both methods.

Gaps are considered differently in Rate4Site and in MP-ConSurf. The scoring method used in MP-ConSurf, heavily penalizes positions exchanged with deletions. Thus, biologically variable regions such as loops, which contain many gapped positions, frequently appear as excessively variable. Also, the sequence homology within protein families is often very low at the N- and C-terminal segments. Methods used for detecting homologous sequences often overlook it and provide homologous sequences that are somewhat shorter than the target sequence. The result is that the N- and C-terminal segments in MSAs are overly populated with gaps and appear extremely variable in MP-ConSurf. In contrast, Rate4Site disregards such gapped positions and calculates the conservation scores based only on the available data for these positions. Such a liberal scoring scheme, in which the conservation grade at each position is calculated using a different number of taxa, may also not always supply truly reliable information. Taking gapped positions into account is problematic because it contradicts the assumption of independency among positions. In practice, by treating gapped positions as missing data, Rate4Site determines the conservation score for only a subset of the sequences.

One of the main problems in calculating conservation grades is distinguishing between amino acids that are conserved due to their functionality and those that appear to be conserved due to shortness of divergence time. This problem is pronounced when dealing with close homologues of limited diversity. Again, Rate4Site appears to be superior to MP-ConSurf. For example, the Rate4Site conservation signal (Figure 3D) is less expanded than the MP-ConSurf signal (Figure 3C) and appears to correlate more closely with the boundaries of the inter-domain interfaces (supplementary information). Thus, Rate4Site appears to differentiate better between amino acids that are conserved due to their functionality and those that appear to be conserved due to shortness of divergence time. This is most probably due to the incorporation of the branch lengths in the calculations.

Multi-domain proteins are also challenging, since the conservation signal from one domain often overrides secondary signals from others. For example, in Src, the dominant signal from the tyrosine kinase domain prevents the detection of the peptide binding-grooves at the SH2 and SH3 domains (data not shown). Thus, it is recommended to analyse the domains composing such proteins one at a time.

The choice of the domain boundaries is important as changes in the MSA can produce conflicting conservation

patterns, especially when the termini residues are of interest. The domain boundaries are difficult to define and in fact even their classification is questionable. For example, the SCOP database (Murzin *et al.*, 1995) defines the SH2 domain from residue Glu146 to Ser248. Xu *et al.* (1997, 1999) who deciphered the X-ray crystal structures of Src (PDB entries 1fmk and 2src), defined the SH2 domain from Ile143 to Cys245, whereas the definition in the CATH database (Orengo *et al.*, 1997; Pearl *et al.*, 2000) is from Glu146 to Pro246, and in the Pfam database (Bateman *et al.*, 2000) from residue 148 to 230. Analysis using the different domain boundaries yielded contradictory conservation patterns for the inter-domain interfaces in Src. These secondary signals appeared in some cases conserved, and in others variable, depending on where the sequence began (data not shown). Here we followed a recent definition of the SH2 domain boundaries (from Trp148 to Pro246), obtained from a combination of multiple sequence and structural alignment (Al-Lazikani *et al.*, 2001).

CONCLUSION

Recently, a radically different variant of the ET approach was developed, in which contributions from residues at spatial proximity to each other were explicitly taken into account (Landgraf *et al.*, 2001). In this approach, the rate is not inferred for a single site, but rather to a spatial 'environment'. In the future, we will adapt Rate4Site to calculate the collective evolutionary rates of such spatially proximal sites. Further improvements, such as implementation of a superior tree building method based on the maximum likelihood process, and different methods of handling the gapped positions (Hein, 2001; Holmes and Bruno, 2001; Mitchison, 1999) will be examined.

To conclude, Rate4Site is a very accurate and sensitive method for detecting functionally important regions in proteins of known 3D-structure. It is likely to be useful in the context of structural genomic effort, in which the 3D-structures of many proteins with yet unknown function has been determined.

ACKNOWLEDGMENTS

We thank Don Graur for stimulating discussions and the Bioinformatics unit of Tel Aviv University for providing us with help and infrastructure. This work was supported by a grant from the Israel Cancer Association. T.P. is supported by a JSPS fellowship.

REFERENCES

Al-Lazikani,B., Sheinerman,F.B. and Honig,B. (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.

Armon,A., Graur,D. and Ben-Tal,N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface-mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.

Bateman,A., Birney,E., Durbin,R., Eddy,R.S., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Branden,C. and Tooze,J. (1999) *Introduction to Protein Structure*, 2nd edition, Garland Publishing, New York.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, **266**, 418–427.

Friedman,N., Ninio,M., Pe'er,I. and Pupko,T. (2002) A Structural EM Algorithm for Phylogenetic Inference. *J. Comput. Biol.*, (in press).

Gonfloni,S., Willliams,J.C., Hattula,K., Weijland,A., Wierenga, R.K. and Superti-Furga,G. (1997) The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src. *EMBO J.*, **16**, 7261–7271.

Graur,D. and Li,W.H. (2000) *Fundamentals of Molecular Evolution*, 2nd edition, Sinauer Associates, Sunderland, MA.

Hein,J. (2001) An algorithm for statistical alignment of sequences related by a binary tree. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 179–190.

Holmes,I. and Bruno,W.J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803–820.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Landgraf,R., Fischer,D. and Eisenberg,D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.*, **12**, 943–951.

Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Mitchison,G.J. (1999) A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.*, **49**, 11–22.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nicholls,A., Sharp,K.A. and Honig,B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.*, **11**, 281–296.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pearl,F.M., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic

sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Superti-Furga,G., Fumagalli,S., Koegl,M., Courtneidge,S.A. and Draetta,G. (1993) Csk inhibition of c-Src activity requires both the SH2 and SH3 domains of Src. *EMBO J.*, **12**, 2625–2634.

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.

Uzzell,T. and Corbin,K.W. (1971) Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096.

Xu,W., Doshi,A., Lei,M., Eck,M.J. and Harrison,S.C. (1999) Crystal structures of c-Src reveal features of its auto-inhibitory mechanism. *Mol. Cell*, **3**, 629–638.

Xu,W., Harrison,S.C. and Eck,M.J. (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, **385**, 595–602.

Yang,Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.

Young,M.A., Gonfloni,S., Superti-Furga,G., Roux,B. and Kuriyan,J. (2001) Dynamic coupling between the SH2 and SH3 Domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell*, **105**, 115–126.