W110–W115 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl203

TreeDet: a web server to explore sequence space

Angel Carro, Michael Tress, David de Juan, Florencio Pazos, Pedro Lopez-Romero[1], Antonio del Sol[2], Alfonso Valencia and Ana M. Rojas*

Protein Design Group, CNB-CSIC, C/Darwin n.3 28049, Madrid, Spain, [1]Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain and [2]Bioinformatics Research Project, Fujirebio Inc., Tokyo, Japan

Received February 10, 2006; Revised and Accepted March 22, 2006

ABSTRACT

The TreeDet (Tree Determinant) Server is the first release of a system designed to integrate results from methods that predict functional sites in protein families. These methods take into account the relation between sequence conservation and evolutionary importance. TreeDet fully analyses the space of protein sequences in either user-uploaded or automatically generated multiple sequence alignments. The methods implemented in the server represent three main classes of methods for the detection of family-dependent conserved positions, a tree-based method, a correlation based method and a method that employs a principal component analyses coupled to a cluster algorithm. An additional method is provided to highlight the reliability of the position in the alignments. The server is available at http://www.pdg.cnb.uam.es/servers/treedet.

INTRODUCTION

Sequence analysis is the first step in predicting functionally important residues in a given protein family. Although conserved regions in optimal multiple alignments are important, additional residues showing alternative family-dependent conservation patterns can also be detected and might reveal different features related to the function.

One particular type of family-dependent conservation aims to detect residues showing conservation trends within subfamilies but differing between subfamilies, the so-called 'tree determinant' positions. These methods have been extensively tested in various studies using non-redundant sets (1,2) and a number of groups (3–6), including our own (1,7), have used these family-dependent conservation patterns to predict specific binding sites and/or substrate/co-factor binding sites. Although these methods aim to find trends in variability, the approaches are quite different. For instance, the method developed by Kalinina *et al*. (8) looks for conservation patterns within orthologues and variation in paralagous sequences at the same positions in a multiple sequence alignment. Other methods rely on the existence of 3D structures to map the residues (3), and other methods explore the variability using correlated mutations and entropy in large protein families (9).

The specific methods implemented in the TreeDet server include the fully-automated sequence space method (FASS) (7), level entropy method (S-method) (1) and the mutational behaviour method (MB) (1). In fact, the three different implementations were developed as distinct concepts to deal with different aspects of the same problem. A detailed explanation of the methods has been published previously (1), and this description includes a comparison of the results in diverse situations.

These methods have been tested independently for different biological systems and the predictions obtained have been subjected to further experimental verification such as the analyses of the *ras* superfamily (10), the ras and ral proteins (11), and the chemokine receptor dimerization residues (12,13). For a comprehensive compilation of applications see Lopez-Romero *et al*. (14). Other groups have also published collaborations based on the application of the same basic ideas (15). In each case the methods provided useful biological insights regarding the function of the proteins.

We have integrated the three tree determinant prediction methods into a server because of the importance of these methods for the future development of this type of project and the need to facilitate access to the various applications.

Given that these methods are particularly useful in the prediction of functionally important residues in families of sequences, the target users of the system will be biologists interested in exploring the potential localization of functional sites.

METHODS

Three tree determinant methods are currently available in the TreeDet server. They are automatic and search for key regions

*To whom correspondence should be addressed at Spanish National Cancer Center (CNIO), Structural Bioinformatics Group, Centro Nacional de Investigaciones Oncológicas (CNIO), C/Melchor Fernandez Almagro 3, 28029 Madrid, Spain. Tel: +34 91 585 46 69; Fax: +34 91 585 45 06; Email: arojas@cnb.uam.es

of functional specificity in protein families. At the same time they can detect key residues responsible for the subfamily structure. SQUARE (16), the fourth method implemented in the server, is an alignment evaluation tool. The specific implementation of SQUARE in TreeDet adds measurement of the reliability of the alignment provided for each position based on the strength of the conservation and the type of conserved residues (1). This is a very important issue, as methods to predict functional important sites are very sensitive to the alignment quality.

The level entropy method

The S-method searches for different levels of splitting of a protein family into subfamilies. Several cuts of the family phylogenetic tree are analysed in order to evaluate the relative entropy (mutual information) for each division level. The mutual information expresses the distance between the probability distribution of tree determinants at a certain level and the product of probability distributions of conserved positions in each subfamily at that level. The method searches for the cut level with the greatest value of relative entropy (1).

The mutational behavior method

The MB-method searches for positions in the alignment whose mutational behaviour is similar to the variation pattern of the whole family. Therefore, these residues will be representative of the overall sequence distance distribution in different subfamilies. The full family and the individual positions in the alignment are represented by distance matrices, which are compared using rank correlation criteria.

The automated sequence space method

The version of FASS incorporated in the server is a new fully-automated implementation. In FASS each sequence is represented as a vector in a multi-dimensional space (the sequence space), and residue types at each position of the alignment as vectors in the reciprocal amino acid space. A principal component analysis renders a reduction of the dimensionality allowing the study of the sequence–sequence, residue–residue and sequence–residue relationships.

The original method required intensive human expert post-processing and manipulation of results, which is circumvented in the current implementation by first statistically defining the dimensionality N of the sequence space that accounts for the maximum data variability with the minimum number of components, and second by clustering the results in the N-dimensional space (14).

SQUARE

This section of the server produces a measure of per residue reliability for the alignments between the sequences in a pre-generated multiple alignment. In contrast to the stand-alone server (16) which bases its calculation of alignment reliability on sequences with known structure, SQUARE @TreeDet calculates the reliability of each pairwise alignment around profiles generated for the first sequence in the multiple alignment (the query sequence). Reliability scores for the residues aligned against each of the positions in the query sequence are calculated from the profile matrix generated by PSI-BLAST (17) and a smoothing function. The higher the score at each position in the alignment, the more likely the two sequences are correctly aligned at this position. Testing has shown that regions defined as reliably aligned by SQUARE are much more likely to be correctly aligned in the evolutionary sense. SQUARE also calculates the score for the optimally aligned residue at each residue position in the query sequence. The scores generated by SQUARE can provide an important insight into the quality of the alignment and the associated predictions, particularly when used in conjunction with the optimal scores.

The software implementing the individual methods is also available upon request.

QUERY PAGE

Input file

The TreeDet server is designed to work with sequence alignments of protein families. Protein sequence alignments in CLUSTAL, FASTA, PIR and MSF formats are accepted as valid input files.

The methods implemented in TreeDet are sensitive to alignment quality, therefore for optimal performance user-optimized alignments are strongly encouraged (TreeDet is not an aligning tool). Regardless, the server also accepts as input a single unaligned protein sequence. In this case, a BLAST search (cutoff E-value 10^{-3}) is conducted to retrieve homologous sequences. Then, the sequences are aligned using clustalW (18). To enhance the alignments, sequences with >90% identity are removed from the alignment. In addition, sequences <25% identical to the query sequence are also eliminated. Finally, sequences that align to <50% of the total length of the query sequence are removed.

The methods can be run using default parameters or alternative parameters can be chosen in the 'advanced run' interface. The parameters are explained in detail in the work of del Sol *et al*. (1) and in the 'more information' section of the server.

Each run can include any or all the methods. Each method has its own separate description.

The results

Once the job is completed, the user receives a URL by Email. The URL contains the results along with thorough explanations. As an example, we show here the multiple alignment from an analysis of protein homologues to the p21 ras oncogen. The results are all shown in one page and divided into two sections: a summary of selected parameters and the results section.

The results section consists of two parts: a table and an alignment (Figure 1). The table indicates the positions predicted as tree determinants by each method. If all three methods have been selected, the sequences shown will be ordered by the groupings obtained by the FASS method. If FASS has not been selected, the sequences will be ordered by the grouping of S-method.

When clicking on the predicted positions in the table, the alignment moves towards the selected position. The predictions from all methods can be mapped onto the alignment at the same time. For instance, in the alignment of ras homologues (Figure 1), position 37 was identified by two

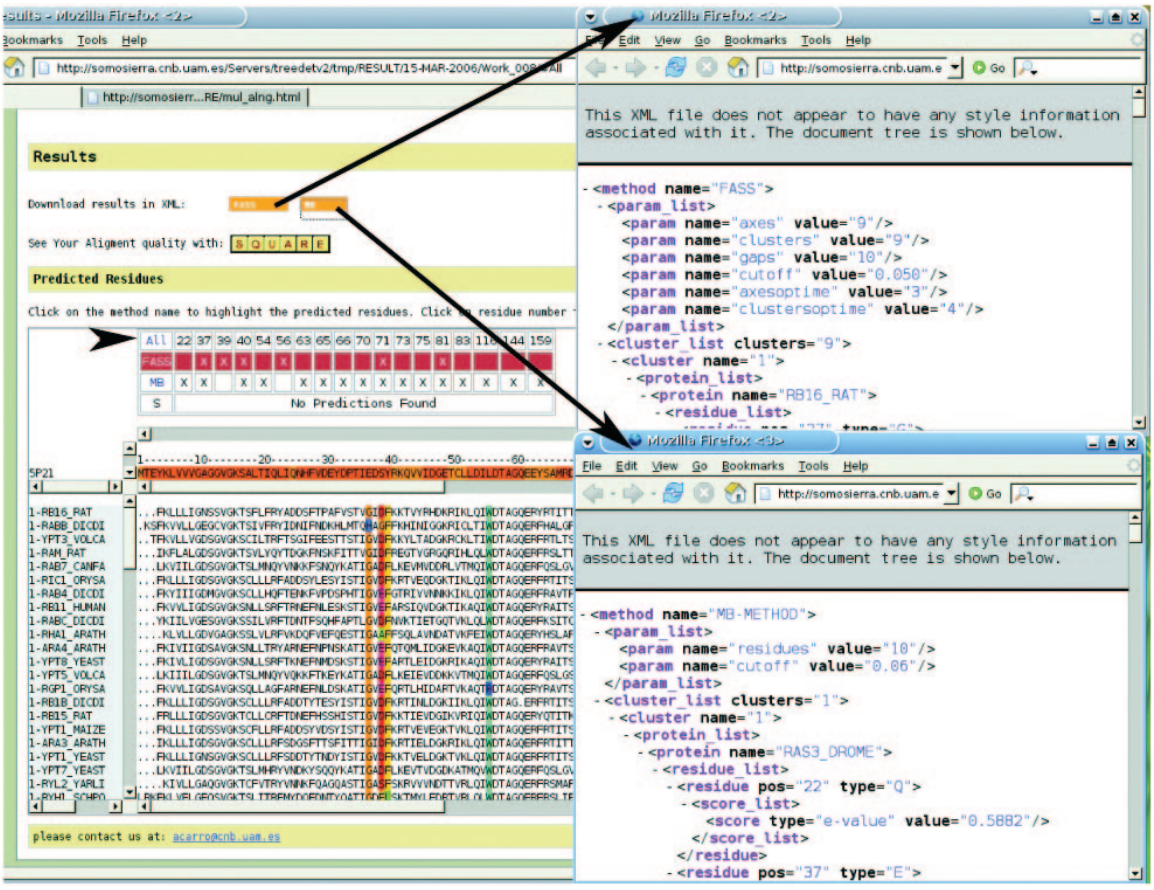

Figure 1. TreeDet results page. This section provides a table with predicted positions associated to an input alignment. By clicking on the numbered positions the alignment moves towards these positions. The sequences in the alignment are re-ordered according either FASS or S-method and the resulting sequence clusters are highlighted in various shades of cyan. The scroll bar above the alignment includes the query sequence. The residues are highlighted by the optimal score from SQUARE, and indication of the conservation of each position within the broader sequence family. The predicted positions are highlighted according to Taylor's schema (19). For instance, the position 37 predicted by FASS and MB is critical for functional specificity (11). Additional XML files are provided to easy automated processing of the files. If a method does not provide results it is also indicated.

methods and has been shown to be critical for function specificity (11).

The predicted residues are highlighted following Taylor's colour schema (19). Additional results in XML formats are also provided (Figure 1) for easy data extraction and automatic post-processing of the results. The results remain available in the server for 7 days. Additional tools are also provided.

If SQUARE has been selected, the first sequence of the input alignment (the one submitted by the user) will appear above the main alignment coloured by the SQUARE optimal score (see below). This gives you an approximation of the conservation of each position within the broader sequence family. Furthermore, to analyze in detail SQUARE results, a link is provided (Figure 2).

The alignment is identical to the input alignment, the one submitted by the user, but in SQUARE all sequences are compared against the first sequence of the input alignment.

The SQUARE optimal link shows the conservation of each residue position within the sequence family and the highest scoring residue (the one that best fits describes the sequence family) in each position. The other links show the individual scores of each sequence against the first sequence in the alignment. Note that all columns that are gapped in the first sequence are removed in all SQUARE alignments.

The 'multiple alignment' option allows the user to visualize the reliability of the pairwise alignment of each sequence in multiple alignment format. The reliability scores in the multiple alignment format range from dark orange (the highest reliability), through various shades of yellow (the more intense, the higher the reliability score) to white (unreliable or evolutionarily distant). The residues in the query sequence are coloured under the same colour scheme, but use the score for the optimally aligned residue at each position (Figure 3). Here, the darker the shade, the more conserved the position within the sequence family.

Help pages and information

The homepage of TreeDet contains detailed information regarding how to use the server and example files are available. An additional information tab provides extensive information, including literature and related services that are available on the web. A performance test conducted on the server is also shown. This table reflects some statistics for hard and easy cases.

Figure 2. SQUARE results output. The main results page provides a link to SQUARE. The multiple alignment shows the reliability of the alignment and detailed graphics show the score distribution along each sequence of the alignment.

The server is quite fast when using the multiple alignment option: a 165 residue alignment of 97 sequences will take 2 min for MB and FASS, 4 min for the S-method, and 6 min for SQUARE.

Specifications: TreeDet's web interface is written in HTML/PHP. The programs for export and import data have been implemented in Perl. The individual methods are written in Fortran, C and Perl. Additional tools used by the methods are ClustalW and the BLAST suite programs from NCBI.

SCOPE AND LIMITATIONS

Aim

TreeDet has been designed to be used by experimental scientists to obtain reliable and interesting predictions of functionally important residues in protein alignments. Three different methods and a method to measure reliability are provided and integrated in a single output interface where the predictions from all three methods can be mapped onto the original alignment.

Features

- Availability of three different methods to predict protein functional sites via a user friendly interface.
- A choice of multiple alignment or single sequence inputs.
- Direct visualization of results over the input alignment. Methods are distinguished in a standard colour-based schema for clarity.
- XML files are also provided for easy data manipulation.
- An additional tool to evaluate alignment reliability is provided: SQUARE.
- Results are stored for a limited period of time in the server.

The server is not designed to create optimal multiple alignments, but if an unaligned protein sequence is provided, automatic alignments are generated. However, it bears repeating that this is not the best option as it can slow the process down considerably.

In addition, in common with many sequence analysis tools, the methods in TreeDet are sensitive to alignment errors and biases stemming from the over-representation of sequences. SQUARE has been included in the package of tools in the server because it provides a means of flagging up errors in the multiple alignments.

The simultaneous use of the three methods tends to improve the results as shown previously (1). Combining predictions reduces the final number of predictions but gives a more significant set of functionally important residues. The three methods are capable of capturing different subsets of functional residues.

In order to provide the user with biological examples of these analyses, we have included the results obtained from a

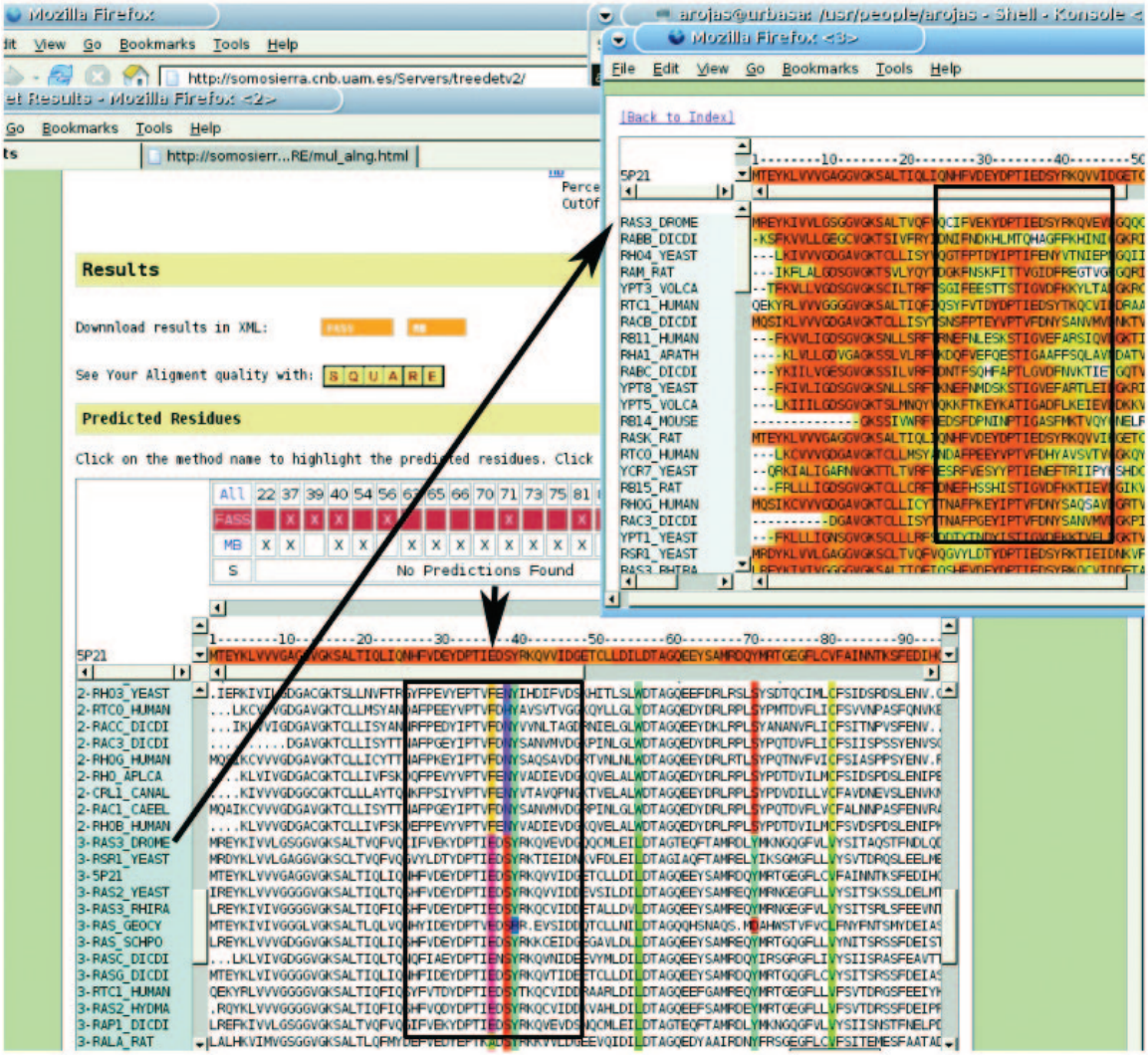

Figure 3. Overall reliability. This figure shows the reliability of the alignment from which the tree determinants for the sequence RAS3_DROME have been predicted. As seen from the darker colours in the figure, the alignment from which the position 37 (arrowhead) has been predicted by FASS and MB is highly reliable according to SQUARE. The regions in both alignments are marked with boxes.

multiple alignment of ras homologues proteins (Figures 1–3), both here and in the server.

Tree determinant methods have aided the analysis of protein families in the example of the ras homologues (11) and in other examples such as the chemokine receptor proteins (9,13). TreeDet now makes it possible for a larger community to use these methods in an integrated platform.

ACKNOWLEDGEMENTS

We would like to thank the remarkably valuable input and suggestions from Ildefonso Cases in terms of web usability and design. We also thank the contribution and suggestions of Eduardo Leon, the Department of Immunology and Oncology, and all the people who have tested the server. F.P. is the recipient of a 'Ramón y Cajal' Contract from the Spanish Ministry for Education and Science. This work was supported by grants: Biosapiens BIO2004-00875, GeneFun LSHG-CT-2004-503567, fundación BBVA, MCyT. Funding to pay the Open Access publication charges for this article was provided by BioSapiens EU projects (BIO2004-00875).

Conflict of interest statement. None declared.

REFERENCES

1. del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
2. Mihalek,I., Res,I. and Lichtarge,O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
3. 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.
4. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

5. 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.
6. Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.
7. Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nature Struct. Biol.*, **2**, 171–178.
8. Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, W424–W428.
9. Oliveira,L., Paiva,P.B., Paiva,A.C.M. and Vriend,G. (2003) Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. *Proteins*, **52**, 553–560.
10. Stenmark,H., Valencia,A., Martinez,O., Ullrich,O., Goud,B. and Zerial,M. (1994) Distinct structural elements of rab5 define its functional specificity. *EMBO J.*, **13**, 575–583.
11. Bauer,B., Mirey,G., Vetter,I.R., Garcia-Ranea,J.A., Valencia,A., Wittinghofer,A., Camonis,J.H. and Cool,R.H. (1999) Effector recognition by the small GTP-binding proteins Ras and Ral. *J. Biol. Chem.*, **274**, 17763–17770.
12. Hernanz-Falcon,P., Rodriguez-Frade,J.M., Serrano,A., Juan,D., del Sol,A., Soriano,S.F., Roncal,F., Gomez,L., Valencia,A., Martinez,A.C. *et al.* (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. *Nature Immunol.*, **5**, 216–223.
13. de Juan,D., Mellado,M., Rodriguez-Frade,J.M., Hernanz-Falcon,P., Serrano,A., Del Sol,A., Valencia,A., Martinez,A.C. and Rojas,A.M. (2005) A framework for computational and experimental methods: identifying dimerization residues in CCR chemokine receptors. *Bioinformatics*, **21**, ii13–ii18.
14. Lopez-Romero,P., Gomez,M., Gomez-Puertas,P. and Valencia,A. (2004) Prediction of functional sites in proteins by evolutionary methods. In Kamp,R., Calvete,J. and Choli,T. (eds), *Principles and Practice. Methods in Proteome and Protein Analyses.* Springer-Verlag, Berlin, Heidelberg, pp. 319–340.
15. Lichtarge,O., Yao,H., Kristensen,D.M., Madabushi,S. and Mihalek,I. (2003) Accurate and scalable identification of functional sites by evolutionary tracing. *J. Struct. Funct. Genomics*, **4**, 159–166.
16. Tress,M.L., Grana,O. and Valencia,A. (2004) SQUARE—determining reliable regions in sequence alignments. *Bioinformatics*, **20**, 974–975.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. 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.
19. Taylor,W.R. (1997) Residual colours: a proposal for aminochromography. *Protein Eng.*, **10**, 743–746.