W242–W248 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks559

SPEER-SERVER: a web server for prediction of protein specificity determining sites

Abhijit Chakraborty[1], Sapan Mandloi[1], Christopher J. Lanczycki[2], Anna R. Panchenko[2] and Saikat Chakrabarti[1,*]

[1]Structural Biology and Bioinformatics Division, Council for Scientific and Industrial Research (CSIR)—Indian Institute of Chemical Biology (IICB), Kolkata, West Bengal 700032, India and [2]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received January 27, 2012; Revised May 16, 2012; Accepted May 18, 2012

ABSTRACT

Sites that show specific conservation patterns within subsets of proteins in a protein family are likely to be involved in the development of functional specificity. These sites, generally termed specificity determining sites (SDS), might play a crucial role in binding to a specific substrate or proteins. Identification of SDS through experimental techniques is a slow, difficult and tedious job. Hence, it is very important to develop efficient computational methods that can more expediently identify SDS. Herein, we present Specificity prediction using amino acids' Properties, Entropy and Evolution Rate (SPEER)-SERVER, a web server that predicts SDS by analyzing quantitative measures of the conservation patterns of protein sites based on their physico-chemical properties and the heterogeneity of evolutionary changes between and within the protein subfamilies. This web server provides an improved representation of results, adds useful input and output options and integrates a wide range of analysis and data visualization tools when compared with the original standalone version of the SPEER algorithm. Extensive benchmarking finds that SPEER-SERVER exhibits sensitivity and precision performance that, on average, meets or exceeds that of other currently available methods. SPEER-SERVER is available at http://www.hpppi.iicb.res.in/ss/.

INTRODUCTION

Recognition of sequence variations that lead to functional diversification within a protein family is not a trivial task. Functional specificity signals must be separated from strong background signals resulting from the phylogenetic differences between the protein subfamilies (subgroups). Earlier methods to identify protein sites that are important to functional specificity used algorithms such as principal component analysis (1) and phylogenetic tree-based partitioning into protein subgroups (1–4). Entropy or mutual information (5–17)-based algorithms have also been widely used to distinguish the distribution of amino acids within and between protein subfamilies to determine specificity determining sites (SDS). Divergence at functional sites can be also inferred from the changes in the evolutionary rates (ERs). Some methods used evolutionary rate-based approaches (18–24) in which either 'Type I' (sites conserved for one subfamily and variable in another) or 'Type II' (sites where different types of amino acids are conserved across different subfamilies) SDS were analyzed to better understand the evolutionary basis of functional diversification. Other evolutionary conservation-based schemes (25–31) were also used to distinguish the specific distribution of amino acids within and across the subfamilies.

The Specificity prediction using amino acids Properties, Entropy and Evolutionary Rate (SPEER) algorithm (8), a method that combined contributions computed from (i) the conservation patterns of amino acid types as determined by their physico-chemical (PC) properties and (ii) the heterogeneity of evolutionary changes between and within the subfamilies, performed reasonably well in the identification of SDS (8,31,32). However, the standalone version of the SPEER program has limitations in terms of its input and output options, and its results could be difficult to interpret or incorporate into larger analysis pipelines. To address these issues, we present in this article a web server (SPEER-SERVER) based on the original SPEER program, which we have supplemented with several important and useful new features. Specifically, in the server we have improved how results are reported to the user, greatly augmented the input and

*To whom correspondence should be addressed. Tel: +91 33 2499 5809; Fax: +91 33 2473 5197; Email: saikat273@gmail.com; saikat@csiriicb.in

output options and added a collection of important post-analysis tools to examine the SDS predictions to tailor it to the broader bioinformatics research community. Finally, in this article, we also provide updated benchmarking results that compare nine of the latest available methods with SPEER-SERVER.

MATERIALS AND METHODS

SPEER algorithm

The required input to the original SPEER algorithm is a multiple sequence alignment (MSA) of N sequences that belong to a protein family, along with a partitioning of the sequences into M subfamilies. The algorithm that underlies SPEER-SERVER is based on our earlier program SPEER that uses a scoring scheme which combines three components to identify SDS (8). The first component calculates the weighted Euclidean distance between the vectors of physico-chemical properties of any two positions in the MSA. The average variability in a given subfamily column (relative to the background variability of that column within the whole family) is calculated by summing the Euclidian distances for all residue pairs within the subfamily and normalizing by the average Euclidian distance of all residue pairs in the column of the overall family. The second component of the SPEER score is the ER of the site as computed by the maximum-likelihood method implemented in the rate4site program (33). A low average ER value indicates a slowly evolving (i.e. more strongly conserved) site in subfamilies. The last component is a relative entropy term (or Kullback–Leibler divergence) used to quantitatively distinguish the amino acid-type distributions of two protein subfamilies.

SPEER-SERVER web interface

The web server is built using CGI scripts written in Perl. We have implemented various JavaScript routines that permit the visualization of SPEER-SERVER predictions and to support a wide range of analysis tools not previously available to the original SPEER software. Figure 1 provides a snapshot of the SPEER-SERVER input and output options we discuss in the following sections.

Input options

The server's submission interface is divided into two main parts: 'single submission mode' and 'batch submission mode'. Each submission mode has 'user-defined subgrouping' and 'automated subgrouping' options.

'Single submission mode' is intended to identify SDS in a single protein family. The SPEER program requires the number of subfamilies/subgroups in a protein family, along with the number of sequences assigned to each subfamily. When using the 'user-defined subgrouping' option, the user enters number of subgroups in the form. The input form also accepts a range of weights (from 0.0 to 1.0) for the three scoring components ('relative entropy', 'physico-chemical property distance' and 'ER') used by the SPEER-SERVER. Users should either upload a pre-aligned MSA (in FASTA format) or direct SPEER-SERVER to create a MSA for a specified set of protein sequences using either MAFFT (34) or PROBCONS (35).

'Automated subgrouping mode' must be used when the user has no prior information about the subfamilies present in the query protein family. We have integrated the SECATOR (36) and SCI-PHY (37) algorithms with SPEER-SERVER to automatically identify probable subgroups for which SDS will be identified.

'Batch submission mode' is intended for identifying SDS within a set of protein families in a single run (maximum five alignments).

Output options

When the server completes its analysis, the user is given multiple ways to view their results as described below.

Alignment display. Detailed information is provided for the identified SDS in the protein alignment, specifically the SPEER scores along with Z score and P values associated with SDS and calculated as described in Chakrabarti *et al.* (8). Alignment editing and subsequent adjustments can be performed using the embedded Jalview applet (38). The predicted type of each SDS is provided on the alignment results page where 'Type I' SDS are defined as those conserved for one subfamily and variable in another and 'Type II' sites are those where different types of amino acids are conserved across different subfamilies. Herein, we consider a site to be conserved for one subfamily if any amino acid type is represented $>75\%$ of the time. The sites that failed to satisfy the above criteria are marked as 'marginally conserved' or 'MC' (none of the amino acids within subfamilies is conserved in this site). For families with more than two subfamilies, sites were categorized into different types based on the category assigned to the majority of subfamily pairs.

Structure display. SPEER-SERVER predicted sites are projected onto the representative 3D structure uploaded by the user in PDB format (39) using JMol web applet (40). Structural display and analysis are provided when input MSA contains the sequence of that particular uploaded structure.

Structure versus sequence distance display. Additional structural analyses were performed. In particular, we calculated spatial distances between the predicted SDS and performed clustering of the predicted SDS, which are represented in the form of distance matrices and dendrograms, respectively. A sequence versus structural distance plot has been included to show how the localization of the predicted SDS correlates in terms of their spatial and sequence coordinates.

Coevolutionary display. Our previous study showed that SDS frequently coevolve with other sites (41). Therefore, SPEER-SERVER includes a feature that calculates the coevolutionary connections of predicted SDS. The results are presented in an interactive network display using Cytoscape web (42) so that the users can easily identify those specificity sites that coevolve. We use the MIp program (43) to calculate the coevolution between protein sites.

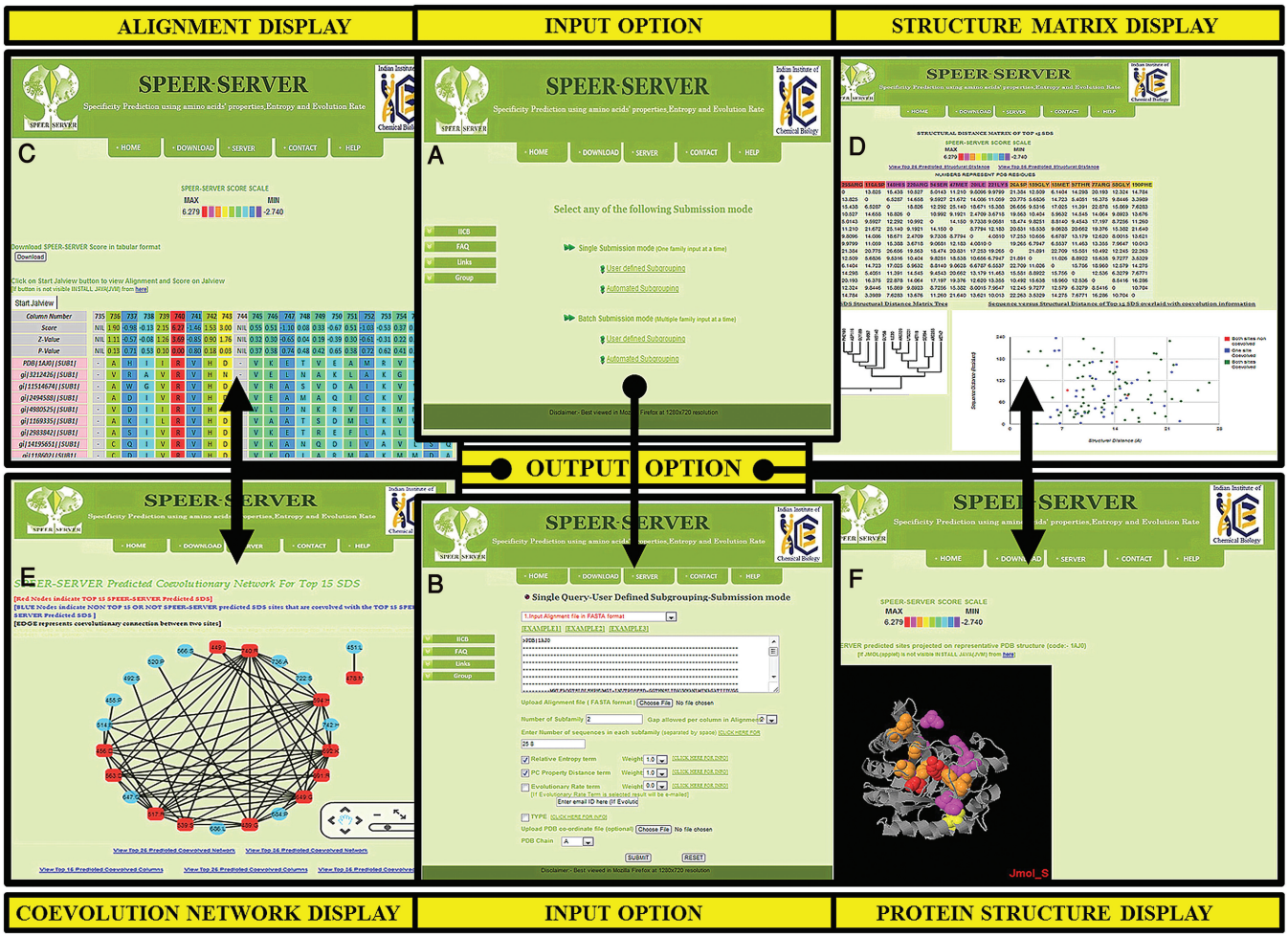

Figure 1. Snapshots of the SPEER-SERVER web interface, displaying its various input options (Panels **A** and **B**) and output analysis tools, such as alignment display (Panel **C**), coevolutionary network display (Panel **E**), structure display (Panel **F**) and structure distance matrix display (Panel **D**).

Benchmarking and validation

Benchmarking of SDS prediction using manually curated alignments with predetermined subgrouping. Standard protocols for evaluating the performance of various prediction methods were used in which SDS prediction methods were applied to 20 manually curated alignments studied previously (8,32). Each of these 20 families possesses human-generated annotations that identify their SDS (total 197; see Supplementary File 1 for details). For a given alignment, scores for each column were collected and the sensitivity and error rate were estimated based on the number of true positives (TPs; correctly predicted known SDS) and false positives (FPs; incorrectly predicted known SDS) found above each score cutoff. Score cutoff list was generated from the output of each SDS prediction method. The sensitivity (TP/TP + FN) was defined as the number of TPs found at each score threshold divided by the sum of TPs and false negatives (FNs), where FNs were defined as actual specificity sites below the score threshold. An error rate (FP/FP + TN) was estimated as the number of FPs divided by the sum of FPs and true negatives (TNs; non-specificity sites below the score threshold).

Receiver operating characteristics (ROC) and precision–recall (PR) plots were generated by (i) averaging sensitivity and specificity per protein family to provide an overall performance and/or (ii) concatenating column scores from all the families and calculating performance indicators at each score threshold. The former, 'average-per-family' approach provides an idea about the methods' performance on per-family basis, whereas the 'concatenation' approach estimates the performance based on the number of sites. Note that the first approach weighs each family equally in the final results independent of the number of known SDS in the family, whereas the second approach attempts to normalize by the number of SDS in a family but necessarily assumes that scores produced by each method are comparable across families. More details regarding the calculation of different performance measures can be found in Supplementary File 2.

Globally conserved sites are less likely to be involved in determining subgroup-specific functions. Hence, to distinguish SDS from globally conserved sites in our benchmarking, we do not consider those sites that are highly conserved in the overall family alignment; 'highly

conserved' columns have a single amino acid type in >80% of the sequences. We also ignore sites that have gaps in >20% sequences. Similar filters regarding conservation and gap content for alignment columns were applied for other SDS prediction methods as well. However, exclusion of highly conserved and invariant positions within the alignment does not impact the performance of SDS prediction significantly (Table SM3 in Supplementary File 2).

The performance of SPEER-SERVER was compared against nine other SDS prediction methods: GroupSim (26), MultiRELIEF (31), MultiRELIEF-3D (31), SDPpred v1.0 (9), XDet (unsupervised and supervised) (7), SPEL (23), SDPfox (44), Multi-Harmony (45) and ProteinKeys (16). For SPEER-SERVER, all the benchmarking data were obtained using weights of 1.0, 1.0 and 1.0 for 'relative entropy', 'PC property distance' and 'ER' components, respectively. Performance results for SPEER-SERVER using few other weight combinations of three components are provided in Supplementary File 2 (Table SM4). All other SDS prediction programs were used with default parameters. In the case of XDet supervised program, we used functional similarity (invoked by –M option of the program) option to provide pre-determined subgrouping information using a binary similarity matrix (1 for the same subfamily sequence pair and 0 for different subfamily sequence pair) for each input alignment. In the case of MultiRELIEF-3D, we used a representative protein structure for the 19 families for which at least one representative protein structure was available (one was not available for the CNmyc family; see Supplementary File 1). SPEL did not produce the results for CNmyc family; hence, the ROC and PR results of SPEL were obtained from scores of 19 families. Inconsistencies in the Multi-Harmony (MR) Z scores were found in the case of nucleotidyl cyclase family; hence, ROC and PR results for Multi-Harmony (MR) Z score approach were provided using the data from rest of the 19 families.

Benchmarking of SDS prediction using alignments generated by automated methods. SPEER-SERVER was run for each alignment generated by MAFFT (34) and PROBCONS (35). To establish the TP set in the case of alignments generated by automated methods, annotated SDS from the manually generated alignment were mapped onto the common representative sequence in the automatically generated alignments. Otherwise, benchmarking of the SDS prediction using these alignments proceeded as described above.

Benchmarking of SDS prediction using automated subgrouping methods. The performance of SPEER-SERVER was also benchmarked when using the automated subgrouping option with manually curated family alignments. SECATOR (36) and SCI-PHY (37) algorithms were separately used to automatically identify probable subgroups. Sequences that were not grouped (unclustered and/or singletons) with any of the automatically generated subgroups were filtered from the MSA before submission to SPEER-SERVER.

RESULTS

Performance evaluation using manually curated alignments

197 SDS from 20 families (Supplementary File 1) comprise the set of TPs for our evaluation of different specificity site prediction methods. The prediction accuracies of these methods are shown as ROC curves (Figure 2A) and table (Table SM1 in Supplementary File 2) suggesting better sensitivity (per family) for SPEER-SERVER, especially at low error rate (≤10%), whereas the PR statistics exhibit precision (per family) for SPEER-SERVER that is as good or better than most of the SDS detection methods over the range of recall values (Figure SM1 and Table SM2 in Supplementary File 2). Performance measures calculated on per-site basis using the 'concatenation' approach suggest that SPEER-SERVER's performance remains better relative to many other programs, although GroupSim (26) and MultiRELIEF (31) appear to perform better than SPEER-SERVER by this metric (Figures SM2 and SM3 in Supplementary File 2).

Dependence of performance on the quality of MSAs

In SPEER-SERVER, two different MSA programs (MAFFT and PROBCONS) are provided to compute a sequence alignment from a set of protein sequences specified by the user. The performance of SDS prediction by the SPEER-SERVER using the alignments generated by these MSA programs has slightly negative impact on sensitivity and precision compared with that achieved using the manually curated alignments (Figure 2B and Figure SM4 in Supplementary File 2, respectively). For example, at 1, 5 and 15% error rates, the sensitivities of SPEER-SERVER using manually curated alignments with predetermined subgroupings are 21, 51 and 72%, respectively, when compared with 18, 45 and 67% for MAFFT-derived alignments and 16, 42 and 66 for PROBCONS-derived alignments, respectively (Figure 2B).

Dependence of performance on the quality of subgrouping

De novo subfamily identification using automated methods enables improved understanding of functional inference and facilitates prediction of functional diversification. Sensitivities and accuracies of SPEER-SERVER in identifying SDS using automated subgrouping methods were calculated using 'average-per-family' approach (Figure 3). Prediction of SDS using the SECATOR (36)-derived subgrouping tends to perform better than SCI-PHY (37) at lower error rates (<12%), whereas at moderate error rates (12–40%) SCI-PHY-derived subgroupings yield better sensitivity (Figure 3). Overall, the prediction sensitivities of SPEER-SERVER using automated subgrouping of manually curated alignments are slightly lower than those observed for predetermined subgroupings on the same alignments. However, they are quite comparable to that achieved using predetermined subgrouping on automatically derived alignments (MAFFT and PROBCONS alignments). For example, at 1, 5 and 15% error rates, the sensitivities of SPEER-SERVER using SECATOR-derived subgroupings are 18, 37 and 56%,

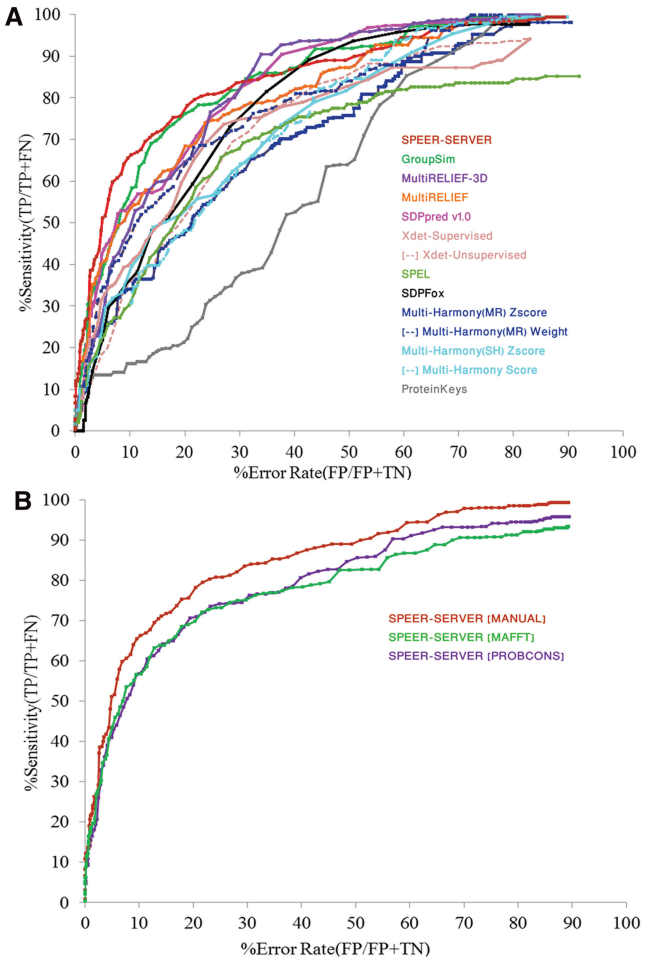

Figure 2. Comparison of prediction performance. ROC curves (**A**) for SDS prediction as performed by various SDS prediction programs. ROC curves (**B**) for the prediction of SDS based on manually curated input alignments and input alignments derived by the MAFFT (35) and PROBCONS (36) programs. Error rate and sensitivity values were calculated by averaging the equivalent error rate and sensitivity values using the 'average-per-family' approach.

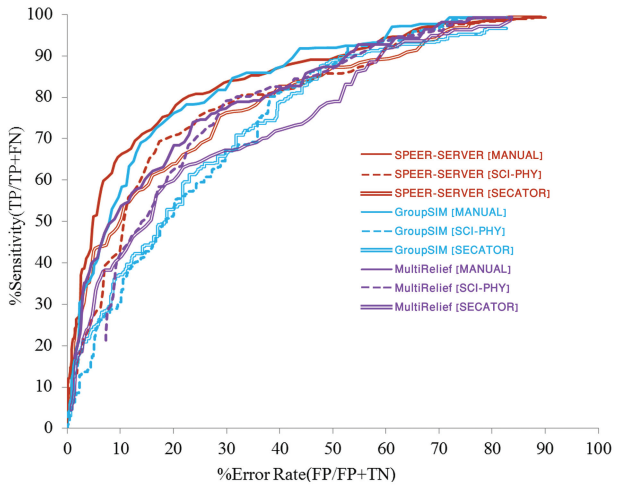

Figure 3. Comparison of performance on specificity site prediction. ROC curves for SPEER-SERVER, GroupSim (26) and MultiRELIEF (31) using manual, predetermined subgrouping and subgroupings computed using the automated methods, SECATOR (36) and SCI-PHY (37). Error rate and sensitivity values were calculated by averaging the equivalent error rate and sensitivity values using the 'average-per-family' approach.

respectively (Figure 3), and 16, 40 and 60%, respectively, for SCI-PHY-derived subgrouping. Two other programs, GroupSim (26) and MultiRELIEF (31), that performed very well in predicting SDS using pre-determined subgrouping were also tested for automated subgrouping-based SDS prediction. However, GroupSim (26) and MultiRELIEF (31) perform comparatively lower than SPEER-SERVER when automated subgrouping was used (Figure 3), indicating the dependency of different algorithms on the quality of subfamily division. Sensitivity and precision values for individual families using automated subgrouping are provided in Supplementary File 3.

DISCUSSION

Recognition of the structural and functional differences that lead to functional diversification in proteins is difficult, especially when only sequence-derived information is used. SPEER-SERVER provides a user-friendly, web-based platform for SDS identification that is built around an enhanced version of our previously reported algorithm SPEER (8). New features such as the projection of predicted SDS onto the alignment, visualizing the SDS in their molecular context using a representative 3D structure and flexible input options enable the SPEER-SERVER to be useful to a broad biological audience. Calculations of structural distances and coevolutionary networks of predicted SDS are other valuable new analysis tools which SPEER-SERVER makes available. Thorough benchmarking and validation tests performed in this study confirm better performance of SPEER-SERVER relative to most of the existing SDS prediction methods at lower error rates ($\leq$10%) and it exhibits similar performance to other methods at higher error rates.

The performance of any SDS-predicting method may depend on the quality of the input protein sequence alignment as suggested by slightly lower performance of SPEER-SERVER using automatically generated alignments. Unfortunately, for general, non-specialist users a manually curated alignment may not be always at hand. Therefore, SPEER-SERVER addresses this issue by incorporating successful MSA programs, such as MAFFT (34) and PROBCONS (35), that users can leverage to create a reliable starting MSA for analysis.

In addition, two tools to perform automated subgrouping of a protein family, SECATOR (36) and SCI-PHY (37), allow users to access SPEER-SERVER for SDS prediction. Both methods are widely used and effectively identify probable subgroups that exist within protein families. However, one concern is that the automated subgrouping of sequences based on phylogeny and alignment could be in disagreement with the subsequent functional

classification. This disagreement could arise for a subset of highly specific families, but in general sequence-based clustering agrees with functional annotations and has been the foundation of many widely used methods and techniques for the functional annotation of proteins. Indeed, we have shown that the sensitivities for SDS prediction using an automated subgrouping are on average good, even if they are suboptimal when compared with the situation when a predetermined subgrouping is available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files 1–3.

ACKNOWLEDGEMENTS

S.C. acknowledges financial and infrastructural support from CSIR-IICB. S.C. also acknowledges the Department of Biotechnology for the Ramalingaswami Fellowship. The work of C.J.L. and A.R.P. was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS.

FUNDING

Funding for open access charge: Council for Scientific and Industrial Research (CSIR) – Indian Institute of Chemical Biology (IICB).

Conflict of interest statement. None declared.

REFERENCES

1. Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.
2. 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.
3. Mihalek,I., Res,I. and Lichtarge,O. (2006) Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. *Bioinformatics*, **22**, 1656–1657.
4. Ward,R.M., Venner,E., Daines,B., Murray,S., Erdin,S., Kristensen,D.M. and Lichtarge,O. (2009) Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates. *Bioinformatics*, **25**, 1426–1427.
5. 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.
6. Mirny,L.A. and Gelfand,M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.
7. del Sol,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
8. Chakrabarti,S., Bryant,S.H. and Panchenko,A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.
9. 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.
10. Pirovano,W., Feenstra,K.A. and Heringa,J. (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.*, **34**, 6540–6548.
11. Feenstra,K.A., Pirovano,W.A., Krab,K. and Heringa,J. (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**, W495–W498.
12. Donald,J.E. and Shakhnovich,E.I. (2005) Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Res.*, **33**, 4455–4465.
13. Ye,K., Lameijer,E.W., Beukers,M.W. and Ijzerman,A.P. (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins*, **63**, 1018–1030.
14. Abhiman,S. and Sonnhammer,E.L. (2005) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.*, **33**, D197–D200.
15. Abhiman,S. and Sonnhammer,E.L. (2005) Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **60**, 758–768.
16. Reva,B., Antipin,Y. and Sander,C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.*, **8**, R232.
17. Kolesov,G. and Mirny,L.A. (2009) Using evolutionary information to find specificity-determining and co-evolving residues. *Methods Mol. Biol.*, **541**, 421–448.
18. Gu,X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.*, **16**, 1664–1674.
19. Gu,X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.*, **18**, 453–464.
20. Gaucher,E.A., Gu,X., Miyamoto,M.M. and Benner,S.A. (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.*, **27**, 315–321.
21. Abhiman,S. and Sonnhammer,E.L. (2005) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.*, **33**, D197–D200.
22. Abhiman,S. and Sonnhammer,E.L. (2005) Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **60**, 758–768.
23. Pei,J., Cai,W., Kinch,L.N. and Grishin,N.V. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
24. Gu,X. and Vander Velden,K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, **18**, 500–501.
25. Yu,G.X., Park,B.H., Chandramohan,P., Munavalli,R., Geist,A. and Samatova,N.F. (2005) In silico discovery of enzyme–substrate specificity-determining residue clusters. *J. Mol. Biol.*, **352**, 1105–1117.
26. Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.
27. Marttinen,P., Corander,J., Toronen,P. and Holm,L. (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.
28. Edwards,R.J. and Shields,D.C. (2005) BADASP: predicting functional specificity in protein families using ancestral sequences. *Bioinformatics*, **21**, 4190–4191.
29. Wuster,A., Venkatakrishnan,A.J., Schertler,G.F.X. and Babu,M.M (2010) Spial: analysis of subtype-specific features in multiple sequence alignments of proteins. *Bioinformatics*, **26**, 2906–2907.
30. Yu,G.X., Park,B.H., Chandramohan,P., Munavalli,R., Geist,A. and Samatova,N.F. (2005) In silico discovery of enzyme–substrate specificity-determining residue clusters. *J. Mol. Biol.*, **352**, 1105–1117.
31. Ye,K., Feenstra,K.A., Heringa,J., Ijzerman,A.P. and Marchiori,E. (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. *Bioinformatics*, **24**, 18–25.
32. Chakrabarti,S. and Panchenko,A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **10**, 207.
33. Mayrose,I., Graur,D., Ben-Tal,N. and Pupko,T. (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

34. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
35. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
36. Wicker,N., Perrin,G.R., Thierry,J.C. and Poch,O. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol.*, **18**, 1435–1441.
37. Brown,D.P., Krishnamurthy,N. and Sjolander,K. (2007) Automated protein subfamily identification and classification. *PLoS Comput. Biol.*, **3**, e160.
38. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
39. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
40. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
41. Chakrabarti,S. and Panchenko,A.R. (2009) Coevolution in defining the functional specificity. *Proteins*, **75**, 231–240.1.
42. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
43. Dunn,S.D., Wahl,L.M. and Gloor,G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.
44. Mazin,P.V., Gelfand,M.S., Mironov,A.A., Rakhmaninova,A.B., Rubinov,A.R., Russell,R.B. and Kalinina,O.V. (2010) An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms Mol. Biol.*, **5**, 29.
45. Brandt,B.W., Feenstra,K.A. and Heringa,J. (2010) Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.*, **38**, W35–W40.