Published online 16 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W529–W533***
doi:10.1093/nar/gkq399

ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids

Haim Ashkenazy[1,2], Elana Erez[1,2], Eric Martz[3], Tal Pupko[1] and Nir Ben-Tal[2,*]

[1]Department of Cell Research and Immunology, [2]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel and [3]Department of Microbiology, University of Massachusetts, Amherst, MA 01003, USA

Received March 8, 2010; Revised April 25, 2010; Accepted April 29, 2010

ABSTRACT

It is informative to detect highly conserved positions in proteins and nucleic acid sequence/structure since they are often indicative of structural and/or functional importance. ConSurf (http://consurf.tau.ac.il) and ConSeq (http://conseq.tau.ac.il) are two well-established web servers for calculating the evolutionary conservation of amino acid positions in proteins using an empirical Bayesian inference, starting from protein structure and sequence, respectively. Here, we present the new version of the ConSurf web server that combines the two independent servers, providing an easier and more intuitive step-by-step interface, while offering the user more flexibility during the process. In addition, the new version of ConSurf calculates the evolutionary rates for nucleic acid sequences. The new version is freely available at: http://consurf.tau.ac.il/.

INTRODUCTION

The degree to which an amino (or nucleic) acid position is evolutionarily conserved is strongly dependent on its structural and functional importance. Thus, conservation analysis of positions among members from the same family can often reveal the importance of each position for the protein (or nucleic acid)'s structure or function. ConSurf (1,2) and ConSeq (3) are web servers for calculating the evolutionary rate of each position of the protein and for identifying structurally and functionally important regions within proteins. The degree of conservation of each position is the inverse of the site's evolutionary rate; rapidly evolving positions are variable while slowly evolving positions are conserved. In ConSurf, the evolutionary rate is estimated based on the evolutionary relatedness between the protein and its homologues and considering the similarity between amino acids as reflected in the substitutions matrix (4,5). One of the advantages of ConSurf in comparison to other methods is the accurate computation of the evolutionary rate by using either an empirical Bayesian method or a maximum likelihood (ML) method (5). The differences between the two methods are explained in detail in reference (4). The strength of those methods is that they explicitly account for the stochastic process underlying the evolution of the analyzed sequences, and that they rely on the phylogeny of the sequences. Thus, they can correctly discriminate between conservation due to short evolutionary time and genuine sequence conservation. In addition, the Bayesian based method provides reliability estimates for the site-specific conservation scores.

METHODS

A short description of the methodology is provided below. More detailed description is available at http://consurf.tau.ac.il/, under 'OVERVIEW', 'QUICK HELP' and 'FAQ'.

ConSurf protocol

A flowchart of the ConSurf web server is shown in Figure 1 and detailed below.

(1) The sequence is extracted from the 3D structure (if given).
(2) Homologous sequences are collected using a BLAST (or PSI-BLAST) (6,7) search against a selected database. The user may specify criteria for defining homologues. The user can also manually select the desired sequences from the BLAST results.
(3) The sequences are clustered and highly similar sequences are removed using CD-HIT (8).

*To whom correspondence should be addressed. Tel: +972 3 640 6709; Fax: +972 3 640 6834; Email: nirb@tauex.tau.ac.il

The authors wish it to be known that, in their opinion the first two authors should be regarded as joint First Authors

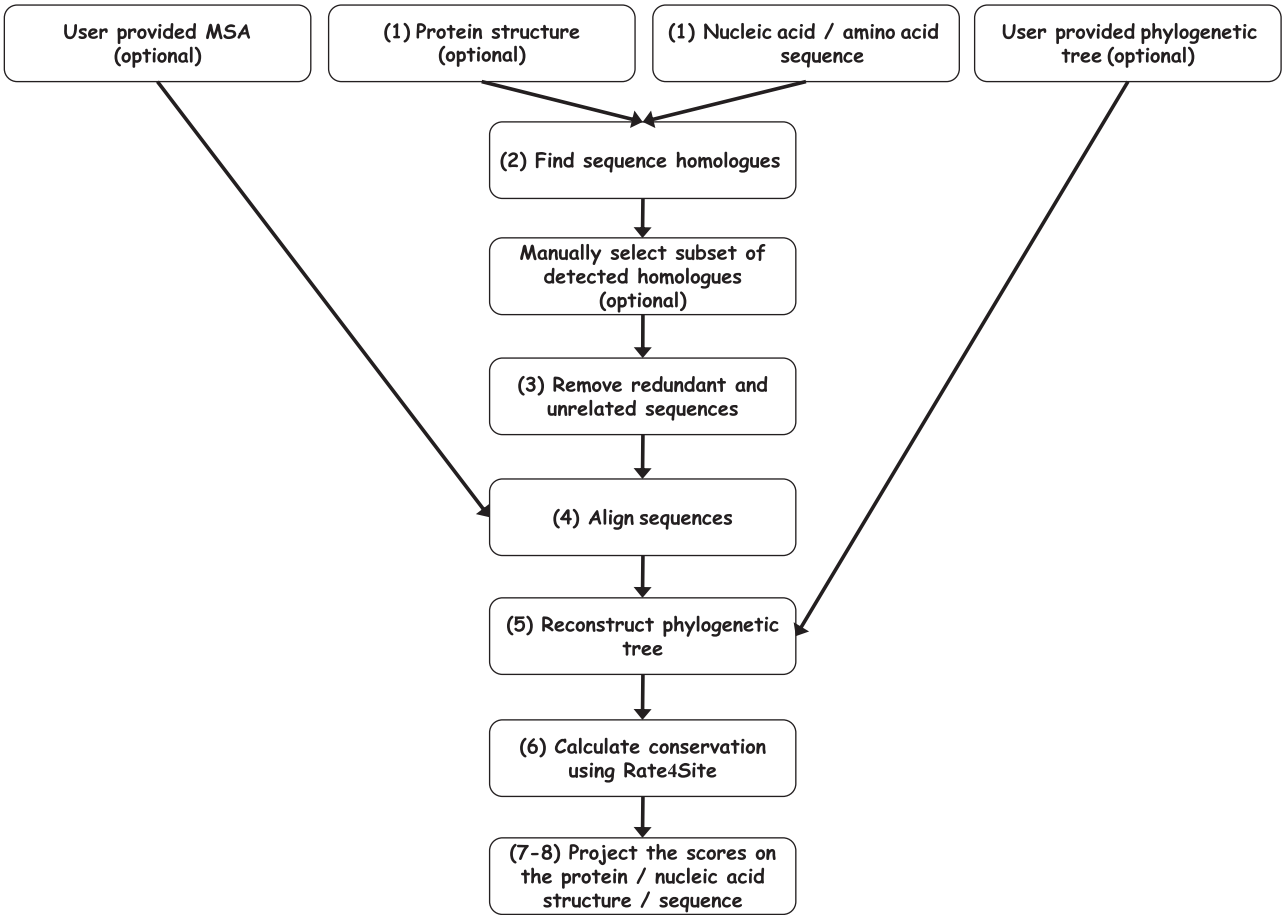

Figure 1. A flowchart of ConSurf protocol.

(4) A multiple sequence alignment (MSA) of the homologous sequences is constructed using MAFFT, PRANK, T-COFFEE, MUSCLE or CLUSTALW.
(5) A phylogenetic tree is reconstructed based on the MSA, using the neighbor-joining algorithm as implemented in the Rate4Site program (4,5).
(6) Position-specific conservation scores are computed using the empirical Bayesian or ML algorithms (4,5).
(7) The continuous conservation scores are divided into a discrete scale of nine grades for visualization, from the most variable positions (grade 1) colored turquoise, through intermediately conserved positions (grade 5) colored white, to the most conserved positions (grade 9) colored maroon.
(8) The conservation scores are projected onto the protein/nucleotide sequence and on the MSA.

Outputs

If a protein 3D structure is provided:

(1) The nine-color conservation scores are projected onto the 3D structure of the query protein and the colored protein structure is shown by FirstGlance in Jmol (http://firstglance.jmol.org).
(2) Scripts for visualizing the protein colored with ConSurf scores are generated for PyMol (http://www.pymol.org; 9), Chimera (10), Jmol (http://www.jmol.org/; 11) and RasMol (12).

For all cases, ConSurf creates the following outputs:

(1) The sequence and MSA colored by ConSurf conservation scores.
(2) A text file that summarizes for each position the normalized score calculated, the assigned color, the reliability estimation (for the Bayesian method) and the amino acids/nucleotides observed in the respective MSA column.
(3) The sequences selected for the MSA and the MSA constructed (unless those files were uploaded by the user).
(4) A file with the frequency of each amino acid/nucleotide observed in each column of the MSA.
(5) The evolutionary tree, which was calculated by the server or uploaded by the user, is shown using an interactive Java applet written for that purpose.

For proteins in which the 3D structure was not provided by the user, an up-to-date version of the Protein Data Bank (13) is searched for relevant homologues. If a structure of at least one homologous protein is available, the user may map the conservation scores on the structure. This option should ease the procedure for the non-expert users, who may be unfamiliar with the 3D structure

homologue. This option can also be useful for analyzing proteins that share the same sequence but differ in their 3D structure (for example, two structures solved in different conformations or with different ligands).

As an example we provide the main output of a ConSurf run for the N-terminal region of the GAL4 transcription factor in yeast (PDB ID: 3COQ, chain A and B) in complex with its DNA recognition site (Figure 2). The analysis revealed, as expected, that the functional regions of this protein are highly conserved. For example, all the cysteines that form the Zn(2)-C6 DNA binding domain (CYS11, CYS14, CYS21, CYS28, CYS31, CYS38; 14) were assigned the highest conservation scores. Likewise, PRO26, which is known to be central for DNA binding (15) is also highly conserved according to our analysis. In addition, other amino acid residues, which are in contact with the DNA (i.e. GLN9, LYS17, LYS18, LYS20, ARG15, LYS23; 16) are relatively conserved.

ConSurf was also applied to nucleic acid sequences from yeast, which are the known binding sites of GAL4 and their adjacent neighborhood (Figure 2). As anticipated, the analysis revealed that the consensus pattern CGG-N_{11}-CCG typical to GAL4 binding site is highly conserved. An extended full ConSurf analysis of this example is available in the 'GALLERY' section on the ConSurf web site.

NEW ADDITIONS AND IMPROVEMENTS IN ConSurf 2010

Analyzing nucleic acid sequences

Despite increasing interest in the non-coding fraction of transcriptomes, the number, the level of conservation, and functions, if any, of many non-protein-coding transcripts remain to be discovered. However, it has already been shown that many of the non-coding sequences are connected to regulatory processes. The new version of ConSurf offers estimations of the evolutionary rate for each position of nucleic acid sequences in the same manner used for amino acid residues. For that purpose, four evolutionary models were implemented in the Rate4Site program: (i) the Juke and Cantor 69 model (JC69), which assumes equal base frequencies and equal substitution rates (17). (ii) The Tamura 92 model that uses only one parameter, which captures variation in G-C content (18). (iii) The HKY85 model, which distinguishes between transitions and transversions and allows unequal base frequencies (19). (iv) The General Time Reversible (GTR) model, which is the most general time-reversible model. The GTR parameters consist of an equilibrium base frequency vector, giving the frequency at which each base occurs at each site, and the rate matrix (20). When enough data (i.e. sequences) are available, the GTR model is superior over the more simplified Tamura

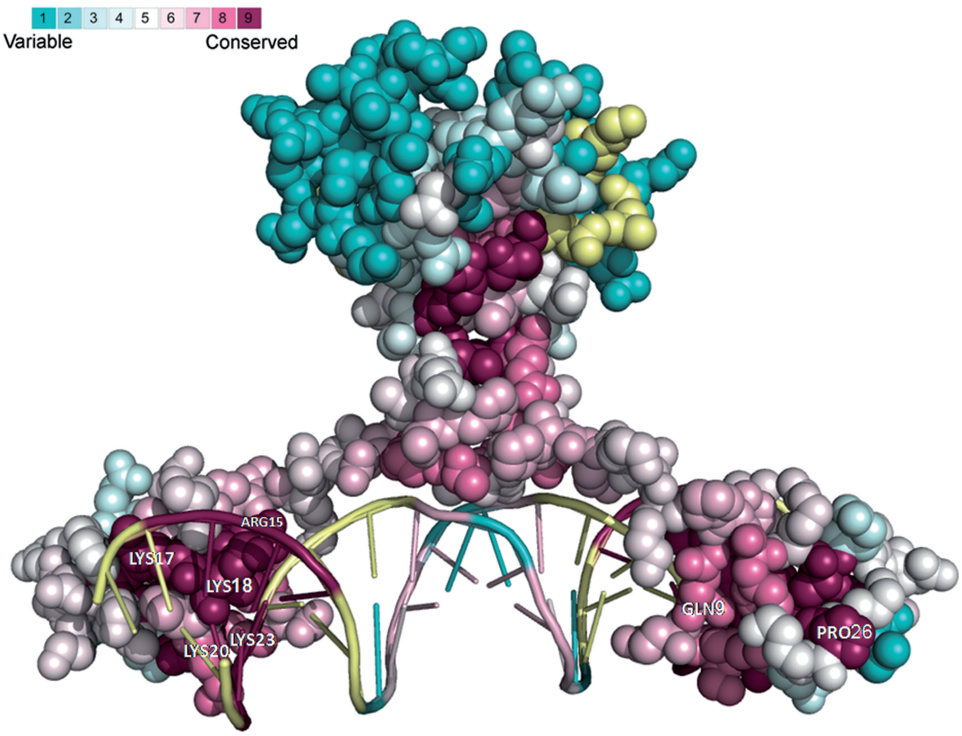

Figure 2. A ConSurf analysis for the GAL4 transcription factor and its DNA binding site. The 3D structure of the N-terminal region of the GAL4 transcription factor in yeast bound to the DNA is presented using a space-filled model. The amino-acids and the nucleotides are colored by their conservation grades using the color-coding bar, with turquoise-through-maroon indicating variable-through-conserved. Positions, for which the inferred conservation level was assigned with low confidence, are marked with light yellow. The figure reveals that the functionally important regions on both the DNA and the protein are highly conserved. The run was carried out using PDB code 3COQ and the figure was generated using the PyMol (10) script output by ConSurf.

92 model. However, the Tamura 92 model is recommended in cases in which the data are not sufficient for reliable estimation of the model parameters and thus it is the default option for analyzing nucleic acid sequences in ConSurf.

Improved substitution matrix for protein sequences

The LG substitution matrix, which incorporates variability of evolutionary rates across sites in the matrix estimation was shown to outperform other substitutions matrices for proteins (21). The LG matrix was added to Rate4Site and is offered in the new version of ConSurf in addition to the previous substitution models: JTT (22), Dayhoff (23), WAG (24), mtREV (25) and cpREV (26).

Improved selection of homologous proteins

The accuracy of conservation scores is directly influenced by the amount and quality of sequence data available in the MSA and the relatedness between the homologous sequences themselves and the sequence of interest. For example, using homologous sequences with different functions might blur the signal. One of the important changes in the new version of ConSurf is the addition of a clear and intuitive interface that helps controlling which of the sequences are included in the analysis. These improvements include:

(1) A variety of sequence databases. The server offers the user the option to search for relevant sequences in several automatically updated sequence databases including: (i) SWISS-PROT (default) (27); (ii) A filtered version of the uniprot database (28); (iii) uniprot (29) (iv) UniRef90 in which redundant sequences were removed at level of 90% identity (30); (v) the NCBI non-redundant (nr) database.
(2) Manual selection of sequences for the analysis. After searching for homologous sequences, the user can manually select the relevant sequences to be included in the analysis using a simple form that provides all the relevant data for the sequences found and links to external web resources.
(3) Removing redundant sequences. The user can specify the level of redundant sequences for removal. The sequences found are clustered by their level of identity using CD-HIT (8) and the cutoff specified by the user (default level is 95% identity). Only one sequence (the longest) from each cluster is used for the analysis.
(4) Automatic removal of remote homologues. The user can control the level of sequence identity for which a hit sequence is still considered a homologue. Filtration according to the sequence identity between the sequences found and the sequence of interest enables the user to filter out sequences that share significant alignment with the protein of interest, however, might have different function or structure. The default level is set to 35% identity, which is the upper bound of the 'twilight zone' for protein structures (31).
(5) Better alignments. The user can choose to align the sequences using one of the following leading alignment algorithms: MAFFT (32), T-COFFEE (EXPRESSO mode) (33), PRANK (34) MUSCLE (35) and CLUSTALW (36). The EXPRESSO mode of T-COFFEE uses structural information (if available) and structural alignment methods to construct structure-based MSA. MAFFT and PRANK were shown to be among the leading sequence alignment algorithms (34,37). MAFFT-LINSi is much faster than PRANK and thus was chosen to be the default alignment algorithm in ConSurf.

Improved user interface

In this new version of ConSurf, we put great emphasis on the user interface. ConSurf now presents an easier and more intuitive step-by-step interface, while still offering the user great flexibility during the process as described above. Each step is accompanied by built-in detailed help.

IMPLEMENTATION

The new version of the ConSurf web server runs on a Linux cluster of 2.6GHz AMD Opteron processors, equipped with 4 GB RAM per quad-core node. The server runs with up to date versions of the supported MSA programs, and regularly updated databases. Running time depends on the dataset size (number and length of sequences) and the server load. The ConSurf server is implemented in PHP and Perl using the support of BioPerl modules (38). Rate4Site is implemented in C++ (4). For proteins with available 3D structure the conservation scores are projected on the structure and visualized using version 1.44 of FirstGlance in Jmol.

CONCLUSIONS

ConSurf and ConSeq have an established reputation in the identification of functional regions in proteins using evolutionary information. In addition, these methods are a focal point that facilitates the development of more useful tools in our group and in other groups. For example, they are the basis for the development of the PatchFinder tool for the automatic detection of clusters of highly conserved amino acids (39), and the detection of DNA-binding proteins (40). Along with the massive growth of sequence and structure databases we believe that this new version of the ConSurf server will be highly useful to a growing number of molecular biology researchers and allow them to perform complex analyses using sophisticated algorithms accurately, easily and comprehensively.

ACKNOWLEDGEMENTS

The authors are grateful to Nimrod Rubinstein, Adi Doron-Faigenboim, Eyal Privman, Itay Mayrose, Fabian Glaser, Maya Schushan, Guy Nimrod, Ofir Goldenberg, Yana Gofman, Uri Zonens, Gilad Wainreb

and Matan Kalman for technical help, useful comments and helpful discussions.

FUNDING

BLOOMNET ERA-PG; Israeli Science Foundation (878/09 to T.P.). Funding for open access charge: BLOOMNET ERA-PG.

Conflict of interest statement. None declared.

REFERENCES

1. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
2. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
3. Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.
4. Mayrose,I., Graur,D., Ben-Tal,N. and Pupko,T. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
5. 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(Suppl 1)**, S71–S77.
6. Altschul,S.F., Wootton,J.C., Gertz,E.M., Agarwala,R., Morgulis,A., Schaffer,A.A. and Yu,Y.K. (2005) Protein database searches using compositionally adjusted substitution matrices. *FEBS J.*, **272**, 5101–5109.
7. 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.
8. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
9. DeLano,W.L. (2008) *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, Palo Alto, CA, USA.
10. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
11. Herráez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
12. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
13. 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.
14. Pan,T. and Coleman,J.E. (1990) GAL4 transcription factor is not a "zinc finger" but forms a Zn(II)2Cys6 binuclear cluster. *Proc. Natl Acad. Sci. USA*, **87**, 2077–2081.
15. Johnston,M. (1987) Genetic evidence that zinc is an essential co-factor in the DNA binding domain of GAL4 protein. *Nature*, **328**, 353–355.
16. Marmorstein,R., Carey,M., Ptashne,M. and Harrison,S.C. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.
17. Jukes,T.H. and Cantor,C.R. (1969) *Evolution of Protein Molecules*. Academic Press, New York.
18. Tamura,K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.*, **9**, 678–687.
19. Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.
20. Tavare,S. (1986) Some probabilistic and statisical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.*, **17**, 57–86.
21. Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
22. 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.
23. Dayhoff,M.O., Hunt,L.T., Barker,W.C., Schwartz,R.M. and Orcutt,B.C. (1978) In Young,C.L. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC.
24. Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
25. Adachi,J. and Hasegawa,M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, **42**, 459–468.
26. Adachi,J., Waddell,P.J., Martin,W. and Hasegawa,M. (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.*, **50**, 348–358.
27. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
28. Goldenberg,O., Erez,E., Nimrod,G. and Ben-Tal,N. (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.
29. The UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
30. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
31. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
32. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.
33. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
34. Loytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.
35. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
36. 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.
37. Nuin,P.A., Wang,Z. and Tillier,E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
38. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
39. Nimrod,G., Schushan,M., Steinberg,D.M. and Ben-Tal,N. (2008) Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure*, **16**, 1755–1763.
40. Nimrod,G., Szilagyi,A., Leslie,C. and Ben-Tal,N. (2009) Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. *J. Mol. Biol.*, **387**, 1040–1053.