W344–W350 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 10 May 2016*
doi: 10.1093/nar/gkw408

ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules

Haim Ashkenazy[1], Shiran Abadi[2], Eric Martz[3], Ofer Chay[1,2,4], Itay Mayrose[2,*], Tal Pupko[1,*] and Nir Ben-Tal[4,*]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [2]Department of Molecular Biology and Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [3]Department of Microbiology, University of Massachusetts, Amherst, MA 01003, USA and [4]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Received February 20, 2016; Revised April 21, 2016; Accepted May 3, 2016

ABSTRACT

The degree of evolutionary conservation of an amino acid in a protein or a nucleic acid in DNA/RNA reflects a balance between its natural tendency to mutate and the overall need to retain the structural integrity and function of the macromolecule. The ConSurf web server (http://consurf.tau.ac.il), established over 15 years ago, analyses the evolutionary pattern of the amino/nucleic acids of the macromolecule to reveal regions that are important for structure and/or function. Starting from a query sequence or structure, the server automatically collects homologues, infers their multiple sequence alignment and reconstructs a phylogenetic tree that reflects their evolutionary relations. These data are then used, within a probabilistic framework, to estimate the evolutionary rates of each sequence position. Here we introduce several new features into ConSurf, including automatic selection of the best evolutionary model used to infer the rates, the ability to homology-model query proteins, prediction of the secondary structure of query RNA molecules from sequence, the ability to view the biological assembly of a query (in addition to the single chain), mapping of the conservation grades onto 2D RNA models and an advanced view of the phylogenetic tree that enables interactively rerunning ConSurf with the taxa of a sub-tree.

INTRODUCTION

ConSurf is a widely used tool for revealing functional regions in macromolecules by analysing the evolutionary dynamics of amino/nucleic acids substitutions among homologous sequences (1–4). ConSurf estimates the evolutionary rates of the amino/nucleic acids and maps them onto the sequence and/or structure of the query macromolecule. Slowly evolving sites on the query surface are usually important for function and thus, ConSurf analysis can pinpoint critically important sites within the query macromolecule. This is particularly true when the structure of the query macromolecule is known, allowing to differentiate between slowly evolving positions at the core, which are usually important for structural stability/folding (e.g. 5), and clusters of slowly evolving surface positions, important for function (6–14). In the absence of structure, the evolutionary data are presented on the query sequence together with site-specific predictions of the buried/exposed status of each position, i.e. ConSeq mode (15). The power of ConSurf, in comparison to other popular alternatives based on consensus and relative entropy approaches, is that the evolutionary rates are estimated based on the phylogenetic relationships among the homologues and the specific dynamics of the analysed sequences using advanced probabilistic evolutionary models (16,17). This statistically robust approach makes it easier to differentiate between apparent conservation due to short evolutionary time and genuine conservation reflecting the action of purifying selection. Notably, ConSurf also assigns confidence intervals around the calculated evolutionary rates, which estimate the credibility of the results.

The superiority of ConSurf's estimation of evolutionary conservation over entropy based methods in accurate pre-

*To whom correspondence should be addressed. Tel: +972 3 640 6709; Fax: +972 3 640 6834; Email: bental@tauex.tau.ac.il
Correspondence may also be addressed to Itay Mayrose. Email: itaymay@post.tau.ac.il
Correspondence may also be addressed to Tal Pupko. Email: talp@post.tau.ac.il

diction of protein active sites, as well as the identification of biologically active peptides, was previously demonstrated (18–20); elaborate comparison of ConSurf to alternatives that do not explicitly account for the phylogenetic relations among sequences is provided in the OVERVIEW section of the ConSurf web server. The best established alternative to ConSurf for the detection of functional regions is the Evolutionary Trace method and variants thereof (6–8,21,22). These are also based on phylogenetic analysis, but lack the mathematical rigour of ConSurf, and do not provide any credibility interval around the inferred scores. Recently, the Golding lab introduced a rigorous model, based on phylogenetic Gaussian process, that accounts for spatial correlation of substitution rates in different positions according to the protein tertiary structure (23,24). Unlike ConSurf, this sophisticated approach requires knowledge of the protein structure.

Here we report the introduction of several new features into ConSurf, designed to improve the performance and the interface of the web server in the detection of functional regions in proteins and nucleic acids.

MATERIALS AND METHODS

In a typical ConSurf application, the query protein is first BLASTed (25) against the UNIREF-90 database (26). Redundant homologous sequences are then removed using the CD-HIT clustering method (27,28). The resulting sequences are next aligned using MAFFT (29) and the generated multiple sequence alignment (MSA) is used to reconstruct a phylogenetic tree. Given the tree and the MSA, the Rate4Site algorithm (16) is used to calculate position-specific evolutionary rates under an empirical Bayesian methodology (17). The rates are normalized and grouped into nine conservation grades 1-through-9, where 1 includes the most rapidly evolving positions, 5 includes positions of intermediate rates, and 9 includes the most evolutionarily conserved positions. It is important to notice that structural data are not used up to this point, and the rates are estimated based on sequence data alone. Finally, the conservation grades are mapped onto the query sequence and/or structure using the ConSurf colour-code, with cyan-through-purple corresponding to variable (grade 1)-through-conserved (grade 9) positions. The analysis is conducted only if there are at least five homologous proteins, otherwise the degree of uncertainty is too high.

The protocol for the selection of homologous sequences was shaped while massive amounts of sequence data are becoming available (2,3). A balance between the number of sequences used for analysis and their evolutionary or functional relationship to the query molecule should be maintained. Thus, we try to adjust the default parameters used for homologues collection to maintain this balance. This includes (i) using CS-BLAST, suggested to be more sensitive and accurate in searching for remote homologues, compared to the commonly used BLAST algorithm (25); (ii) for proteins, only sequences sharing at least 35% sequence identity with the query sequence are considered. This was suggested to be the upper boundary of the 'twilight zone' for protein structures (30); (iii) the MAFFT-LINSi procedure, suggested to be one of the most accurate MSA methodologies (31), is used to align the homologous sequences. Many alternatives to this typical outline are provided in the server. For example, ConSurf is also applicable to nucleotide sequences, it can be used with an external pre-built MSA, and users can control many details of the default algorithmic flow described above. The ConSurf methodology and these advanced options are described in detail in the 'OVERVIEW', 'QUICK HELP' and 'FAQ' sections of the ConSurf website (http://consurf.tau.ac.il).

RECENT ADDITIONS AND IMPROVEMENTS

Selecting the evolutionary model that best fits the data

In the previous ConSurf version, the user was allowed to select one of several evolutionary models which differ from each other in their biological assumptions and in the number of free parameters. For nucleotide sequences the following models have been implemented: the Jukes and Cantor model (JC69), which assumes equal base frequencies and equal substitution rates (32); the Tamura 92 model that uses only one parameter, which captures variation in G-C content (33); the HKY85 model, which distinguishes between transitions and transversions and allows for unequal base frequencies (34); and the General Time Reversible model, which includes free parameters for each transition type and base frequency (35). For protein sequences several models were implemented: LG (36), JTT (37), Dayhoff (38), WAG (39), mtREV for mitochondrial proteins (40) and cpREV for chloroplast proteins (41). Different models can result in different estimations of the phylogeny and evolutionary rate (42,43). ConSurf now allows automatic selection of the model that best fits the analysed sequences, as determined by the Akaike information criterion (AIC) (44–46). Users who prefer this option need to select it in the menu.

Predicting RNA secondary structures

For RNA sequence queries, ConSurf now offers the possibility to predict the secondary structure. Structures are predicted using the RNAfold program of the Vienna package (47,48), and the structure with the lowest free energy is selected. The ConSurf conservation grades are mapped onto the predicted secondary structure. Correlating the evolutionary data with the structural model offers the means to quickly detect functional regions within the RNA query. To exemplify this feature, we analyse the well-studied Phe-tRNA molecule (Figure 1A). The calculation is based on RFAM homologous sequences (49) of the Phe-tRNA molecule (RFAM RF00005 family) clustered by CD-HIT to the level of 80% sequence identity and aligned using MAFFT. The results show that some bases in the TΨC and D loops are assigned particularly high conservation grades. Some of these positions are known to be of structural and functional importance (50,51). Figure 1B shows the conservation grades on the 3D structure of the Phe-tRNA molecule (PDB ID: 1EHZ chain A), further emphasizing the importance of the evolutionarily conserved positions.

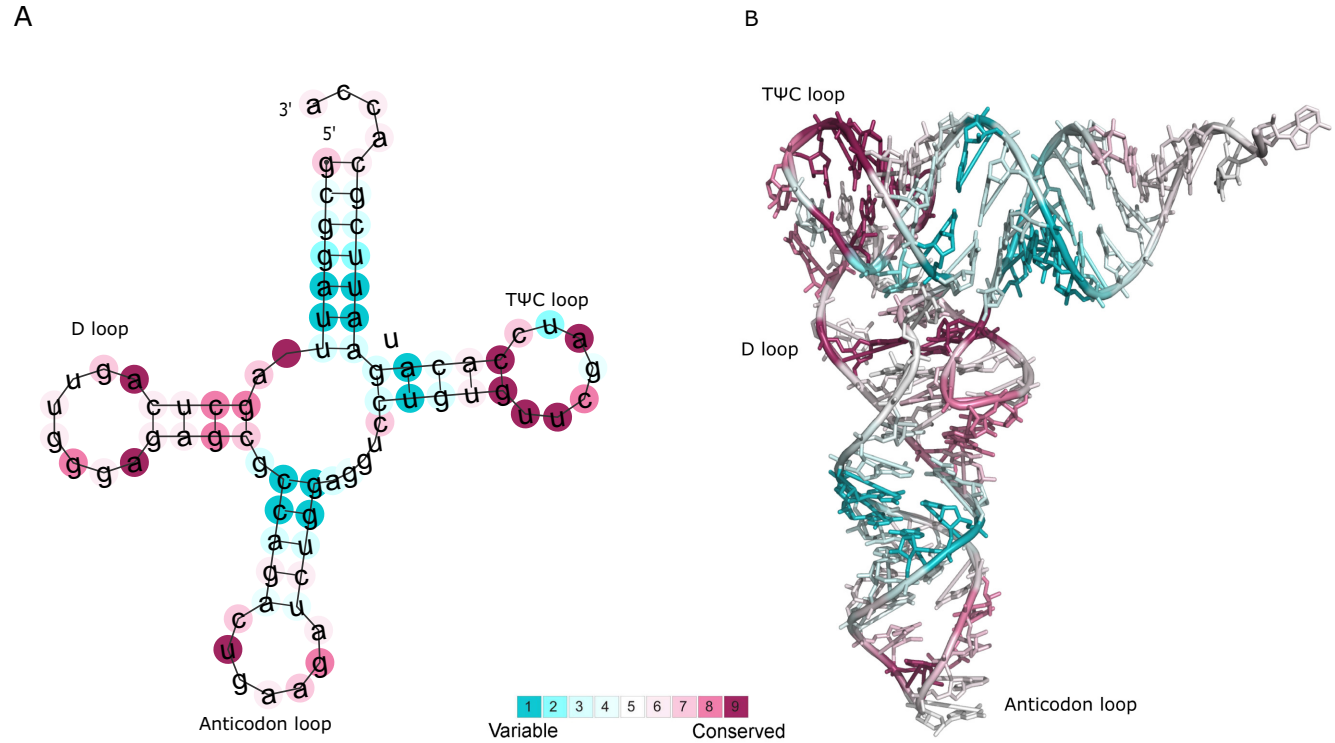

Figure 1. ConSurf analysis of yeast Phe-tRNA. (**A**) Secondary structure prediction of the molecule coloured by conservation using the colour-code bar. (**B**) Same analysis using the X-ray crystal structure of the molecule (PDB ID: 1EHZ, chain A).

Predicting a template-based structure for protein sequences

In the previous version, when only the sequence (rather than the structure) of the query protein was provided as input (i.e. 'ConSeq mode'), ConSurf searched the PDB (52,53) using BLAST (54) to suggest probable homologues of known structure. If the search was productive, the conservation grades were mapped on each of the homologous structures. In the new version, we go one step further and use HHPred (55) and MODELLER (56) to produce a homology-model of the query. Briefly, HHPred uses a hidden Markov model to search for potential templates of known 3D structure in the PDB (57). The MODELLER algorithm (56) is then used to predict a 3D model for the query sequence. The ConSurf conservation grades are subsequently mapped on the predicted model.

In addition, the homology model is used to predict the solvent accessibilities of the amino acids. To this end, we use the relative solvent accessible surface areas of the amino acids, calculated using NACCESS (58) and the predicted structure. The derivation of solvent accessibility from the 3D model is expected to be more accurate compared to the buried/exposed prediction, made solely using the protein sequence (59). The latter option is still offered in cases where a template is not available.

Refining ConSurf results using a subset of sequences

Occasionally, purifying selective forces may be strong in one part of the phylogeny yet relaxed (or different) in the remaining parts, indicative of gain or loss of function in some taxonomic clades or in protein subfamilies. In a typical ConSurf analysis, the whole set of homologues (either user-supplied or automatically collected using the default setting) is analysed as a single group, masking this important functional signal. The new ConSurf version provides the means to refine an initial ConSurf analysis by allowing users to select a subtree containing a fraction of the homologous sequences and conduct a follow-up analysis of these selected sequences. To this end, in the ConSurf Results page, the MSA and tree are now visualized using the WASABI platform (60). Users can thus choose any internal node on the phylogenetic tree and open a WASABI menu using a right mouse click (see an example in Supplementary Figure S1). Selecting the option 'run ConSurf on subtree' will issue a new window with a follow-up ConSurf run for the selected sequences of the subtree.

Improved visualizations

This new version of ConSurf suggests three major visualization improvements:

(i) **Accounting for protein assembly.** Many proteins function together as complexes, or biological units (5). Therefore, accounting for the full assembly can shade further light on the importance of residues located at the interfaces between the subunits. The new version of ConSurf automatically suggests the possibility to map the calculated evolutionary conservation grades of the amino acids not only onto a single chain, taken from the asymmetric unit of the crystal, which is often deposited in the PDB, but also on all the appearances of the chain in the biological assembly as pre-

Table 1. The main recent improvements in ConSurf

Feature	ConSurf—2010	ConSurf—2016
Selecting evolutionary model	Only according to user selection	New option for automatic selection of the model showing the best fit to MSA
RNA secondary structure	Not available	Predicting RNA secondary structure using Vienna package and projecting ConSurf grades on the structure
Projecting scores on identical chains and protein assemblies	Scores projected only on single chain, and do not support protein assemblies	Projecting scores on all identical chains and the most probable assembly downloaded from PISA
FirstGlance in Jmol	Version 1.44 supporting only Jmol viewer	Version 2.42 supporting JSmol which is Java free viewer (see additional features and improvements of the new version at: http://bioinformatics.org/firstglance/fgij/versions.htm)
Phylogenetic tree viewer	Only the tree is shown using a Java applet	The MSA is shown together with the phylogenetic tree using the WASABI platform (Java free)
Rerun ConSurf using sequences from sub-tree	Not available	Interactive selection of sub-tree sequences using WASABI, and rerun ConSurf with these sequences
Structural information for proteins query sequence (no PDB provided)	Suggesting highly similar homologues sequences to the protein query sequence and projecting ConSurf scores on them	In addition to the suggested homologues, template based structure prediction is performed using HHPred and MODELLER
Solvent accessibility information when protein's PDB structure is not available	Predicted from sequence information only	When possible, extracted using NACESS from the 3D structure modelled by HHPred

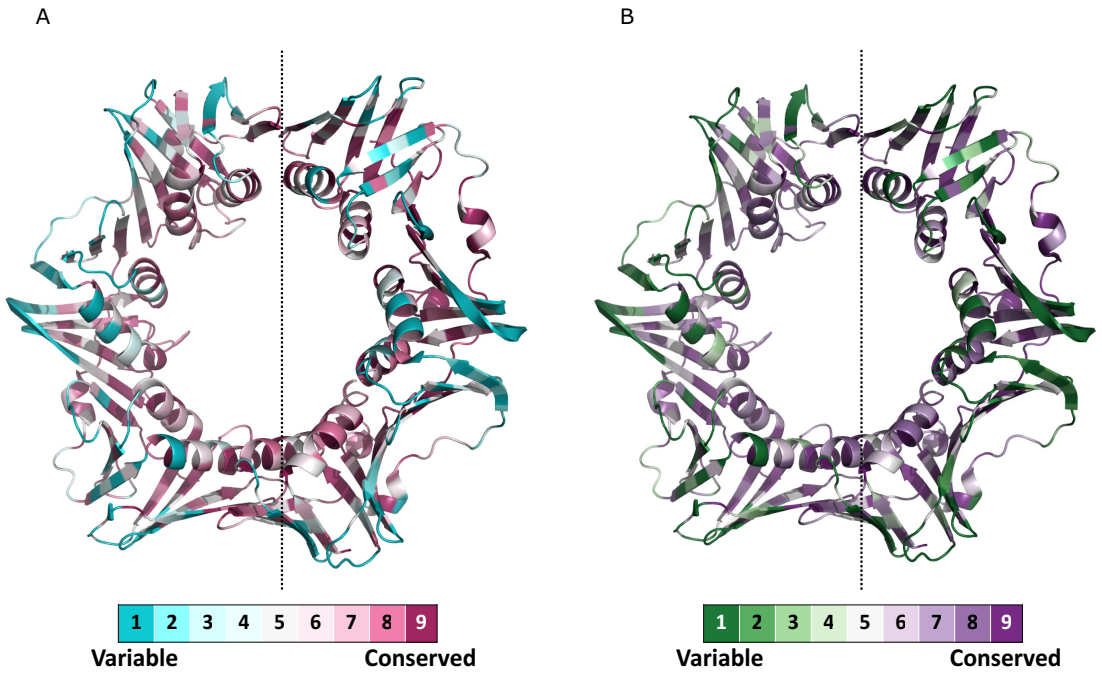

Figure 2. ConSurf analysis of the β subunit of DNA polymerase III from *Escherichia coli* (PDB ID: 2POL). The interfaces between the two subunits of the homodimer (on both sides of the dotted line) are highly conserved, as well as the internal face of the ring, which interacts with the DNA. (**A**) Molecule coloured by the traditional ConSurf scale. (**B**) Molecule coloured by the new colour-blind friendly scale.

dicted by PISA (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html) (61). Figure 2A demonstrates this new feature using the 3D structure of the β subunit of DNA polymerase III from *Escherichia coli*. The protein functions as a homo-dimer, and as anticipated, most of the residues at the inter-subunit interfaces (62) are highly evolutionarily conserved (Leu108, Lys74, Ile272, Leu273, Glu300, Glu304). The full ConSurf analysis of this structure is available for interactive exploration under the 'GALLERY' section of the web server.

(ii) **Supporting non-Java-based visualization**. To enable interactive visualization of 3D molecular structures on devices with no Java installed, or for which Java is not available (e.g. smart phones), a new version of FirstGlance in Jmol (http://bioinformatics.org/firstglance/fgij/) was implemented with JSmol (63). Briefly, JSmol uses HTML5 and JavaScript to implement the functionality offered by the Jmol application, which is implemented in Java (http://www.jmol.org/). It allows 3D visualization on modern web browsers (e.g. Chrome, Edge) which, in attempt to avoid security threats, no longer support Java applets (such as Jmol) running from the browser.

(iii) **New colour-blind friendly pallet**. In addition to the traditional cyan-through-purple pallet corresponding to variable (grade 1)-through-conserved (grade 9) scores, the new version also suggests a more colour-blind friendly pallet of green-through-purple scale (see example in Figure 2B).

CONCLUSIONS AND PROSPECTS

We presented improvements to the ConSurf method and web server for the detection of functional regions in protein and nucleotide sequences. The ConSurf calculation is conducted using sequence data, but the results are particularly enlightening when viewed on the 3D structure of the macromolecule, or model thereof. The main changes compared to the previous version of ConSurf are summarized in Table 1.

Our understanding of the physicochemical interactions underlying the selective forces responsible for evolutionary-rate differences among sites is very partial. Quantitatively, it was estimated that only 60% of the data can be explained (64). Nevertheless, exploiting evolutionary rate differences is useful in various biological studies, including structure analysis (e.g. 65,66) and prediction (e.g. 67), interpretation (68) and design of mutations (69), identification of natural peptides (20), systems and genome-wide studies (70) and studies of the last common ancestor (71). Until a few years ago, the bottleneck of the analysis was to obtain a sufficient number of homologous sequences of the query and to design an algorithm that makes the best use of these homologues to estimate the evolutionary rates. While efforts to improve the rates estimates are constantly undergoing (23,24,72,73), with the flood of sequence data, the main challenge now is to cope with thousands of homologous sequences by clustering the data and finding a large and diverse set of true homologues that faithfully represents the diversity. We plan to implement such a clustering method in ConSurf in the very near future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors are thankful to Aya Narunsky for the help with MODELLER integration into ConSurf and to Andres Veidenberg for the help in integrating the WASABI platform.

FUNDING

I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation [1775/12 to N.B-.T.]; Israeli Science Foundation [1092/13 to T.P.]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University (to H.A., O.C., in part). Funding for open access charge: I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation [1775/12 to N.B-.T.].
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. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
4. Celniker,G., Nimrod,G., Ashkenazy,H., Glaser,F., Martz,E., Mayrose,I., Pupko,T. and Ben-Tal,N. (2013) ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.*, **53**, 199–206.
5. Kessel,A. and Ben-Tal,N. (2010) *Introduction to proteins: structure, function, and motion*. CRC Press, Boca Raton.
6. 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.
7. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 7507–7511.
8. Lichtarge,O., Yamamoto,K.R. and Cohen,F.E. (1997) Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.*, **274**, 325–337.
9. Gallet,X., Charloteaux,B., Thomas,A. and Brasseur,R. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, **302**, 917–926.
10. 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.
11. Valdar,W.S.J. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
12. del Sol,A., del Sol Mesa,A., Pazos,F. and Valencia,A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
13. Nimrod,G., Glaser,F., Steinberg,D., Ben-Tal,N. and Pupko,T. (2005) In silico identification of functional regions in proteins. *Bioinformatics*, **21**(Suppl. 1), i328–i337.

14. 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.
15. 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.
16. 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.
17. 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.
18. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
19. Johansson,F. and Toh,H. (2010) A comparative study of conservation and variation scores. *BMC Bioinformatics*, **11**, 388.
20. Toporik,A., Borukhov,I., Apatoff,A., Gerber,D. and Kliger,Y. (2014) Computational identification of natural peptides based on analysis of molecular evolution. *Bioinformatics*, **30**, 2137–2141.
21. 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.
22. Sankararaman,S., Kolaczkowski,B. and Sjölander,K. (2009) INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. *Nucleic Acids Res.*, **37**, W390–W395.
23. Huang,Y.-F. and Golding,G.B. (2014) Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Comput. Biol.*, **10**, e1003429.
24. Huang,Y.-F. and Golding,G.B. (2015) FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. *Bioinformatics*, **31**, 523–531.
25. Biegert,A. and Söding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 3770–3775.
26. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
27. 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.
28. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
29. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
30. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
31. Nuin,P.A., Wang,Z. and Tillier,E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
32. Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In: Munro,HN (ed). *Mammalian Protein Metabolism*. Academic Press, NY, pp. 21–132.
33. 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.
34. 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.
35. Tavare,S. (1986) Some probabilistic and statisical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.*, **17**, 57–86.
36. Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
37. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, **8**, 275–282.
38. Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In: Dayhoff,M (ed). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, D.C., pp. 345–352.
39. 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.
40. Adachi,J. and Hasegawa,M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, **42**, 459–468.
41. 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.
42. Posada,D. and Crandall,K.A. (2001) Selecting the best-fit model of nucleotide substitution. *Syst. Biol.*, **50**, 580–601.
43. Posada,D. (2003) Selecting models of evolution. In: Salemi,M and Vandamme,AM (eds). *The Phylogenetic Handbook: a Practical Approach to DNA and Protein Phylogeny*. Cambridge University Press, Cambridge, pp. 256–282.
44. Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723.
45. Darriba,D., Taboada,G.L., Doallo,R. and Posada,D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, **9**, 772.
46. Darriba,D., Taboada,G.L., Doallo,R. and Posada,D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164–1165.
47. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
48. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7287–7292.
49. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
50. Zagryadskaya,E.I., Doyon,F. and Steinberg,S. (2003) Importance of the reverse Hoogsteen base pair 54-58 for tRNA function. *Nucleic Acids Res.*, **31**, 3946–3953.
51. Guy,M.P., Young,D.L., Payea,M.J., Zhang,X., Kon,Y., Dean,K.M., Grayhack,E.J., Mathews,D.H., Fields,S. and Phizicky,E.M. (2014) Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. *Genes Dev.*, **28**, 1721–1732.
52. 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.
53. Rose,P.W., Prlić,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB protein data bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
54. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
55. Meier,A. and Söding,J. (2015) Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. *PLoS Comput. Biol.*, **11**, e1004343.
56. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
57. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
58. Hubbard,S.J. and Thornton,J.M. (1993) *NACCESS, computer program*.
59. Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.
60. Veidenberg,A., Medlar,A. and Löytynoja,A. (2016) Wasabi: an integrated platform for evolutionary sequence analysis and data visualisation. *Mol. Biol. Evol.*, **33**, 1126–1130.
61. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
62. Kong,X.P., Onrust,R., O'Donnell,M. and Kuriyan,J. (1992) Three-dimensional structure of the beta subunit of E. coli DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell*, **69**, 425–437.

63. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to Proteopedia. *Isr. J. Chem.*, **53**, 207–216.
64. Echave,J., Spielman,S.J. and Wilke,C.O. (2016) Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.*, **17**, 109–121.
65. Guo,Y., Kalathur,R.C., Liu,Q., Kloss,B., Bruni,R., Ginter,C., Kloppmann,E., Rost,B. and Hendrickson,W.A. (2015) Structure and activity of tryptophan-rich TSPO proteins. *Science*, **347**, 551–555.
66. Gu,Y., Li,H., Dong,H., Zeng,Y., Zhang,Z., Paterson,N.G., Stansfeld,P.J., Wang,Z., Zhang,Y., Wang,W. *et al.* (2016) Structural basis of outer membrane protein insertion by the BAM complex. *Nature*, **531**, 64–69.
67. Mulligan,C., Fenollar-Ferrer,C., Fitzgerald,G.A., Vergara-Jaque,A., Kaufmann,D., Li,Y., Forrest,L.R. and Mindell,J.A. (2016) The bacterial dicarboxylate transporter VcINDY uses a two-domain elevator-type mechanism. *Nat. Struct. Mol. Biol.*, **23**, 256–263.
68. Kondapalli,K.C., Hack,A., Schushan,M., Landau,M., Ben-Tal,N. and Rao,R. (2013) Functional evaluation of autism-associated mutations in NHE9. *Nat. Commun.*, **4**, 2510.
69. Wong,T.S., Roccatano,D. and Schwaneberg,U. (2007) Steering directed protein evolution: strategies to manage combinatorial complexity of mutant libraries. *Environ. Microbiol.*, **9**, 2645–2659.
70. Levy,E.D., De,S. and Teichmann,S.A. (2012) Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 20461–20466.
71. Jékely,G. (2006) Did the last common ancestor have a biological membrane? *Biol. Direct*, **1**, 35.
72. Mayrose,I., Friedman,N. and Pupko,T. (2005) A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, **21**(Suppl. 2), ii151–ii158.
73. Mayrose,I., Mitchell,A. and Pupko,T. (2005) Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J. Mol. Evol.*, **60**, 345–353.