D308–D312 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 20 November 2015*
doi: 10.1093/nar/gkv1279

UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures

Rhonald C. Lua[1,†], Stephen J. Wilson[2,†], Daniel M. Konecki[3], Angela D. Wilkins[1,4], Eric Venner[3], Daniel H. Morgan[3] and Olivier Lichtarge[1,2,3,4,5,*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA, [2]Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA, [3]Department of Structural and Computational Biology and Molecular Biophysics, Houston, TX 77030, USA, [4]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX 77030, USA and [5]Department of Pharmacology, Baylor College of Medicine, Houston, TX 77030, USA

Received August 21, 2015; Revised October 8, 2015; Accepted November 2, 2015

ABSTRACT

The structure and function of proteins underlie most aspects of biology and their mutational perturbations often cause disease. To identify the molecular determinants of function as well as targets for drugs, it is central to characterize the important residues and how they cluster to form functional sites. The Evolutionary Trace (ET) achieves this by ranking the functional and structural importance of the protein sequence positions. ET uses evolutionary distances to estimate functional distances and correlates genotype variations with those in the fitness phenotype. Thus, ET ranks are worse for sequence positions that vary among evolutionarily closer homologs but better for positions that vary mostly among distant homologs. This approach identifies functional determinants, predicts function, guides the mutational redesign of functional and allosteric specificity, and interprets the action of coding sequence variations in proteins, people and populations. Now, the UET database offers pre-computed ET analyses for the protein structure databank, and on-the-fly analysis of any protein sequence. A web interface retrieves ET rankings of sequence positions and maps results to a structure to identify functionally important regions. This UET database integrates several ways of viewing the results on the protein sequence or structure and can be found at http://mammoth.bcm.tmc.edu/uet/.

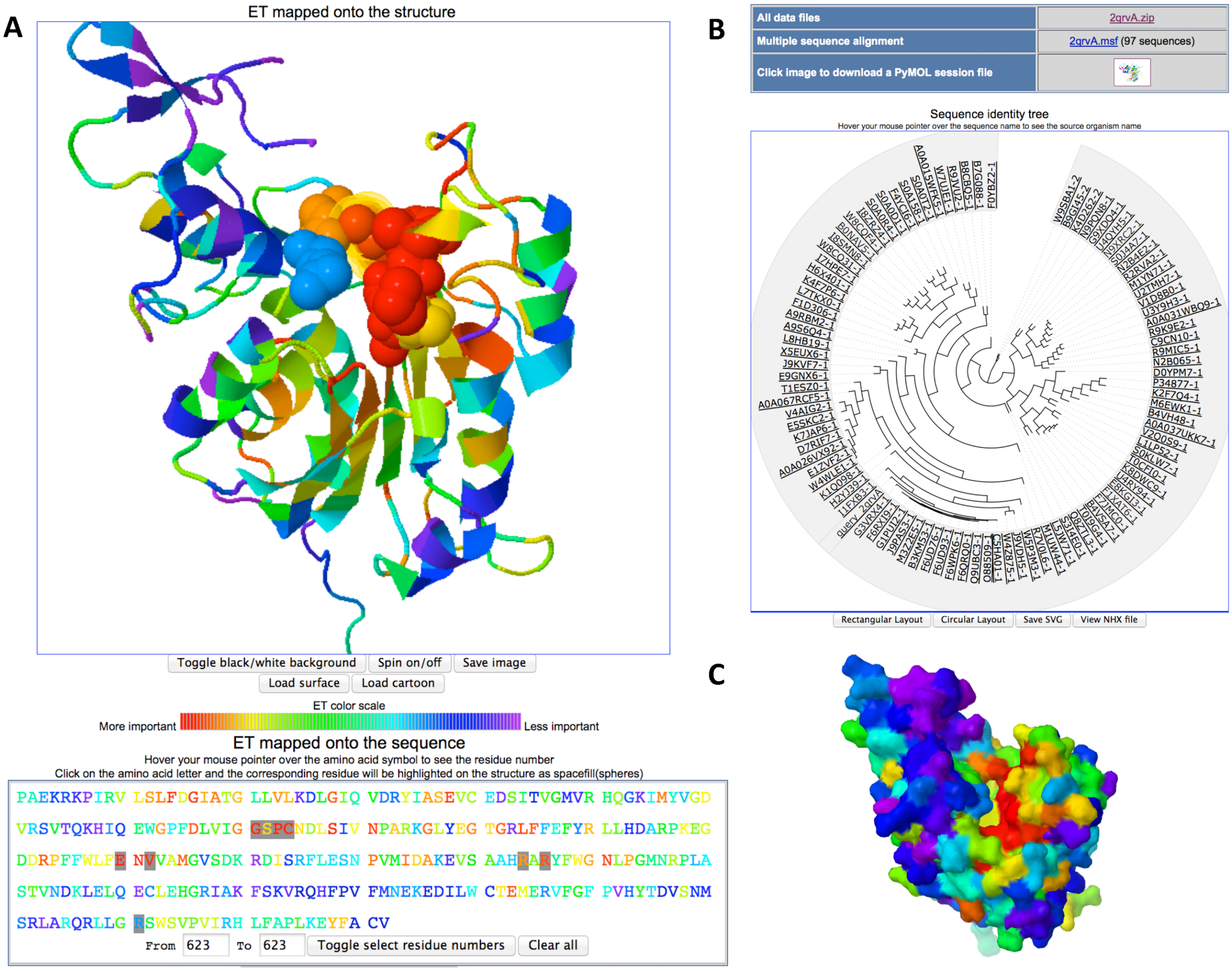
INTRODUCTION

The Evolutionary Trace (1,2) (ET) was developed as a scalable computational method to identify functionally and structurally important sequence positions. In turn, knowing the molecular determinants of protein structure and function has many critical applications across biology and medicine. For example, to guide efficient mutagenesis (3,4); interpret patient mutations (5–7); design potential therapeutic peptides (8–10); engineer separation of function in animal models (11); extract functional motifs that predict functions and substrates over the structural proteome (12–14); and measure the molecular, clinical and population-wide action of human coding variations (5,15).

ET uses the 'evolutionary record,' to establish a relative rank among sequence positions. Those positions that vary mostly among distant homologs rank ahead of positions that vary mostly among evolutionarily close homologs. Critically, top-ranked ET residues consistently exhibit useful structural and functional features: they form statistically significant clusters in native protein structures (16), they overlap extensively with known functional sites (17) and they guide mutational studies that predictably alter function as well as form general 3D functional motifs (14).

Previous public tools for ET analysis of sequence, structure and function included, first, a Java ET Viewer (18), followed by the ET report_maker (19), JEvTrace (20), TraceSuite II (21) and PyETV (22). Both the Java ET Viewer and JEvTrace combine an interactive molecular view of the structure with the multiple sequence alignment and phylogenetic tree. TraceSuite II compiles the trace results together with snapshots of the structure, sequence alignment and tree in a webpage. The report_maker presents ET analysis superimposed on information about sequence, struc-

[*]To whom correspondence should be addressed. Tel: +1 713 798 5646; Fax: +1 713 798 7773; Email: lichtarge@bcm.edu
[†]These authors contributed equally to the paper as first authors.



All data files	2qrvA.zip
Multiple sequence alignment	2qrvA.msf (97 sequences)
Click image to download a PyMOL session file	

ET mapped onto the sequence

Hover your mouse pointer over the amino acid symbol to see the residue number

Click on the amino acid letter and the corresponding residue will be highlighted on the structure as spacefill(spheres)

```
PAEKRKPIRV LSLFDGIATG LLVLKDLGIQ VDRYIASEVC EDSITVGMVR HQGKIMYVGD
VRSVTQKHIQ EWGPFDLVIG GSPCNDLSIV NPARKGLYEG TGRLFFEFYR LLHDARPKEG
DDRPFFWLFE NVVAMGVSDK RDISRFLESN PVMIDAKEVS AAHRARYFWG NLPGMNRPLA
STVNDKLELQ ECLEHGRIAK FSKVRQHFPV FMNEKEDILW CTEMERVFGF PVHYTDVSNM
SRLARQRLLG RSWSVPVIRH LFAPLKEYFA CV
```

From 623 To 623 Toggle select residue numbers Clear all

Figure 1. Example UET web browser output. (**A**) ET analysis of the DNA-binding domain of mouse DNMT3A (PDB code + chain identifier 2qrvA) (30) can be seen in the structure view with the DNA-binding site selected via the sequence view. Residues highlighted were within four Angstroms of the cytosine targeted by methylation (identified from superposition of PDB 1MHT (37)). (**B**) Sequence identity tree view and links to data files. (**C**) When the surface view is selected, a surface rendering is visible that can help highlight important surface regions, such as binding sites.

ture and elementary annotation, in a human-readable static PDF document. PyETV is an ET analysis plugin for the PyMOL molecular visualization platform (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.). However, these tools are now at least 5 years old. Modern platforms along with additional data demand an update for usability and wider applications. Moreover, some tools, like the combined PyMOL and PyETV methods of viewing ET information are more technical; they may require significant computer knowledge to implement and understand.

To facilitate access and broad use of ET analysis, we now present a new website and database called UET (Universal Evolutionary Trace). This is a repository of pre-computed ET analyses performed on the protein structure databank (PDB) (23). It can be accessed via a web interface using a given protein structure to retrieve an ET ranking of sequence positions and to identify functionally important regions in that structure. In UET, seamless integration of structure and phylogenetic tree viewers in the web browser means that a user can examine protein structures and sequences with their ET analyses, without any prior software installation (assuming the user's computer has a web browser). It also avoids access, update and digital signing issues that often plague viewers based on Java applets that run on browsers. Furthermore, tight integration with a web browser enables ET analysis to be accessible to ubiquitous mobile devices such as tablets and smart phones.

FEATURES

Inputs

UET stores ET analyses of unique protein chains in a PDB entry. In order to retrieve pre-computed ET analyses, the user is prompted for a PDB code plus a chain identifier (e.g. 2qrvA).

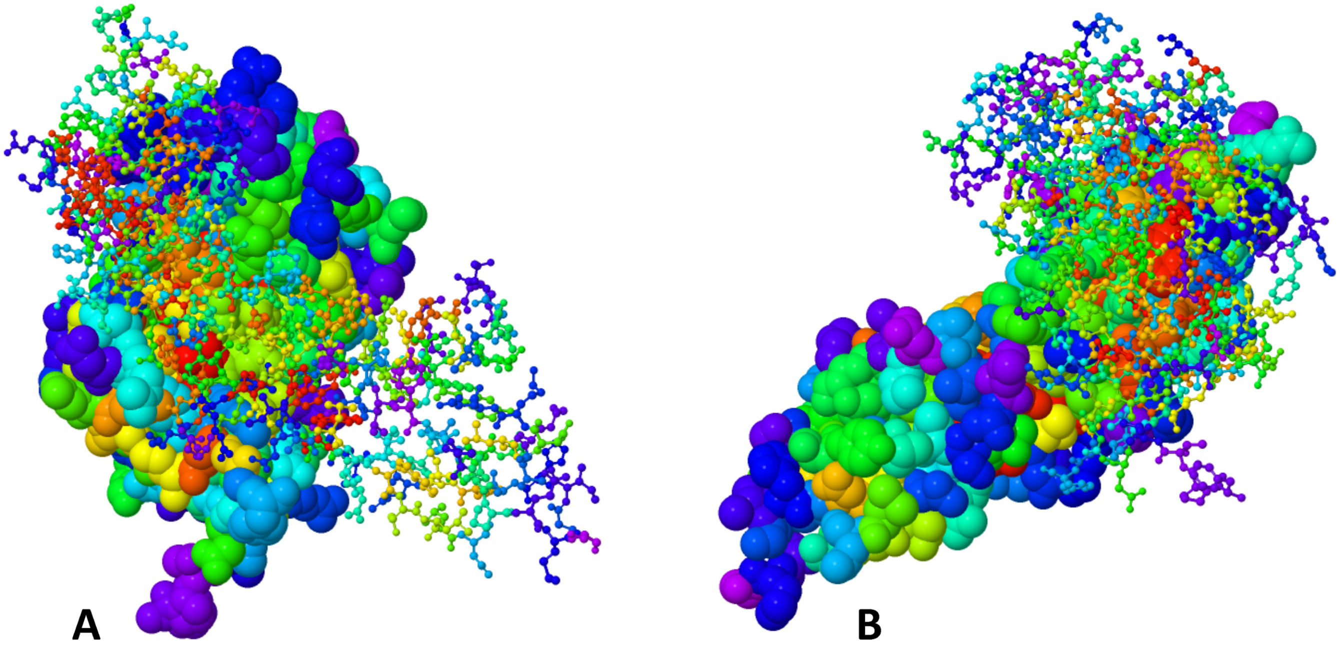


Figure 2. The human growth hormone in complex with the growth hormone receptor (PDB code: 1a22 (31)) with ET analysis. (**A**) Human growth hormone is shown in spacefill mode, while the human growth hormone receptor is shown as ball and stick. (**B**) The human growth hormone receptor is shown as a spacefill, while the human growth hormone is displayed as ball and stick.

UET accepts several other inputs for *de novo* ET analysis. This significantly expands the coverage of ET analyses to include protein sequences without representative structures in the PDB, or structures that are custom produced (such as models). To be clear, these inputs may consist of a protein sequence (specified by a UniProt (24) accession number or explicitly in FASTA format) or of a novel structure (PDB coordinates file supplied by the user, in confidentiality). Of note, the user can also tailor the multiple sequence alignment and other parameters of the ET analysis.

Structure view

To identify functional sites, the structure view shows a cartoon representation of the PDB structure, with prismatic colors that indicate the relative evolutionary importance of each residue according to its ET percentile rank (Figure 1A, red is most important and magenta is least so). The structure view exploits JSmol (25). The structure can be examined and manipulated in the usual intuitive way (left-mouse-click or double tap on a touchscreen, then drag to rotate, etc.). Placing the mouse-pointer over a residue will show its amino acid type and sequence number. The 'Load surface' option displays the protein surface making it easier to spot functional or binding sites (Figure 1C). 'Save image' lets the user save the current view of the structure into an image file. A right-mouse-click on the viewer reveals the JSmol menu with more visualization options.

Sequence view

To promptly find the most evolutionarily important residues, the sequence view presents the chain of amino acids in one-letter code. As before, the color key indicates relative evolutionary importance according to the ET percentile rank of each position (Figure 1A).

The sequence view is coupled to the structure view. Selecting an amino acid letter code by a click of the mouse (or tap on the touchscreen) causes the corresponding residue in the structure to be highlighted with a spacefill representation of the residue. An option to select a series of residues at once is also available.

Sequence identity tree view

As a guide in assessing the specificity and applicability of the predicted functional sites, the sequence identity tree used in the ET analysis is shown in a circular layout (the default, which can be switched to a rectangular layout)(Figure 1B). The tree view is provided by jsPhyloSVG (26), enhanced by a description of tree nodes using phyloXML (27). Hovering the mouse pointer over the sequence name will show the associated source organism or species. The tree view may also be saved as an SVG file, as well as in the raw tree data NHX format.

ET analysis output data files

The ET analysis data files, including files necessary to view the results in PyMOL, can be downloaded through a link on the web output. The ET analysis pipeline is described elsewhere (28).

Documentation

Multiple videos are online and show how to carry out an ET analysis with UET and with other tools. The URL is at http://www.youtube.com/user/EvolutionaryTrace.

EXAMPLES

ET has been extensively tested both in case studies and on a large scale. It identifies statistically significant clusters and functional sites within protein structures (16), and it guides the redesign of functional and allosteric sites (29). Such analyses often lead to new insights, now made more readily accessible with the release of the UET database and website interface. For example, UET of the DNA-binding domain of the DNA methyltransferase DNMT3A (PDB ID: 2qrv chain A) (30) from mouse, reveals a cluster of critically important residues immediately adjacent (4 Angstroms) to the cytosine targeted by methylation (Figure 1A). The most important residues are highlighted in red in the structure image and in the sequence mapping. Selecting these residues in the cartoon view shows that they tend to be central to the molecule, while switching to the surface view (Figure 1C) makes it apparent that they highlight a functional site. Likewise, clusters of evolutionarily important residues map the binding site between the human growth hormone and its receptor (PDB ID: 1a22 chains A and B, Figure 2) (31). Of note, ET performance can sensitively depend on the choice of parameters. Thus, a database with bulk ET analyses of all PDB structures is meant to provide a starting point for more detailed analyses, which is made possible by providing direct access to all ET parameters. Still, as is, this integrated web interface will allow other users to quickly determine a baseline generic importance of sequence positions, and often to immediately narrow their search for functional residues to target for mutational analysis of their functional roles or for redesign purposes.

CONCLUSION

UET complements existing computational and biophysical approaches (32–34) and provides simple and universal access to interpret protein structures and sequences in light of their evolutionary variations and divergences. Unlike simpler measures of residue conservation, ET explicitly correlates evolutionary substitutions with functional divergences estimated by evolutionary distances. This explicit coupling between sequence variations and fitness variations means that ET is best interpreted as a formal gradient of the evolutionary function between genotype and phenotype in the fitness landscape, an observation with important consequences (5,15,35,36). The fundamental role of this evolutionary gradient explains the myriad uses of ET in guiding predictions and rational engineering of protein functional sites, activity and binding.

ACKNOWLEDGEMENT

The authors thank Rong Yao, David M Kristensen and David Mittelman.

FUNDING

National Institutes of Health [GM079656, GM066099]; National Science Foundation [DBI-1062455]; Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia Fellowship [T15 LM007093 to S.J.W.]. Funding for open access charge: National Institutes of Health [GM079656, GM066099]; National Science Foundation [DBI-1062455].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Wilkins,A.D., Venner,E., Marciano,D.C., Erdin,S., Atri,B., Lua,R.C. and Lichtarge,O. (2013) Accounting for epistatic interactions improves the functional analysis of protein structures. *Bioinformatics*, **29**, 2714–2721.
3. Rodriguez,G.J., Yao,R., Lichtarge,O. and Wensel,T.G. (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 7787–7792.
4. Gaj,T., Mercer,A.C., Gersbach,C.A., Gordley,R.M. and Barbas,C.F. III (2011) Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 498–503.
5. Katsonis,P. and Lichtarge,O. (2014) A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res.*, **24**, 2050–2058.
6. 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.
7. Oliveira,L., Paiva,A.C. and Vriend,G. (2002) Correlated mutation analyses on very large sequence families. *Chembiochem*, **3**, 1010–1017.
8. Shoji-Kawata,S., Sumpter,R., Leveno,M., Campbell,G.R., Zou,Z., Kinch,L., Wilkins,A.D., Sun,Q., Pallauf,K., MacDuff,D. *et al.* (2013) Identification of a candidate therapeutic autophagy-inducing peptide. *Nature*, **494**, 201–206.
9. Preza,G.C., Ruchala,P., Pinon,R., Ramos,E., Qiao,B., Peralta,M.A., Sharma,S., Waring,A., Ganz,T. and Nemeth,E. (2011) Minihepcidins are rationally designed small peptides that mimic hepcidin activity in mice and may be useful for the treatment of iron overload. *J. Clin. Invest.*, **121**, 4880–4888.
10. Mitra,P., Shultis,D. and Zhang,Y. (2013) EvoDesign: de novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res.*, **41**, W273–W280.
11. Peterson,S.M., Pack,T.F., Wilkins,A.D., Urs,N.M., Urban,D.J., Bass,C.E., Lichtarge,O. and Caron,M.G. (2015) Elucidation of G-protein and beta-arrestin functional selectivity at the dopamine D2 receptor. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7097–7102.
12. Suel,G.M., Lockless,S.W., Wall,M.A. and Ranganathan,R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, **10**, 59–69.
13. Aguilar,D., Oliva,B. and Marino Buslje,C. (2012) Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. *PLoS One*, **7**, e41430.
14. Amin,S.R., Erdin,S., Ward,R.M., Lua,R.C. and Lichtarge,O. (2013) Prediction and experimental validation of enzyme substrate specificity in protein structures. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E4195–E4202.
15. Katsonis,P., Koire,A., Wilson,S.J., Hsu,T.K., Lua,R.C., Wilkins,A.D. and Lichtarge,O. (2014) Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci.*, **23**, 1650–1666.
16. Yao,H., Kristensen,D.M., Mihalek,I., Sowa,M.E., Shaw,C., Kimmel,M., Kavraki,L. and Lichtarge,O. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.
17. Wilkins,A.D., Lua,R., Erdin,S., Ward,R.M. and Lichtarge,O. (2010) Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Sci.*, **19**, 1296–1311.
18. Morgan,D.H., Kristensen,D.M., Mittelman,D. and Lichtarge,O. (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, **22**, 2049–2050.

19. 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.
20. Joachimiak,M.P. and Cohen,F.E. (2002) JEvTrace: refinement and variations of the evolutionary trace in JAVA. *Genome Biol.*, **3**, RESEARCH0077.
21. Innis,C.A., Shi,J. and Blundell,T.L. (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.*, **13**, 839–847.
22. Lua,R.C. and Lichtarge,O. (2010) PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions in protein complexes. *Bioinformatics*, **26**, 2981–2982.
23. 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.
24. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
25. 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.
26. Smits,S.A. and Ouverney,C.C. (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, **5**, e12267.
27. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
28. Wilkins,A., Erdin,S., Lua,R. and Lichtarge,O. (2012) Evolutionary trace for prediction and redesign of protein functional sites. *Methods Mol. Biol.*, **819**, 29–42.
29. Kang,H.J., Wilkins,A.D., Lichtarge,O. and Wensel,T.G. (2015) Determinants of endogenous ligand specificity divergence among metabotropic glutamate receptors. *J. Biol. Chem.*, **290**, 2870–2878.
30. Jia,D., Jurkowska,R.Z., Zhang,X., Jeltsch,A. and Cheng,X. (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, **449**, 248–251.
31. Clackson,T., Ultsch,M.H., Wells,J.A. and de Vos,A.M. (1998) Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.*, **277**, 1111–1128.
32. Meireles,L.M., Domling,A.S. and Camacho,C.J. (2010) ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. *Nucleic Acids Res.*, **38**, W407–W411.
33. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Computational alanine scanning mutagenesis–an improved methodological approach. *J. Comput. Chem.*, **28**, 644–654.
34. Sheinerman,F.B., Norel,R. and Honig,B. (2000) Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **10**, 153–159.
35. Neskey,D.M., Osman,A.A., Ow,T.J., Katsonis,P., McDonald,T., Hicks,S.C., Hsu,T.K., Pickering,C.R., Ward,A., Patel,A. *et al.* (2015) Evolutionary Action Score of TP53 Identifies High-Risk Mutations Associated with Decreased Survival and Increased Distant Metastases in Head and Neck Cancer. *Cancer Res.*, **75**, 1527–1536.
36. Osman,A.A., Neskey,D.M., Katsonis,P., Patel,A.A., Ward,A.M., Hsu,T.K., Hicks,S.C., McDonald,T.O., Ow,T.J., Alves,M.O. *et al.* (2015) Evolutionary Action Score of TP53 Coding Variants Is Predictive of Platinum Response in Head and Neck Cancer Patients. *Cancer Res.*, **75**, 1205–1215.
37. Klimasauskas,S., Kumar,S., Roberts,R.J. and Cheng,X. (1994) HhaI methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357–369.