D246–D250 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 4 November 2013*
doi:10.1093/nar/gkt1014

LenVarDB: database of length-variant protein domains

Eshita Mutt[1,2], Oommen K. Mathew[2,3] and Ramanathan Sowdhamini[2,*]

[1]International Institute of Information Technology-Hyderabad, Gachibowli, Hyderabad 500032, Andhra Pradesh, India, [2]National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road, Bangalore 560065, Karnataka, India and [3]SASTRA University, Tirumalaisamudram, Thanjavur 613401, Tamil Nadu, India

Received August 15, 2013; Revised October 5, 2013; Accepted October 7, 2013

ABSTRACT

Protein domains are functionally and structurally independent modules, which add to the functional variety of proteins. This array of functional diversity has been enabled by evolutionary changes, such as amino acid substitutions or insertions or deletions, occurring in these protein domains. Length variations (indels) can introduce changes at structural, functional and interaction levels. LenVarDB (freely available at http://caps.ncbs.res.in/lenvardb/) traces these length variations, starting from structure-based sequence alignments in our Protein Alignments organized as Structural Superfamilies (PASS2) database, across 731 structural classification of proteins (SCOP)-based protein domain superfamilies connected to 2 730 625 sequence homologues. Alignment of sequence homologues corresponding to a structural domain is available, starting from a structure-based sequence alignment of the superfamily. Orientation of the length-variant (indel) regions in protein domains can be visualized by mapping them on the structure and on the alignment. Knowledge about location of length variations within protein domains and their visual representation will be useful in predicting changes within structurally or functionally relevant sites, which may ultimately regulate protein function. Non-technical summary: Evolutionary changes bring about natural changes to proteins that may be found in many organisms. Such changes could be reflected as amino acid substitutions or insertions–deletions (indels) in protein sequences. LenVarDB is a database that provides an early overview of observed length variations that were set among 731 protein families and after examining >2 million sequences. Indels are followed up to observe if they are close to the active site such that they can affect the activity of proteins. Inclusion of such information can aid the design of bioengineering experiments.

INTRODUCTION

Protein domains are functional and compact structural units of protein, which evolve with time by incorporating various changes in the form of amino acid substitutions and insertions. Insertion/deletion of contiguous stretch of amino acids (indels) can induce structural modifications, which can eventually impart functional diversity or affect stability or differences in quaternary arrangements of these protein domains. Analysis of indels among sequence homologues of domain superfamilies will provide an early understanding of the evolutionary changes of protein function from a handful of protein folds (1).

Occurrence of indels is a continuous process and found frequently in loop regions, especially as 'nested forms' into previously inserted regions (2) or observed as substructural regions in the minimum core of protein domain scaffold, termed as 'structural embellishments' (3). Analysis of such embellishments in HIGH-signature proteins, UspA, and PP-ATPase (HUP) domain superfamily showed that indels are usually located sequentially far apart, but are spatially proximate and form subdomains, which further fine-tune the diverse functions of different members of the same superfamily (4). Recent studies by our group and others had focussed on the role played by indels in affecting structure, function (5,6) and oligomeric status (7) of a protein, while emphasizing that fixation of such indels in genome is highly context-dependent (8). Introduction of short indels within active-site loops caused emergence of other enzymatic capabilities (9,10) or led to conformational switches, as in the case of C2A domain in Piccolo protein (11). Although most of the studies have been carried out on full-length proteins, it will be interesting to trace these length variations (indels) within protein domain boundaries, given the fact that protein domains are functionally independent modules (12).

*To whom correspondence should be addressed. Tel: +91 80 23666250; Fax: +91 80 23636421; Email: mini@ncbs.res.in

LenVarDB is a database of length-variant protein domains that documents length variation statistics within homologues of each SCOP (Structural Classification of Proteins) (13) entry, considered within our PASS2 (Protein Alignments organized as Structural Superfamilies) database (14). Alignments of structural and sequence domains are annotated for their length-variant regions (indel) or 'structurally conserved blocks' (SSB) (15) in LenVarDB. Starting from 731 multi-membered PASS2, consisting of 8394 individual protein domains, non-redundant sequence database (NCBI-NR) was queried to obtain 2 730 625 sequence homologues, whose analysis in turn led to the detection of 192 742 indels (Supplementary Table S1). Unlike other indel databases (such as IndelFR (16), IndelPDB (17), IndelScan (18), which deal only with pairwise alignment to detect indels), we introduce sequences in a structure-based multiple sequence alignment, thus extracting evolutionary insight from the sequence space in detecting indel-prone regions. LenVarDB is a useful compendium to gain insight about the effects of length variations (indels) on the structure and function of a protein domain.

MATERIALS AND METHODS

Inclusion of sequence homologues and their alignment

In all, 731 superfamilies (with >2 members) from PASS2 version 4 (14) were used as an initial dataset, whose sequences (members of each superfamily) were queried against 'NCBI NR (June 2012)' database by using an in-house standardized pipeline explained elsewhere (19). This pipeline had stage-specific filters and had been optimized to encourage the collection of length-variant homologues without encountering false positives. Superfamilies were classified into groups (as length-deviant, length-rigid or length-normal) based on their extent of length variation (details in Supplementary Methods). Non-redundant homologues for each superfamily member were included in an alignment, which were importantly guided by structure-based sequence alignments obtained from PASS2, as it was more accurate than using any sequence-based alignment (20) and to reduce the random insertion of gaps within secondary structures. Length-variant (indel) regions, determined from these alignments, along with SSB obtained from conserved units of structure in proteins (CUSP) algorithm (15), were mapped on the protein domain structures (details in Supplementary Methods). Statistical tests were carried out using *R* package (21).

Database design

Web interface of LenVarDB was made interactive and efficient by using HTML, CSS, JavaScript, AJAX and JQuery. The backend services were developed using PERL, PERL-CGI and MySQL in Apache2 environment, and data analysis was performed by BioPERL modules (22) and atm2seq (23) package. Data visualization was aided by PERL-GD::Graphs module, and representation of structure and alignments was implemented using Jmol (Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/) and Jalview (24).

RESULTS AND DISCUSSION

In this study, we have traced the length variations within a protein domain using SCOP (13)-derived PASS2 and their structure-based alignments. To obtain comprehensive pointers about the role and position of length variations across superfamilies, we have introduced sequence homologues apart from their structural entries, in the respective superfamilies. LenVarDB database has been organized in a user-friendly manner, and various features that have been implemented will be helpful to a user in determining the length variation status of their query protein.

Organization of the data (browse option)

Superfamilies, along with their respective members, are organized in the following two ways: first, according to groups of length variation, and second, in accordance to SCOP classes.

Information about each superfamily

The superfamily page summarizes the length variation features computed for each superfamily and its members at the time of database creation. Length variation is calculated at two levels, one with only the structural members of superfamily [average length variation (structure)] and the other with all the sequence homologues collected [average length variation (sequence)]. The length-variant group to which the particular superfamily belongs to (according to structural entries) is noted, and information about any shuffling event (within length-variant groups) is mentioned. A separate panel enlists each member and its length, which is also hyperlinked to 'Member page'. Dynamically created graphs are provided for easy visualization of the data provided. Length variation status of each homologue and its contribution in maintaining or changing the length variation status of the superfamily is featured in an interactive scatter plot. The relation between length variation and sequence identity (%) of every homologue collected for every superfamily member is represented by a heat map. Fluctuations in length, captured in the form of amplification/shrinkage from the domain length, are shown for all members in the form of a stacked bar graph. An estimate of hypothetical proteins and splice variants (found in the homologue sets) is also provided (Figure 1). To view the distribution of homologues from a taxonomy perspective, a 'Taxonomy tab' has been created to classify the homologues as per their TaxID (NCBI) and annotate average and range of length variation accumulated for the major domains of life, namely, archaea, bacteria and eukaryotes (Supplementary Data).

Descriptions of every superfamily member

Length variations, recorded across the homologues of a certain member in the form of indels (please see Supplementary Methods), are mapped on the multiple sequence alignment and 3D structure. SSBs (from CUSP algorithm) are also marked in a similar manner (Figure 1B). Users can interactively visualize the length-variant regions on the 3D structure of a protein domain.

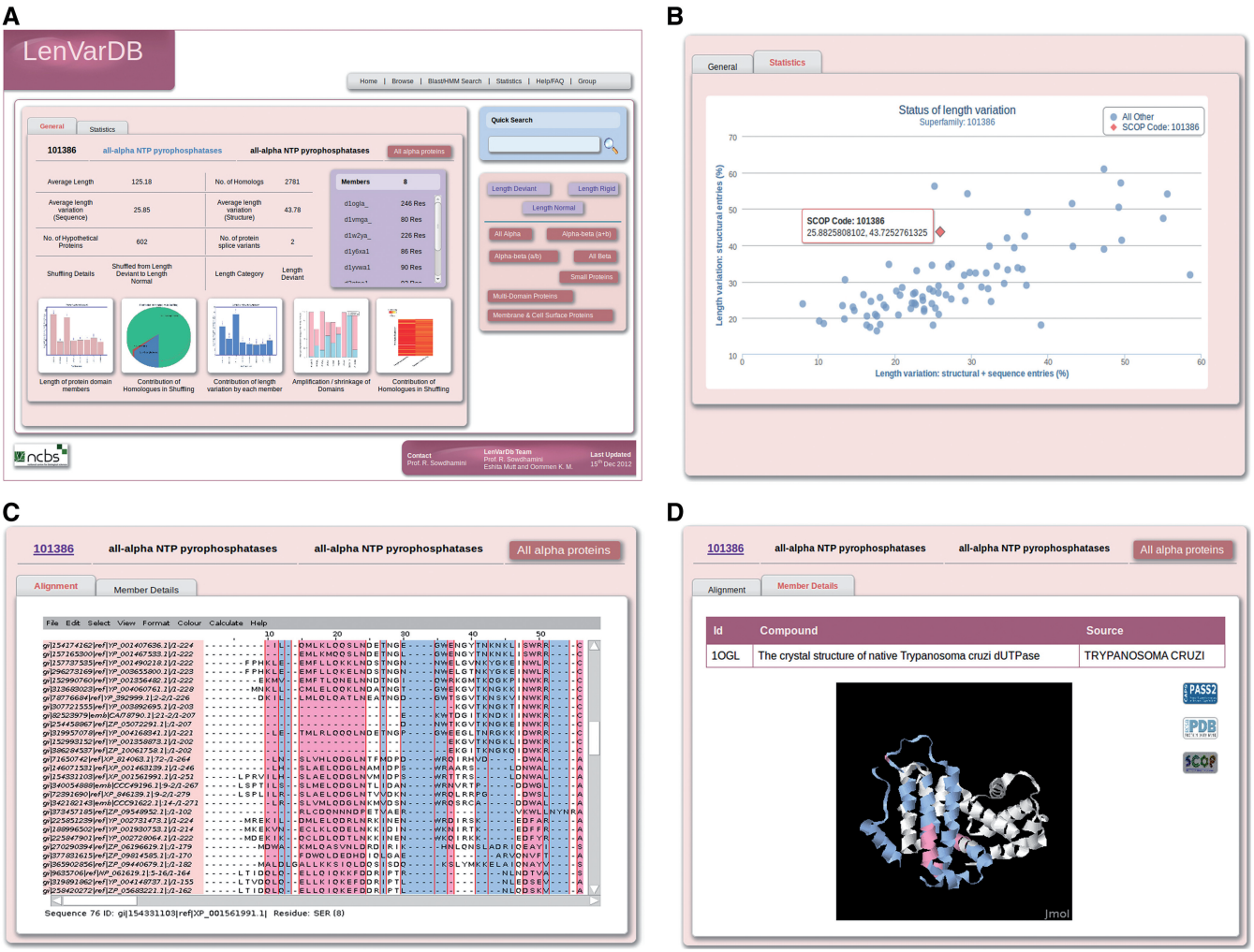


Figure 1. Illustration of the features of the database. (**A**) Length variation details at superfamily level, which have been comprehensively described by the graphs. (**B**) Correlation plot for finding length-variant status of a particular superfamily in the length-deviant group (as shown here). (**C**) Multiple sequence alignment annotated by length-variant regions (indels in blue) and SSB (in pink). (**D**) Details of superfamily member and the length-variant regions and SSBs marked over the member structure to aid visualization in spatial format. Cross-references to PDB, PASS2 and SCOP for each entry are provided.

Each superfamily member is also cross-linked to widely known databases like Protein Data Bank (25), PASS2 and SCOP for detailed structural and domain level information.

Searching and aligning query of interest to superfamily

Keyword search option has been integrated in the home page for easy access to any superfamily or its member. An amino acid sequence of a query protein can be used to search for the nearest superfamily by the integrated PSI-BLAST and HMMER programs. Further, the query can be aligned with the pre-aligned set of homologues corresponding to superfamily members.

Accessing information as downloads

Length-variance information about each superfamily member and their homologues is available for download. For each member, multiple sequence alignment with their homologues, Genbank identifier list of hypothetical proteins and listing of length variation (%) and sequence identity (%) of every homologue with respect to the superfamily member can be accessed.

Assistance for users

A user-friendly help page has been created to guide the access and features of the database. A video tour of the database has also been embedded in the home page for the same purpose [Supplementary Movie S1 (http://caps.ncbs.res.in/download/lenvardb/)]. Frequently asked questions enlisted will enable users in analysing their query proteins from a length variation perspective.

Statistical analysis of LenVarDB

Length variation parameter has been used for statistical analyses (as shown in Supplementary Data) and integrated as 'Statistics' page in the database.

Presence of insert domain has been analysed and found to occupy only 2% of the PASS2 starting dataset (8970

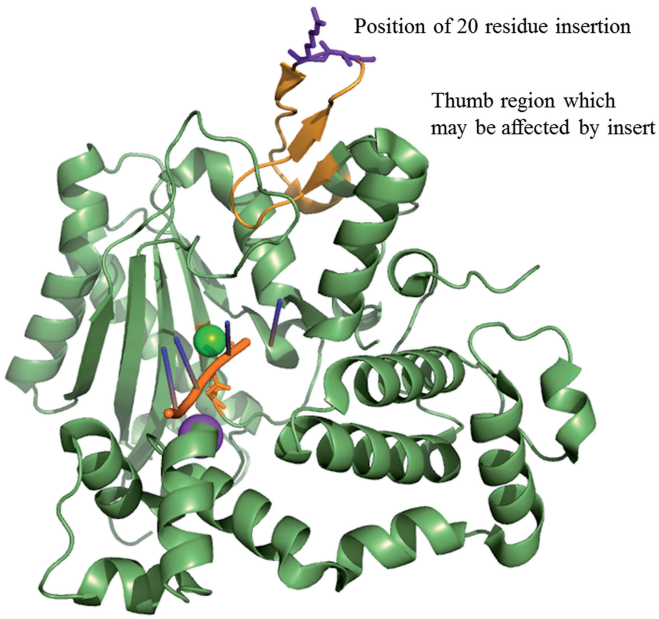
superfamily members), which gives the user an insight about one of the causes of length variations in LenVarDB (Details in Supplementary Table S2 and Supplementary Figure S6). The range of length variation [shown as box plots to represent the range of length variation (%)] in eukaryotic homologues is slightly higher than bacterial and archaeal homologues (Supplementary Figure S8).

APPLICATION OF THE DATABASE

Detection of multiple length variations (indels) in their spatial orientation within a protein domain structure can be useful in focusing on the deviations caused from its usual structure and function. Study of length variations in a homologue (*Geobacillus S. Y412MC61* dUTPase [gi|261418089] from all-alpha NTP pyrophosphatase SCOP superfamily) of *Campylobacter jejuni* dUTPase reveals the presence of a 20 residue-insert near the substrate binding domain and dimerization interface, which can regulate its functionality or oligomerization status (Supplementary Data). Further biochemical characterization can show if the presence of these indels (in the homologue) interferes with the pathogenic nature of this dUTPase in *C. jejuni*, which is known to be a gastric pathogen.

Another use of LenVarDB can be in tracking length-variant regions in newly sequenced proteins. An example of aligning and structure-mapping of an isoform of terminal deoxynucleotidyl transferase (gi|112734847 from deoxynucleotidyl transferase superfamily) is illustrated (Figure 2). Literature suggests that the indel found near the carboxy-terminal end of this domain significantly alters its function (26).

LenVarDB is a comprehensive resource designed for researchers interested in protein domain biology. This database systematically and automatically gathers sequences, aligns them to pre-existing structure-guided PASS2 alignments and derives indel (length-variant)-prone regions of protein domain superfamilies. The web interface has been made interactive and readily available for the scientific community. The key advantage of LenVarDB over other indel databases is that it introduces evolutionary insights from all the closely and distantly related sequence homologues. LenVarDB allows users to identify the indel-prone regions, which can enable rational design of bioengineering experiments.



Terminal deoxynucleotidyl transferase [PDB: 1jms]

Figure 2. Position of insertion and deletion traced on mouse terminal deoxynucleotidyl transferase homologue can aid in structural and modelling studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [27–32].

ACKNOWLEDGEMENTS

E.M. is supported by DBT-BINC grant funded by Department of Biotechnology, India. O.K.M. thanks the Vice Chancellor of SASTRA University for encouragement and support. The authors also thank NCBS (TIFR) for infrastructural support and Dr N. Srinivasan and Dr Abhijit Mitra for helpful discussions.

FUNDING

Funding for open access charge: Centre for Excellence [BT/01/09] supported by Department of Biotechnology, India.

Conflict of interest statement. None declared.

REFERENCES

1. Chothia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
2. Jiang,H. and Blouin,C. (2007) Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics*, **8**, 444.
3. Reeves,G.A., Dallman,T.J., Redfern,O.C., Akpor,A. and Orengo,C.A. (2006) Structural diversity of domain superfamilies in the CATH Database. *J. Mol. Biol.*, **360**, 725–741.
4. Dessailly,B.H., Redfern,O.C., Cuff,A.L. and Orengo,C.A. (2010) Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification. *Structure*, **18**, 1522–1535.
5. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
6. Sandhya,S., Rani,S.S., Pankaj,B., Govind,M.K., Offmann,B., Srinivasan,N. and Sowdhamini,R. (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS One*, **4**, e4981.
7. Hashimoto,K. and Panchenko,A.R. (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl Acad. Sci.*, **107**, 20352–20357.
8. Studer,R.A., Dessailly,B.H. and Orengo,C.A. (2013) Residue mutations and their impact on protein structure and function:

detecting beneficial and pathogenic changes. *Biochem. J.*, **449**, 581–594.
9. Afriat-Jurnou,L., Jackson,C.J. and Tawfik,D.S. (2012) Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry*, **51**, 6047–6055.
10. Davies,G.J., Brzozowski,A.M., Dauter,M., Varrot,A. and Schülein,M. (2000) Structure and function of humicola insolens family 6 cellulases: structure of the endoglucanase, Cel6B, at 1.6 A resolution. *Biochem. J.*, **348**, 201.
11. Garcia,J., Gerber,S.H., Sugita,S., Südhof,T.C. and Rizo,J. (2004) A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat. Struct. Mol. Biol.*, **11**, 45–53.
12. Song,B., Gold,B., O'hUigin,C., Javanbakht,H., Li,X., Stremlau,M., Winkler,C., Dean,M. and Sodroski,J. (2005) The B30.2(SPRY) domain of the retroviral restriction factor TRIM5? Exhibits lineage-specific length and sequence variation in primates. *J. Virol.*, **79**, 6111–6121.
13. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
14. Gandhimathi,A., Nair,A.G. and Sowdhamini,R. (2011) PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic Acids Res.*, **40**, D531–D534.
15. Sandhya,S., Pankaj,B., Govind,M.K., Offmann,B., Srinivasan,N. and Sowdhamini,R. (2008) CUSP: an algorithm to distinguish structurally conserved and unconserved regions in protein domain alignments and its application in the study of large length variations. *BMC Struct. Biol.*, **8**, 28.
16. Zhang,Z., Xing,C., Wang,L., Gong,B. and Liu,H. (2011) IndelFR: a database of indels in protein structures and their flanking regions. *Nucleic Acids Res.*, **40**, D512–D518.
17. Hsing,M. and Cherkasov,A. (2008) Indel PDB: A database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics*, **9**, 293.
18. Chen,F.-C., Chen,C.-J. and Chuang,T.-J. (2007) INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events. *Nucleic Acids Res.*, **35**, W633–W638.
19. Mutt,E., Mitra,A. and Sowdhamini,R. (2011) Search for protein sequence homologues that display considerable domain length variations. *Int. J. Knowl. Discov. Bioinformatics*, **2**, 55–77.
20. Zhang,Z., Wang,Y., Wang,L. and Gao,P. (2010) The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One*, **5**, e14316.
21. Development Core Team,R. (2005) *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
22. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
23. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
24. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
25. 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.
26. Bentolila,L.A., Fanton d'Andon,M., Nguyen,Q.T., Martinez,O., Rougeon,F. and Doyen,N. (1995) The two isoforms of mouse terminal deoxynucleotidyl transferase differ in both the ability to add N regions and subcellular localization. *EMBO J.*, **14**, 4221–4229.
27. Altschul,S.F., Madden,T.L., Schäffer,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.
28. Eddy,S.R. (1996) Hidden markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
29. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
30. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
31. Toth-Petroczy,A. and Tawfik,D.S. (2013) Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol. Biol. Evol.*, **30**, 761–771.
32. Moroz,O.V., Harkiolaki,M., Galperin,M.Y., Vagin,A.A., González-Pacanowska,D. and Wilson,K.S. (2004) The crystal structure of a complex of campylobacter jejuni dUTPase with substrate analogue sheds light on the mechanism and suggests the 'Basic Module' for dimeric d(C/U)TPases. *J. Mol. Biol.*, **342**, 1583–1597.