D512–D518 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 29 November 2011*
doi:10.1093/nar/gkr1107

IndelFR: a database of indels in protein structures and their flanking regions

Zheng Zhang[1], Cheng Xing[2], Lushan Wang[1], Bin Gong[2,*] and Hui Liu[2]

[1]State Key Laboratory of Microbial Technology and [2]School of Computer Science and Technology, Shandong University, Jinan 250100, China

Received August 14, 2011; Revised October 1, 2011; Accepted November 6, 2011

ABSTRACT

Insertion/deletion (indel) is one of the most common methods of protein sequence variation. Recent studies showed that indels could affect their flanking regions and they are important for protein function and evolution. Here, we describe the Indel Flanking Region Database (IndelFR, http://indel.bioinfo.sdu.edu.cn), which provides sequence and structure information about indels and their flanking regions in known protein domains. The indels were obtained through the pairwise alignment of homologous structures in SCOP superfamilies. The IndelFR database contains 2 925 017 indels with flanking regions extracted from 373 402 structural alignment pairs of 12 573 non-redundant domains from 1053 superfamilies. IndelFR provides access to information about indels and their flanking regions, including amino acid sequences, lengths, locations, secondary structure constitutions, hydrophilicity/hydrophobicity, domain information, 3D structures and so on. IndelFR has already been used for molecular evolution studies and may help to promote future functional studies of indels and their flanking regions.

INTRODUCTION

Insertions/deletions (indels) and amino acid substitutions are two of the most common forms of protein sequence variations, which enables the evolution of protein structures (1,2). Although the frequency of indels is lower than that of substitutions in the genome (3), indels are considered to lead to most of the differences among species (4) and to possibly be related to many human diseases (5–8). Conserved indels in signature proteins could be considered as phylogenetic markers to discern the course of evolutionary history (9,10). However, there are few studies of the sequence–structure–function relationships of indels.

Indels usually occur as reverse turns or coils within loops in a domain, most often on the surface of proteins (11–13). Compared to deletions, a succession of insertions and rapid evolution is considered to be a reasonable process that could produce novel protein structures (14,15). Ancient domain families show some bias toward insertions which grow in size in evolution (16). Indels likely occurred more often in essential proteins and proteins that highly interacted with others (17). These indels located on the interaction interfaces are significant for protein–protein interactions (18–20). The functional divergence between homologous proteins may also be caused by indels that occurred in the periphery of the conserved protein structure core (21,22).

Recent studies showed a significantly increased degree of nucleotide divergence between indel-flanking regions in a genome, indicating that indels can induce a rise in the substitution rate in the flanking regions (23,24). The impacts of indels on their flanking regions may play an important role in molecular evolution. In addition, owing to the solvent accessibility environments and pairwise amino acid interactions, protein tertiary structure also was considered to impact to a certain degree on molecular evolution (25,26). Our previous study also showed the impact of indels on their flanking regions in a domain, including the shift of flanking structure, the destruction of secondary structure elements, the increased amino acid sequence substitution rate, etc. (27). Indels in regions under lower selection pressure more commonly survive within a domain. Additionally, the occurrence of an indel can further lower the selection pressure on its flanking regions. Therefore, the information contained in indel-flanking regions is also important to the study of indels.

Here, we constructed IndelFR (Indel Flanking Region Database, http://indel.bioinfo.sdu.edu.cn), a database of

*To whom correspondence should be addressed. Tel: +86 531 88391261; Fax: +86 531 88390059; Email: gb@sdu.edu.cn

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

information regarding indels and their flanking regions within homologous domains. IndelFR contains sequence and structure information of 2 925 017 indels and their flanking regions, including their position, length, amino acid composition, secondary structure information, hydrophilicity/hydrophobicity, etc. The classification of homologous domains in IndelFR is based on SCOP superfamilies (28). Structure files were obtained from the ASTRAL95 non-redundant structural database (29). For IndelFR, indels and their flanking regions were extracted from structure-based sequence alignments between homologous non-redundant domains by an online alignment program, PDBeFold (30).

The IndelFR database enables rapid searching for information about indels and their flanking regions between any two known homologous non-redundant domains. This will facilitate studies involving the structural and functional analyses of protein indels. IndelFR may promote future studies of the functions of indels in protein structural evolution and functional divergence, as well as homology modeling and functional site analyses toward improving protein structures.

DATA COLLECTION AND DATABASE CONSTRUCTION

To create the IndelFR database, we collected information about indels and their flanking regions between any two non-redundant proteins within a protein structure superfamily. Proteins in the same superfamily are usually considered to be homologous; the data about superfamilies were obtained from the structural classification database SCOP 1.73 (28). The IndelFR database contains all of the superfamilies that have two or more non-redundant structures in the first five SCOP classes. The five SCOP classes include: all alpha proteins, all beta proteins, alpha and beta proteins (a/b), alpha and beta proteins (a+b) and multi-domain proteins. Due to its enormous size, the immunoglobulin superfamily is temporarily excluded from the current version of the IndelFR database. The data regarding non-redundant protein domains in each superfamily was obtained from the ASTRAL95 database, in which the percent sequence identity between any two structures is always <95% (29).

The indels and their flanking regions within a superfamily were obtained through pairwise alignment between non-redundant protein domains (Figure 1). Sequence alignment based on structure was performed using the online alignment program PDBeFold (formerly SSM, http://www.ebi.ac.uk/msd-srv/ssm/) (30). PDBeFold is currently considered to be one of the best structure alignment programs (31). All alignment results were downloaded, and each alignment was called a match. Statistical significance of a match can be evaluated by *P*-score and *Z*-score. *P*-score is minus logarithm of the *P*-value, which is an estimate of the probability of achieving the same or better quality of match at random picking of a structure from the non-redundant database. While *Z*-score represents the statistical significance of a match in terms of Gaussian statistics. The higher *P*-score and *Z*-score is, the higher statistical significance of a match will be.

Some matches contain one or more sections that cannot be aligned on amino acid sequences between two domains. Here, gaps only refer to those that are consecutive sequences (containing one or more residues) in one domain and consecutive spaces of equal length in the other domain, located between two sections of aligned residues. All of the gaps and their flanking alignment regions were extracted from the matches using our own program. To avoid duplication, the alignment region between two adjacent gaps was equally divided into two parts. The upper limit for the number of residues extracted from the unilateral flanking region of a gap was customized to 10.

There are often missing sequence sections in some protein domains. Those gaps caused by the missing sections in a match were eliminated. All of the gaps after screening are considered as indels in domain. In total, we conducted 373 402 structural alignments between 12 573 non-redundant domains from 1053 superfamilies using the above-mentioned method, and extracted 2 925 017 indels with flanking regions. Furthermore, every indel with flanking regions was presented with a series of annotations, including the location, length and composition of indels, the length and composition of flanking regions, the status and evaluation of structural alignments, etc.

The IndelFR database was created based on the data mentioned above. The detailed process of database creation included nine steps (see Supplementary Figure S1). A single entry was established for every match and every indel in the IndelFR database. We introduced the superfamily-match-indel relationship, and included this into the tree-graph catalog according to SCOP classification. In addition, protein PDB files were also included in the IndelFR database for three-dimensional structural displaying. All of the information is stored and managed by an open-source database management system, MySQL, which allows rapid data retrieval. An Apache tomcat web server was set up on a node of the Langchao TS10000 cluster, and GridSphere portal framework was utilized to establish and provide data access.

USER INTERFACE DESIGN

Browse

In the IndelFR database, indels with flanking regions are classified according to their SCOP superfamilies. In the 'SCOP Tree' interface, users can explore any of the superfamilies through the entire SCOP Tree or five subtrees corresponding to the five structure classes (Figure 2A). In each leaf node of the tree, two links are provided along with the name of the superfamily: one for the 'Indel Information' interface, and the other for the 'Match Information' interface. Users can browse all of the indels or matches in a superfamily through the two links. The number of matches or indels in a superfamily is shown in parentheses (Figure 2A).

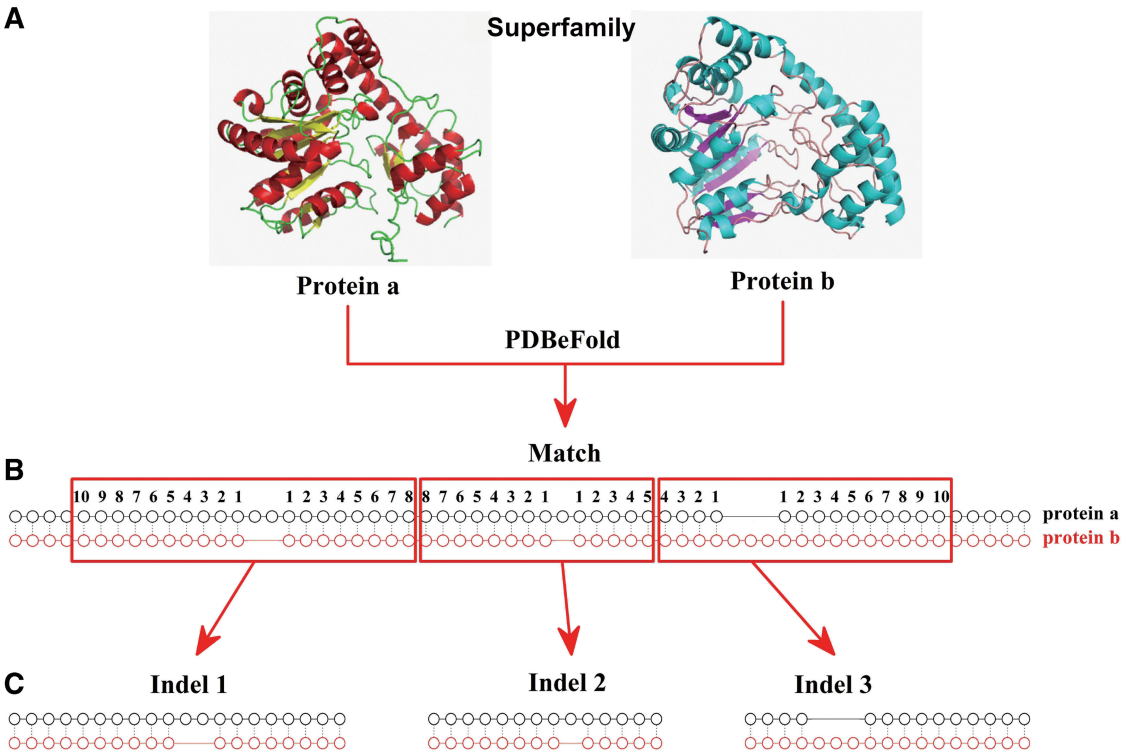

Figure 1. Data collection for IndelFR database. (**A**) Selection of superfamilies and non-redundant protein domains. (**B**) Structural alignment by PDBeFold and generation of match files. (**C**) Locating and extracting indels and their flanking regions from matches.

In the 'Indel Information' interface, each indel is shown as an independent entry (Figure 2B). Users can adjust the number of entries displayed on each page. These entries are arranged with an alternately blue or gray background according to the different matches that the indels belong to. Each entry contains the following information (see Supplementary Table S1): the serial number of the superfamily, the SCOP sids (SCOP domain identifier) and common names (selectable by users) of the query and target structures, the corresponding position of the indel on the query and target structures, the length of the indel, the length of its flanking regions, *P*-score and *Z*-score. Mouseover will give out corresponding explanation for each table heading. Detailed information about an indel and its flanking regions can be browsed online or downloaded as a text file from the right side of an entry (Figure 2D). This information includes amino acid composition, secondary structure composition, hydrophilicity/hydrophobicity, status of alignment, etc. In addition, the 3D query and target structures can be browsed online through Jmol.

In the 'Match Information' interface, each match is displayed as an independent entry (Figure 2C). Every entry includes the following information (see Supplementary Table S2): the serial number of the superfamily, the information about two structures, and the alignment analysis. The information about the query and target structures includes their respective SCOP sids and common names, lengths and the numbers of secondary structure elements. The alignment analysis consists of the alignment length, RMSD (root mean square deviation), the number of secondary structure elements aligned (Aligned SSEs), *P*-score, *Z*-score and the number of indels. Users can access the 'Indel Information' interface and obtain information about all of the indels in a match by clicking on the link behind the figure in the 'Number of Indels' column. Detailed information about match files can be browsed online by clicking the link on the right side.

Search

In addition to browsing indels and matches in superfamilies through the 'SCOP Tree' interface, users are allowed to retrieve indels and matches. The IndelFR database provides two indel retrieval approaches. One is to search directly by the indel information in the 'Indel Search' interface; the other is to search by the information about the matches that contain indels in the 'Match Search' interface. In order to make it more convenient for users to search the database, several searching methods are given for each retrieval approach.

In the 'Indel Search' interface, four searching methods are provided for the first approach (Figure 2E). The first one is Basic Search, which allows users to input one or more PDB IDs (four characters) or SCOP sids (seven characters) separated by commas or semicolons and search for all corresponding indels. Moreover, in Basic Search, three selectable parameters (Indel length, *P*-score and *Z*-score) are given and users can set their own required value ranges. The second one is Advanced Search, which requires users to input two or more PDB IDs or SCOP sids to search for indels between specified structures. The third one is Common Name Search.

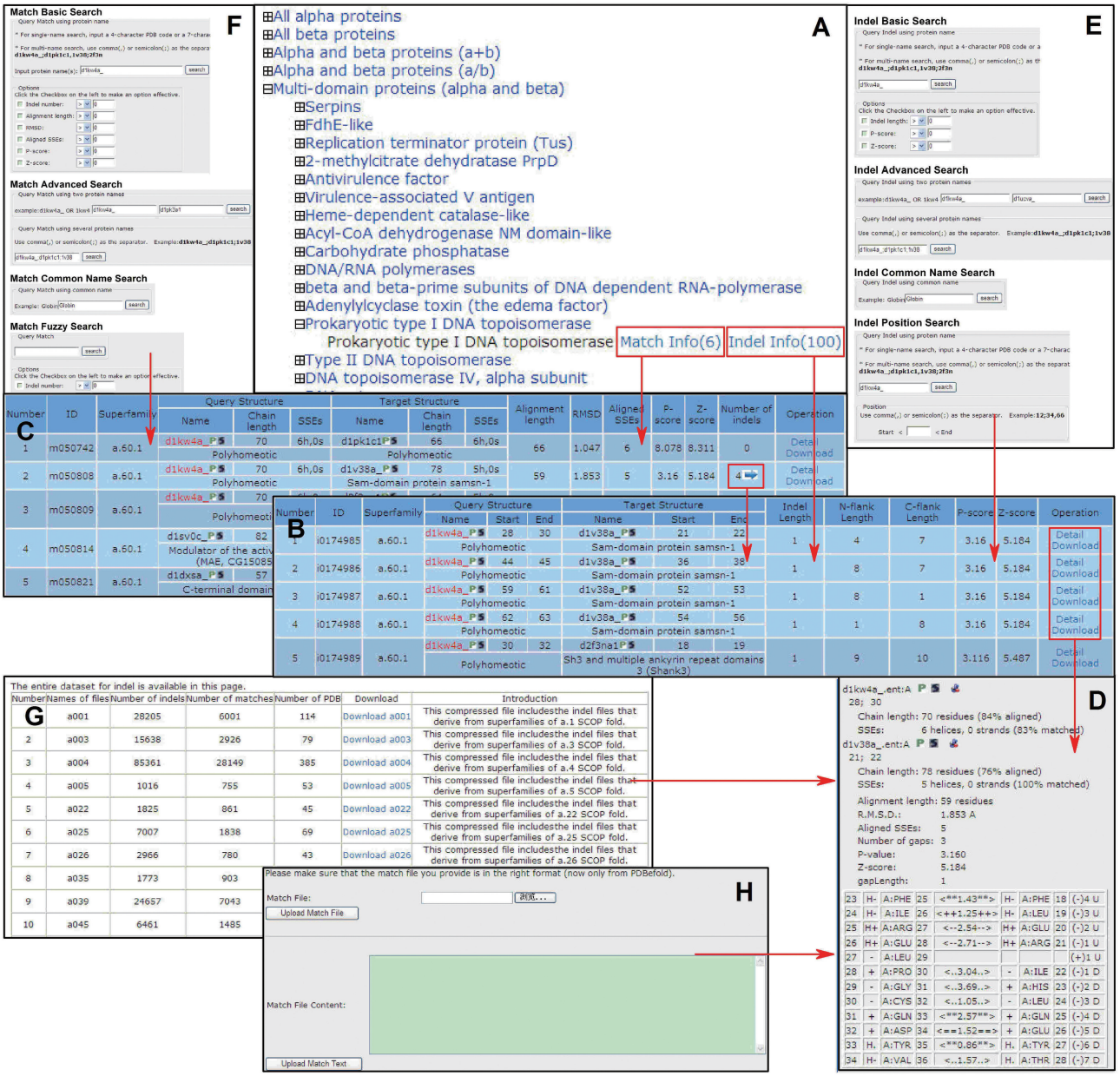

Figure 2. Samples of the IndelFR interfaces. (**A**) The 'SCOP Tree' interface. All indels and matches in the superfamilies can be browsed through this interface. (**B**) The 'Indel Information' interface. The indel files can be browsed online and downloaded. (**C**) The 'Match Information' interface. The match files can be browsed online and downloaded. (**D**) The format of indel files, containing detailed information about indels and their flanking regions. (**E**) The 'Indel Search' interface. Target indels can be searched in three ways. (**F**) The 'Match Search' interface. Target matches can be searched in four ways. (**G**) The 'Download' interface. Users can download the entire dataset stored in the database in this interface. (**H**) The 'Online Indel Creation' interface. Users can submit match files and extract indels and their flanking regions in this interface.

Users can input the full or part of a protein common name and search corresponding non-redundant domains for indel information in these domains. The forth one is Position Search. In this method, users can input a bunch of PBD IDs or SCOP sids and positions to search for indels located in these positions if there are any. For each method, the search result is displayed in the 'Indel Information' interface. Additionally, users can sort the results according to different parameters to satisfy their specific demand. If the search results include only one item, the indel file will be automatically shown together with the search result, which facilitates users to directly view corresponding information.

In the 'Match Search' interface, we also provided four methods (Figure 2F). The first three methods are similar to those in the first approach. Differently, the Basic Search using match information has six selectable parameters: Indel number, Alignment length, RMSD, Aligned SSEs, *P*-score and *Z*-score. The fourth method is Fuzzy Search, in which any keyword including a null value can be

acceptable, and the users can retrieve the match records by combining different parameters. The results will be shown in the 'Match Information' interface.

Downloads and online indel creation

Batch download of the entire dataset stored in IndelFR database can be achieved in the 'Download' interface (Figure 2G). Each indel together with its flanking regions is saved as a text file. All the files are compressed to different packages according to SCOP classification.

If information about unsaved indels and their flanking regions are required, users can obtain them in the 'Online Indel Creation' interface (Figure 2H). Both files and text are accepted for submitting matches to IndelFR.

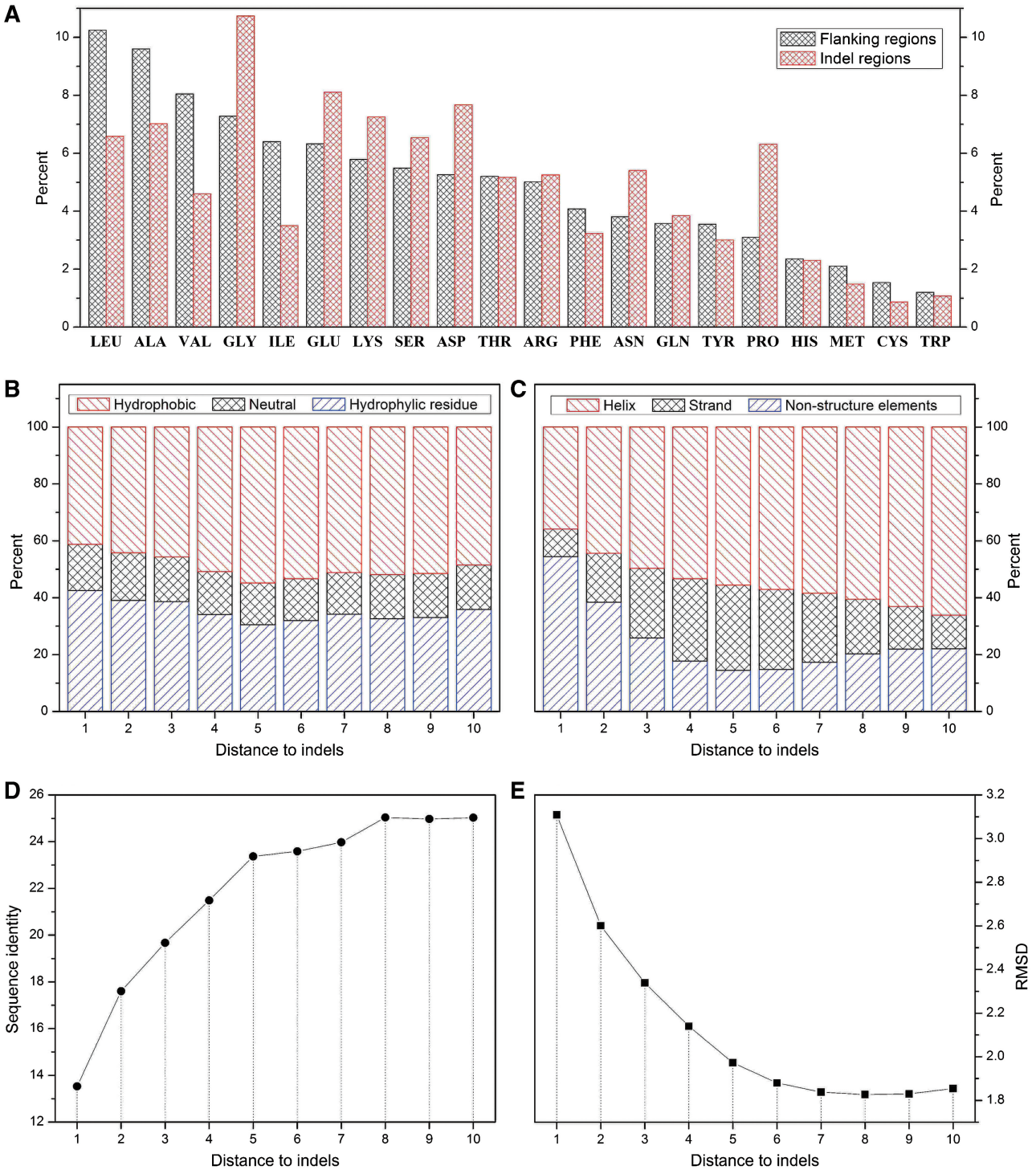

Figure 3. Display of some special qualities of indels and their flanking regions in a protein domain using data in IndelFR. (**A**) Comparison of amino acid composition between indel regions and flanking regions. (**B**) Amino acid hydrophilicity/hydrophobicity of indel flank sites. (**C**) Secondary structure element composition of indel flank sites. (**D**) Sequence identity of indel flank sites. (**E**) Tertiary structure shift of indel flank sites.

The search results, which include information about all of the indels in the match, will be successively shown in the result interface and can be downloaded. Currently, this function can only parse the match files created by PDBeFold, so users first need to conduct sequence alignment based on structure by PDBeFold with two specified domains.

DISCUSSION

It is already known that the indel region of a protein domain has a bias for amino acid usage and secondary structure element composition (11–13). Intriguingly, the flanking regions of an indel also have some special qualities that are different from other regions. Utilizing the data stored in IndelFR database, we can display some of these qualities (Figure 3). Since longer flanking regions can reduce the superposition influence of two adjacent indels, the indels utilized in those analyses mentioned above include 10 aligned sites on either side of the flanking regions.

In indel regions, the usage of amino acid is significantly different from that of their flanking regions (Figure 3A). Residues with hydrophilic side chains are more likely to occur in indel regions, such as ASP, ASN, GLU, LYS, SER, GLN, ARG, THR and HIS. Besides, compared to flanking regions, the occurrence rate of four kinds of amino acids i.e. PRO, GLY, ASP and ASN, in the indel regions increase the most, all of which have the potential to damage the secondary structure of proteins.

For flanking regions, some interesting qualities are more likely to occur in the sites nearer indels. Some of these qualities are stated below. (i) The nearer a flank site to an indel is, the larger the probability of hydrophilic amino acid usage is (Figure 3B). (ii) The nearer a flank site to an indel is, the lower the rate of the site occurring in an α-helix or a β-strand while the higher the rate of the site occurring in a non-secondary structure element is (Figure 3C). (iii) The nearer a flank site to an indel is, the lower the sequence identity is, which indicates more amino acid substitution (Figure 3D). (iv) The nearer a flank site to an indel is, the larger the tertiary structure shift between the site and its homologous non-indel flank site is. Here the structure shift is indicated by RMSD between Cα atoms (Figure 3E).

The special qualities occurred in the indel flanking regions are obviously related to indels. These qualities may consist of three parts: those that were regional-inherent, those that ocurred accompanying an indel and those that occurred after an indel. In our previous study, we analyzed and estimated their respective effects (27). The IndelFR database may contribute to promote future study on this problem.

FUTURE DIRECTIONS

The IndelFR database is a free network-based resource, storing vast information about indels and their flanking regions of protein domains. This large amount of data requires a large amount of storage space and complicated management, and it also makes data transfer difficult. The publication of the current version is an important first step.

The data set in current version of IndelFR is based on pairwise alignment, which contains all of the indels between any known non-redundant homologous domains. In the future, we plan to provide conserved indels within one SCOP superfamily by multiple sequence alignment based on structure. However, one superfamily may contain many distant-related homologous proteins, so they may be quite different from each other in both sequence and structure. Consequently, current algorithms for multiple sequence alignment based on structure may not be able to conduct multiple sequence alignment in every superfamily.

In the future, we will keep updating our database following SCOP updates. In order to facilitate users, we are considering adding Sequence BLAST Search in the future and enabling Indel Fuzzy Search that will allow users to search for target indels from millions of indels in the IndelFR database. In addition, we plan to open a parameter selection for the online indel-extraction program, so that users can extract indels and their flanking regions according to their own requirements.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S2, Supplementary Figure S1.

ACKNOWLEDGEMENTS

The authors are grateful to Xiaoyun Duan, Jinlan Wang, Yangyang Liu and Jie Huang for assistance with IndelFR database design and construction. They also thank three anonymous reviewers for their comments and suggestions.

FUNDING

Independent Innovation Foundation of Shandong University (2009JC006); National Natural Science Foundation of China (30970092 and 61070017) and Scientific Research Reward Fund for excellent Young and Middle-Aged scientists in Shandong Province (20090451326). Funding for open access charge: National Natural Science Foundation of China (61070017).

Conflict of interest statement. None declared.

REFERENCES

1. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
2. 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.
3. Chen,J.Q., Wu,Y., Yang,H., Bergelson,J., Kreitman,M. and Tian,D. (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.*, **26**, 1523–1531.

4. Britten,R.J., Rowen,L., Williams,J. and Cameron,R.A. (2003) Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl Acad Sci. USA*, **100**, 4661–4665.
5. Ashley,C.T. Jr and Warren,S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.
6. Zielenski,J. and Tsui,L.C. (1995) Cystic fibrosis: genotypic and phenotypic variations. *Annu. Rev. Genet.*, **29**, 777–807.
7. Zoghbi,H.Y. and Orr,H.T. (2000) Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.*, **23**, 217–247.
8. Duval,A. and Hamelin,R. (2002) Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. *Cancer Res.*, **62**, 2447–2454.
9. Gupta,R.S. (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.*, **62**, 1435–1491.
10. Gupta,R.S. and Mathews,D.W. (2010) Signature proteins for the major clades of Cyanobacteria. *BMC Evol. Biol.*, **10**, 24.
11. Hsing,M. and Cherkasov,A. (2008) Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinform.*, **9**, 293.
12. Pascarella,S. and Argos,P. (1992) Analysis of insertions/deletions in protein structures. *J. Mol. Biol.*, **224**, 461–471.
13. Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.*, **229**, 1065–1082.
14. Aravind,L., Mazumder,R., Vasudevan,S. and Koonin,E.V. (2002) Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.*, **12**, 392–399.
15. Blouin,C., Butt,D. and Roger,A.J. (2004) Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. *Protein Sci.*, **13**, 608–616.
16. Wolf,Y., Madej,T., Babenko,V., Shoemaker,B. and Panchenko,A.R. (2007) Long-term trends in evolution of indels in protein sequences. *BMC. Evol. Biol.*, **7**, 19.
17. Chan,S.K., Hsing,M., Hormozdiari,F. and Cherkasov,A. (2007) Relationship between insertion/deletion (indel) frequency of proteins and essentiality. *BMC Bioinformatics*, **8**, 227.
18. Wagner,A. (2003) How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.*, **270**, 457–466.
19. Akiva,E., Itzhaki,Z. and Margalit,H. (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl Acad. Sci. USA*, **105**, 13292–13297.
20. 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. USA*, **107**, 20352–20357.
21. 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.
22. 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.
23. Tian,D., Wang,Q., Zhang,P., Araki,H., Yang,S., Kreitman,M., Nagylaki,T., Hudson,R., Bergelson,J. and Chen,J.Q. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, **455**, 105–108.
24. Zhu,L., Wang,Q., Tang,P., Araki,H. and Tian,D. (2009) Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol. Biol. Evol.*, **26**, 2353–2361.
25. Robinson,D.M., Jones,D.T., Kishino,H., Goldman,N. and Thorne,J.L. (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, **20**, 1692–1704.
26. Choi,S.C., Hobolth,A., Robinson,D.M., Kishino,H. and Thorne,J.L. (2007) Quantifying the impact of protein tertiary structure on molecular evolution. *Mol. Biol. Evol.*, **24**, 1769–1782.
27. Zhang,Z., Huang,J., Wang,Z., Wang,L. and Gao,P. (2011) Impact of indels on the flanking regions in structural domains. *Mol. Biol. Evol.*, **28**, 291–301.
28. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
29. Chandonia,J.M., Hon,G., Walker,N.S., Lo,C.L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
30. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta. Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
31. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.