A SYSTEMATIC STUDY OF HOMOLOGOUS PROTEIN STRUCTURES WITH INSERTIONS/DELETIONS

RyangGuk Kim, Jon McCafferty, and Jun-tao Guo*

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC 28223 USA
rkim6@uncc.edu, jmccaff2@uncc.edu, jguo4@uncc.edu

Sequence insertions/deletions (indels) represent one of the mechanisms of protein evolution. Alternative splicing (AS), considered as the major means of expanding structural and functional diversity in eukaryotes, can generate protein isoforms with indels when compared to the reference splicing variant. Knowledge of the effect of indels on the structural changes of the isoform structures is essential to our understanding of the functionality of splicing isoforms and protein evolution. Very little is known about how the indels, especially the ones that involve the core secondary structures, affect protein structures as only a few genes (<10) have two solved isoform structures. Here we show a systematic analysis on the structural changes due to indels through mining the Protein Data Bank (PDB) for highly homologous proteins. We found that more than 30% of indel residues adopt disordered "conformation", which is significantly higher than that in the control dataset. In addition, protein structures tend to be conserved and can tolerate structural insertions and deletions, suggesting the plasticity of protein structures. We also presented examples to show how structural core conservation and sequence/structure flexibility can help accurately predict isoform structures with indels, which has been shown to be extremely difficult with current comparative modeling techniques. To our knowledge, this is the first systematic study of the effects of indels on structural changes.

1. INTRODUCTION

As protein evolves, insertions and deletions (indels) can be introduced to create protein variants for survival needs. A recent large-scale indel analysis at sequence level revealed that up to 5-10% of all proteins contained indels when using human homologs as references[1]. Alternative splicing (AS), a major mechanism in eukaryotes for increasing the proteome size and functional diversity, is a primary source of generating many protein isoforms with indels[2, 3]. It has been shown that alternatively spliced protein isoforms are involved in a variety of biological processes and deviant splicing could have serious implications[4, 5].

While high-throughput data analysis suggested that up to 94% of human genes undergo alternative splicing, and provided a genome-wide view of the evolution and regulation of alternative splicing[3, 6-8], our general knowledge of the isoform protein structures is very limited. Little is known about how alternative splicing affects protein structures. Currently, fewer than 10 alternatively spliced isoforms with documented structures are in the Protein Data Bank (PDB)[9, 10] though there are over 13,000 annotated protein isoforms in human alone from Swissprot protein database (Release 14.2, September 23, 2008). This clearly represents a major knowledge gap as structures hold key information for the function of protein isoforms. More importantly, it is interesting how different isoforms with largely identical sequences perform different functions. In addition, isoform structure prediction represents a great challenge to homology modeling techniques for accurately modeling structures with indels[11-13].

The lack of such information has prompted several recent studies on AS isoform structures by mapping the sequence fragment affected by the alternative splicing events onto known isoform or homologous structures[14-18]. While there are several types of splicing events that result in different splice isoforms when compared to the primary sequences, such as truncation, substitution, insertion and deletion, the internal insertion/deletion cases are the dominant form of alternative splicing variants and are of great interest due to its potential impact on the folding and stability of isoform structures[6, 16]. Conflicting results have been reported in studying the effects of indels on the isoform structures. Tress *et al.* concluded that internal insertion/deletions may have larger structural impact and AS isoform is an unlikely route to increase functional diversity[16]. However, three other large scale analyses offered a different

*Corresponding author

view and suggested that protein structures have some degree of "plasticity" to tolerate insertions and deletions[14, 15, 17]. Based on threading analysis, Wang *et al*. found that most of the splicing isoforms probably adopt the same structural folds of their full-length counterparts and the boundaries of AS events generally happen in coil regions and involve exposed residues [14]. Romero *et al*. revealed an association between protein disorder and alternative splicing events [15]. These conflicting views will not be resolved unless we have more experimentally determined structures of the alternatively spliced isoforms, which might take a long time[19].

To gain more insight into the structural changes of AS isoforms with internal indels (or gaps) and to shed light on protein evolution, we performed a systematic structural analysis of protein structural pairs that have high sequence similarity and contain internal indels. The basic idea behind this study is the analogy between the AS variants with internal insertions/ deletions and the homologous protein pairs diverged over the course of evolution through insertions and deletions. Analysis of sequence or structural indels/gaps in similar proteins has been attempted since early 90's[20-24]. However, our study is significantly different from previous studies in several aspects. First, the major goal of our study is to systematically study *structural* changes in highly homologous proteins with indels/gaps based on *sequence* alignment and to gain valuable information for AS isoform structure prediction. Previous studies on indels primarily focused on some of the statistics of the indels/gaps in the proteins that are structurally similar but do not necessarily have high sequence homology[20, 22]. The very recent "Indel PDB" database reported the secondary structure composition and solvent accessibility of indel sequences, but its focus is not to address the structural changes affected by indel sequences[24]. Secondly, we compiled a non-redundant dataset of *highly homologous* protein pairs with internal gaps (see Methods section) while other analyses used sequentially or structurally similar protein pairs with sequence similarity ranging from very low to very high[20, 22-24]. More importantly, we applied an extra filter to ensure high similarity of sequences flanking the indels, which dramatically reduced the possibility of having "random" indel positions and sequences due to low local sequence similarity.

Another unique feature of our approach is that we considered disordered segments in our structural comparison analysis. It has been shown that intrinsically disordered or unstructured regions are responsible for many important cellular functions[25, 26]. A recent study by Dunker's group revealed the link between alternative splicing and protein intrinsic disorder, suggesting structural and functional diversity through alternative splicing[15]. However data generated from previous studies did not include protein pairs in which the indels or flanking regions are disordered. In some cases, proteins with disordered indels were intentionally filtered out for the purpose of assigning secondary structures for the indel fragments[24].

Here we report our findings from a systematic analysis of a non-redundant dataset with highly homologous protein pairs. We found that the indels tend to have less regular secondary structures (both α-helices and β-strands), but are rich in disordered "conformation" when compared to a non-redundant reference protein dataset. Proteins with indels occurring in the middle of regular secondary structures generally preserve the structural fold and at the same time go through local structure rearrangement and refolding for structural stability. In addition, we found that the immunoglobulin (Ig) family is heavily overrepresented in the indel dataset. Therefore we generated a new dataset by removing indels derived from the Ig family members to avoid bias in statistical analysis. We believe this study can serve as a useful resource for modeling homologous structures as well as the alternatively spliced protein isoform structures, and shed light on protein evolution.

2. METHODS

2.1. Datasets and method overview

Three different datasets are used in this study. The first dataset (Dataset I) contains a list of 25674 protein chains culled from the PISCES "pdbaanr" dataset that includes representative protein chains based on the resolution and R-values among a group of protein chains having up to 100% sequence identity[27]. The selection criteria for Dataset I from "pdbaanr" are: experimental method = X-ray crystallography, maximum resolution = 3.5 Å, and the sizes range from 50 to 1000 amino acids. The second dataset (Dataset II) is a non-redundant data set with 4731 protein chains, in which no pair of protein chains has more than 25% sequence identity, each structure has a resolution better than 2.5 Å, and the size is in the 50-1000 amino acids range. The statistics from this dataset, such as amino acid frequencies and secondary structure types, is used as background distribution for comparison purpose.

The third one (Dataset III) is a dataset of human alternative splicing isoforms with indels culled from the UnitProtKB database (http://www.uniprot.org). Since many deletion sequences resulted from alternative splicing of the same primary protein have large overlapping regions, we constructed a non-redundant dataset with human AS indel sequences (termed AS indels) through filtering out the near identical sequences from the same primary protein.

The flowchart for identification of homologous protein pairs with indels is shown in Figure 1. Briefly, protein chains in Dataset I were clustered into 9,513 groups using BLASTCLUST, a part of BLAST package at NCBI[28], with a sequence similarity of at least 50% and an alignment coverage of at least 40%. After four filtering steps, a non-redundant list of indels were subjected to statistical analysis, such as amino acids composition, secondary structure types, and local/global structural changes induced by the indels. We describe the details of these steps in the following sections.

2.2. Filtering steps for a non-redundant indel dataset

To make the indel statistics more meaningful for highly homologous protein pairs, we performed the following filtering procedures. The first step is to remove redundant protein chains in each cluster that has at least two members using a similar approach as described by Pascarella and Argos[20]. Briefly, if two sequences are highly similar without any internal gaps when aligned, the one with lower resolution is filtered out. The sequence comparison was done using a global sequence alignment program NEDDLE in the EMBOSS package with default parameters: gap open 10, gap extension 5, and the Blosum62 substitution matrix[29].

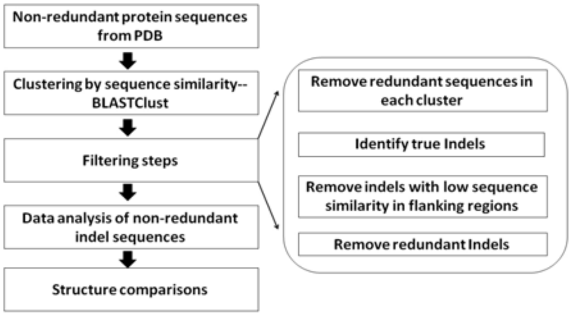

Non-redundant protein sequences from PDB
Clustering by sequence similarity-- BLASTClust
Filtering steps
Data analysis of non-redundant indel sequences
Structure comparisons
Remove redundant sequences in each cluster
Identify true Indels
Remove indels with low sequence similarity in flanking regions
Remove redundant Indels

Figure 1. Flowchart of the identification of homologous protein structures with indels

The second filtering step is to check the sequence similarities of regions flanking the indel site. Even though the overall sequence similarity of two proteins is high (we used a cutoff of 75%), it is possible that the protein pair have very diverse local sequences. For example, proteins 1R6ZA-1Y4CA showed over 80% sequence identity in the global alignment, in which the first 367 residues of these two proteins are the same. However, the alignment in the C-terminal portion that shows five indels/gaps has very low sequence similarity (Figure 2, the alignment for the N-terminal 250 residues that are 100% identical is not shown), suggesting the indels/gaps derived from this alignment are not reliable. It does not make much sense to include these indel sequences in analysis as the indel positions may change dramatically with a minor change of alignment parameters. To avoid the uncertainty of the

```
1R6ZA 251 FKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPL 300
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1Y4CA 250 FKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPL 299

1R6ZA 301 GAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVIN 350
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1Y4CA 300 GAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVIN 349

1R6ZA 351 AASGRQTVDEALKDAQTNAAAEFVDISHKSFPISMPMIEYLERFSLKAKI 400
          ||||||||||||||||||::                        |.....
1Y4CA 350 AASGRQTVDEALKDAQTNSS------------------------SNNNNN 375

1R6ZA 401 NNTTNLDYSRRFLEPFLRGINVVYTPPQSFQSAPRVYRVNG--------L 442
          ||..||....|..|..|:  ..:....::.|.|..:.:..|        |
1Y4CA 376 NNNNNLGIEGRSSEELLK--IALQEAQKTLQQAQELAKKGGGEEQLKRAL 423

1R6ZA 443 SRAP----ASSETFEHDGKKVTIASYFHSRNYPLKFPQLHCLNVGSSIKS 488
          .||.    |:.|..:..|.                         |..:..
1Y4CA 424 KRADRNLWAAQELAKKGGG-------------------------GEELLK 448

1R6ZA 489 ILLPIELCSIEEGQALNRKDG------------------------- 509
          ..|......:.:.|.|.:|.|
1Y4CA 449 QALQQAQQLLRQAQELAKKGGGEELLKQALQQAQQLLQQAQELAKK 494
```

Figure 2. Sequence alignment with high global sequence similarity and low similarity of sequences flanking the gaps

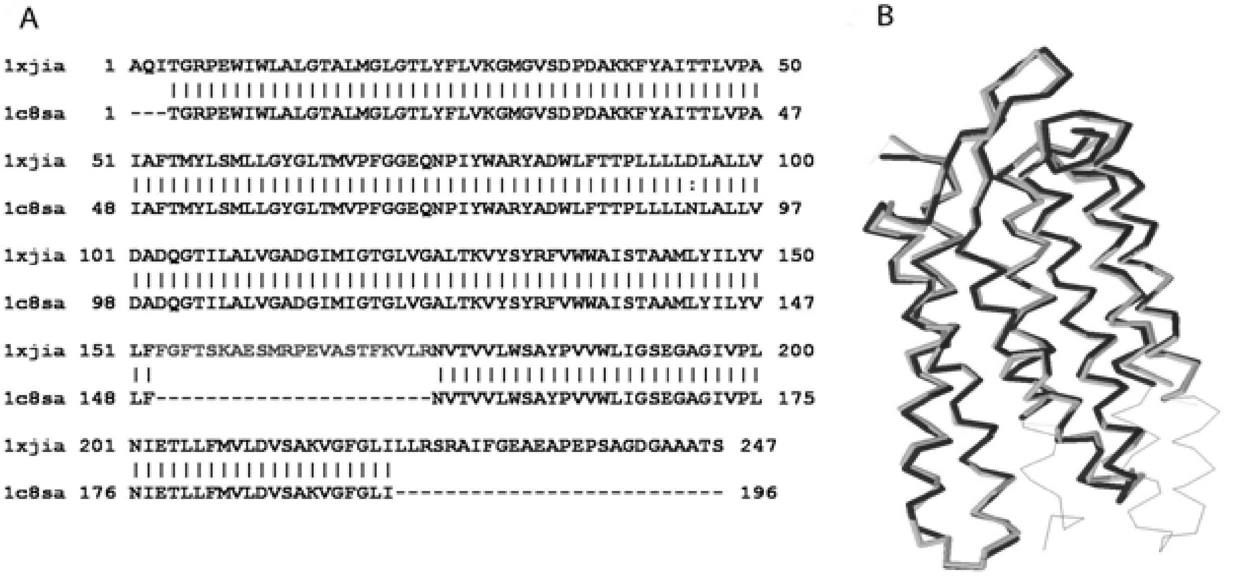
indel or gap positions, we calculate the sequence similarity of the flanking regions of the indels/gaps (20 AA on each side) and only indel sequences with highly similar flanking regions (above a cutoff value of 75%) are considered.

The third filter applied in our systematic analysis is to detect and remove false gaps/indels based on sequence alignment. The major goal of this study is to investigate the impact of sequence insertion or deletion on the AS isoform structures. As mentioned earlier, we consider the indels that adopt disordered "conformation" (missing coordinates in PDB files) since protein intrinsic disorder has been associated with alternatively spliced isoforms and functional diversity[15]. To do this, we first read the SEQRES records to get the protein sequences since sequences derived from the ATOM record of PDB files don't have disordered regions. However, due to the discrepancies of deposition of SEQRES, we applied an additional filtering step to remove the "false gaps/indels" from the sequence alignments. For example, the sequence alignment between 1XJIA-1C8SA (from SEQRES record) shows an internal gap even though the structures are from the same protein sequence of bacteriorhodopsin[30]. The difference is that the fragment 154-175 in 1C8SA is disordered and its sequence was not reported in the SEQRES record while the corresponding fragment in 1XJIA has coordinates and appears in the SEQRES report (Figure 3). This type of fragments that adopt ordered conformation in at least one structure and are disordered in other structure(s) are called "Dual Personality" fragments[31] and are not true indels (false gaps/indels). To identify these false gaps/indels, we calculated the Cα distance between the two immediate residues flanking the gap/indel in the sequence alignment (residues F and N in Figure 3). If the distance is more than 4.5 Å, it would suggest that the two residues are not directly connected and a false gap is flagged.

The last step in generating non-redundant indel sequence dataset is to filter out redundant indel sequences. If two protein pairs are from the same family and have the same indel sequences with very similar secondary structures at approximately the same residue positions, we consider these two indel sequences redundant. Only one indel sequence will be selected for further statistical and structural analysis. The highly homologous protein pair with and without an indel is termed an "indel pair" in this study.

2.3. Secondary structure assignment and solvent accessibility

The secondary structure type and the solvent accessibility of each residue were determined using the DSSP program[32]. The relative solvent accessibility is then calculated by dividing the absolute value by the maximum accessibility of each residue. In this study, we use four secondary structure



```
1xjia    1 AQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPA  50
              |||||||||||||||||||||||||||||||||||||||||||||||
1c8sa    1 ---TGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPA  47

1xjia   51 IAFTMYLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTTPLLLLDLALLV  100
           ||||||||||||||||||||||||||||||||||||||||||||:|||||
1c8sa   48 IAFTMYLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTTPLLLLNLALLV  97

1xjia  101 DADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYV  150
           ||||||||||||||||||||||||||||||||||||||||||||||||||
1c8sa   98 DADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYV  147

1xjia  151 LFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPL  200
           ||                      ||||||||||||||||||||||||||
1c8sa  148 LF----------------------NVTVVLWSAYPVVWLIGSEGAGIVPL  175

1xjia  201 NIETLLFMVLDVSAKVGFGLILLRSRAIFGEAEAPEPSAGDGAAATS  247
           |||||||||||||||||||||
1c8sa  176 NIETLLFMVLDVSAKVGFGLI--------------------------  196
```

Figure 3. An example of false gaps/indels due to different SEQRES reporting practices

types, H (helix), E (strand), C (coil), and U (unstructured/disordered) and three-state solvent classification, buried (B), intermediate (I), and exposed (E) with 7% and 37% as the thresholds to define these three states, that is, ≤7%, 7%< and ≥37%, and >37% [33]. The disordered residues or fragments were defined by comparing the "ATOM" and "SEQRES" records in PDB file. If a residue or a fragment appears in "SEQRES", but is missing from the "ATOM" record in a PDB file, this residue or fragment is considered as disordered or unstructured[34].

2.4. Structure comparison and modeling

To examine the structural changes caused by the indels, the two protein structures of each indel pair were compared using two structure alignment programs, FAST[35] for global structure alignment and CE[36] for local structure alignment. The differences between the structure- and sequence-alignments of each pair were then evaluated. A webserver was developed at http://bioinfozen.uncc.edu/scindel for a convenient visualization of both the sequence and structure alignments. All the analyses were done with Python scripts developed in our lab. The comparative modeling was done using MODELLER [37].

3. RESULTS

3.1. Non-redundant indel dataset

A total of 25,674 protein chains that meet the selection criteria as described in Methods section were clustered into 9,513 groups using BLASTCLUST with 50% sequence identity cutoff and 40% coverage cutoff. After filtering out redundant protein chains in each cluster, 1,607 clusters have at least two protein chains. Except for the largest cluster that contains 499 protein chains belonging to the immunoglobulin family, no other clusters have more than 20 protein chains. Based on sequence alignments, there are a total of 1,296,086 indels from 179,262 distinct pairs with internal indels/gaps. The number of indels at this step is higher than that in Indel PDB (488,038) as we used a different coverage in BLASTCLUST[24]. We then applied the four filters to generate a dataset of 454 non-redundant indel sequences (called Indel NR): 1) at least 75% sequence identity between the pair in aligned regions and the flanking regions of indels/gaps (20 AA on each side); 2) false gaps/indels removal; 3) indel length of 40 AA or shorter; and 4) redundant indel sequence removal as described in Methods.

3.2. Statistical analysis of non-redundant indel sequences

Based on SCOP protein structure classification using the latest 1.73 release[38], the protein chains that harbor the 454 indel sequences belong to at least 97 different families, 110 superfamiles, and 127 different folds (some newly solved structures have yet to be annotated in SCOP). These protein chains on average have good fold coverage (~4 protein chains/fold). However, one protein family (b.1.1.1) dominates the indel sequences with 219 sequences. More specifically, these indel fragments are generally from the third complementarity-determining region of the immunoglobulin (Ig) heavy chain (CDR-H3), which is the most diverse region and plays a crucial role in antigen recognition and binding specificity[39, 40]. The CDR-H3 loops are dominated by residues tyrosine (Y), glycine (G), and serine (S), which can heavily skew the amino acid frequencies towards the composition of CDR-H3[40,41]. Due to the over-representation of indels from Ig proteins and the fact that the indels derived from these proteins are irrelevant to the AS analogy of our interest, we compiled three different indel datasets for statistical analysis: Ig indels, Non-Ig indels (152 protein pairs), and Ig+Non-Ig indels. Our data confirmed that tyrosine, glycine, and serine residues dominate the indel sequences from immunoglobulin proteins (Ig) (Figure 4A). The Ig dataset has about five times more tyrosine residues over the background level while several other amino acid types are underrepresented. Figure 4A shows that inclusion or exclusion of Ig indel sequences can result in major differences in amino acid compositions, which has not been reported in previous indel sequence studies[24].

Dataset Non-Ig is enriched in residues G, E, D, K, and S, but is depleted in residues F, I, L, V, W, and Y when compared with the background frequencies, suggesting that indels have more residues with high propensity to coil structure (G, D,

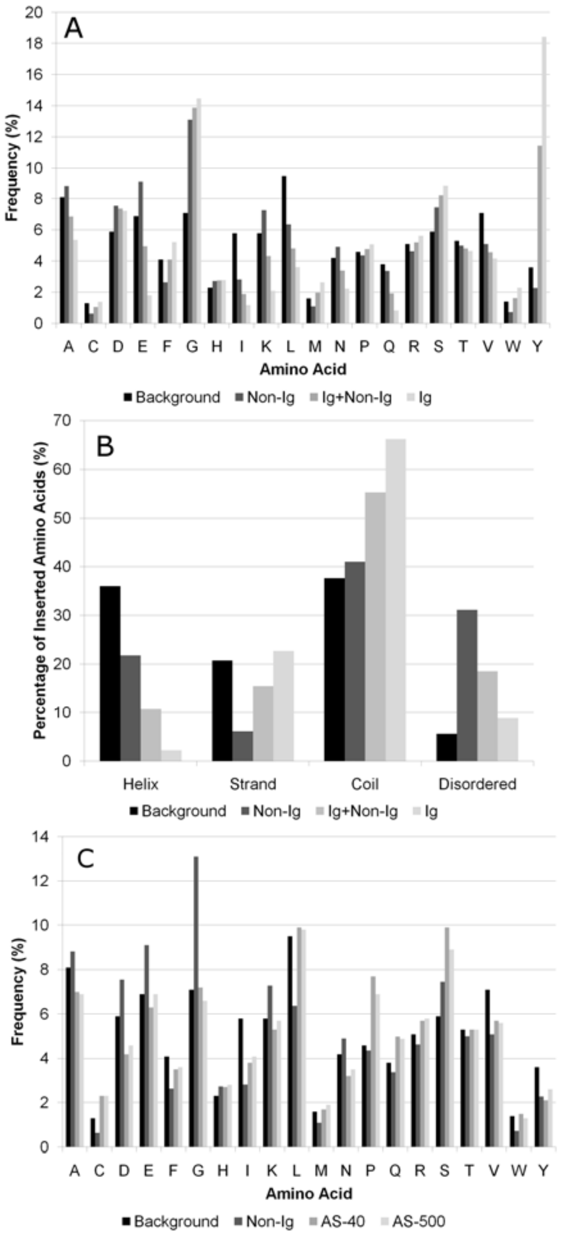


Figure 4. Frequencies of amino acids (A and C) and secondary structure types (B) of the indel sequences.

and S) and less residues that prefer an α-helix or β-sheet conformation (F, I, L, V, W, and Y) (Figure 4A)[13]. Analysis of secondary structure types is consistent with the amino acid composition analysis of indel sequences. While there is a dramatic decrease in the number of residues that adopt regular secondary structures, especially the sheet conformations, the number of coil residues is only slightly more than that from the background distribution (Figure 4B). Instead, relative to the background frequencies, indel sequences have markedly increased number of residues in disordered state (over five-fold increase) (Figure 4B). Taken together, the majority of the indel sequences adopted either coil or disordered "conformation", consistent with previous observations that insertions/deletions are most likely to occur in loop regions or between regular secondary structure elements and thus preserve the overall structural fold [19]. Similar observations have been reported for alternative splicing events, which by and large prefer coil regions and exposed residues[14, 15].

It is interesting to see if there is any similarity between the above indel statistics and that from AS indel sequences. We compiled a non-redundant human AS indel sequences from UnitProtKB. The distribution of the human AS indel sequences showed that majority of the sequences (~93%) is shorter than 500 amino acids. We constructed a human AS dataset (AS-500) by excluding very long indel sequences. In addition, we generated a second set (AS-40) with human AS indels that have 40 or less amino acids since our indel sequences are generally shorter than 40 amino acids. As shown in Figure 4C, there are essentially no differences between AS-500 and AS-40 in terms of amino acid composition. While several amino acids displayed similarities to the background distributions but were different from the Non-Ig set (G, E, F, L, K, and W), amino acids I, V, S and Y, on the other hand, have similar frequencies to those in the Non-Ig set. The decreased frequencies of isoleucine (I) and valine (V) suggests that the isoform may adopt less β-sheet structures. Another interesting observation is that proline (P) and glycine (G) showed different patterns in Non-Ig and AS datasets. Glycine is dramatically increased in the Non-Ig set while more proline residues are seen in the AS datasets. It is well known that both proline and glycine have high propensity to coil conformations. Changes in serine and tyrosine might have functional implications in alternatively spliced isoforms. In addition to its ability to serve as functional residue, serine is often observed in loops. Therefore despite the differences, both the Non-Ig and human AS datasets are rich in residues that prefer coil or loop conformations and are depleted in β-sheet forming residues.

3.3. Structural changes by indels

Global structural changes by indels in the Non-Ig dataset were examined using FAST. Figure 5 shows the histogram of the root-mean-square-deviations (RMSDs) of the structure alignments. Most of the pairs have small structural changes induced by the indels (about 87% pairs with less than 2Å RMSDs), suggesting that protein structure in general can tolerate and accommodate the indels[17, 19]. Although a small number of pairs show large RMSDs, we found that all the 9 pairs with RMSD more than 4Å are the results of indels acting as "pivots", causing changes in the relative orientations of the domains rather than a fold change. For example, though the pair 1UX5A-1UX4A (with a four-residue indel sequence REDL folding as a helical structure) has the largest global RMSD of 22.85 Å (Figure 6A), they have almost identical structures separated by the indel sequence, with RMSDs of 0.95 Å and 1.11 Å, respectively (Figure 6B and 6C).

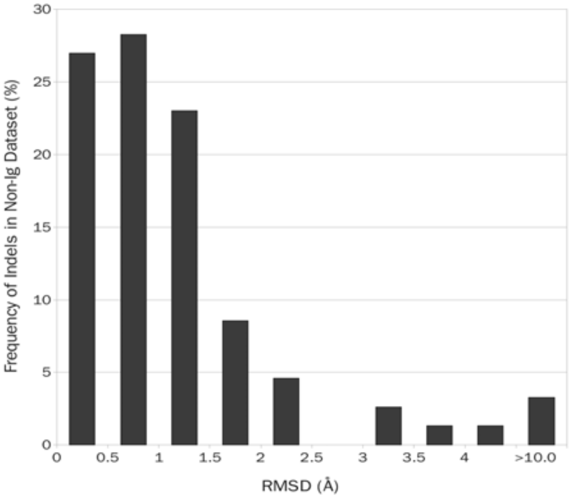


Figure 5. Global structural changes due to indels

The insertion of indels could result in several major types of structural changes (Figure 7). One is that the indel sequence folds as a separate domain as seen in 1AD2A-20V7A (Figure 7A). The second type is that the indel is disordered (Figure 7C) or adopts a longer loop (Figure 7B and 7F). It is not surprising that the overall structures are conserved well as in general insertions/deletions tend to occur in the loop regions, which are relatively flexible.

One of the most interesting questions concerns the structural change if the indel events occupy or happen in the middle of a secondary structure elements, especially when the deletion of internal strands from a β-sheet as reported in previous studies[13, 17, 42]. The deletion of β-strands of a β-sheet presents a tremendous challenge and is problematic for comparative modeling approaches. In our Non-Ig dataset, about 15% (23 out of 152) of its indel sequences were flanked at each side by two or more consecutive amino acids with helix or strand conformations. We found in these cases, the core secondary structures tend to be conserved even though one strand in the longer form is deleted compared to the short form (Figure 7D and 7E). This is accomplished by folding the neighboring sequences as the structural conformation and filling the "hole" left with the strand deletion.

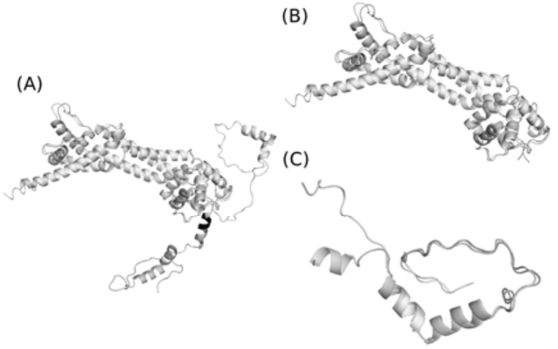
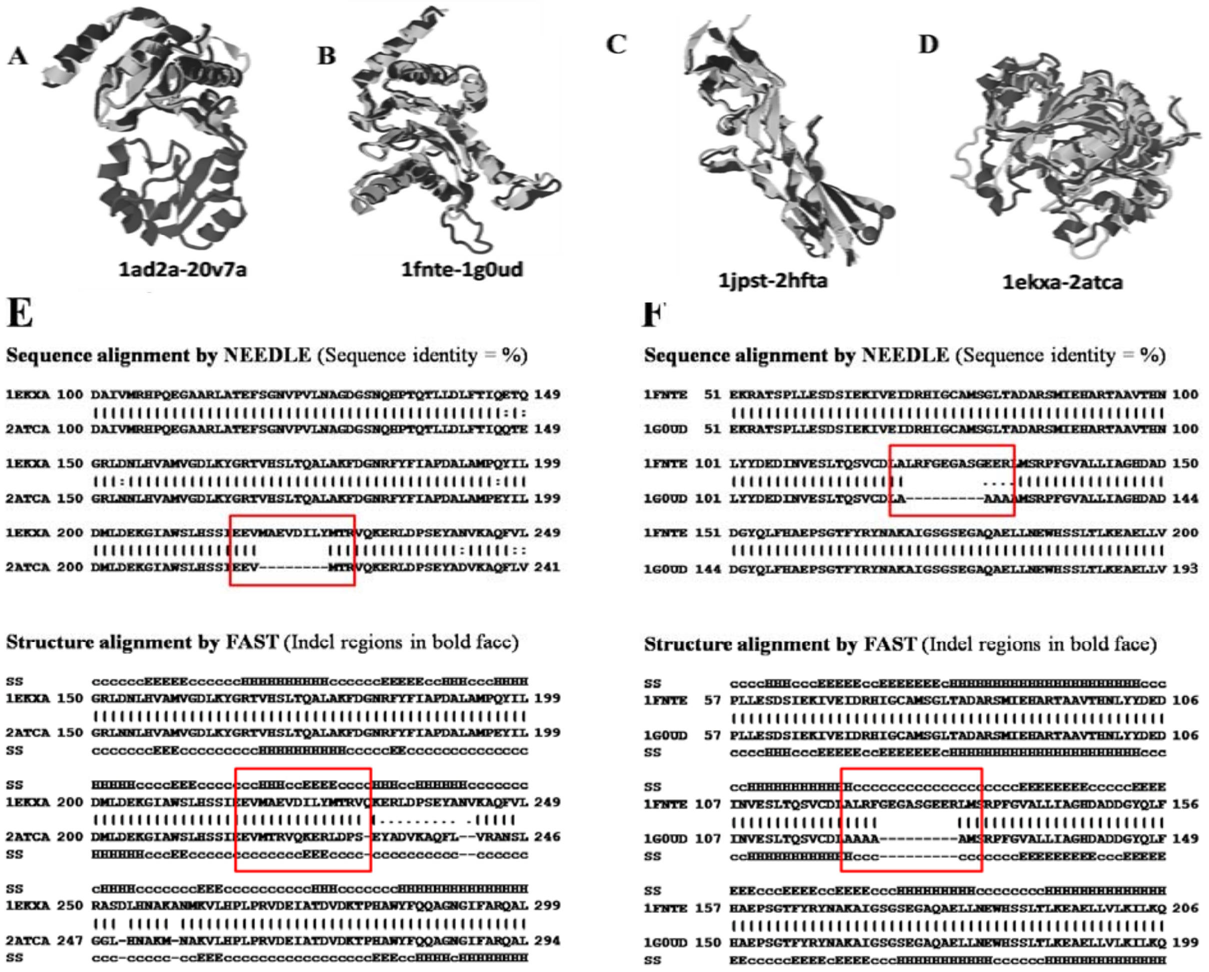


Figure 6. Structure comparison between 1UX5A and 1UX4A. Dark color represents the indel sequence.

3.4. Homology modeling of protein structures with indels

Homology modeling of proteins to see the effect of indels has been proven difficult. One (in)famous case is the modeling of a protein called Piccolo[11]. In the short isoform, a nine-residue fragment that is missing in the alternatively spliced form of Piccolo C2A domain folds as a β-strand in the long isoform. The short and long isoforms have different calcium binding affinity. Surprisingly, the short variant maintains the structural fold by moving a short fragment that flanks the strand and folds as a strand in the short isoform. Figure 8 shows two more examples that current modeling techniques would fail to accurately predict the structure of one protein using the other one as template (Figure 8ABC:1EKXA-2ATCA Figure 8DEF: 2HKDA-2AF5A). Assuming we only have the short form (Figure 8B, 8E) or the long form (Figure 8C, 8F)



Sequence alignment by NEEDLE (Sequence identity = %)

```
1EKXA 100 DAIVMRHPQEGAARLATEFSGNVPVLNAGDGSNQHPTQTLLDLFTIQETQ 149
          |||||||||||||||||||||||||||||||||||||||||||||||:|:
2ATCA 100 DAIVMRHPQEGAARLATEFSGNVPVLNAGDGSNQHPTQTLLDLFTIQQTE 149

1EKXA 150 GRLDNLHVAMVGDLKYGRTVHSLTQALAKFDGNRFYFIAPDALAMPQYIL 199
          |||:||||||||||||||||||||||||||||||||||||||||||:|||
2ATCA 150 GRLNNLHVAMVGDLKYGRTVHSLTQALAKFDGNRFYFIAPDALAMPEYIL 199

1EKXA 200 DMLDEKGIAWSLHSSIEEVMAEVDILYMTRVQKERLDPSEYANVKAQFVL 249
          |||||||||||||||||||        |||||||||||||||:|||||::
2ATCA 200 DMLDEKGIAWSLHSSIEEV--------MTRVQKERLDPSEYADVKAQFLV 241
```

Structure alignment by FAST (Indel regions in bold face)

```
SS        ccccccEEEEEccccccHHHHHHHHHHccccccEEEEEccHHHcccHHHH
1EKXA 150 GRLDNLHVAMVGDLKYGRTVHSLTQALAKFDGNRFYFIAPDALAMPQYIL 199
          ||||||||||||||||||||||||||||||||||||||||||||||||||
2ATCA 150 GRLNNLHVAMVGDLKYGRTVHSLTQALAKFDGNRFYFIAPDALAMPEYIL 199
SS        cccccccEEEccccccccccHHHHHHHHHHHcccccEEcccccccccccc

SS        HHHHHccccEEEcccccccHHHccEEEEccccHHHccHHHHHHccccccc
1EKXA 200 DMLDEKGIAWSLHSSIEEVMAEVDILYMTRVQKERLDPSEYANVKAQFVL 249
          ||||||||||||||||||||||||||||||| |.........  .|||||
2ATCA 200 DMLDEKGIAWSLHSSIEEVMTRVQKERLDPS-EYADVKAQFL--VRANSL 246
SS        HHHHHHcccEEcccccccccccccEEEcccc-cccccccccc--cccccc

SS        cHHHHcccccccEEEccccccccccHHHcccccccHHHHHHHHHHHHHHH
1EKXA 250 RASDLHNAKANMKVLHPLPRVDEIATDVDKTPHAWYFQQAGNGIFARQAL 299
          ||| ||||| ||||||||||||||||||||||||||||||||||||||||
2ATCA 247 GGL-HNAKM-NAKVLHPLPRVDEIATDVDKTPHAWYFQQAGNGIFARQAL 294
SS        ccc-ccccc-ccEEEcccccccccccccccccccEEEccHHHHcHHHHHH
```

Sequence alignment by NEEDLE (Sequence identity = %)

```
1FNTE  51 EKRATSPLLESDSIEKIVEIDRHIGCAMSGLTADARSMIEHARTAAVTHN 100
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1G0UD  51 EKRATSPLLESDSIEKIVEIDRHIGCAMSGLTADARSMIEHARTAAVTHN 100

1FNTE 101 LYYDEDINVESLTQSVCDLALRFGEGASGEERLMSRPFGVALLIAGHDAD 150
          ||||||||||||||||||||         ....|||||||||||||||||
1G0UD 101 LYYDEDINVESLTQSVCDLA----------AAAAMSRPFGVALLIAGHDAD 144

1FNTE 151 DGYQLFHAEPSGTFYRYNAKAIGSGSEGAQAELLNEWHSSLTLKEAELLV 200
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1G0UD 144 DGYQLFHAEPSGTFYRYNAKAIGSGSEGAQAELLNEWHSSLTLKEAELLV 193
```

Structure alignment by FAST (Indel regions in bold face)

```
SS        ccccHHHcccEEEEEccEEEEEEEcHHHHHHHHHHHHHHHHHHHHHHccc
1FNTE  57 PLLESDSIEKIVEIDRHIGCAMSGLTADARSMIEHARTAAVTHNLYYDED 106
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1G0UD  57 PLLESDSIEKIVEIDRHIGCAMSGLTADARSMIEHARTAAVTHNLYYDED 106
SS        ccccHHHcccEEEEEccEEEEEEEcHHHHHHHHHHHHHHHHHHHHHHccc

SS        ccHHHHHHHHHHHHcccccccccccccccccccEEEEEEEEcccccEEEE
1FNTE 107 INVESLTQSVCDLALRFGEGASGEERLMSRPFGVALLIAGHDADDGYQLF 156
          |||||||||||||||||         ||||||||||||||||||||||||
1G0UD 107 INVESLTQSVCDLAAAA---------AMSRPFGVALLIAGHDADDGYQLF 149
SS        ccHHHHHHHHHHHHccc---------cccccccEEEEEEEEEcccEEEEE

SS        EEEcccEEEEccEEEEcccHHHHHHHHHHcccccccHHHHHHHHHHHHHH
1FNTE 157 HAEPSGTFYRYNAKAIGSGSEGAQAELLNEWHSSLTLKEAELLVLKILKQ 206
          ||||||||||||||||||||||||||||||||||||||||||||||||||
1G0UD 150 HAEPSGTFYRYNAKAIGSGSEGAQAELLNEWHSSLTLKEAELLVLKILKQ 199
SS        EEccccccEEEEcEEEEcccHHHHHHHHHHHHHccccccHHHHHHHHHHH
```

Figure 7. Structural comparisons of protein pairs with indels

structures and use them to model the long form (short form as template) and short form (long form as template) structures based the sequence alignments. As seen in Figure 8BC and 8EF, both the longer (with insertion) and shorter (with deletion) structures are not correct. When the *real* longer and shorter forms were superimposed, the location of structural difference was not at where the indel is located (Figures 8A and 8D). The same structural conformations (dark in both short and long forms) are from different sequences. However, in the homology models the longer forms were generated merely by inserting surface loops (Figures 8B and 8E) and the shorter forms were made by deleting the strands (indels) and connecting the flanking regions (Figures 8C and 8F). Our structural analysis by aligning the homologous structure through multiple structure alignments showed that the indel structures are conserved in homologous proteins while the variable regions are in the downstream of the indel site (data not shown), suggesting we can make a better model by refining the sequence alignment guided by structural information rather than relying only on the optimal sequence alignment.

4. DISCUSSION

We performed a systematic study to investigate the structural changes caused by indel sequences, by mining the highly homologous protein pairs with internal gaps/indels. The goal is to gain insights into the mechanism of protein evolution and provide guides to model protein structures with indels compared with the homologous templates. In addition to protein evolution, indels can be also the results of alternative splicing. Considering the contribution of alternative splicing in expanding the protein

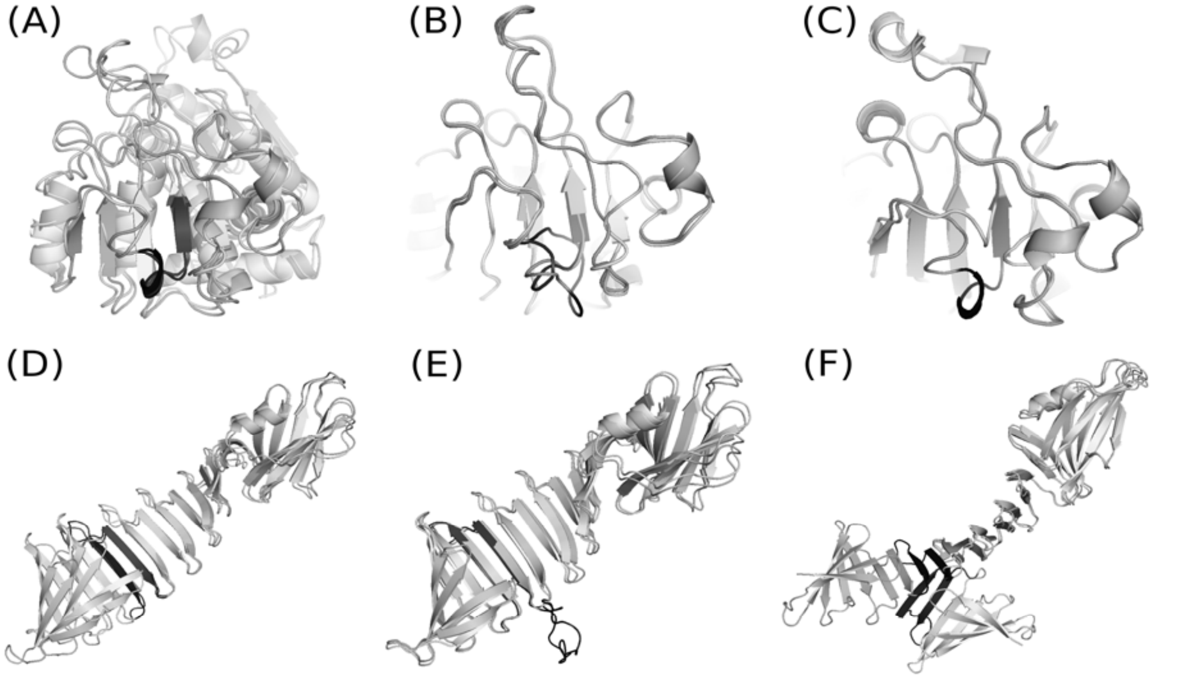


Figure 8. Homology modeling of proteins with indels

functionality, the importance of studying the effect of indels on structural change cannot be overstated. We found that the indels tend to occur between secondary structure elements and a significant number of indels are disordered, which is consistent with the earlier study that demonstrated the associations among indels/disordered/ function[15]. We consider this as one of the major contributions from this study as previous studies did not take disordered information into account. In addition, protein structures have inherent capability to tolerate structural deletions and insertions[13, 17]. Despite the interruption of regular secondary structures, structural folds are conserved through local structure rearrangements and refolding (Figure 7DE and Figure 8).

The rationale of choosing highly homologous protein pairs (both for the overall and indel flanking sequences) is two-fold: 1) to provide a better approximation to the AS isoforms of interest with internal gaps; and 2) to avoid the positioning of "random gaps" due to low local sequence similarity even though the overall sequence similarity is high (Figure 2). These steps ensure the unique positions of the indels and the unambiguous indel sequences, reducing the possibility of including those sequences due to sequence alignment error. An interesting finding in the analysis of indels is the abundance of tyrosine, glycine and serine[24]. We reported here, for the first time, that the heavy amino acid bias in indel sequences is due to the overrepresentation of one fold family, the immunoglobulin proteins. To make the statistics of indels' amino acid composition and secondary structure content meaningful, we constructed a dataset without immunoglobulin proteins. Although these indels showed differences from those of human AS datasets in terms of amino acid frequencies, some key features are very similar (Figure 4C).

Our analysis retrieved all AS isoform pairs that exist in the PDB except for 1Q56A-1PZ9A, structures of the C-terminal agrin domain. It is not surprising that our procedure missed this pair as 1Q56A was solved by NMR method, which we did not include in our initial data selection. The pair can be easily detected when we add the NMR structure to the dataset.

The very question about modeling isoform structures or structural changes due to indels is to improve the sequence alignment used for comparative modeling. No matter how good a comparative modeling program is, it cannot recover from the alignment error. The pitfall of current homology modeling techniques is that they heavily rely on the sequence similarity. We believe this systematic analysis, along with earlier reports on individual or a small number of case studies will serve as the tip of the iceberg in our understanding of the structural plasticity of proteins and how the indels are accommodated by the structure and at the same time deliver a new functionality.

Acknowledgements

This research was supported by the UNC Charlotte startup fund to JTG.

References

1. Cherkasov A, *et al.* Large-scale survey for potentially targetable indels in bacterial and protozoan proteins. *Proteins* 2006; 62: 371-80.
2. Pennisi E. Why do humans have so few genes? *Science* 2005; **309**: 80.
3. Xing Y and Lee C. Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* 2006; **7**: 499-509.
4. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003; **72**: 291-336.
5. Nakao M, *et al.* Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res* 2005; **33**: 2355-63.
6. Wang ET, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**: 470-6.
7. Blencowe BJ. Alternative splicing: new insights from global analyses. *Cell* 2006; **126**: 37-47.
8. Takeda J, *et al.* H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res* 2007; **35**: D104-9.
9. Berman HM, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000; **28**: 235-42.
10. Stetefeld J and Ruegg MA. Structural and functional diversity generated by alternative mRNA splicing. *Trends Biochem Sci* 2005; **30**: 515-21.
11. Garcia J, *et al.* A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat Struct Mol Biol* 2004; **11**: 45-53.
12. Ginalski K. Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* 2006; **16**: 172-7.
13. Guo JT, *et al.* Analysis of chameleon sequences and their implications in biological processes. *Proteins* 2007; **67**: 548-58.
14. Wang P, *et al.* Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc Natl Acad Sci U S A* 2005; **102**: 18920-5.
15. Romero PR, *et al.* Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* 2006; **103**: 8390-5.
16. Tress ML, *et al.* The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci U S A* 2007; **104**: 5495-500.
17. Birzele F, *et al.* Alternative splicing and protein structure evolution. *Nucleic Acids Res* 2008; **36**: 550-8.
18. Birzele F, *et al.* ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res* 2008; **36**: D63-8.
19. Laskowski RA and Thornton JM. Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet* 2008; **9**: 141-51.
20. Pascarella S and Argos P. Analysis of insertions/deletions in protein structures. *J Mol Biol* 1992; **224**: 461-71.
21. Benner SA, *et al.* Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol* 1993; **229**: 1065-82.
22. Wrabl JO and Grishin NV. Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. *Proteins* 2004; **54**: 71-87.
23. Chang MS and Benner SA. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *J Mol Biol* 2004; **341**: 617-31.
24. Hsing M and Cherkasov A. Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics* 2008; **9**: 293.

25. Dunker AK, *et al.* Intrinsically disordered protein. *J Mol Graph Model* 2001; **19**: 26-59.
26. Iakoucheva LM, *et al.* Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002; **323**: 573-84.
27. Wang G and Dunbrack RL, Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003; **19**: 1589-91.
28. Rice P, *et al.* EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000; **16**: 276-7.
29. Luecke H, *et al.* Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution. *Science* 1999; **286**: 255-61.
30. Zhang Y, *et al.* Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure* 2007; **15**: 1141-7.
31. Kabsch W and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983; **22**: 2577-637.
32. Kim D, *et al.* PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Eng* 2003; **16**: 641-50.
33. Vucetic S, *et al.* DisProt: a database of protein disorder. *Bioinformatics* 2005; **21**: 137-40.
34. Zhu J and Weng Z. FAST: a novel protein structure alignment algorithm. *Proteins* 2005; **58**: 618-27.
35. Shindyalov IN and Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998; **11**: 739-47.
36. Sali A and Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993; **234**: 779-815.
37. Andreeva A, *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004; **32 Database issue**: D226-9.
38. Xu JL and Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 2000; **13**: 37-45.
39. Zemlin M, *et al.* Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol* 2003; **334**: 733-49.
40. Birtalan S, *et al.* The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol* 2008; **377**: 1518-28.
41. Grishin NV. Fold change in evolution of protein structures. *J Struct Biol* 2001; **134**: 167-85.