BMC
Structural Biology

RESEARCH ARTICLE   Open Access

Systematic analysis of short internal indels and their impact on protein folding

RyangGuk Kim, Jun-tao Guo[*]

Abstract

Background: Protein sequence insertions/deletions (indels) can be introduced during evolution or through alternative splicing (AS). Alternative splicing is an important biological phenomenon and is considered as the major means of expanding structural and functional diversity in eukaryotes. Knowledge of the structural changes due to indels is critical to our understanding of the evolution of protein structure and function. In addition, it can help us probe the evolution of alternative splicing and the diversity of functional isoforms. However, little is known about the effects of indels, in particular the ones involving core secondary structures, on the folding of protein structures. The long term goal of our study is to accurately predict the protein AS isoform structures. As a first step towards this goal, we performed a systematic analysis on the structural changes caused by short internal indels through mining highly homologous proteins in Protein Data Bank (PDB).

Results: We compiled a non-redundant dataset of short internal indels (2-40 amino acids) from highly homologous protein pairs and analyzed the sequence and structural features of the indels. We found that about one third of indel residues are in disordered state and majority of the residues are exposed to solvent, suggesting that these indels are generally located on the surface of proteins. Though naturally occurring indels are fewer than engineered ones in the dataset, there are no statistically significant differences in terms of amino acid frequencies and secondary structure types between the "Natural" indels and "All" indels in the dataset. Structural comparisons show that all the protein pairs with short internal indels in the dataset preserve the structural folds and about 85% of protein pairs have global RMSDs (root mean square deviations) of 2Å or less, suggesting that protein structures tend to be conserved and can tolerate short insertions and deletions. A few pairs with high RMSDs are results of relative domain positions of the proteins, probably due to the intrinsically dynamic nature of the proteins.

Conclusions: The analysis demonstrated that protein structures have the "plasticity" to tolerate short indels. This study can provide valuable guides in modeling protein AS isoform structures and homologous proteins with indels through placing the indels at the right locations since the accuracy of sequence alignments dictate model qualities in homology modeling.

Background

Sequence insertions/deletions (indels) occur during evolution and alternative splicing (AS) process in eukaryotes. The generation of various protein isoforms through alternative splicing has been considered as one of the major evolutionary mechanisms for increasing the proteome size and functional diversity [1,2]. Recent high-throughput analysis based on mRNA-SEQ data from diverse human tissue

and cell lines suggested that alternative splicing is almost universal (up to 94%) in human multi-exon genes [3]. 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 [3,4]. In addition, genes containing "switch-like" exons are more likely to have isoforms with indels [3]. It is critical to our understanding of the function of alternatively spliced protein isoforms if we know how sequence changes, especially sequence

* Correspondence: jguo4@uncc.edu
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte 9201 University City Blvd, Charlotte, NC 28223 USA
Full list of author information is available at the end of the article

BioMed Central

insertions and deletions, affect the structure of the splice variants as structures hold key information for the function of proteins.

Our current knowledge about how alternative splicing affects protein structures is very limited. While there are about 28,000 annotated protein isoforms from recent UniProt release 15.11 (November 24, 2009) [5] and over 60,000 protein structures deposited in Protein Data Bank (PDB) [6], fewer than 10 pairs of alternatively spliced isoforms have documented structures [7]. Prediction of isoform structures generally falls into the category of homology modeling. However, homology modeling of proteins with indels is not a trivial task. The key to the success in homology modeling with indels is alignment accuracy, especially the positioning of the insertion or deletion sequences. For example, several groups at CASP8 (the 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) used the same protein 2G39 as the template to model target protein T0438, but only three of nine models placed the insertion sequence (12 amino acids) in the right place [8]. Another infamous/famous example in indel positioning is the modeling of the long AS isoform of Piccolo C_2A domain that has a nine-residue insertion in a loop. Instead of folding as part of the loop, the nine-residue insert displaces a β-strand that is pushed into the calcium-binding region through local rearrangement, leading to a dramatic change in calcium binding affinity [9].

While it is generally believed that insertions and deletions are well tolerated in loops [10,11], insertions and deletions within secondary structures (α-helices and β-sheets) may have a dramatic effect on the overall structure and are considered deleterious and unfavorable during evolution [4,12]. Tress *et al*. argued that AS isoform is probably an unlikely route to increase functional diversity due to probably large structural impact induced by indels [4]. Yet in a number of studies with genetically engineered insertions and deletions on T4 lysozyme, Matthews' group showed that the protein has structural plasticity to tolerate indels within secondary structures [13-15]. Three recent large scale analyses also offered a similar view that protein structures have some degree of "plasticity" to tolerate insertions and deletions through maintaining the same structural fold [16-18].

The major goal of this paper is to investigate the impact of short internal indels (less than 40 amino acids) on protein structures, especially for indels within secondary structures. Large indels may fold as an individual domain or the protein pairs may adopt different folds due to the large differences between two sequences [8]. Terminal indels are not considered in this study as terminal fragments are relatively flexible and terminal deletion/truncation have become a standard protocol in
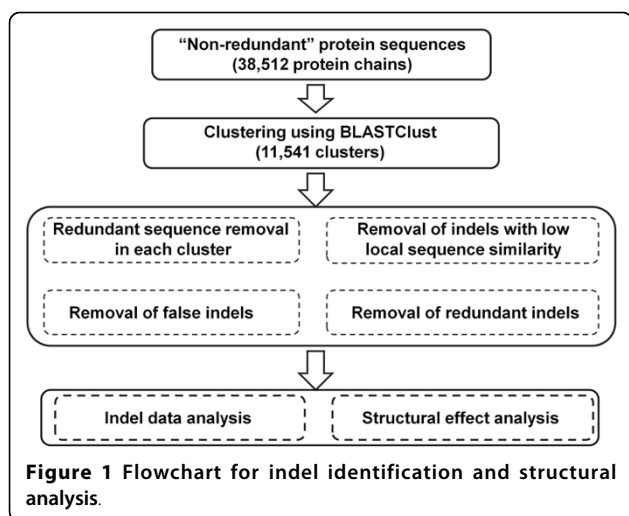
recombinant protein expression and protein crystallization [19-21]. In addition, the widely-used cloning vectors with His-tags introduce sequence artifacts that are included in the PDB SEQRES records and determining the exact tag sequences is problematic [22,23]. Though there are several similar surveys about indel statistics since 1992 [10,24-27], our approach is different as our goal is to study the impact of indels on protein structure and to provide guidance for isoform structure prediction. Therefore it is critical that the locations of the indels are unique and unambiguous. Otherwise the structural changes would be less well defined. For example, it is not uncommon that two proteins with high global sequence identity have a low local sequence similarity [28]. To address this issue, we take local sequence similarity into account and only consider protein pairs with both high global and local (sequences flanking the indels) sequence similarity (>75%) to ensure the uniqueness of indel sequences and positions. In addition, we include the "disordered conformation" in our structural analysis. It has been demonstrated that intrinsically disordered or unstructured regions are responsible for many important cellular functions and a link between alternative splicing and protein intrinsic disorder has been recently reported [17,29,30].

In this paper, we report a systematic analysis of a large non-redundant indel dataset with highly homologous protein pairs. Previously we found that the immunoglobulin (Ig) family, rich in certain amino acids including tyrosine, glycine, and serine in the third complementarity-determining region of the Ig heavy chain (CDR-H3), was overrepresented in an indel dataset [28,31,32]. Therefore those Ig-related indel sequences are not considered in our current analysis. Our results show that internal indels tend to have less regular secondary structures (α-helices and β-strands), but are rich in "disordered conformation", which is in line with the work by Romero *et al* [17]. Our data also show that proteins with short indels, including the ones within regular secondary structures, generally preserve the structural fold with some local structure rearrangement and refolding presumably for structural stability and functionality. The source of the indel, either naturally occurring or experimentally engineered, are described and the statistical significance of the features from natural indels is discussed. A webserver SCINDEL http://bioinfozen.uncc.edu/scindel was developed for convenient visualization of indel induced structural changes.

Methods
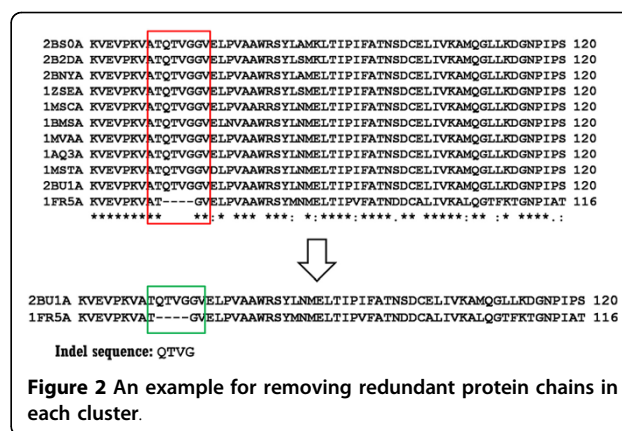### Generation of a non-redundant dataset of short internal indels
The flowchart for identification of indels from highly homologous protein pairs is shown in Figure 1. We

Figure 1 Flowchart for indel identification and structural analysis.

Figure 2 An example for removing redundant protein chains in each cluster.

started with a list of 38,512 protein chains from PISCES "pdbaanr" dataset (June, 2009) that contains representative protein chains selected based on the resolution and R-values [33]. The initial list of protein chains were further processed to remove NMR structures and X-ray structures with low resolution (>3.5 Å), which resulted in 31,206 protein chains. BLASTClust was used to group the protein chains into 11,541 clusters with cut-offs set at 50% and 40% for sequence similarity and alignment coverage respectively [34]. The sequence of each protein chain is derived from the SEQRES record in PDB. The purpose of this clustering step is to reduce the number of pairwise comparisons needed to find the indel sequences. A non-redundant indel dataset was generated after four filtering steps and was then subjected to statistical analysis, such as amino acids composition, secondary structure types, relative solvent accessibility, and local/global structural changes induced by the indels. Here we briefly describe the details of these filtering steps for generation of a non-redundant indel sequence dataset.

It is possible that two or more sequences in each cluster are redundant as shown in Figure 2 in which there is only one unique indel sequence instead of 10 redundant ones from 10 pairwise alignments. Therefore the first step is to remove such redundant protein chains in each cluster that has at least two members using a similar approach as described by Pascarella and Argos [10]. Briefly, if two sequences are highly similar with no internal gaps in the alignment, the one with lower resolution is removed from the cluster. The Needle program in the EMBOSS package, an implementation of the Needleman-Wunsch global alignment algorithm, was used for sequence alignment with default parameters [35,36].

The second step of the procedure is to ensure the uniqueness of indel sequences and locations by checking the local sequence similarity. Two proteins with high global sequence identity may have regions that show low local sequence similarity [28]. If an indel happens to be in the low sequence similarity area, the placement of this indel may change dramatically with minor changes of alignment parameters, resulting in different indel sequences and locations. In our approach, similarities of the sequences flanking the indels (20 amino acids on each side) were calculated. We only consider indel sequences from protein pairs with both high global sequence similarity and highly similar flanking regions (above a cutoff value of 75%).

Due to the discrepancies of deposition of SEQRES in PDB by experimental structural biologists, some indels derived from the SEQRES sequence comparisons are not true indels, especially in the cases of disordered fragments. For example, the sequence alignment between proteins 1XJIA and 1C8SA of the same protein, bacteriorhodopsin, shows an internal gap based on the SEQRES sequences (Figure 3). However, the fragment 154-175 in 1C8SA is disordered in the X-ray structure

Figure 3 An example of a false indel sequence derived from 1XJIA-1C8SA. Dashed red line represents the disordered fragment. The Cα distance between the two residues (F and N) that flank the disordered fragment in 1C8SA is 8.54 Å.
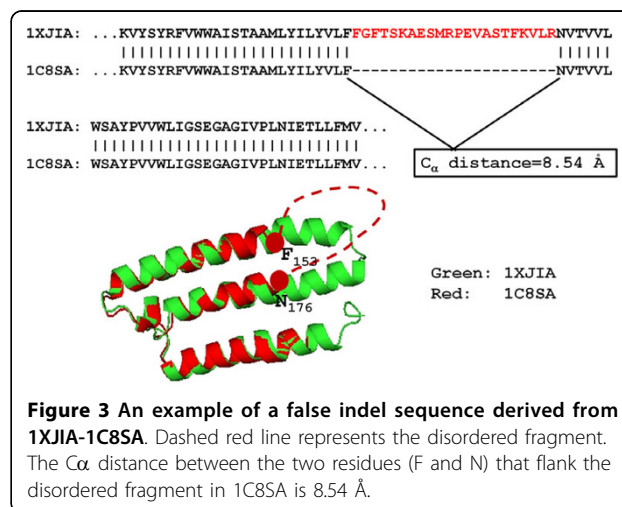
and the sequence of this fragment was not reported in the SEQRES. The same fragment adopts an α-helical structure in 1XJIA and appears in the SEQRES record. These types of fragments that adopt ordered conformation in one structure and are disordered in another structure have been dubbed as "dual personality" fragments [37]. If the disordered fragments were not reported in SEQRES section, such "dual personality" fragment would introduce false indels. A simple way to identify the false indels is to check the C_α distance between the two indel flanking residues in the short form. If the two residues are connected, the C_α distance should be around 3.85Å (data not shown). An indel is flagged as false if the C_α distance is more than a cutoff value (4.5Å in this study). In the above example, the C_α distance between F_{153} and N_{176} in 1C8SA is 8.54Å. Therefore this indel of 22 residues was flagged as a false one and was removed from the dataset (Figure 3).

Lastly, indel sequences were further processed to generate a non-redundant dataset of indels ranging from 2 to 40 amino acids. Two indel sequences are considered redundant 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.

To check if an internal indel is a result from engineered mutants or from natural variants, we combined the information from the SEQADV record of PDB files with manual inspection of related publications. The PDB SEQADV record describes conflicts between residue sequences in the ATOM/HETATM records and those in sequence databases [6]. Since there are several possible reasons for these conflicts, including engineered mutants, natural variants, disordered fragments, or cloning artifact, careful manual inspection is needed.

Secondary structure types and relative solvent accessibility of indel residues

Each indel residue was assigned to one of four secondary structure states, helix, strand, coil and disordered. DSSP program was used to assign the first three secondary structure states: helix, strand and coil [38]. Following the widely used convention, H (α-helix), G (3_{10}-helix) and I (π-helix) from DSSP are classified as helix type while E (extended strand) and B (residue in isolated β-bridge) states are classified as strand type. All the other states from DSSP are considered as coil. The disordered residues are 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 [39]. The relative solvent accessibility was calculated by dividing the

absolute value from DSSP by the maximum accessibility of each residue [40]. We employ a three-state classification for relative solvent accessibility: buried ($\leq7\%$), intermediate ($>7\%$ and $\leq37\%$), and exposed ($>37\%$) [41]. The disordered/unstructured residues were considered as exposed in solvent accessibility analysis.

For comparison purpose, 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 of better than 2.5 Å, and the size is in the 50-1000 amino acids range, was used as background analysis for amino acid composition, secondary structure types, and residue solvent accessibility.

Protein sequence and structure comparisons
Sequence alignment is done with the Needle program with default parameters [35,36]. Two different structure alignment programs, FAST [42] and CE [43], were used for global and local structure alignment respectively. The structural difference/similarity is measured by the $C\alpha$ RMSD of aligned residues between two structures. The structural changes induced by indels were evaluated by comparing the structure and sequence alignments of each pair. A webserver SCINDEL http://bioinfozen.uncc.edu/scindel was developed for convenient visualization of both the sequence and structure alignments and structural changes caused by indels.

Results and discussion
A non-redundant dataset of short internal indels from highly homologous protein pairs
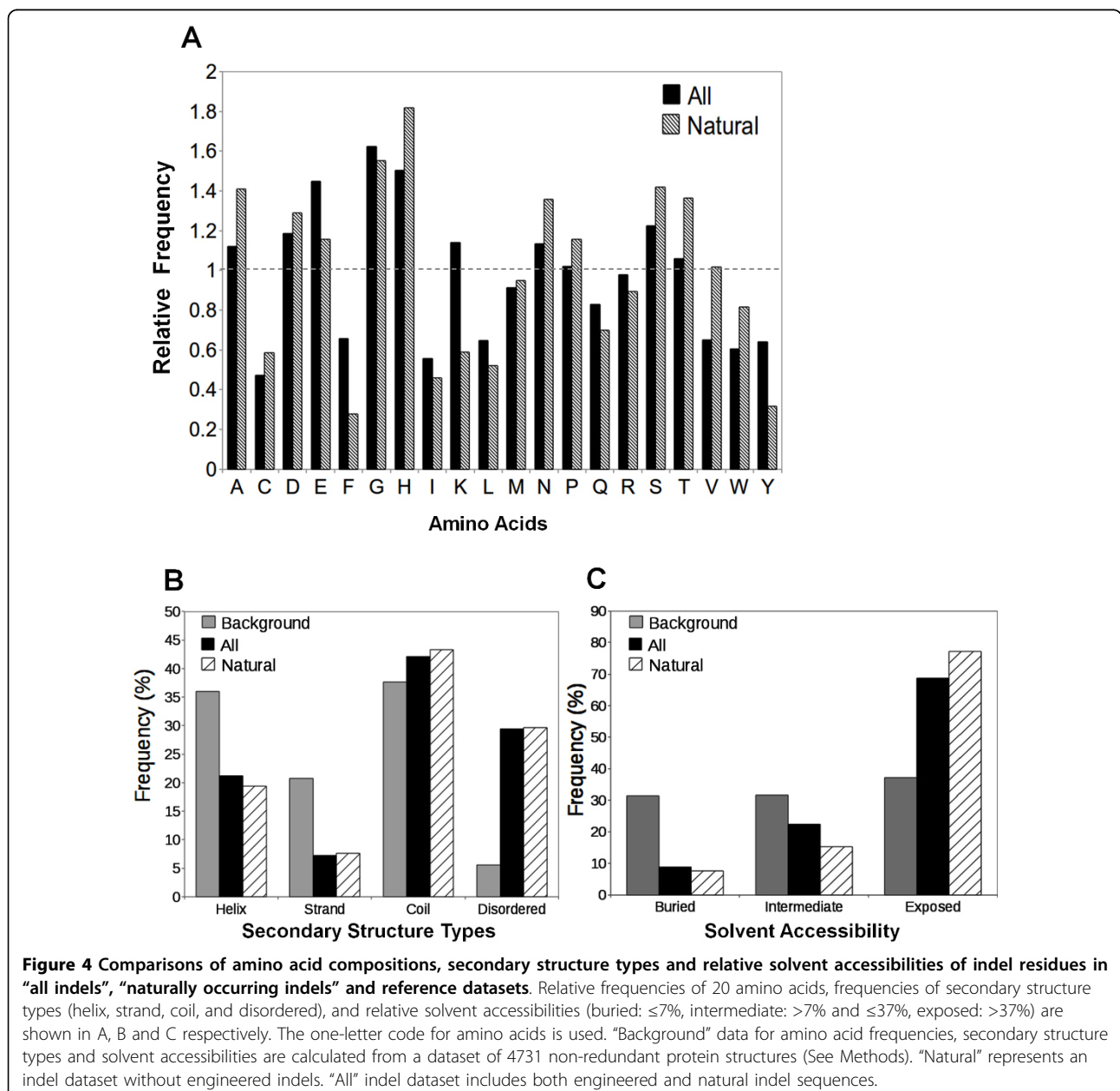The protein chains were first clustered into 11,541 groups using BLASTClust as described in the Methods section. The first filtering step of removing redundant protein chains in each cluster resulted in 1,932 clusters that have two or more members. A total of 1,237,062 internal indels were identified from 445,552 distinct protein pairs. A dataset of 1,114 non-redundant indels were generated after removing indels that are false, redundant, or the results of low local sequence similarity. As described earlier [28], the dataset is rich in indels (931 of 1114) derived from one specific family member, immunoglobulin variable domain (b.1.1.1) based on the latest SCOP release 1.75 [44]. These indels are generally located in the CDR-H3 loops that play crucial roles in antigen recognition and binding specificity [32,45]. The CDR-H3 loops are dominated by residues tyrosine, glycine, and serine, but have fewer lysine, glutamine, and glutamic acid [28,31,32]. Due to the over-representation of indels from the immunoglobulin family, these indels were removed and the final non-redundant indel dataset includes 183 indels. The detailed information for each indel, including length, amino acid sequence, host

proteins, start and end positions of the indel, and SCOP classification is available at http://bioinfozen.uncc.edu/scindel/nonredundant_indels.html.

Statistical analysis of the indel sequences and structures

The dataset with all short internal indels has 1301 total residues and is rich in residues glycine, histidine, glutamic acid, aspartic acid, and serine, but are depleted in residues cysteine, phenylalanine, isoleucine, leucine, tryptophan, and tyrosine when compared with the background residue frequencies ("All" in Figure 4A and Additional file 1, Figure S1). The relative frequency of each amino acid in Figure 4A is normalized with its background frequency. Residues isoleucine and leucine have high propensity to adopt α-helix or β-sheet conformations while glycine, aspartic acid, and serine prefer to be in loops. The amino acids compositions of indels suggest that indel sequences assume less regular secondary structures and prefer to be in more flexible regions, which are supported by the analysis of secondary structure types (Figure 4B). 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

Figure 4 Comparisons of amino acid compositions, secondary structure types and relative solvent accessibilities of indel residues in "all indels", "naturally occurring indels" and reference datasets. Relative frequencies of 20 amino acids, frequencies of secondary structure types (helix, strand, coil, and disordered), and relative solvent accessibilities (buried: ≤7%, intermediate: >7% and ≤37%, exposed: >37%) are shown in A, B and C respectively. The one-letter code for amino acids is used. "Background" data for amino acid frequencies, secondary structure types and solvent accessibilities are calculated from a dataset of 4731 non-redundant protein structures (See Methods). "Natural" represents an indel dataset without engineered indels. "All" indel dataset includes both engineered and natural indel sequences.

sequences have a markedly increased number of residues in disordered state (over five-fold increase) (Figure 4B). Figure 4C shows that most indel residues (~70%) are exposed to solvent. Similar observations have been reported for alternative splicing events that, by and large, prefer coil regions and exposed residues [16,17]. It should be pointed out that some disordered fragments may fold as regular secondary structures under different conditions, such as the existence of ligands or other proteins. This type of fragments that can exist in both ordered and disordered states have been termed as "dual personality" fragments [37]. Taken together, only a small percentage of the indel residues fold into regular secondary structures and are embedded inside of the proteins. χ^2 analysis based on the observed and expected (background frequencies as references) numbers indicates that the differences are statistically significant with very low *p*-values (Additional file 1, Table S1).

We also checked the conformation of five residues flanking the indel sequences on each side and found that more residues are in β-sheet or α-helix states when they are further way from the indel sites (Additional file 1, Figure S2).

Source of the non-redundant indel sequences

A number of studies have been devoted to study the structural and functional consequences of insertions and deletions with artificially engineered constructs [12-15]. For example, Matthews' group has used T4 lysozyme as a model system to investigate the effect of insertions on protein structures, including an insertion in an α-helix [13,14]. The very question about the indel dataset in this study is how many of the indel sequences are from naturally occurring proteins rather than "man-made". Based on the SEQADV records of PDB files and manual inspection, we found that about 70% of the indel sequences in our dataset are the results of engineered insertion/deletion mutants. Besides the T4 lysozyme structural studies, the majority of the engineered insertion/deletion mutants were constructed to investigate the functional importance of a particular fragment. Deletion mutants have also served as one of the rational engineering approaches to better crystallization in protein structure determination [21]. Indels from protein pair 2J4OA-2POPA (19 AAs) and 2RHSB-2RHQB (4 AAs) are such examples. The indel statistics from naturally occurring indels in terms of secondary structure types and solvent accessibilities are significantly different from those of the reference dataset and are highly similar to those derived from all indels (Figures 4B and 4C). Though there are variations in amino acid frequencies between all indels and naturally occurring ones, the general trend is surprisingly similar (Figure 4A). Due to the small size of the naturally occurring indel dataset (55 indels, 263 residues), the

significances of differences between the observed numbers of amino acids, secondary structure types, and relative solvent accessibility and the expected numbers (using either "Background" or "All" indels as references) were calculated with χ^2 test. Not surprisingly, there are significant differences between the observed numbers in the naturally occurring indel dataset and the expected numbers (based on background distributions) with *p*-values of $5.1e^{-4}$, $2.6e^{-68}$, and $9.32e^{-41}$ for amino acid, secondary structure types, and relative solvent accessibility, respectively (Additional file 1, Table S2). More importantly, there are no statistically significant differences between the naturally occurring indels and the all indel dataset in terms of amino acid frequencies ($p = 0.46$) and secondary structure types ($p = 0.91$) (Additional file 1, Table S2), suggesting that the features from our all indel dataset may represent the properties of real, non-engineered indel sequences. Statistical analysis also showed that the levels of solvent accessibility are different at a significance level of 0.01 between all indels and naturally occurring indels ($p = 0.0094$), largely due to the differences in the intermediate and exposed categories (Figure 4C). Unlike secondary structure types, the classification of relative solvent accessibility into three categories is rather broad and arbitrary; so the differences may become minor if different bins or classifications are used.

Impacts on structural changes by indels

To investigate the structural changes caused by indels, we used two structural alignment programs FAST [42] and CE [43] for global and local structural comparison respectively. Figure 5A shows the distribution of the global structure differences in terms of RMSD. Most of the structural pairs are highly similar with about 85% of the protein pairs having less than 2Å RMSDs, suggesting that these protein structures in general can tolerate and accommodate the indels [12,18]. Intuitively, the RMSDs that measure the structural differences would increase with the length of indels. However, there is no clear relationship between the RMSDs and the length of the indels (Figure 5B). Many long indels lead to minimal conformational changes, which is not surprising as majority of the indel sequences adopted either coil or disordered "conformation". This observation is consistent with the previous reports that insertions/deletions are most likely to occur in loop regions or between regular secondary structure elements and thus preserve the overall structural fold [12].

On the other hand, several large RMSDs are associated with relatively short indel sequences (Figure 5B). We found that all the nine pairs with RMSDs of 4Å or more are the results of indels in the hinge area, causing changes in the relative orientations of the domains connected by the indels, rather than changing to

Figure 5 Structural comparisons of protein pairs with indel sequences using FAST [42]. (A) Distribution of global RMSDs; (B) relationship between global RMSDs and indel lengths.
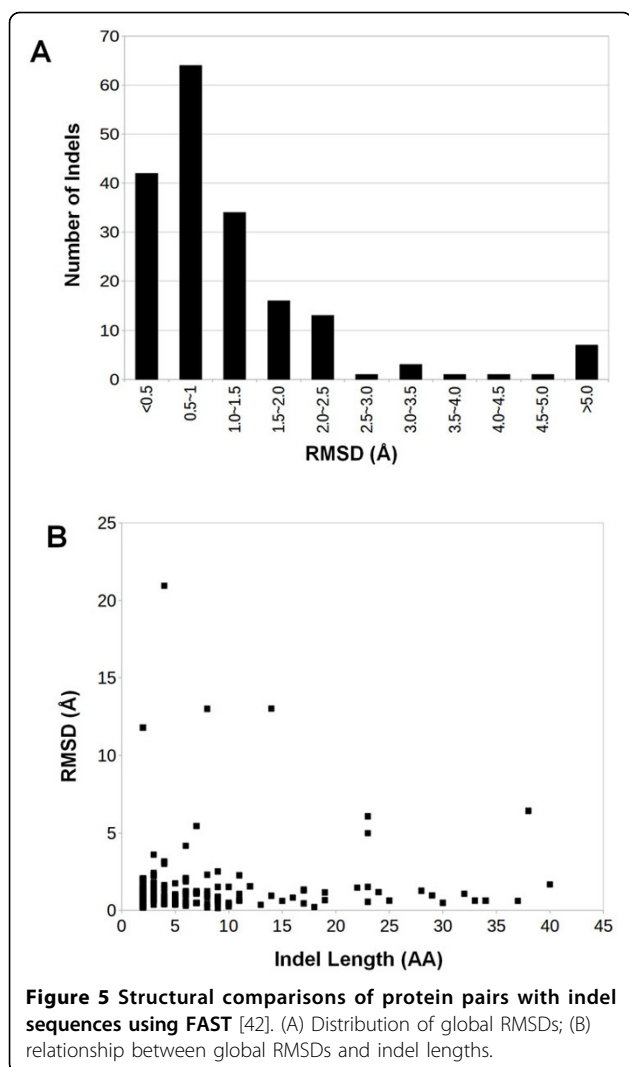
different folds. For example, 1Y64B-1UX4A (with a four-residue indel sequence "REDL" folding into a helical structure) has the largest global RMSD of 20.94Å (Figure 6A). However, the N- and C-terminal domains separated by the indel sequence have almost identical structures in 1Y64B and 1UX4A, with RMSDs of 0.95Å and 1.11Å respectively (Figure 6B and 6C). The obvious question in these cases is whether the short indels induce such domain movements. It turned out that the intrinsic flexibility and dynamics of proteins may play a bigger role in the structural changes. For example, 1UX5A and 1Y64B have different domain positions though both crystal structures are from the same protein (Figure 6D).

While it is generally accepted that insertion/deletion in loops introduces minimal structural changes, the effects of indels on regular secondary structures,

especially β-strands, are greatly debated [18,46,47]. The deletion of β-strands in a β-sheet also presents a challenging problem for comparative modeling approaches. In our indel dataset, there are fifteen cases in which indels fold into α-helices or β-strands and are less exposed. In almost all cases, the core secondary structures tend to be conserved even though the shorter form lacks part of the sequence that folds as a strand or a helix in the longer form (Figure 7A-H). In these protein pairs, part of a strand (1RJ8A-1RJ7A, Figure 7A&7B), a helix (5PGMA-3PGMA, Figure 7C&7D), or a combination of strand and helix conformations (1EKXA-2ATCA, Figure 7E-H) at the sites of indels are conserved with small changes in the neighboring loop areas through local structure rearrangement and refolding. For example, the indel sequence with eight residues (MAEVDILY) from 1EKXA folds as a helix-loop-strand. In the short form of 2ATCA, the eight residues after the indel (MTRVQKER) assume the same structural conformation of the indel sequence while the downstream loop becomes shorter and deviates from the loop conformation in the long form (Figure 7E-H). The slight conformational change in the loop is not surprising as the loop is on the surface area and is flexible. Nevertheless, it shows the inherent capability of proteins to tolerate short structural deletions and insertions [18,47].

We shall point out that even though all the short indels (engineered or natural) in our dataset do not show big impact on protein structures, it does not necessary mean that short internal indels have no deleterious effect. First of all, all the proteins with natural indels in our dataset are probably the ones that survived from evolution events while the others with dramatic structural effect might have disappeared during evolution. Secondly, even in the cases where indels do not induce structural change, a disastrous loss of function may occur. Nevertheless, these data (natural and engineered indels) strongly suggest the inherent structural plasticity of protein structures [16-18].

All the protein pairs with indels http://bioinfozen.uncc.edu/scindel/nonredundant_indels.html can be visualized at both sequence and structural level using our SCINDEL (Structural Comparison of Similar Proteins with Insertion and Deletion) webserver at http://bioinfozen.uncc.edu/scindel. Figure 8 shows a snapshot of the comparisons between 1GSAA and 1GLVA with an indel of 13 amino acids.

Conclusions

We performed a systematic study to investigate the impact of short internal indels on protein structures through mining the highly homologous protein pairs in

Figure 6 Global and local structure alignments. (A) Structural alignment between 1Y64B and 1UX4A; (B) and (C) N-terminal and C-terminal structural alignments, respectively; and (D) structural alignment between 1Y64B and 1UX5A. Green: 1Y64B. Red: 1UX4A in A, B, and C and 1UX5A in D.

Figure 7 Structural comparisons of proteins with indels adopting α-helix and/or β-strand conformations. (A, B): 1RJ7A-1RJ8A; (C, D): 5PGMA-3PGMA; (E, F): 1EKXA-2ATCA. (A, C, E): whole structure alignments; (B, D, F): highlights of alignments in the indel region; (G): part of the sequence alignment between 1EKXA and 2ATCA involving the indel sequence MAEVDILY; (H): part of the structural alignment between 1EKXA and 2ATCA involving the indel sequence. Green: long protein; Red: short protein; Blue: indel sequence.

PDB. In addition to protein evolution, indels can be the results of alternative splicing. We found that short internal indels tend to occur between secondary structure elements and a significant number of indels are disordered, which is in agreement with the earlier studies that demonstrated the associations among indels, structural disorder, and functional diversity [17]. The rationale of choosing highly homologous protein pairs (with high sequence identity for both overall and indel flanking sequences) is two-fold: 1) to avoid "random" positioning of indel(s) in a protein pair due to low local sequence similarity even though overall sequence similarity is high; and 2) to provide a better approximation to the AS isoforms with internal gaps (100% identical in indel flanking regions). These steps ensure unambiguous indel sequences and their unique positions, reducing the possibility of including false indels due to sequence alignment errors.

One important observation from this study is that most of the indels in the dataset are not derived from evolution events. Indels have been engineered into proteins for various purposes, including structural and functional studies of short peptides and better protein crystallization. Our statistical analysis showed that there are significant differences between naturally occurring indels and the control dataset. On the other hand, there are no statistically significant differences between naturally occurring indels and all indels in terms of amino acid frequencies and secondary structure types. These data suggest that the indel properties derived from our all indel dataset are very useful.

The very question about modeling isoform structures or structural changes due to indels is how to improve
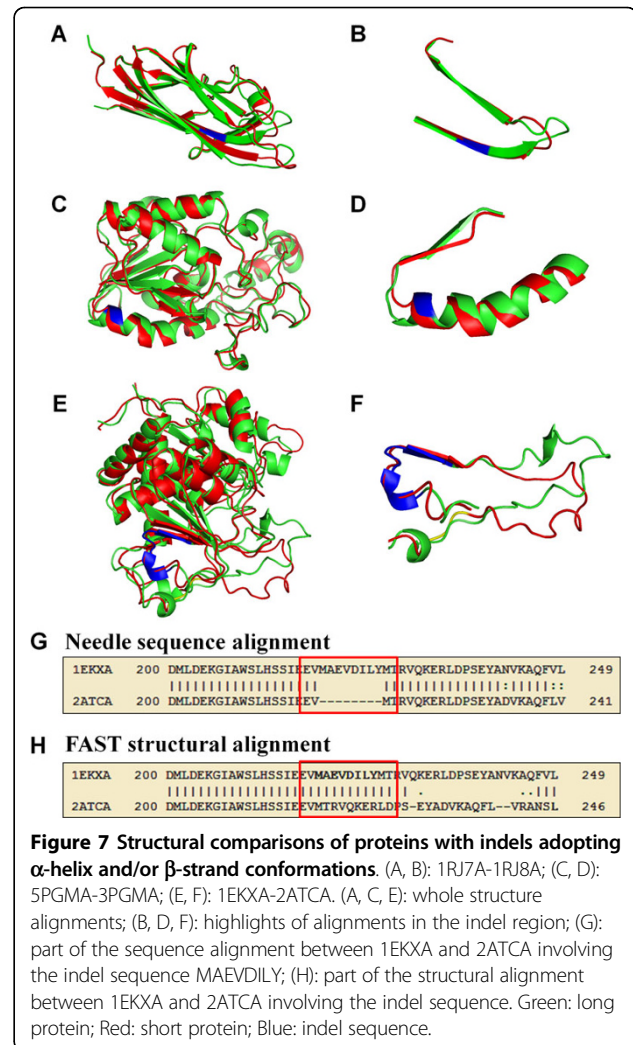
the sequence alignment for comparative modeling since the performance of current comparative modeling techniques rely heavily on accurate alignments [8]. Very rarely, sequence alignment errors can be recovered by current comparative modeling programs. We believe this systematic analysis, along with earlier reports on the case studies with individual or a small number of indel pairs, will help us in this regard as well as in our understanding of the structural plasticity of proteins.

Additional material

Additional file 1: Figure S1: Comparison of amino acid frequencies of indel sequences in "all indels" (All), "naturally occurring indels" (Natural) and reference (Background) datasets. Figure S2: Frequencies of secondary structure types for residues flanking indel sequences. Table S1: Statistical significance analysis of the observed numbers in "all indels" dataset. Table S2: Statistical significance analysis of the observed numbers in naturally occurring indels.

Figure 8 A snapshot of SCINDEL webserver. Sequence and structure alignments between 1GSAA and 1GLVA with an indel sequence of 13 amino acids.

List of abbreviations used
PDB: Protein Data Bank; RMSD: Root Mean Square Deviation; AS: Alternative Splicing; Indel: Insertion/deletion;

Authors' contributions
JTG designed and supervised the study and wrote the manuscript. JTG was also involved in data analysis. RGK developed the webserver and the programs for data generation, and participated in data analysis. Both authors read and approved the final manuscript.

References
1. Pennisi E: **Why do humans have so few genes?** *Science (New York, NY)* 2005, **309(5731)**:80..
2. Xing Y, Lee C: **Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes.** *Nature reviews* 2006, **7(7)**:499-509.
3. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456(7221)**:470-476.
4. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, Lopez G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Storling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, *et al*: **The implications of alternative splicing in the ENCODE protein complement.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104(13)**:5495-5500.
5. UniProt-Consortium: **The Universal Protein Resource (UniProt) 2009.** *Nucleic acids research* 2009, , **37 Database**: D169-174.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic acids research* 2000, **28(1)**:235-242.
7. Stetefeld J, Ruegg MA: **Structural and functional diversity generated by alternative mRNA splicing.** *Trends in biochemical sciences* 2005, **30(9)**:515-521.
8. Keedy DA, Williams CJ, Headd JJ, Arendall WB, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC, Richardson JS: **The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models.** *Proteins* 2009, **77(Suppl 9)**:29-49.
9. Garcia J, Gerber SH, Sugita S, Sudhof TC, Rizo J: **A conformational switch in the Piccolo C2A domain regulated by alternative splicing.** *Nature structural & molecular biology* 2004, **11(1)**:45-53.
10. Pascarella S, Argos P: **Analysis of insertions/deletions in protein structures.** *Journal of molecular biology* 1992, **224(2)**:461-471.
11. Wright CF, Christodoulou J, Dobson CM, Clarke J: **The importance of loop length in the folding of an immunoglobulin domain.** *Protein Eng Des Sel* 2004, **17(5)**:443-453.
12. Laskowski RA, Thornton JM: **Understanding the molecular machinery of genetics through 3 D structures.** *Nature reviews* 2008, **9(2)**:141-151.
13. Heinz DW, Baase WA, Dahlquist FW, Matthews BW: **How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme.** *Nature* 1993, **361(6412)**:561-564.
14. Sagermann M, Baase WA, Matthews BW: **Sequential reorganization of beta-sheet topology by insertion of a single strand.** *Protein Sci* 2006, **15(5)**:1085-1092.
15. Vetter IR, Baase WA, Heinz DW, Xiong JP, Snow S, Matthews BW: **Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme.** *Protein Sci* 1996, **5(12)**:2399-2415.
16. Wang P, Yan B, Guo JT, Hicks C, Xu Y: **Structural genomics analysis of alternative splicing and application to isoform structure modeling.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(52)**:18920-18925.
17. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK: **Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(22)**:8390-8395.
18. Birzele F, Csaba G, Zimmer R: **Alternative splicing and protein structure evolution.** *Nucleic acids research* 2008, **36(2)**:550-558.
19. Halle B: **Flexibility and packing in proteins.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(3)**:1274-1279.
20. Li X, Romero P, Rani M, Dunker AK, Obradovic Z: **Predicting Protein Disorder for N-, C-, and Internal Regions.** *Genome Inform Ser Workshop Genome Inform* 1999, **10**:30-40.
21. Dale GE, Oefner C, D'Arcy A: **The protein as a variable in protein crystallization.** *Journal of structural biology* 2003, **142(1)**:88-97.
22. Carson M, Johnson DH, McDonald H, Brouillette C, Delucas LJ: **His-tag impact on structure.** *Acta Crystallogr D Biol Crystallogr* 2007, **63(Pt 3)**:295-301.
23. David FP, Yip YL: **SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase.** *BMC bioinformatics* 2008, **9**:391.
24. Benner SA, Cohen MA, Gonnet GH: **Empirical and structural models for insertions and deletions in the divergent evolution of proteins.** *Journal of molecular biology* 1993, **229(4)**:1065-1082.
25. Wrabl JO, Grishin NV: **Gaps in structurally similar proteins: towards improvement of multiple sequence alignment.** *Proteins* 2004, **54(1)**:71-87.
26. Chang MS, Benner SA: **Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments.** *Journal of molecular biology* 2004, **341(2)**:617-631.
27. Hsing M, Cherkasov A: **Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins.** *BMC bioinformatics* 2008, **9**:293.
28. Kim RM, J Guo J-T: **A systematic study of homologous protein structures with insertions/deletions.** *Computational Systems Bioinformatics: 2009* Stanford, CA, USA: Life Sciences Society 2009, **000**:103-113.
29. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z: **Intrinsically disordered protein.** *J Mol Graph Model* 2001, **19(1)**:26-59.
30. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK: **Intrinsic disorder in cell-signaling and cancer-associated proteins.** *Journal of molecular biology* 2002, **323(3)**:573-584.
31. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS: **The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies.** *Journal of molecular biology* 2008, **377(5)**:1518-1528.
32. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder HW Jr, Kirkham PM: **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.** *Journal of molecular biology* 2003, **334(4)**:733-749.
33. Wang G, Dunbrack RL Jr: **PISCES: a protein sequence culling server.** *Bioinformatics (Oxford, England)* 2003, **19(12)**:1589-1591.
34. BLASTClust. [http://www.ncbi.nlm.nih.gov/].
35. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *Journal of molecular biology* 1970, **48(3)**:443-453.
36. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6)**:276-277.
37. Zhang Y, Stec B, Godzik A: **Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins.** *Structure* 2007, **15(9)**:1141-1147.
38. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22(12)**:2577-2637.
39. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK: **DisProt: a database of protein disorder.** *Bioinformatics (Oxford, England)* 2005, **21(1)**:137-140.

40. Miller S, Janin J, Lesk AM, Chothia C: **Interior and surface of monomeric proteins.** *Journal of molecular biology* 1987, **196(3)**:641-656.
41. Kim D, Xu D, Guo JT, Ellrott K, Xu Y: **PROSPECT II: protein structure prediction program for genome-scale applications.** *Protein engineering* 2003, **16(9)**:641-650.
42. Zhu J, Weng Z: **FAST: a novel protein structure alignment algorithm.** *Proteins* 2005, **58(3)**:618-627.
43. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein engineering* 1998, **11(9)**:739-747.
44. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic acids research* 2004, , **32 Database**: D226-229.
45. Xu JL, Davis MM: **Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities.** *Immunity* 2000, **13(1)**:37-45.
46. Grishin NV: **Fold change in evolution of protein structures.** *Journal of structural biology* 2001, **134(2-3)**:167-185.
47. Guo JT, Jaromczyk JW, Xu Y: **Analysis of chameleon sequences and their implications in biological processes.** *Proteins* 2007, **67(3)**:548-558.