BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 5 2010, pages 625–631
doi:10.1093/bioinformatics/btq012

Structural bioinformatics

Advance Access publication January 16, 2010

Protein secondary structure appears to be robust under *in silico* evolution while protein disorder appears not to be

Christian Schaefer[1,2,*], Avner Schlessinger[3] and Burkhard Rost[1,2,4]

[1]Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics (C2B2), Columbia University, 1130 St Nicholas Ave. Rm. 802, New York, NY 10032, USA, [2]Department of Computer Science, Institute of Advanced Studies (IAS), NorthEast Structural Genomics Consortium (NESG), TUM Bioinformatics, TUM Munich, Boltzmannstr. 3, 85748 Garching, Germany, [3]Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, 1700, 4th Street, San Francisco, CA 94158 and [4]NorthEast Structural Genomics Consortium (NESG) and New York Consortium on Membrane Protein Structure (NYCOMPS), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The mutation of amino acids often impacts protein function and structure. Mutations without negative effect sustain evolutionary pressure. We study a particular aspect of structural robustness with respect to mutations: regular protein secondary structure and natively unstructured (intrinsically disordered) regions. Is the formation of regular secondary structure an intrinsic feature of amino acid sequences, or is it a feature that is lost upon mutation and is maintained by evolution against the odds? Similarly, is disorder an intrinsic sequence feature or is it difficult to maintain? To tackle these questions, we *in silico* mutated native protein sequences into random sequence-like ensembles and monitored the change in predicted secondary structure and disorder.

Results: We established that by our coarse-grained measures for change, predictions and observations were similar, suggesting that our results were not biased by prediction mistakes. Changes in secondary structure and disorder predictions were linearly proportional to the change in sequence. Surprisingly, neither the content nor the length distribution for the predicted secondary structure changed substantially. Regions with long disorder behaved differently in that significantly fewer such regions were predicted after a few mutation steps. Our findings suggest that the formation of regular secondary structure is an intrinsic feature of random amino acid sequences, while the formation of long-disordered regions is not an intrinsic feature of proteins with disordered regions. Put differently, helices and strands appear to be maintained easily by evolution, whereas maintaining disordered regions appears difficult. Neutral mutations with respect to disorder are therefore very unlikely.

Contact: schaefer@rostlab.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2009; revised on December 18, 2009; accepted on January 9, 2010

1 INTRODUCTION

Random, undirected mutation is a major driving force for change in nature. In the protein universe, selection is realized through function: mutations leading to loss of function are rarely observed. As protein structure determines protein function, it is also subjected to evolutionary selection. Most problematic single nucleotide polymorphisms (SNP) that alter the amino acid sequence (non-synonymous SNPs) appear to impact the stability of protein structure (Yue *et al*., 2005; Yue *et al*., 2006).

Helices and strands constitute the major macromolecular building blocks of all ‘well-ordered’ proteins (Benner *et al*., 1997; Kabsch and Sander, 1983; Levitt and Chothia, 1976; Morea *et al*., 1998; Pauling and Corey, 1951a; Pauling and Corey, 1951b). The particular 3D structure of a protein is assumed to correspond to the global minimum free energy and hence defines the unique fold of an amino acid polymer (Anfinsen and Scheraga, 1975; Dill, 1993; Karplus and Petsko, 1990; Levitt and Warshel, 1975; Liwo *et al*., 1999; Reva *et al*., 1995; Sippl, 1993). Another essential feature of protein structure is the unique interplay between well-ordered and flexible regions (Alexov and Gunner, 1997; Cavasotto and Abagyan, 2004; Claussen *et al*., 2001; Daniel *et al*., 2003; Gu *et al*., 2006; Morea *et al*., 2000; Radivojac *et al*., 2004; Schlessinger *et al*., 2006). One particular aspect of this interplay is that between what we may loosely refer to as ‘order’ and ‘disorder’ (Dunker and Obradovic, 2001; Dunker *et al*., 2008; Radivojac *et al*., 2004; Uversky, 2003).

Many proteins have regions that remain ‘unstructured’ unless bound to a substrate: they do not adopt a unique stable conformation in isolation. Such regions are also referred to as *intrinsically disordered* or simply as *disordered*. Our operational definition for this vague term is: *we consider as disorder whatever is predicted as such*. Proteins with long-disorder regions have unique biophysical traits that enable the binding to different substrates, often at different cellular conditions (Wright and Dyson, 2009). Very long regions without regular secondary structure (loosely referred to as ‘loops’) may resemble disorder (Liu *et al*., 2002); nevertheless, we can clearly distinguish between disorder-like and well-structured loops (Schlessinger *et al*., 2007a; Schlessinger *et al*., 2009). Disorder is an important ‘building block’ for the increase in complexity in the evolution from unicellular prokaryotes to multi-cellular eukaryotes.

[*] To whom correspondence should be addressed.

Our two hypotheses were: (i) we assumed that regular secondary structure is difficult to maintain evolutionarily, i.e. single residue mutations are likely to impact helices and strands and that we would lose regular secondary structure and transit into 'loopy' polypeptide chains with increasing random mutations away from the native state. (ii) We assumed, furthermore, that disordered regions provide a means to become robust against mutations because most mutations would rather increase than decrease disorder by increasing the non-regular secondary structure. Here, we present results that falsify both hypotheses as clearly as possible without investing tens of millions of dollars.

2 METHODS

2.1 Datasets

We used protein sequences from two databases for the *in silico* mutation. First, we assessed the robustness of secondary structure through globular proteins from the Protein Data Bank (PDB) (Berman *et al*., 2000). Secondly, we assessed the robustness of disordered regions through proteins from DisProt (Vucetic *et al*., 2005) (version 4.9). We applied UniqueProt (Mika and Rost, 2003) to reduce the redundancy in both sets filtering at a sequence similarity threshold of HVAL > 10 (Rost, 1999; Sander and Schneider, 1991) (this corresponds to ~30% pairwise sequence identity—PIDE—for alignments over 250 residues). The redundancy-reduced sets comprised 1369 (PDB) and 374 (DisProt) proteins.

For each of the two datasets (PDB and DisProt), we also created random sequences that had the same amino acid composition, same length distribution and same number of sequences as the natives. The random sets served as convergence control: if we mutate enough to 'lose all memory' (convergence), the random sets will not differ from the mutated sets.

To shed light on potential biases from the chosen databases, we additionally predicted the secondary structure in 33 812 proteins, representing the entire human proteome as taken from RefSeq 2006.

Finally, we sub-sampled a set of sequences from the PDB set with the same size, amino acid and length distribution as that of the DisProt set to examine the ability of ordered proteins to retain or lose their ordered state.

2.2 Mutation protocol

We gradually mutated native protein sequences into quasi-random strings of amino acids by the following iterative procedure.

2.2.1 One mutation step It consisted of two moves: (i) select a particular residue position, i.e. site in the sequence to mutate, and (ii) mutate the amino acid X at that position with amino acid Y with the probability p_{XY} (X=Y). For technical reasons (lack of CPU because after each step we have to apply several prediction methods), we repeat these two moves $N/10$ times (N number of residues in the protein). Effectively, we thereby touch 10% of all residues in one mutation step.

2.2.2 Sixty-nine mutation steps We carried out 69 mutation steps (with $69\times N/10$ mutations) for each protein. Any other, sufficiently large, number would have worked. We chose 69 because we had reached convergence in all the cases that we looked at in detail after 65 steps.

Effectively, we applied a Markovian-like model for evolution, i.e. assuming that each residue mutates independently of all others and that the mutation depends only on the amino acid type. We applied three alternative substitution schemes: (i) we mutated according to the PAM120 probability (Dayhoff, 1978). (ii) PAM120 is valid for great evolutionary distance. In order to also cover closer relations, we also implemented BLOSUM62 (Henikoff and Henikoff, 1992). (iii) Finally, we took the underlying amino acid distribution in the database (PDB, DisProt—ordered/disordered regions in DisProt not distinguished) as substitution probabilities. Note that for the most PAM120 and BLOSUM62 mutations, the most likely 'mutation step' was the maintenance of the current amino acid as the diagonals are typically highest in these matrices. We did not consider mutations that led to insertions or deletions. BLOSUM62 and PAM120 behaved identically with respect to our results. For readability, we confined the BLOSUM62 results to the Supplementary Material.

2.2.3 Single trajectory versus ensemble The 'mutation path' for each native sequence constitutes a single unique trajectory in the space of all possible mutations. We created five different such single paths (five different mutants) in order to investigate the divergence from the native of an ensemble of evolutionary paths. From these five, we compiled a consensus by per-residue averaging over each of the five predictions (secondary structure/disorder). Note that by default, we reported the results for single trajectories and added the ensemble comparison only where explicitly stated.

2.3 Secondary structure

We predicted secondary structure through PROFsec (Rost, 2005). Secondary structure prediction methods improve when using evolutionary information (Liu and Rost, 2001; Rost, 1996; Rost and Sander, 1993). Without this information, PROFsec reaches a sustained single-sequence level of ~68% three-state per-residue accuracy (Q_3 is the percentage of residues predicted correctly in one of the three states helix, strand and other). We had to use this single-sequence mode to monitor the effect of point mutations. Prediction mistakes might invalidate the generality of our findings. One way in which we addressed this concern was by monitoring the parameters that we plotted for our mutants also for the experimental observations from the native proteins as taken from DSSP (Kabsch and Sander, 1983) with the usual conversion of eight into three 'states' (Andersen *et al*., 2002; Rost, 1996; Rost and Sander, 1993). For each mutation step (i.e. after each step of 10% change), we monitored the sequence similarity compared with the native sequence, the relative content of residues predicted in helix and strand and the average length of predicted helices and strands.

2.4 Disordered regions

We predicted disordered regions by three methods: IUPred (Dosztányi *et al*., 2005), MD (Schlessinger *et al*., 2009) and VSL2 (Obradovic *et al*., 2005; Peng *et al*., 2006) and compared the predictions to the experimental annotations in DisProt. IUPred has three options (*long*, *short* and *glob*); we chose *short* for short and *long* for long disorder. MD (Meta Disorder predictor) combines independent methods through machine learning. We used it without alignments. VSL2 is a collection of eight methods. We used the VSL2B variant that uses only single sequences as input.

The three methods focus on different aspects of disorder and have different strengths and weaknesses. We did not combine methods and, for simplicity, focused only on IUPred. The results from the other methods that were crucial to rule out method-specific findings are given in the Supplementary Material. We chose IUPred because it is accurate, fast and set up to work only with single sequences.

For each mutation step (i.e. after each step of 10% change), we monitored sequence similarity to native, the relative content of residues predicted in short/long-disordered regions and the length of the regions (SOM).

2.5 Box plots to present results

Box plots (McGill *et al*., 1978; Tukey, 1977) present our results concisely. The lower and upper box edges depict the first and third quartile, respectively. The length of a box is the interquartile range of the distribution. The bold bar inside the box represents the median, while dashed lines reach to the most extreme data point that is no more than 1.5 times the interquartile range away from the upper or lower box edge. Average (mean) values are connected through solid lines and intersect with box plots.

Median and mean are related to the protein level, i.e. summarize the specific feature of all sequences that fall within the same interval of PIDE.

3 RESULTS AND DISCUSSION

3.1 Secondary structure surprisingly robust

Comparisons of pairs of evolutionarily related protein structures reveal two major results (Abagyan and Batalov, 1997; Chothia and Lesk, 1986; Chung and Subbiah, 1996; Sander and Schneider, 1991): first, the less similar their sequences, the less similar their 3D structures [as well as their secondary structures (Rost *et al*., 1994; Rost *et al*., 1997)]; and second, the transition from the regime of 'similar structure' to 'non-similar structure' is highly non-linear and characterized by sigmoids indicative of phase transitions in physics. Our mutation protocol yielded a very different outcome.

Secondary structure diverged to almost random levels over the course of our mutation protocol. We compared this divergence to what is observed between naturally occurring homologues. Towards this end, we used the HSSP database (Sander and Schneider, 1991) and compared homologues at the corresponding levels of PIDE (Supplementary Fig. SOM_5). The change of secondary structure on random mutation was much more dramatic than that for homologous proteins (Fig. 1A), e.g. at 30%, PIDE natural homologues still had levels of Q_3 ~63%, while the random mutants reached Q_3 ~45% (Supplementary Fig. SOM_5). This result is not surprising: evolution *feels* the pressure to enrich neutral mutations, i.e. those that do not alter structure, while no such incentive was built into our *in silico* mutation protocol. Nevertheless, secondary structure was surprisingly robust under mutation. The consensus over ensembles of five different mutation trajectories (Fig. 1C and D) diverged much more dramatically from wild type than any single mutant (Fig. 1A and B).

Another important difference between our *in silico* mutation and natural evolution pertained to the shape of the transition: instead of a sigmoidal phase transition, we observed an almost linear transition from native wild-type to almost random mutant. This was true for both the single trajectory (Fig. 1A) and the ensemble (Fig. 1C), although the signal was clearer for the ensemble.

We observed that some regions did not alter secondary structure even at the end of our protocol at which the mutant was as similar to the wild type as to any other sequence in our dataset (Fig. 1B). For the ensemble, in contrast, the consensus secondary structure had changed almost completely from the native (Fig. 1D). Nevertheless, the Q3 levels converged to the same level in both cases.

3.2 Helix and strand intrinsic to random sequences

Our most surprising finding was that neither the overall content (Fig. 2A and B) nor the length (Fig. 2C and D) of *predicted* helices and strands was altered during the course of our mutation protocol. The average helix content remained ~30%, whereas the average strand content around 20%; the average helix was about 10 residues long (2–3 helix turns), and the average strand extended over about five residues. In other words, regular secondary structure was predicted to be robust under extreme mutation. In this respect, we observed no significant difference between choosing mutations according to the background distribution and PAM120, although the latter tends to follow the evolutionarily more accepted mutations (mutations according to BLOSUM62 gave similar results Supplementary Fig. SOM_6).

After the 69 mutation steps (Section 2), we reached a point at which the mutant was as similar to the native as to any other sequence. This was reflected by the similarity in the prediction of helix/strand content/length between the final mutant and randomly created sequences (Fig. 2: two rightmost bars almost identical).

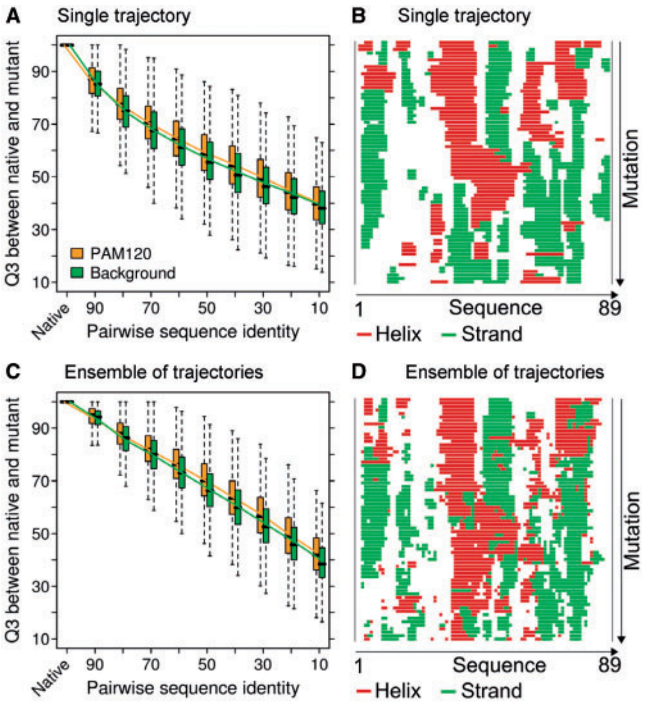

Fig. 1. Secondary structure changes proportional to sequence. (**A** and **C**) For decreasing pairwise percentage sequence identity (*x*-axis, PIDE), we monitored the similarity between secondary structure predictions (Q_3, i.e. percentage of residues identical in one of the three states helix, strand and other) for native and for mutant (yellow: mutations according to PAM120, green: according to background distribution, Section 2). (A and B) show results for a single trajectory, (C and D) the consensus over an ensemble of five trajectories (Section 2). Box plots reflect the range of the distribution (Section 2); median values are marked by horizontal bars and mean values are connected by dotted lines. For instance, at ~90% pairwise sequence identity, ~88% of the residues are predicted in the same secondary structure as the native; for the ensemble, this value is slightly higher (leftmost bars in A and C). The curves converge nearly linearly towards values ~35% corresponding to random. (**B** and **D**) For one particular example (PDB identifier 1a2s chain A), we display the actual secondary structure predictions for each mutant: native on top; each row marks one of the 69 mutation steps (Section 2); mutation by PAM120. The top (B) is for one single mutation trajectory, the bottom (D) for an ensemble of five trajectories. One observation stands out and is representative for all such plots that we looked at: blocks of regular secondary appear to be more robust under mutation than the actual type of secondary structure, i.e. helices flip to strands and vice versa and this happens more often than the transitions helix→other and strand→other. Borders are much more 'fluid' for the ensemble (D) than for a single mutation trajectory (B).

Our results were based on predictions rather than on observations. Prediction methods make mistakes. One might hypothesize that rather than shedding light on protein features, our results are caused by those prediction mistakes. As no large-scale experiments establish structure for random sequences, we cannot refute this view. However, we could provide evidence that prediction mistakes might

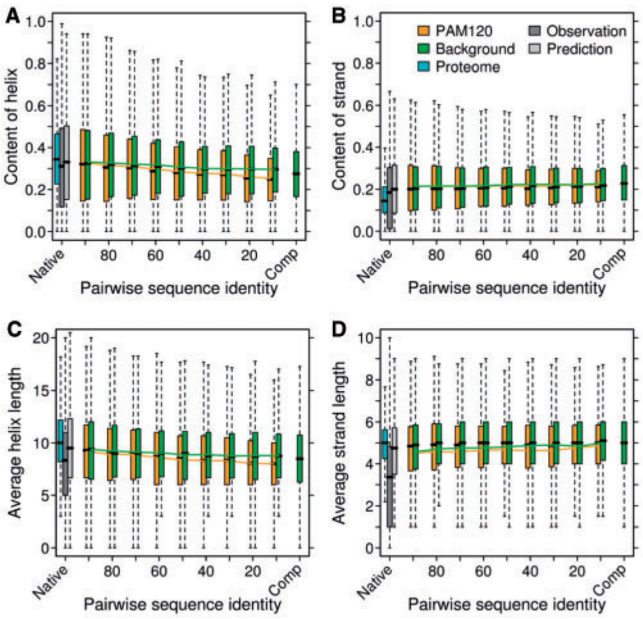

Fig. 2. Content and length of regular secondary structure unchanged. Box plots and coloring as in Figure 1. Change of regular secondary structure on mutation given by the composition of predicted helix (**A**) and strand (**B**), as well as the average lengths of predicted helices (**C**) and strands (**D**). The second and third bar on the left in (A) and (B) compare predictions (light gray) with observations (taken from DSSP, dark gray) for the PDB dataset; the first bar on the left in (A) and (B) indicates the degree to which the predictions differ for the PDB dataset (dark gray) and for a set of all human proteins (light blue). The right-most green bars mark the predictions for randomly assembled sequences (Section 2, labeled as 'Comp'). Overall, neither the length nor the content of regular secondary structure appears to differ between native and random.

not matter for the aspects of structure that we monitored. In fact, by the measures that we used to report our results, predictions and observations were almost identical (Fig. 2: left gray bars in each panel). The precise levels of helix/strand content and length differed indeed more between different datasets (PDB subset versus entire set of human proteins) than between observation and prediction for any set for which we have experimental information. In other words, prediction mistakes appeared not to matter for all the proteins for which we could verify this statement.

Our findings that random and wild-type sequences were predicted to have similar content of regular secondary structure along with the observation that mistakes in predicting this were negligible suggest that the formation of helices and strands is an intrinsic feature of amino acid sequences. Neither helices nor strands were predicted to be significantly shortened during our drastic *in silico* mutation protocol. Note that this is not a consequence of the fact that PROFsec is trained to predict a particular length distribution, because predicted length distributions deviate substantially between all-helical and coiled-coil proteins. The maintenance of such regular secondary structure elements would then appear to come at seemingly low costs, i.e. mutations that are neutral with respect to structure might be more likely than might have been anticipated. Finally, we verified that the reliability of the predictions did not change during mutation (Supplementary Fig. SOM_10).

3.3 Long regions of disorder sensitive, short not

Arguably, there are two different regimes of disorder (Dosztányi *et al*., 2005; Liu *et al*., 2002; Obradovic *et al*., 2005; Peng *et al*., 2006; Schlessinger *et al*., 2007b; Schlessinger *et al*., 2009): very short and very long regions. No threshold distinguishes between these two regimes in a biophysically meaningful way.

In particular, there likely exists an intermediate range that might belong to both regimes. Here, we followed the typical 'convention' in the field and defined as short disorder regions with eight or less consecutive residues and as long disorder regions with 30 or more consecutive residues. Thereby, we ignored the uncertain regime in between these two extremes. In order to establish that our results did not crucially depend on the particular threshold, we also tested other thresholds for long disorder, namely 20, 40 and 50. We found that the trend of loss during *in silico* mutation is independent of the chosen cut-off and is even clearer for larger thresholds (40 and 50) (Supplementary Fig. SOM_09).

First, we observed that regions of short disorder behaved like regular secondary structure in that their content (Fig. 3B, D and F; Supplementary Fig. SOM_2D and E) and length (Supplementary Fig. SOM_2A–C) did not alter on mutation. In stark contrast was the result for long regions with predicted disorder gradually diminished over the course of our mutation protocol (Fig. 3A, C and E; by definition a prediction of 29 disordered residues for some mutant implies that for that mutant the long disordered region seemingly 'disappeared', e.g. Fig. 3E middle; Supplementary Fig. SOM_1). The loss on mutation was much more dramatic for mutations according to PAM120 (yellow in Fig. 3C) than for those according to the background distribution (green in Fig. 3C). This is understandable because disordered regions are abundant in polar residues, and these are more likely to be chosen if mutation probability is 'skewed' toward this abundance. Put differently, PAM120-driven mutations drifted toward sequences that resembled regular well-structured proteins and as such had no disorder, while background-driven mutations yielded sequences that were as abundant in disorder as the native wild types and therefore had many long regions with predicted disorder.

The actual numbers in terms of content of predicted long disorder decreased from ∼18% for the native to ∼9% for the final mutant by using the background mutation protocol (Fig. 3C, green). This reflected the fact that a considerable fraction of the residues in our DisProt dataset was polar: for mutations according to PAM120 (Fig. 3C, yellow) or BLOSUM62 (Supplementary Fig. SOM_7), the content dropped to 0. However, at this level of mutations, almost no single residue predicted as long disorder in the native was predicted as disorder in the mutant (Fig. 3A). For some, this might appear to PAM120.

Studies of particular mutation paths revealed that long disorder might just appear to vanish *suddenly* (Fig. 3E). This was partially a threshold issue: assume a region with 35 consecutive 'disordered' residues and assume the mutant loses three on each side (six in total); we will no longer consider this as long disorder (35–6 <30). This also explains how additional mutations may *recover* the long disorder (Fig. 3E: after solid block of red bars, suddenly one mutant has disorder again as seen by a single bar below this block).

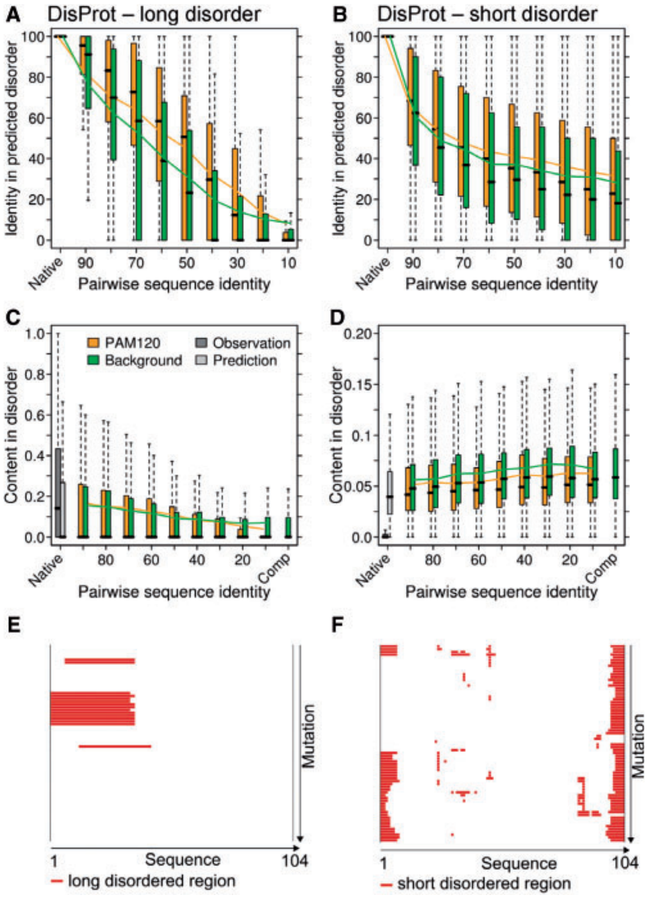

Fig. 3. Predicted long disorder changes rapidly. Panels on the left show results for long regions of disorder (30 or more consecutive residues), those on the right for short regions (less than eight). The top panels (**A** and **B**) demonstrate how much the predictions of disorder changed over the course of mutations (*y*-axis: residues predicted identical as disorder between native and mutant as percentage of disorder predicted in native). Disorder predictions differ much more rapidly from native than do secondary structure predictions, and much more for long (A) than for short (B) disorder. The relative content of residues in predicted long (**C**) and short (**D**) disordered regions diverge differentially. The first two box plots for (C) depict the observed (dark gray) and predicted (light gray) disordered content in native sequences. Right box plots in both (C) and (D) show the disordered situation in the artificially created dataset sequences (Section 2, labeled as 'Comp'). For a representative example (DisProt identifier: DP 00006), the IUPred predictions for long (**E**) and short (**F**) disorder are shown for each mutant: native on top; each row marks 1 of the 69 PAM120 mutation steps (Section 2). Red lines mark predictions that fall into the threshold category ((30 or more/less than eight). Long disordered regions disappear (E) while especially short disorder remains at both termini, while re- and disappearing in the middle region during mutation (**F**).

Another observation reflects one of the important aspects when studying short disorder: a considerable fraction of the short disorder is predicted (and observed) near the protein termini (Fig. 3F). Short disorder 'comes and goes' during mutation (middle region in Fig. 3F). Although this effect is biologically relevant and dominates the study of disorder in otherwise well-ordered proteins (Bordoli *et al*., 2007; Jin and Dunbrack, 2005), it again underlines the problem of not differentiating between long and short disorder.

Our analyses of regular secondary structure and disorder are based on very different datasets. PDB is biased in many ways (Liu and Rost, 2001), one of those pertains to disorder (Liu and Deber, 1999; Peng *et al*., 2004). One reason simply is that proteins with disordered regions pose extreme challenges to structure determination (Burley *et al*., 2008; Dunker *et al*., 2008; Graslund *et al*., 2008; Liu *et al*., 2004; Nair *et al*., 2009; Romier *et al*., 2006). To address this difference, we predicted disorder also for the dataset of well-ordered proteins from the PDB. As expected, the level of both long and short disorder for both of those was very low (Supplementary Figs SOM_3 and 4); given the lack of disorder in these proteins, we could therefore not observe any significant difference between close-to-zero in the wild type and close-to-zero in the mutants.

IUPred is arguably one of the best disorder prediction methods (Bordoli *et al*., 2007; Le Gall *et al*., 2007; Schlessinger *et al*., 2007b; Schlessinger *et al*., 2009; Shimizu *et al*., 2007); however, it is still only one of many and it has specific strengths and weaknesses. Therefore, we also predicted disorder with two other state-of-the-art prediction methods, namely VSL2 (Obradovic *et al*., 2005; Peng *et al*., 2006) and MD (Schlessinger *et al*., 2009). Although the predictions for those two differed slightly from those for IUPred, by the measures we reported here, they revealed exactly the same trend: while predicted long disorder disappeared on mutation, the content and length distribution of predicted short disorder remained largely unaffected by the mutation.

We addressed the impact of incorrect predictions by randomly introducing errors. At any significant error rate, long disorder disappeared in the native. This highlights the high prediction accuracy of today's methods. For short disorder, the added error did not alter the content over the course of our mutation protocol (Supplementary Fig. SOM_8).

As short and long disorders have different physical traits, we need length thresholds. However, we can drop these thresholds while monitoring the disappearance of disorder. Toward this end, we began with all native regions longer than *N* (chosen in steps of between 20 and 50), and monitored the percentage of disorder predicted after mutation irrespective of the length of the predicted regions. We found that long disordered regions indeed get decomposed into shorter ones and that disorder disappears throughout (Supplementary Figs SOM_11 and 12).

4 CONCLUSIONS

We addressed the general question whether or not well-ordered regular secondary structure and disordered regions sustain random mutations. Is it likely or unlikely that any mutation affects this particular coarse-grained feature of protein structure (and through it's function)? Do random sequences have different content in secondary structure and disorder than native proteins that have evolved to satisfy many constraints? Our analysis clearly suggests two different answers for regular secondary structure and long disorder. On the one hand, the maintenance of regular secondary structure might not be too challenging because its formation appears to be an intrinsic feature of random sequences. It, therefore, appears surprisingly likely to transit from helix to strand and back. In fact, this is exactly what we dynamically observed during the course of our mutations (Fig. 4). On the other hand, regions of long disorder do not appear to be robust under mutation. Random changes likely disrupt this feature that thereby appears volatile and unique.

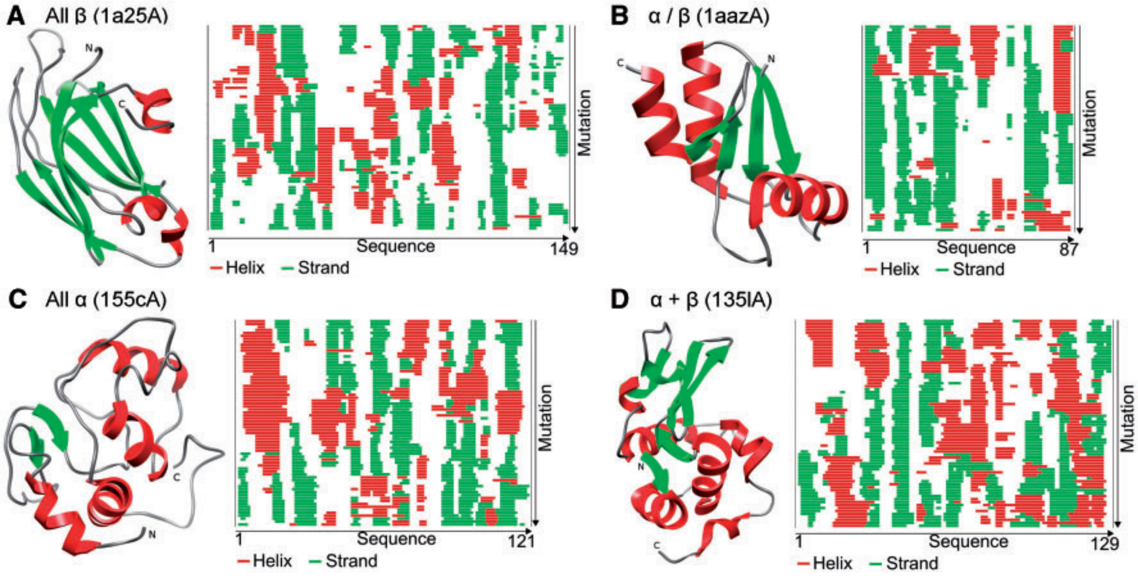

Fig. 4. Examples of proteins with mutation trajectories. For each of the four main SCOP classes (Murzin *et al.*, 1995), we randomly picked one representative short enough to fit into the space here. Ribbon plots were generated by Chimera (Pettersen *et al.*, 2004) [red: helix, green: strand, according to DSSP (Kabsch and Sander, 1983)]. (**A**–**D**) In each of the four panels, the ribbon diagram for the native is on the left, and on the right are the 69 mutation trajectories (top: native, degree of mutation decreases downwards; mutations according to PAM120, Section 2). The sequence runs from the most N-terminal residues (labeled '1') to the most C-terminal ones. Note that although we show only single trajectories, rather than ensemble averages here, almost no helix or strand withstands the mutation protocol to the end.

This has important impact on how we picture the role of long disorder in proteins: it is not 'easy' to acquire. Prokaryotes have only ∼10–25% of the disorder observed in multi-cellular eukaryotes (Dunker *et al.*, 2008; Ekman *et al.*, 2005; Liu *et al.*, 2002; Oldfield *et al.*, 2005; Romero *et al.*, 2004; Schlessinger *et al.*, 2009; Ward *et al.*, 2004). Our observation of how volatile long disorder is provides another evidence for the importance of this feature for the transition from prokaryotes to eukaryotes.

Many SNPs that alter the protein sequence (nsSNPs) appear to be deleterious. Is this a bias in the experimental technique (more likely to be observed/reported if deleterious), or is it a genuine feature of proteins imposed by the sensitivity of protein structure to mutations? Although our work neither addresses nor answers this question, the surprising robustness of regular secondary structure might support the view that protein structure is more flexible and adaptable than the intricate details of the concert of interacting residues in protein 3D structures might suggest.

ACKNOWLEDGEMENTS

The authors would like to thank the following for valuable discussions: Zsuzsanna Dosztanyi (Eötvös Loránd University Budapest and Columbia University in the City of New York), Dietlind Gerloff (UCSC Santa Cruz), Marco Punta (Columbia University in the City of New York and TUM Munich), Reinhard Schneider (EMBL Heidelberg), Anna Tramontano (La Sapienza Rome and KAUST); the anonymous reviewers for very constructive and helpful suggestions that helped shaping this work; and also to all those who deposit their experimental data in public databases and to those who maintain these databases, in particular to those who contribute to PDB and DisProt.

Funding: National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH) (grant number R01-GM079767).

Conflict of Interest: none declared.

REFERENCES

Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.

Alexov,E.G. and Gunner,M.R. (1997) Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.*, **72**, 2075–2093.

Andersen,C.A.F. *et al.* (2002) Continuum secondary structure captures protein flexibility. *Structure*, **10**, 175–184.

Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Prot. Chem.*, **29**, 205–300.

Benner,S.A. *et al.* (1997) Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.*, **97**, 2725–2844.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bordoli,L. *et al.* (2007) Assessment of disorder predictions in CASP7. *Prot. Struct. Funct. Genet.*, **69**(Suppl. 8), 129–136.

Burley,S.K. *et al.* (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. *Structure*, **16**, 5–11.

Cavasotto,C.N. and Abagyan,R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.*, **337**, 209–225.

Chothia,C. and Lesk,A.M. (1986) The use of sequence homologies to predict protein structures. In Robert,F. and Mark,Z. (eds) *Computer Graphics and Molecular Modeling*. Cold Spring Harbor Laboratory, New York, pp. 33–37.

Chung,S.Y. and Subbiah,S. (1996) A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123–1127.

Claussen,H. *et al.* (2001) FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, **308**, 377–395.

Daniel,R.M. *et al.* (2003) The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 69–92.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, pp. 345–358.

Dill,K.A. (1993) Folding proteins: finding a needle in a haystack. *Curr. Opin. Struct. Biol.*, **3**, 99–103.

Dosztányi,Z. *et al*. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity-linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.

Dunker,A.K. *et al*. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.

Ekman,D. *et al*. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Graslund,S. *et al*. (2008) Protein production and purification. *Nat. Methods*, **5**, 135–146.

Gu,J. *et al*. (2006) Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput. Biol.*, **2**, e90.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Jin,Y. and Dunbrack,R.L. Jr (2005) Assessment of disorder predictions in CASP6. *Proteins*, **61**(Suppl. 7), 167–175.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karplus,M. and Petsko,G.A. (1990) Molecular dynamics simulations in biology. *Nature*, **347**, 631–639.

Le Gall,T. *et al*. (2007) Intrinsic disorder in the Protein Data Bank. *J. Biomol. Struct. Dyn.*, **24**, 325–342.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Levitt,M. and Warshel,A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694–698.

Liu,J. *et al*. (2004) Automatic target selection for structural genomics on eukaryotes. *Prot. Struct., Funct., Bioinform.*, **56**, 188–200.

Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.

Liu,J. *et al*. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.

Liu,L.P. and Deber,C.M. (1999) Combining hydrophobicity and helicity: a novel approach to membrane protein structure prediction. *Bioorg. Med. Chem.*, **7**, 1–7.

Liwo,A. *et al*. (1999) Protein structure prediction by global optimization of a potential energy function. *Proc. Natl Acad. Sci. USA*, **96**, 5482–5485.

McGill,R. *et al*. (1978) Variations of box plots. *Am Statistician*, **32**, 12–16.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Morea,V. *et al*. (1998) Protein structure prediction and design. *Biotechnol. Annu. Rev.*, **4**, 177–214.

Morea,V. *et al*. (2000) Antibody modeling: implications for engineering and design. *Methods*, **20**, 267–279.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nair,R. *et al*. (2009) Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics*, **10**, 181–191.

Obradovic,Z. *et al*. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Prot. Struct., Funct., Genet*. **61**(Suppl. 7), 176–182.

Oldfield,C.J. *et al*. (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.

Pauling,L. and Corey,R.B. (1951a) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl Acad. Sci.*, **37**, 729–740.

Pauling,L. and Corey,R.B. (1951b) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl Acad. Sci. USA*, **37**, 251–256.

Peng,K. *et al*. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, **9**, 435–446.

Peng,K. *et al*. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Pettersen,E.F. *et al*. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Radivojac,P. *et al*. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.

Reva,B.A. *et al*. (1995) Constructing lattice models of protein chains with side groups. *J. Comput. Biol.*, **2**, 527–535.

Romero,P. *et al*. (2004) Natively disordered proteins : functions and predictions. *Appl. Bioinform.*, **3**, 105–113.

Romier,C. *et al*. (2006) Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 1232–1242.

Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525–539.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rost,B. (2005) How to use protein 1-D structure predicted by PROFphd. In Walker,J.M. (ed.), *The Proteomics Protocols Handbook*, Humana Press, pp. 875–901.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B. *et al*. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Rost,B. *et al*. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, **270**, 471–480.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Prot. Struct. Funct. Genet.*, **9**, 56–68.

Schlessinger,A. *et al*. (2007a) Natively unstructured loops differ from other loops. *PLoS Comput. Biol.*, **3**, e140.

Schlessinger,A. *et al*. (2007b) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**, 2376–2384.

Schlessinger,A. *et al*. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLOS ONE*, **4**, e4433.

Schlessinger,A. *et al*. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891–893.

Shimizu,K. *et al*. (2007) Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78.

Sippl,M.J. (1993) Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Des.*, **7**, 473–501.

Tukey,J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley Pub. Co., Reading, MA.

Uversky,V.N. (2003) Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol. Life Sci.*, **60**, 1852–1871.

Vucetic,S. *et al*. (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wright,P.E. and Dyson,H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.

Yue,P. *et al*. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.*, **353**, 459–473.

Yue,P. *et al*. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.