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

Schaefer C, Schlessinger A, Rost B.
Bioinformatics. 2010 Mar 1;26(5):625-31. Epub 2010 Jan 16.

Bioinformatics Journal Club
Lauris Kaplinski
30 March 2010

Introduction

- Helices and strands constitute the major macromolecular building blocks of all ‘well-ordered’ proteins
- Many proteins have regions that remain ‘unstructured’ unless bound to a substrate
- As protein structure determines protein function, it is also subjected to evolutionary selection
- Disorder vs. lack of structure
- How are the major blocks of protein structure affected by random or semi-random mutations

Two hypotheses

- Regular secondary structure is difficult to maintain evolutionarily, i.e. single residue mutations are likely to impact helices and strands
- 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

Both were falsified!

In Silico experiment

- Protein sequences
  - Globular proteins from the Protein Data Bank (PDB)
  - Proteins with disordered regions from DisProt
- UniqueProt was applied to reduce the redundancy in both sets. The redundancy-reduced sets comprised 1369 (PDB) and 374 (DisProt) proteins
- Random sequences for convergence control
- The secondary structure in entire human proteome (33 812 proteins) was predicted to shed light on potential biases from the chosen databases
- A set of sequences from the PDB set with the same size, amino acid and length distribution as that of the DisProt set was sub-sampled to examine the ability of ordered proteins to retain or lose their ordered state

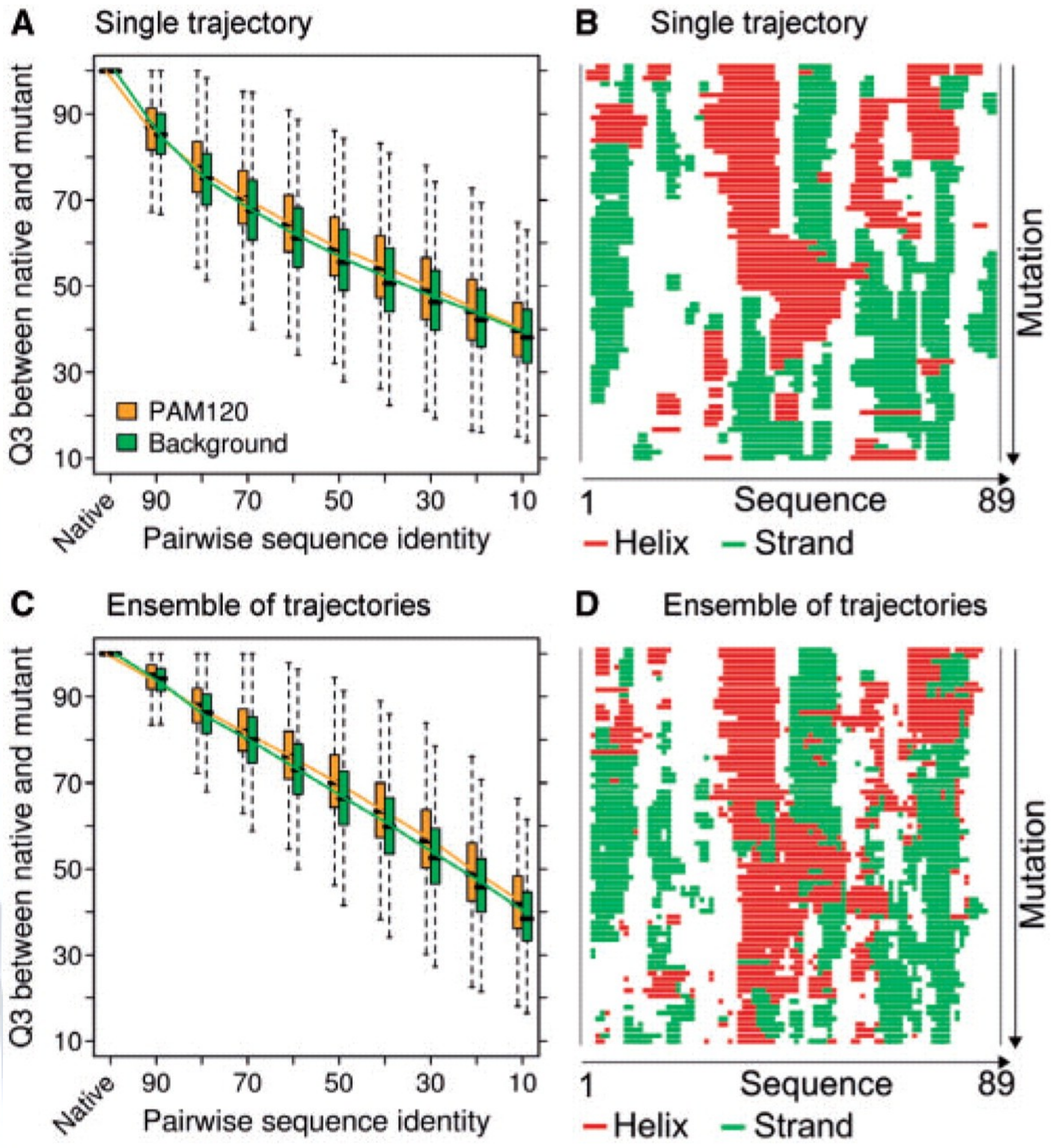
In Silico experiment

- gradually mutated native protein sequences into quasi-random strings of amino acids
  - select a particular residue position
  - mutate the amino acid X at that position with amino acid Y with the probability p_{XY}
  - repeat these two moves N/10 times (N number of residues in the protein)
- 69 mutation steps (with 69 x N/10 mutations) for each protein were carried out. After 65 steps the properties reached convergence in all the cases.

In Silico experiment

- Substitution schemes: PAM120 (big evolutionary distances), BLOSUM62 (short evolutionary distances), amino acid distribution in the databases (PDB, DisProt)
- Five different random mutation paths (five different mutants) in order to investigate the divergence from the native of an ensemble of evolutionary paths
- Secondary structure was predicted with PROFsec. In independent tests it achieved single-sequence level of 68% three-state per-residue accuracy (Q3 is the percentage of residues predicted correctly in one of the three states helix, strand and other)
- Disordered regions were predicted with three methods: IUPred, MD and VSL2 and compared the predictions to the experimental annotations in DisProt
- Cutoffs for short long disordered sequence were 8 and 30

Secondary structure changes proportional to sequence

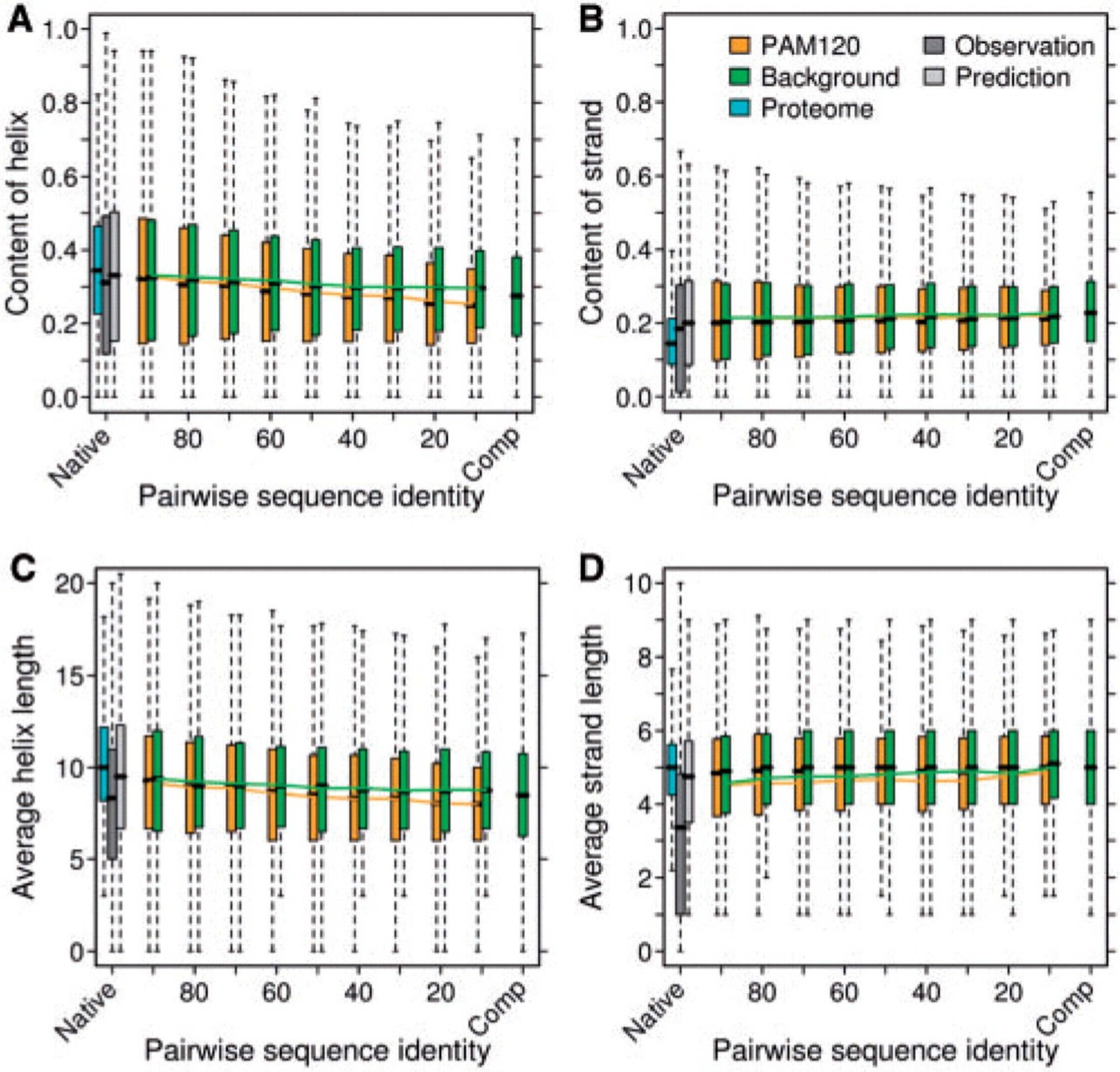


(A and C) For decreasing pairwise percentage sequence identity (x-axis, PIDE), we monitored the similarity between secondary structure predictions (Q3, i.e. percentage of residues identical in one of the three states helix, strand and other) for native and for mutant (yellow: mutations according to PAM120, 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.

The curves converge nearly linearly towards values 35% corresponding to random. (B and D) For one particular example (PDB identifier 1a2s [PDB] 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.

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).

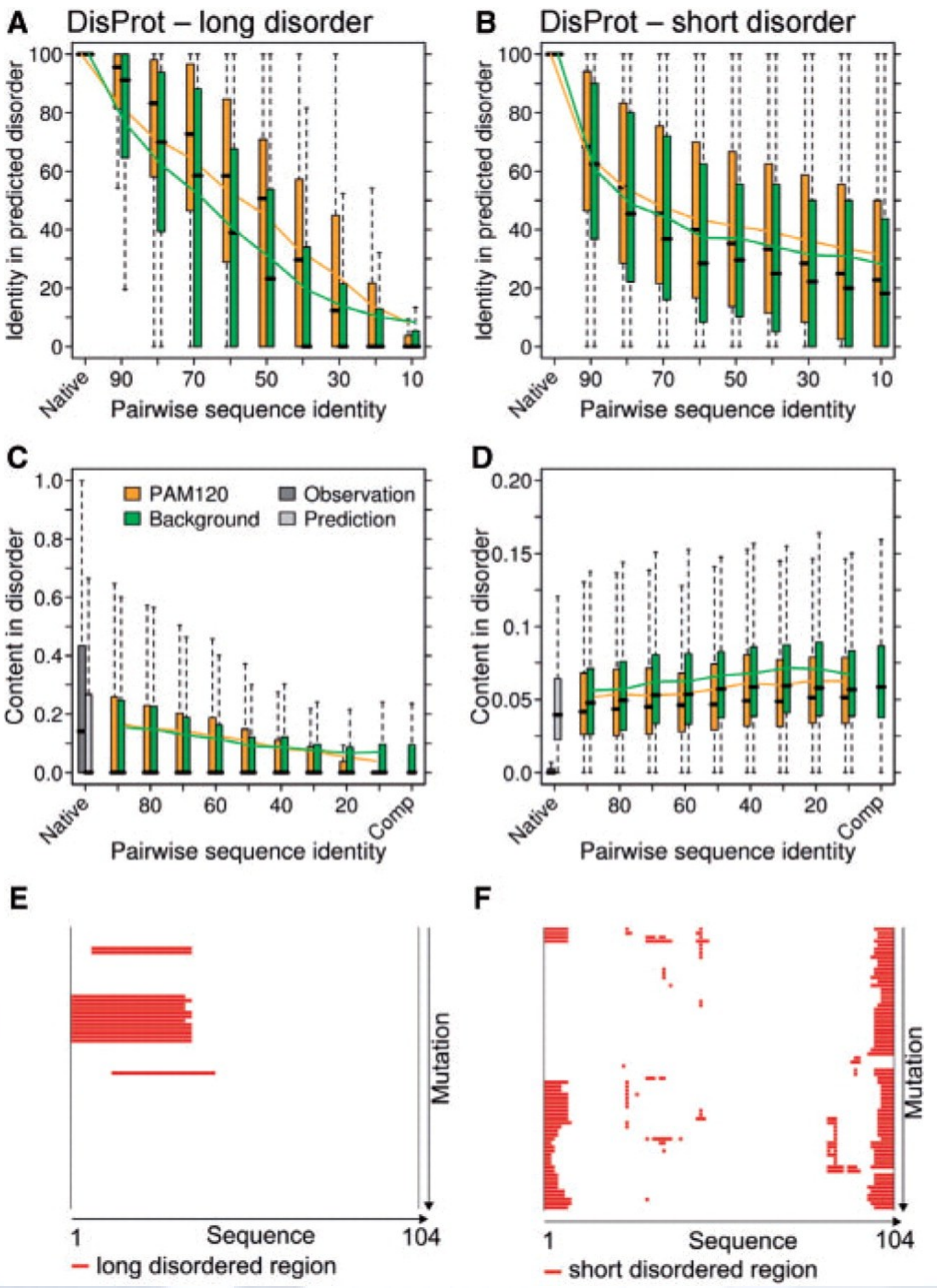
The second and third bar on the left in (A) and (B) compare predictions (light gray) with observations (taken from DSSP, dark gray) for the PDB dataset; the first bar on the left in (A) and (B) indicates the degree to which the predictions differ for the PDB dataset (dark gray) and for a set of all human 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.

Results

- Neither the overall content nor the length of predicted helices and strands was altered during the course of mutation
  - Average helix content was 30%
  - Average strand content was 20%
  - Average helix was about 10 residues long (2–3 helix turns)
  - Average strand extended over about 5 residues
- No significant difference between choosing mutations according to the background distribution and PAM120



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).

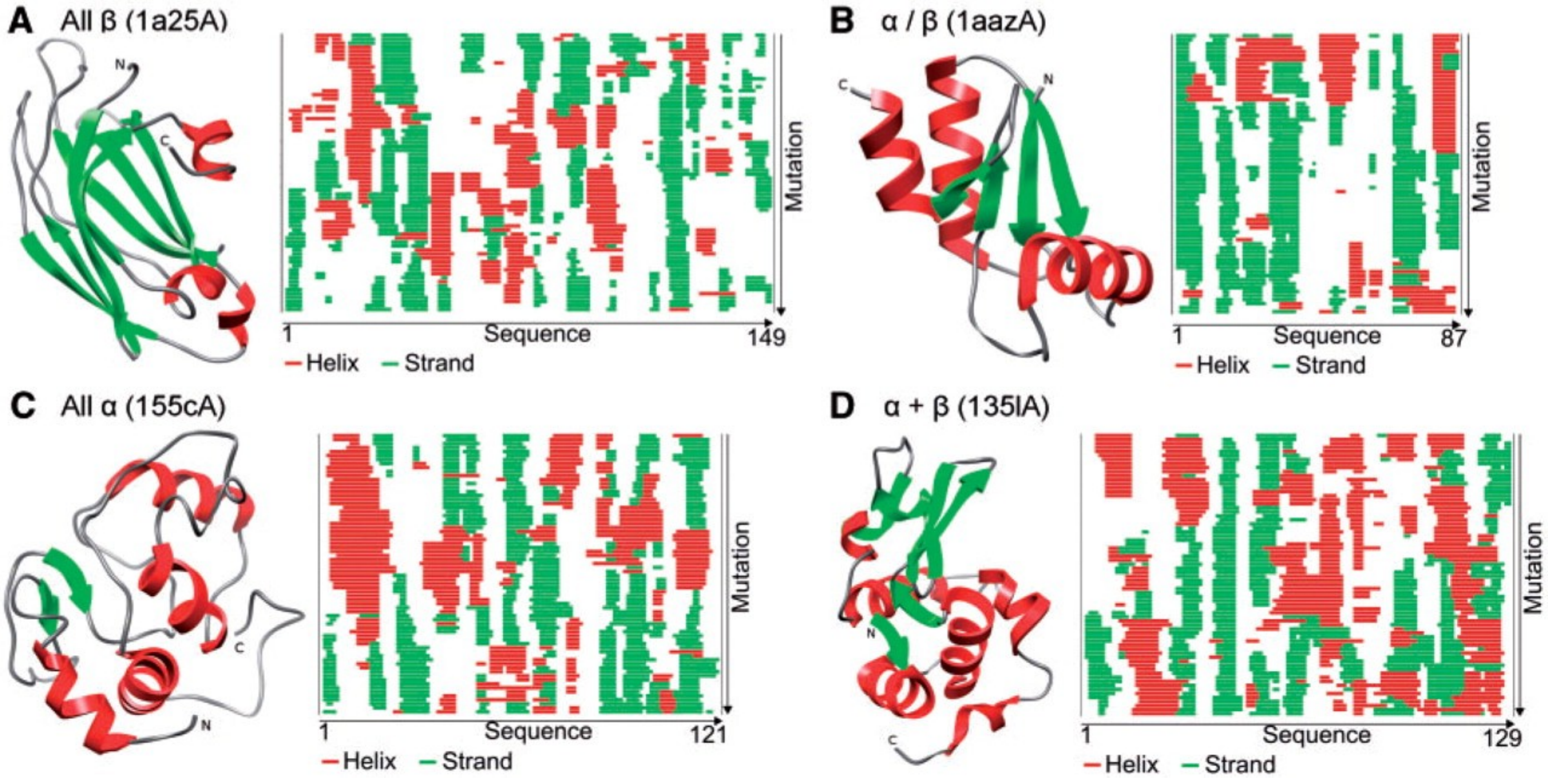
The first two box plots for (C) depict the observed (dark gray) and predicted (light gray) disordered content in native sequences. Right box plots in both (C) and (D) show the disordered situation in the artificially created dataset sequences (Section 2, labeled as ‘Comp’).

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).

Results

- Content of predicted long disorder decreased from 18% for the native to 9% using the background mutation protocol and 0%using PAM120 or BLOSUM62
- This reflects the fact, that considerable fraction of amino acids in DisProt proteins were polar
- Long disorder appears to vanish suddenly (partially may be cutoff problem)
- Short disorder comes and goes during mutation
- There is a bias for short disorder near the ends of peptide chain
- long disordered regions get decomposed into shorter ones and that disorder disappears finally

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)].

Conclusions

- The maintenance of regular secondary structure is not very challenging evolutionally because its formation appears to be an intrinsic feature of random sequences
- Transition from helix to strand is surprisingly likely
- Regions of long disorder do not tolerate mutations, so preserving them may be evolutionally expensive (prokaryotes have only 10–25% of the disorder observed in multi-cellular eukaryotes)