

Supplementary data: Identifying sequence regions undergoing conformational change via predicted continuum secondary structure

Mikael Bodén

School of Information Technology and Electrical Engineering
QLD 4072, The University of Queensland, Australia

Timothy L. Bailey

Institute of Molecular Bioscience
QLD 4072, The University of Queensland, Australia.

May 11, 2006

1 Introduction

This paper contains supplementary data for the study entitled “Identifying sequence regions undergoing conformational change via predicted continuum secondary structure” (currently in submission). More specifically, it outlines and details three tests done with the continuum secondary structure predictor <http://pprowler.itee.uq.edu.au/sspred> to elucidate its applicability for identifying conformational changes in proteins.

Details of the method used as well as the broader scientific background are provided in the aforementioned study.

2 Ras: a detailed look at conformational switching

Schlessinger and Rost used Ras (Human) as a case study for evaluating their protein B-value predictor PROFBVAL (Schlessinger and Rost, 2005). Looking at switch II (which is responsible for binding GTP), they report that PROFBVAL singled out one of the critical residues and three neighbours involved in “switch II” of Ras (residue 60 in the sequence we analysed, see Fig. 1).

```
30          40          50          60          70          80
EYDPTIEDSY RKQVVIDGET CLLDILDITAG QEEYSAMRDQ YMRTGEGFLC VFAINNTKSF
          S EEEEE EEEEEETTEE EEEEEEEEE SS TTTTTT TTTT S EEEEE EEETT HHHH
S   S EEEEE EEEEESSS E EEEEEEEEE SS  EHHHH HHHH S EEEEE EEETT BHHHH
S   S EEEEE EEEEEETTEE EEEEEEEEE STTTGGGHHH HHHH S EEEEE EEETT BHHHH
TTTTT  EEE EEEEEETTEE EEEEEEEEE SS SSSSHH HHHH S EEEEE EEETT HHHH
*****--  -----*****  -----*--  -*****  ****  -----  -----*-----
```

Figure 1: The amino acids for positions 30–89 in PDB:6Q21 are presented using their one-letter codes (top row). The DSSP secondary structure states for the four recorded conformations are listed, aligned with their position (middle four rows). Letter codes are the standard DSSP codes specified previously, except C that is here replaced by white space. The classification of each position as non-variable - or variable * is also supplied (bottom row).

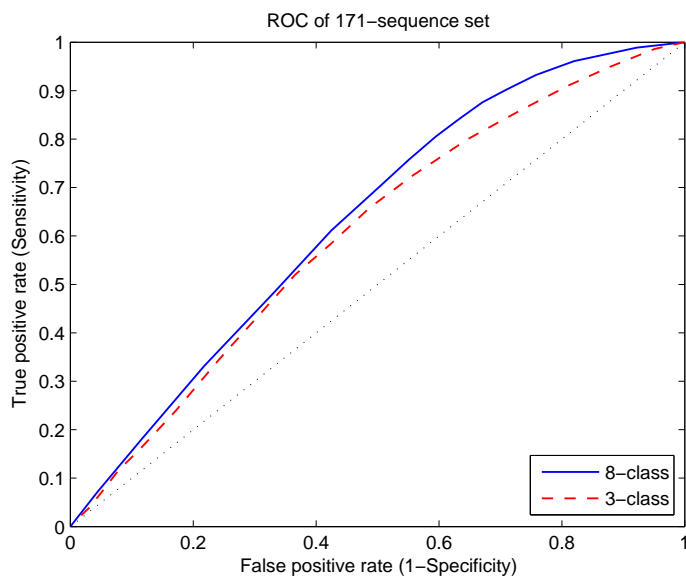


Figure 2: The ROC curve for our method and the PHD/ASP method (using three different window sizes, 1, 5 and 21) when tested on Ras (PDB:6Q21).

Using our method, we note that switch II is recognised well—all the residues that change secondary structure (61-74) are identified as positives. The sequence for Ras (PDB:6Q21) resulted in the entropy shown in Fig. 3 (top). We note that the entropy of the predicted continuum secondary structure is not specific to regions that are known to be flexible. However, each known flexible region results in an entropy “hump”. Notably, the end points of such regions sometimes constitute “near misses” close to transitions between high and low entropy.

For comparison we include the PHD/ASP scores in Fig. 3. It is clear that PHD/ASP misses many of the regions in 6Q21 known to change conformation—even when various window sizes are tried (1, 5 and 21 as prescribed in Young et al., 1999). To quantify the differences, and remove any bias of the threshold selection, we (1) plotted the ROC curve (see Fig. 2) and (2) determined the area under the ROC (AROC) for each method. When using the entropy of the predicted continuum secondary structure the AROC is 0.76. The corresponding values for the ASP-based predictions are 0.71, 0.68, and 0.66 for window sizes 1, 5, and 21, respectively.

The accuracy of secondary structure prediction usually improves as a trivial effect of more (and better) data accessible to predictors. Since the PHD model was trained on fewer sequences than our CPNN models, this may partially explain why the present method provides a better score than PHD/ASP. Another possible explanation of the improvement shown by our method is that it is based on 8-class secondary structure predictions, whereas ASP is based only on helix, sheet and coil.

3 β -propeller folds: a detailed look at conformational stability

The previous tests are biased in that we try to identify known “positives”. All tested sequences involve regions that are known to exhibit a degree of flexibility, i.e. change conformation. Similar to Schlessinger and Rost we now turn to find known “negatives”, we look at a sequence that is known to exhibit rigid regions (Fülöp and Jones, 1999). β propeller folds occur in several structures. The fold has a very symmetrical structure, utilising four-stranded, anti-parallel beta sheets, packing face-to-face, to form a very rigid tunnel. In Fig. 4 the entropy generated from predicting the continuum secondary structure of PDB:1TBG is shown, with the stable beta sheets marked. 1TBG has seven blades, each consisting of four strands. Not all of the residues involved in the tunnel are predicted below the entropy threshold (again the mean entropy is used) but one can clearly discern that these residues are much less likely to change conformation compared to

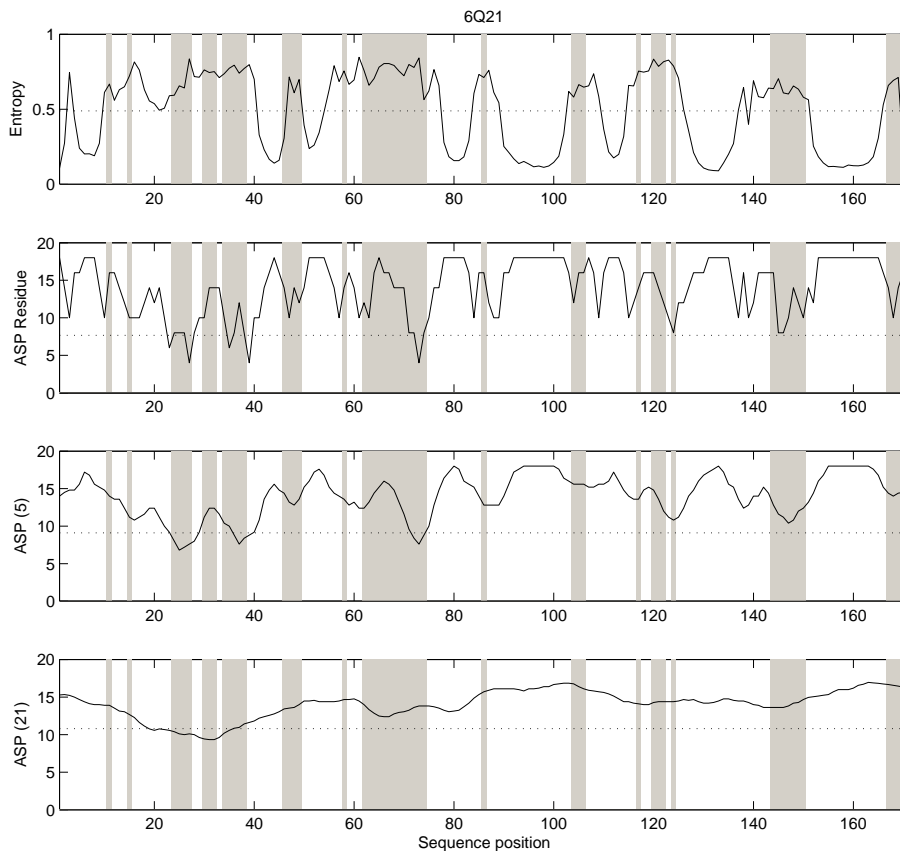


Figure 3: Graph 1: The entropy of the predicted continuum secondary structure of Ras (PDB:6Q21) is plotted as a solid line. The mean entropy predicted by our CPNN model is plotted as a dotted line (0.49). High entropy corresponds to high flexibility. The area under ROC (varying the threshold) is 0.76. Graph 2-4: The scores produced by PHD/ASP for each residue in 6Q21, for a window of 5 residues and for a window of 21 residues, respectively. The recommended cut-off scores are plotted as dotted lines (Young et al., 1999). Low score means high structural ambivalence (or low reliability of secondary structure prediction). The area under ROC for all possible threshold values is 0.71, 0.68, and 0.66, respectively. In all graphs, the regions that are subject to conformational change are marked with a shaded background.

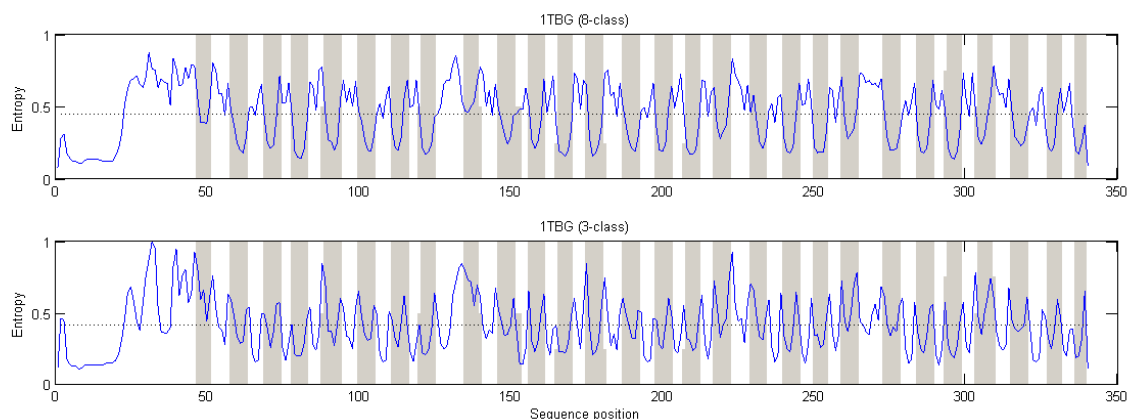


Figure 4: The entropy of the predicted continuum secondary structure of PDB:1TBG, a seven-bladed beta-propeller structure, is plotted as a solid line (on basis of 8-class predictions above, on basis of 3-class predictions below). The mean entropy is plotted as a dotted line. Low entropy corresponds to high rigidity. The beta sheet regions that are known to be rigid are marked with a shaded background. Note that the 28 regions correspond to the seven blades each with four strands.

their neighbours.

We note that the entropy for 3-class and 8-class predictions are different in Fig. 4. As suggested by a reviewer of the original study, this may be due to the resolution of structural classes. 3-class secondary structure provides little room for expressing subtle structural variability, whereas 8-class secondary structure distinguishes between several helical (G and H), sheet-like conformations (E and B) and others (the four remaining classes). In fact, using the 8-class predictor, the caps of the anti-parallel beta sheets are predicted as varying coil structures. This variability is hidden by the 3-class predictor by grouping all four into one class.

4 Endonuclease VII: conformational switching determined from quaternary structure

Endonuclease VII of the T4 bacteriophage is a repair enzyme with a broad substrate specificity (believed to be involved in various stages of infection when transcribed from different promoters). It acts as a dimer, and the interface by which two chains interact is two pairs of anti-parallel helices (helix-2/chain-A interacts with helix-2/chain-B and helix-3/chain-A interacts with helix-3/chain-B). Raaijmakers and colleagues (Raaijmakers et al., 2001) present experimental evidence of conformational flexibility in this and two other regions that are functionally relevant. PDB:1EN7, PDB:1E7D and PDB:1E7L all exemplify structures that exhibit flexible conformations but are based on essentially the same primary sequence (Raaijmakers et al., 2001). We determine the categorical 8-class secondary structures from the crystallised structures and identify the residues that both match the flexible regions identified from the quaternary structure *and* participate in different secondary structures (judging from the six chains). The entropy of the 8-class continuum secondary structure, predicted from the primary sequence of the Endonuclease VII, is shown in Fig. 5.¹ Three regions of interest are indicated below, all resulting in changes in 8-class secondary structure. We exclude three single-residue changes that we failed to relate to any of the three regions. See the Fig. 5 caption for sequence details.

- Helix 2 in the dimer interface can shift by almost a complete helical turn resulting in a major exchange of dimer interaction partners;

¹Two of the six studied chains are different in a single position, N62D. The output of the model did not change with this single-point mutation.

- The orientation of the N-terminus can change relative the helix-based dimer interface;
- The orientation of the C-terminus can change relative the helix-based dimer interface;

Using the 8-class model, aforementioned regions have a mean entropy of 0.66 (exceeding the overall mean 0.53). Sixteen of the 19 aforementioned residues with changed conformation exceeded the mean entropy (corresponding to a sensitivity of 0.84). Fig. 5 illustrates the utility of the method for detecting where motion is unlikely to occur. Qualitatively, errors mainly concern the precise extent of flexible segments.

References

- Fülöp, V. and Jones, D. T. (1999). β propellers: structural rigidity and functional diversity. *Current Opinion in Structural Biology*, 9(6):715–721.
- Raaijmakers, H., Törö, I., Birkenbihl, R., Kemper, B., and Suck, D. (2001). Conformational flexibility in T4 endonuclease VII revealed by crystallography: implications for substrate binding and cleavage. *Journal of Molecular Biology*, 308(2):311–323.
- Schlessinger, A. and Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. *Proteins*, 61:115–126.
- Young, M., Kirshenbaum, K., Dill, K., and Highsmith, S. (1999). Predicting conformational switches in proteins. *Protein Sci*, 8(9):1752–1764.

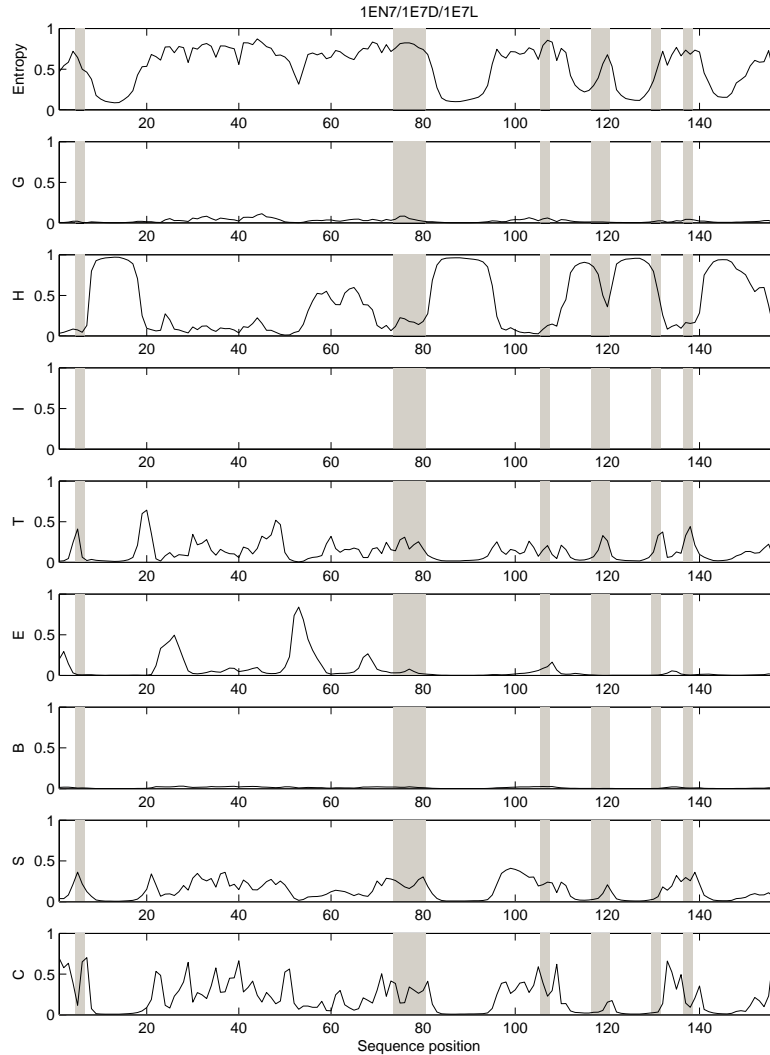


Figure 5: The entropy and class predictions predicted by the 8-class continuum secondary structure model when presented with the Endonuclease VII sequence (top). Observed changes in secondary structure are indicated by a grey bar: The dimer interface concerns 74R(T/H)-75S(T/H/S)-76G(T/S)-77L(T/G/H/S)-78K(T/G/H)-79G(T/G/H)-80Q(T/G/H) where each residue is listed by position, then the amino acid and in parenthesis the 8-class secondary structures that are observed. The determination of N-terminal orientation concerns 5G(H/T)-6K(H/T) and the C-terminal orientation concerns 106P(H/T)-107N(H/T), 117S(H/T)-118R(H/T)-119L(S/T)-120G(T/C), 130Q(H/T)-131R(H/T), 137E(S/T)-138S(S/T). Three residues that changed secondary structure class (10E, 28R, 45L) were excluded since we failed to relate any of them to the three regions of interest and since the change was only evident in a single chain (of six). The predicted 8-class continuum secondary structure is also plotted (below).