

# Ab initio secondary structure prediction using inter-residue contacts

Gabriel Renaud,<sup>1</sup> Brendan J. McConkey<sup>2</sup>

**Keywords:** structure prediction, ab initio, statistical models, computational biochemistry.

## 1 Introduction.

Ab initio protein fold prediction tries to provide a framework to predict the structure based on sequence alone. Such an algorithm would greatly benefit from having an accurate secondary structure predictor that does not rely on multiple sequence alignments like modern algorithms (e.g. [3]), but would achieve an accurate prediction based on sequence alone, as attempted by previous heuristics (e.g. [1]). This would provide a tool for biologist to predict secondary structure elements (SSE) for proteins having few or very distant evolutionary related sequences. We introduce a novel approach for predicting secondary structure based on inter-residue contact scores obtained using a statistical analysis of a non-redundant database of 645 resolved structures [2]. This method is a component of a larger project to predict tertiary structure.

## 2 Methodology

**Alpha helices:** We distinguish 2 types of  $\alpha$ -helices: amphipathic helices, those partially exposed to the solvent, and hydrophobic helices, those completely buried in the core of the protein. For each type, we consider 2 types of interactions, interaction with the solvent and between the side chains of the amino acids forming the helix. For amphipathic helices, we use a helical wheel representation of the region to be predicted (figure 1). For core helices, we use a set of scores homogeneously applied on all the residues in the window being assessed. We merge high scoring regions to obtain our prediction of helical regions.

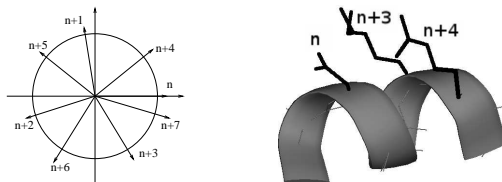


Figure 1: Helical Wheel Representation. To model the solvent interaction with the amphipathic  $\alpha$ -helix, we use a set of amino acid solvent propensities. These are applied using a helical wheel representation of the sequence to be assessed (left). For interactions between residues, we use the fact that any residue at position  $n$  will be in contact with residues at position  $n + 3$  and  $n + 4$  (right)

We also use the fact that  $\alpha$ -helices create an electric dipole that is often counterbalanced by negative and positive residues at the beginning and end of helices respectively. We use

<sup>1</sup>Department of Computer Science, University of Waterloo E-mail: [g2renaud@cs.uwaterloo.ca](mailto:g2renaud@cs.uwaterloo.ca)

<sup>2</sup>Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1 E-mail: [mcconkey@sciborg.uwaterloo.ca](mailto:mcconkey@sciborg.uwaterloo.ca)

the different amino acid propensities to either start or end a helix to adjust the predicted regions made at the previous step. The obtained scores are normalized to a probability using a logistic function.

**Beta Sheets :** We use the hypothesis that  $\beta$ -sheet regions in the native structure are the ones that are the most likely to form favorable interfaces with the rest of the sequence. We compute a scoring function for each window of 4 residues against all the remaining sequence and count the number of times that the score was higher than a set threshold. Since these interfaces are more likely to occur with nearby residues, this feature was also modeled. We used previous equations obtained in statistical physics that model the number of consecutive residues in coils by modeling proteins as a worm-like chain [4]. The observed distribution versus the theoretical model are shown in figure 2. The loop length distribution is integrated with the pairwise interaction score to obtain the overall  $\beta$ -sheet score. We scan the whole sequence to find high scoring regions and return those as our predicted sheet regions.

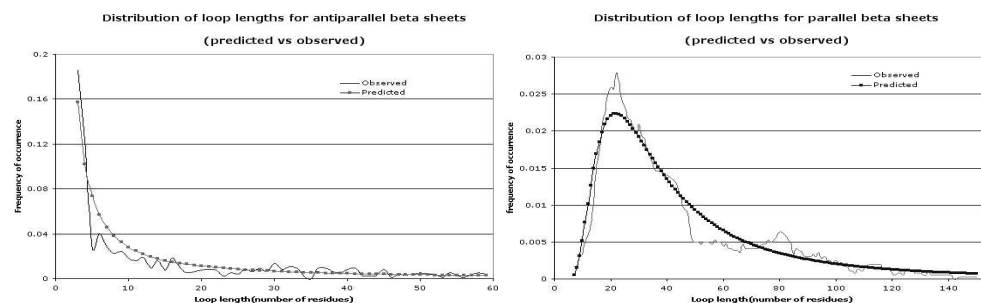


Figure 2: Distribution of loop lengths for  $\beta$ -sheets. We compare empirical data for the distribution of the number of residues between corresponding  $\beta$ -strands against our theoretical model for both antiparallel sheet (left) and parallel ones (right)

### 3 Results

Using various test sets, we achieve accuracies equal to or better than most single sequence methods and comparable accuracy to modern methods that incorporate information from homologous sequences. The presented method is currently being integrated into a tertiary structure prediction algorithm.

### References

- [1] Garnier, J. Gibrat, J.-F. Robson, B. 1996. GOR secondary structure prediction method version IV, *Meth. Enzym.*, R.F. Doolittle Ed., **266**, 540-553.
- [2] McConkey, B.J. Sobolev, V. Edelman, M. 2003. Discrimination of native protein from decoy structures using atom-atom contact frequencies *Proc. Nat. Acad. Sci.*, **100**, 3215-3220.
- [3] Post, B. Sander, C. (1993), Prediction of protein secondary structure at better than 70 % Accuracy, *J Mol. Biol.*, **232**, 584-599.
- [4] Rho, H.-X. 2001. Loops in proteins can be modeled as worm-like chains. *J. Phys. Chem. B* **105**, 6763-6766.