Sequence Motifs and Statistical Features of Distance Constraints Enhance Neural Network Predictions

J. Gorodkin[1], O. Lund[2], C. A. Andersen[3], and S. Brunak[3]

[1]Department of Genetics and Ecology The Institute of Biological Sciences University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C, Denmark

[2]Structural Bioinformatics Advanced Technologies A/S, Agern Alle 3, DK-2970 Hørsholm, Denmark

[3]Center for Biological Sequence Analysis, Department of Biotechnology, The Technical University of Denmark, B208, DK-2800 Lyngby, Denmark

Abstract

We analyze correlations between sequence separation (in residues) and physical distance (in Angstrom) of any pair of amino acids in polypeptide chains where the 3-dimensional structure is known. For each sequence separation we consider the distribution of the corresponding physical distances relative to their respective means. We find that this distribution is most bimodal for a sepration of 3 residues, that the average distance for separation 12 matches the average distance of helices of 12 residues, and that the distribution changes from being bimodal to consisting of a single peak only. Interestingly, as the mean distance keeps growing for increasing distance separation, the distance distribution relative to its mean becomes independent of the sequence separation. In agreement with these changes, similar behavior of the corresponding sequence (logo) motifs are found for increasing sequence separation. The motifs (logos) are made for the residues connecting pairs of amino acids for which the distance between the C^{α} atoms is smaller than the mean value at that separation. For small separations the motif consists of one peak located in between the two residues, then additional peaks at the separated residues appear, and for large separations, the center peak has smeared out. From this analysis we can explain why neural networks perform better, in predicting above and below the mean for given separation, than simple statistical data-driven approaches such as pair probability density functions. When incorporating the statistical knowledge in a new neural network design (compared to an older approach) a significant improvement in performance is obtained in the sequence separation range of 10–30 residues. The statistical properties and sequence motifs account qualitatively for the network performance with increasing sequence separation. In fact, we find the performance curve to be strongly correlated to the corresponding curve for the information content. A WWW server will be available at http://www.cbs.dtu.dk/services/distanceP/.

Introduction

- Statistical analysis directed towards prediction.
- Predicting distance constraints between C^α atoms in proteins.
- Other methods should use the predicted constraints to construct the 3D coordinates.
- Previous methods have mainly been based on statistical approaches.
- Neural networks. Adaptive and use only the information given by the sequence composition.
- Continuation of earlier neural network metodology (Lund et al. 1997).

Results

- Distance distributions. Bimodal for small sequence separations. Universal for large sequence separation.
- Sequence motifs and information content. Sequence logos show one peak for small separations, three peaks for medium size separations, and two peaks for large separations.
- Statistical features can be used two enhance the predictability of distance constraints. They also account for limitations.
- Predictions made by neural networks.
- Networks evalutated by Mathews correlation coefficient:

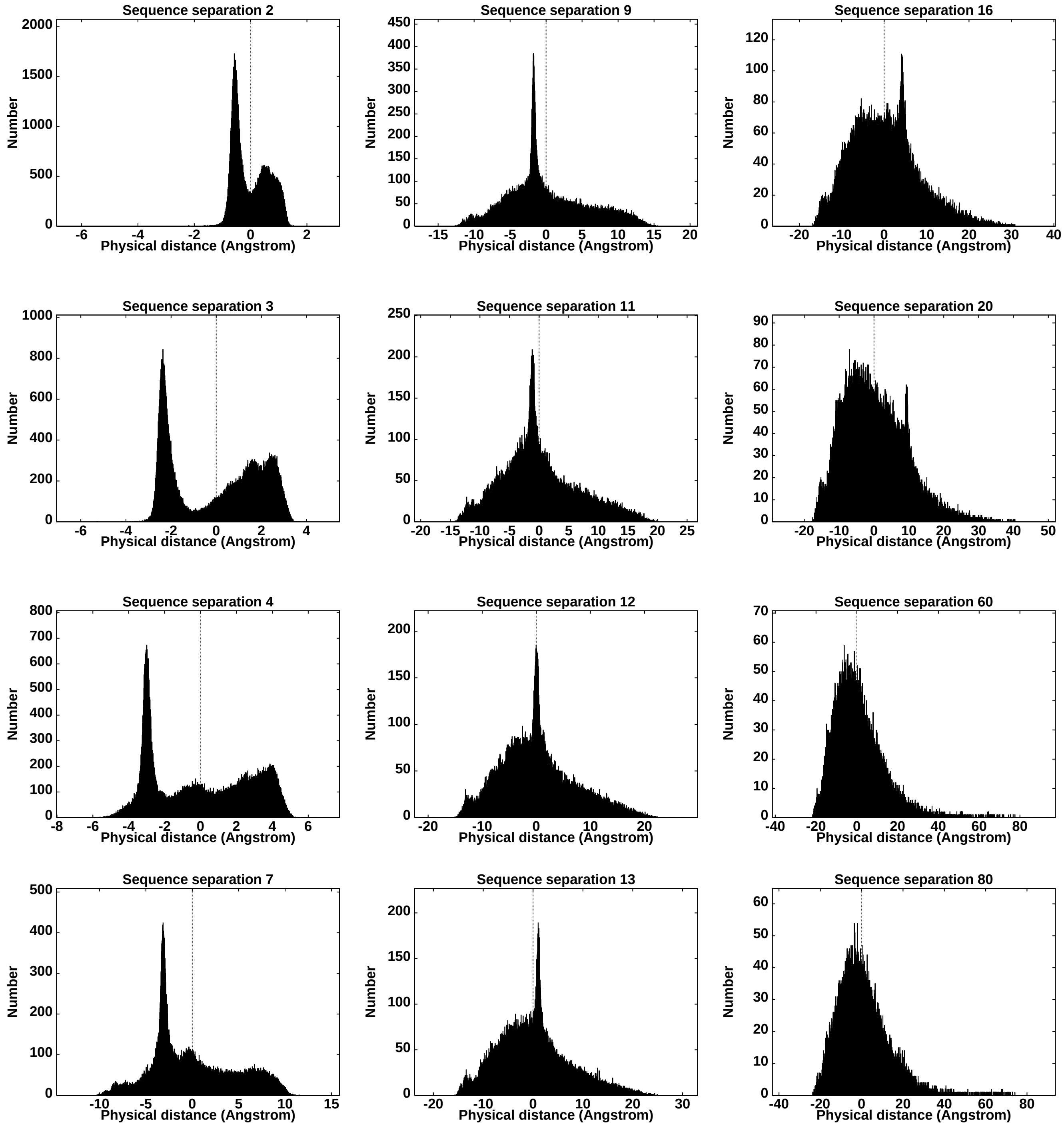
$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(N_t + N_f)(N_t + P_f)(P_t + N_f)(P_t + P_f)}} ,$$

  where P_t is the number of true positives (contact, predicted contact), N_t the number of true negatives (no contact, no contact predicted), P_f the number of false positives (no contact, contact predicted), and N_f is the number of false negatives (contact, no contact predicted).

- Performance evaluation by cross-validation. In total 8000 networks were trained.

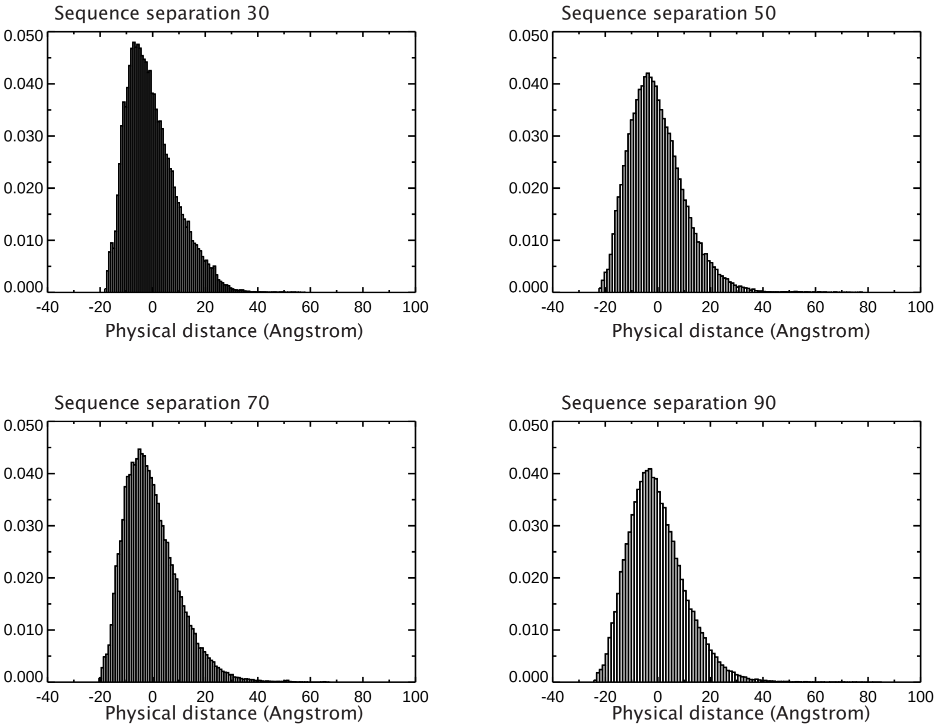
Distance distributions

Distributions relative to their mean values:

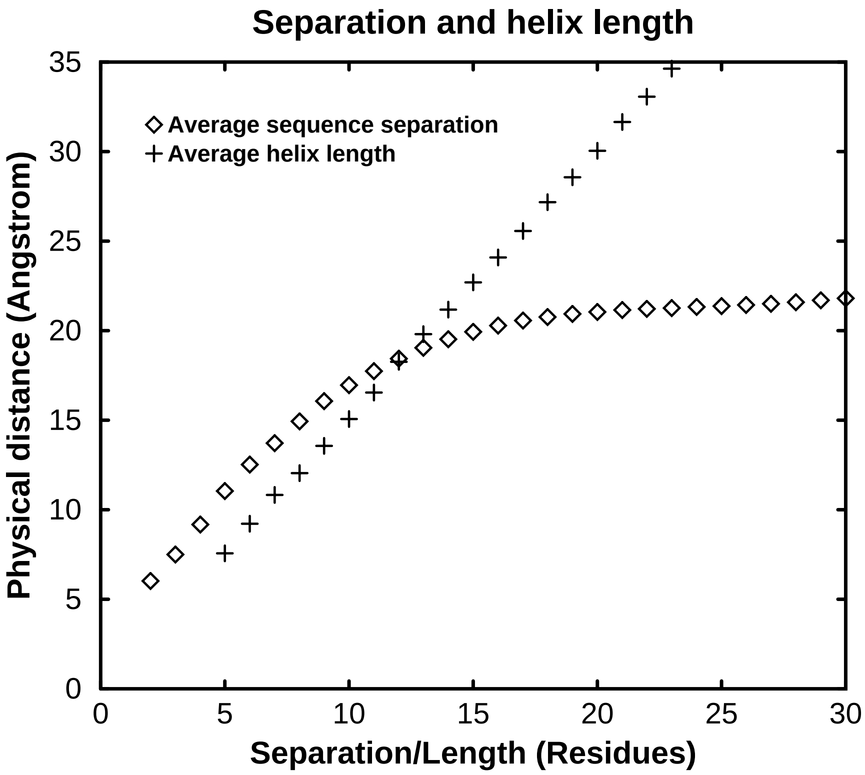


Universality of the distributions:

For large sequence separation the distance distribution becomes independent of the separation. This is evident from the probability density functions:

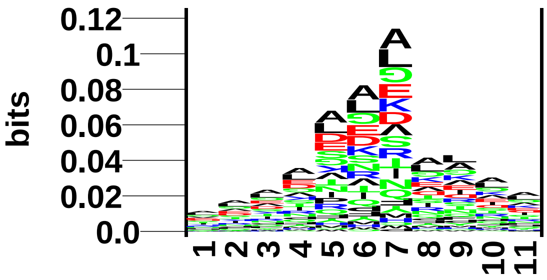


Helix length and distance thresholds:

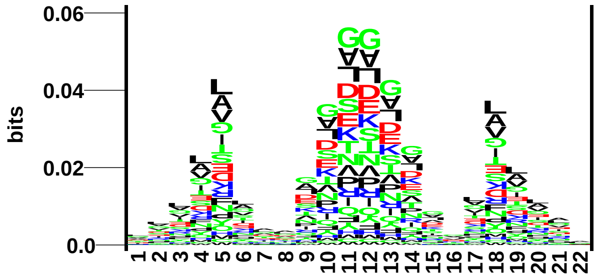


Sequence motifs

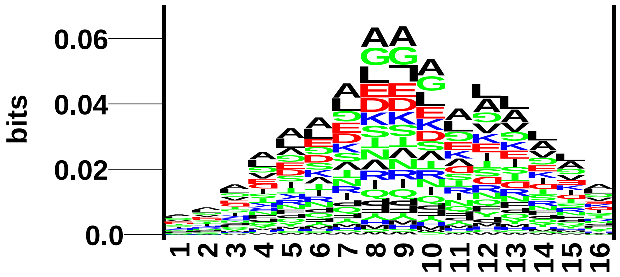
Separation: 2



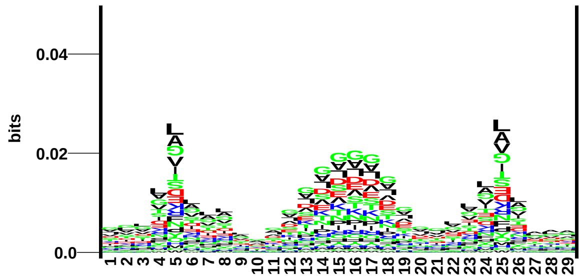
Separation: 13



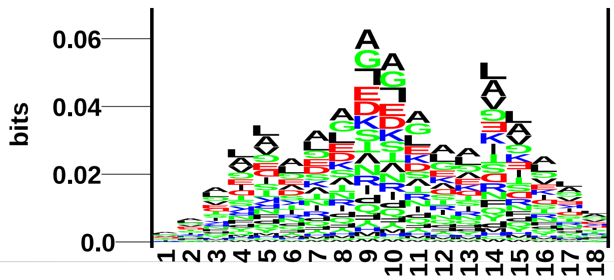
Separation: 7



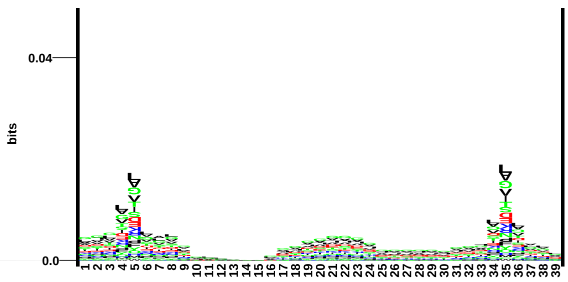
Separation: 20



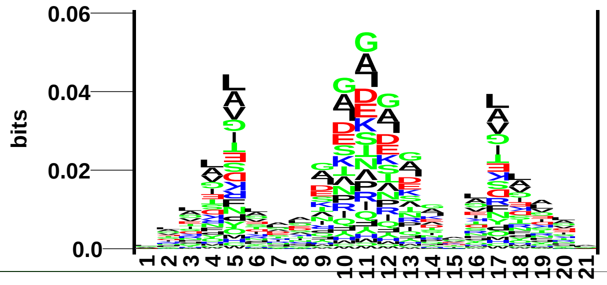
Separation: 9



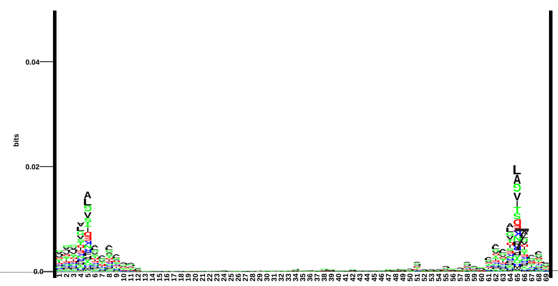
Separation: 30



Separation: 12



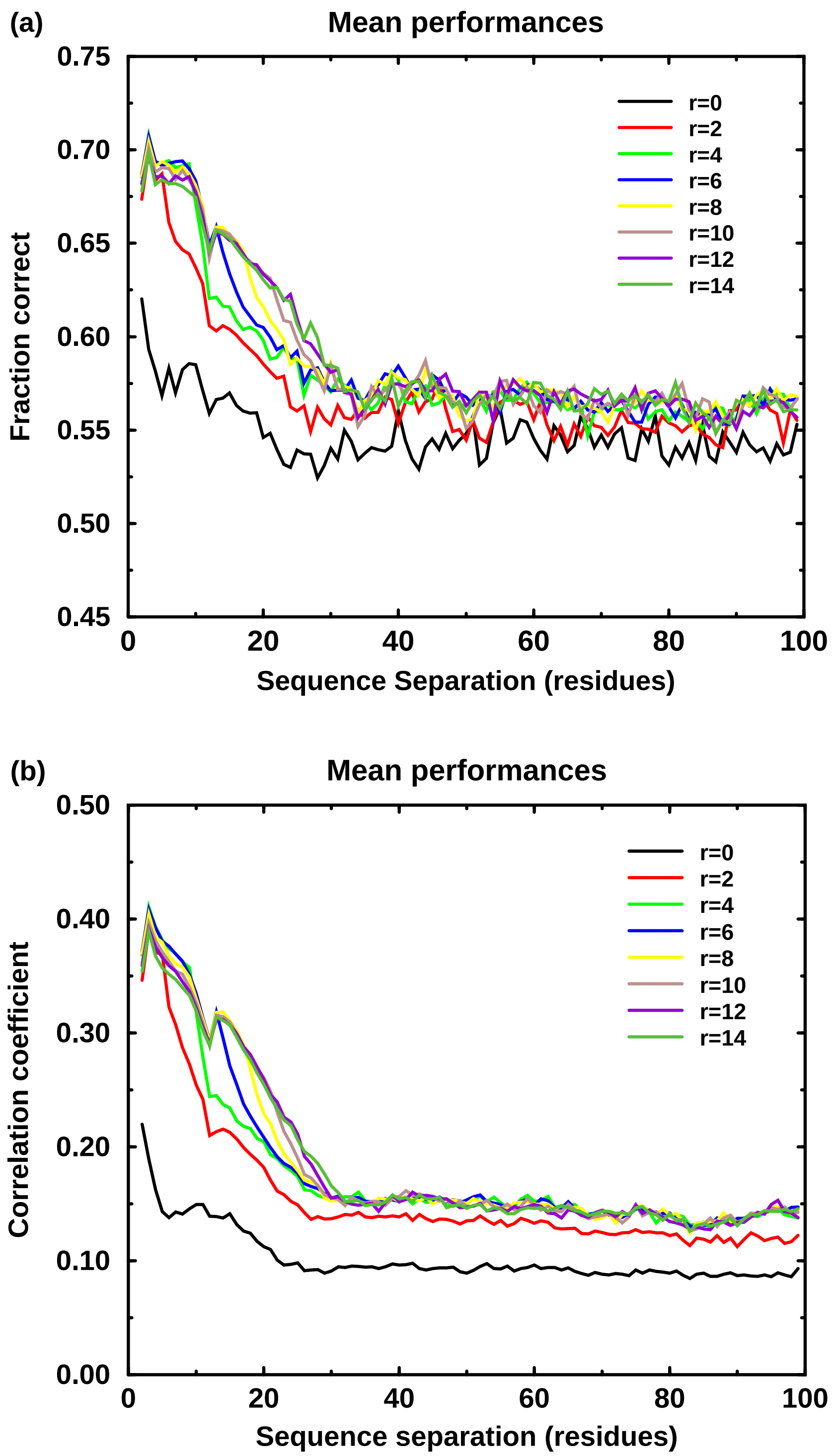
Separation: 60



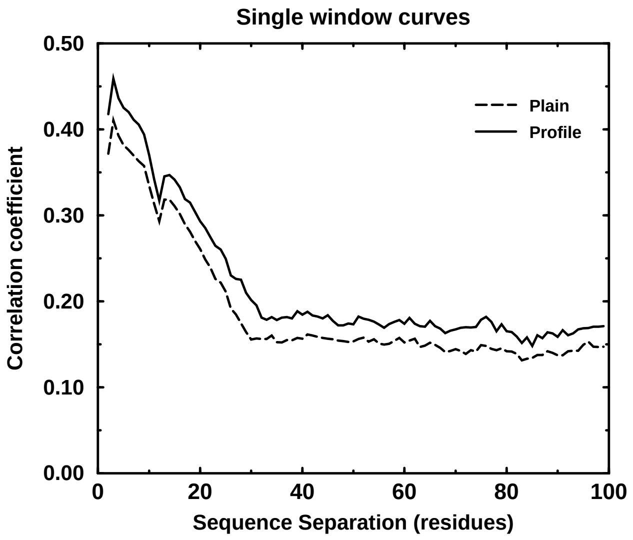
NEURAL NETWORK PREDICTIONS

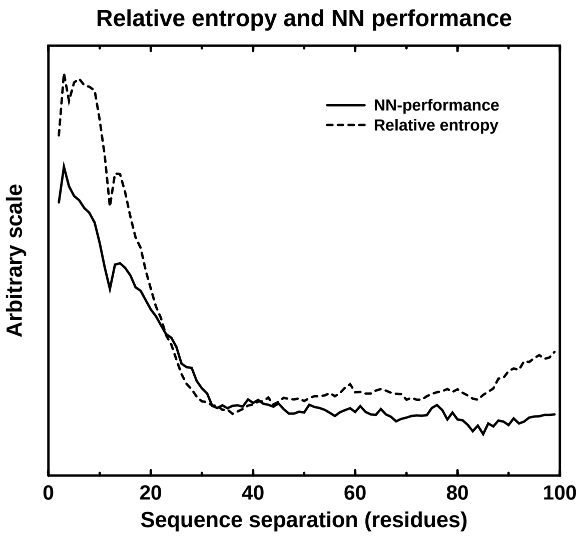
Range between separated residues in sequence window

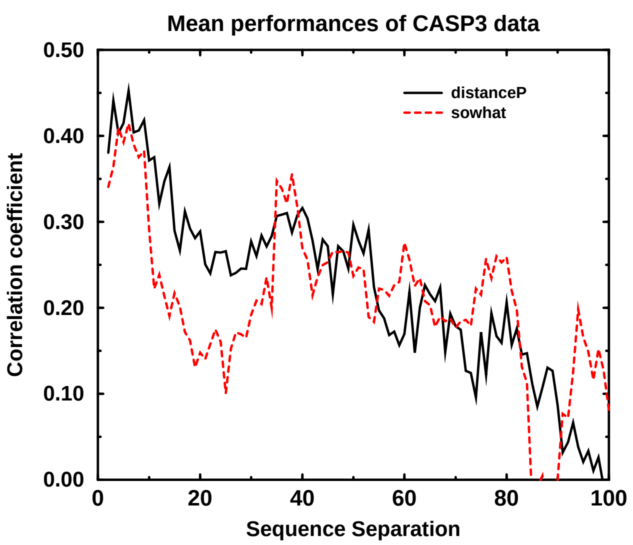
The performance as function of increasing sequence context when predicting distance constraints. The sequence context is the number of additional residues $r = 0, 2, 4, 6, 8, 10, 12, 14$ from the amino acid and toward the other amino acid. The performance is showed in percentage (a), and in correlation coefficient (b).

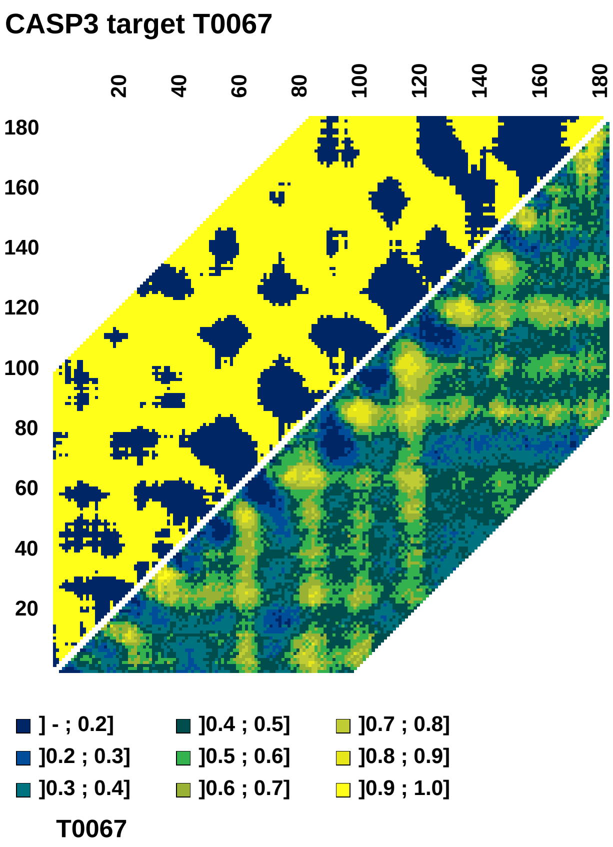


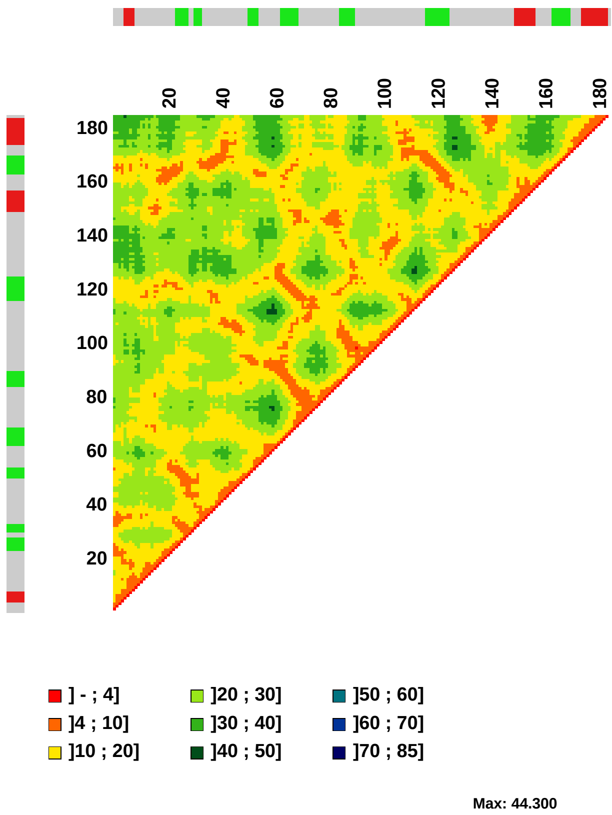
Prediction evaluation









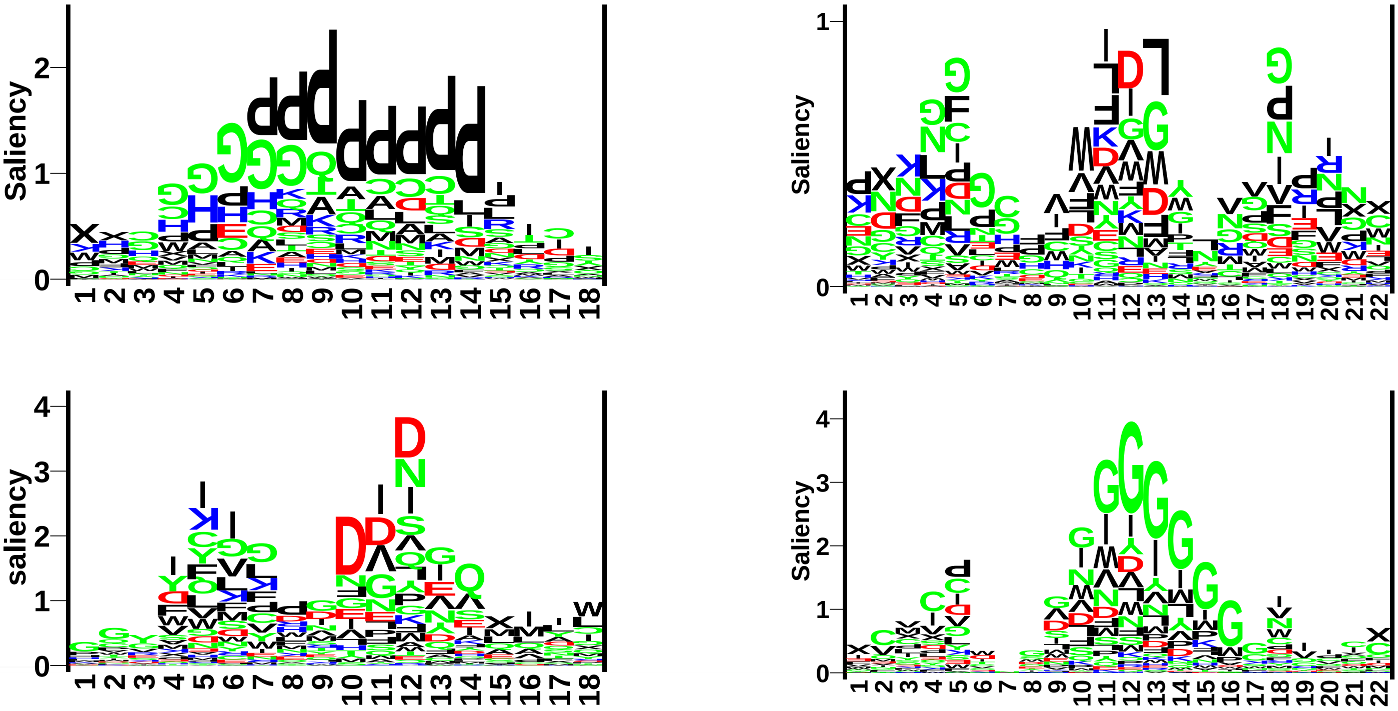


Weight-saliency logos

The cost of removing a weight from the network, the *saliency*, is related the sensitivity of the corresponding input. Saliency logos can provide information about the network and the data considered. The saliencies for the weights between input and hidden layer can be approximated (Gorodkin et al, 1997):

$$s_{ji} = w_{ji}^2 W_j^2 K,$$

where w_{ji} is the weight between input i and hidden unit j, and W_j the weight between hidden unit j and the output. K is a constant. If the signs of w_{ji} and W_j are opposite, we display the corresponding symbol upside down in the weight-saliency logo.



Conclusion

- Statistical analysis of distance constraints gave new insight in protein structure:
  - Distance distributions and their universality.
  - Sequence motifs and information content.
- Improved neural network perfomance over earlier approach.
- Explaining the behavior of the perdiction performance.