Using Sequence Motifs for Enhanced Neural Network Prediction of Protein Distance Constraints

Jan Gorodkin
Center for Biological Sequence Analysis, The Technical University of Denmark
The Institute of Biological Sciences, University of Aarhus, Denmark
Co-workers: Ole Lund, Claus A. Andersen, and Søren Brunak

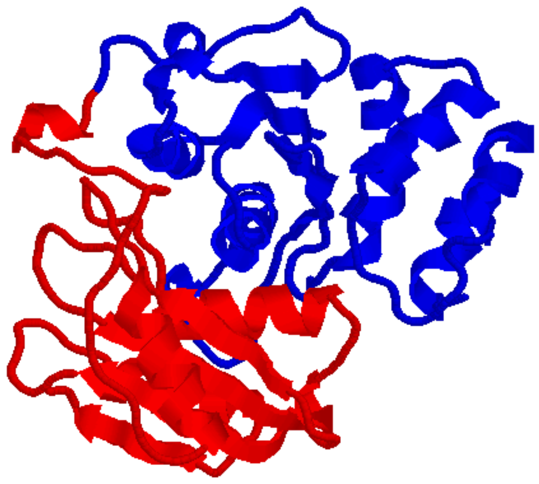
Outline:

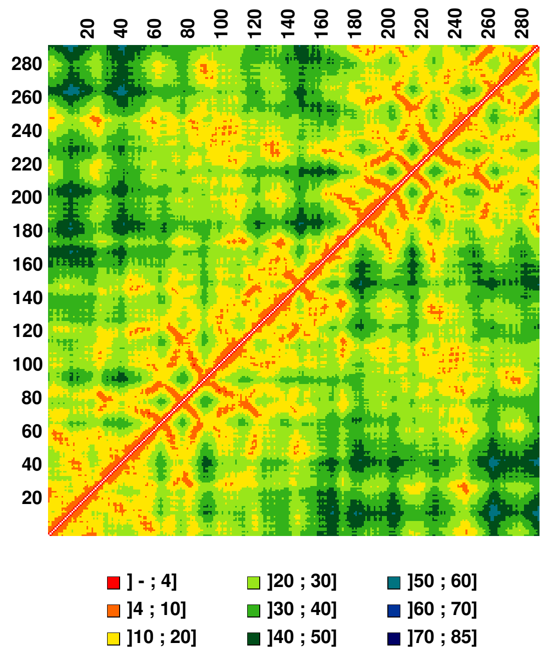
- Introduction
- Statistical analysis: distributions and motifs
- Neural networks: prediction and behavior
- Conclusion

Introduction

- Prediction of 3D distances from sequence.
- 3D structure and distance map:

  PDB entry 1ctt





- Predicting distance constraints between C^{α}

  Sequence separation 17

  atoms. Y W Q R N A A L P P K T R E V E D D C F E G Q

- Other methods to complete the construction of 3D coordinates.

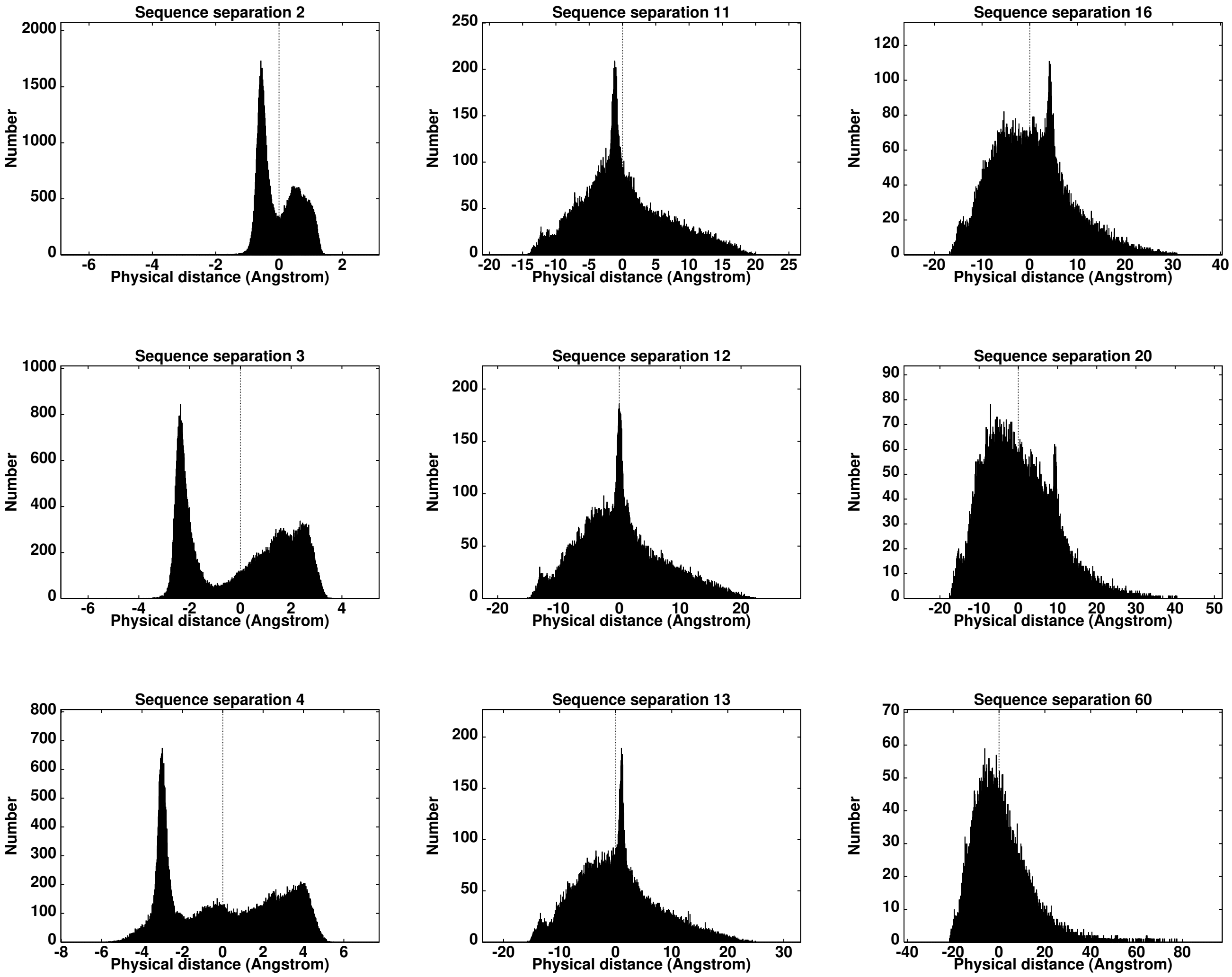
- Previous work: CPHmodels (Lund et al., 1997)
(http://www.cbs.dtu.dk/services/CPHmodels/)
  - NN better than pair probability density functions.
  - Use predictions in homology building (loops).
- Relating NN behavior to statistical features.
- Significant improvement of NN predictions.

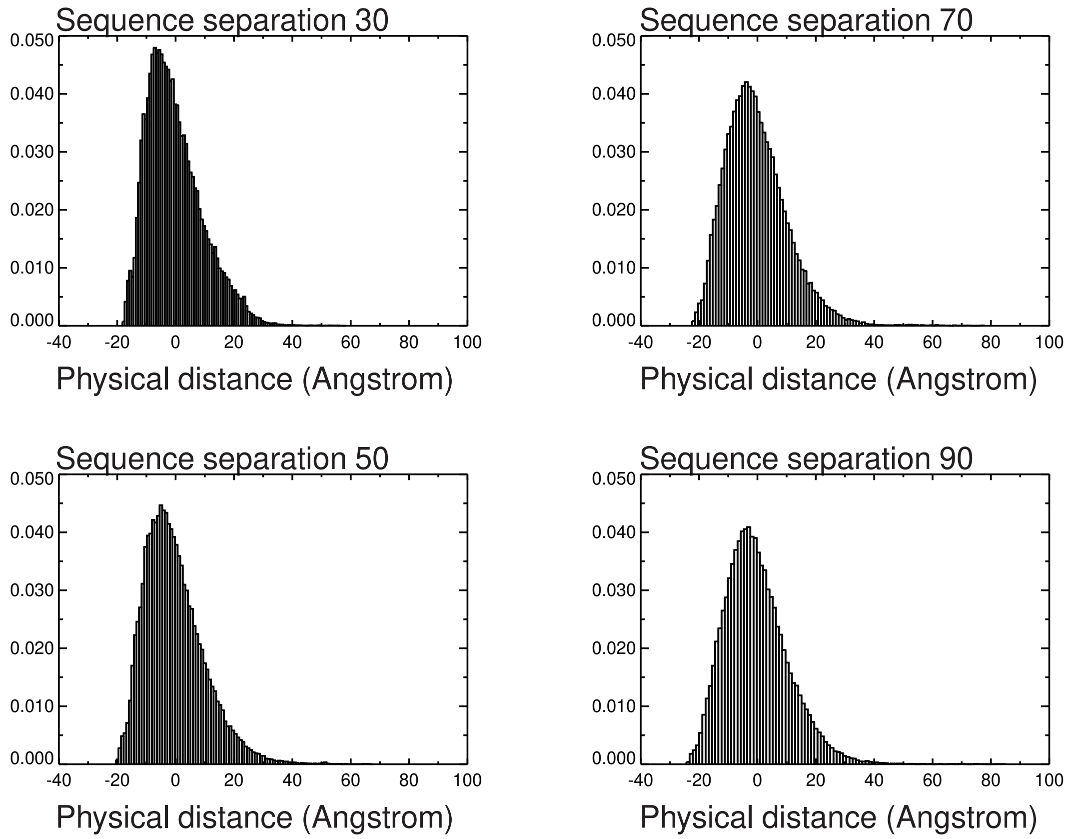
Data set

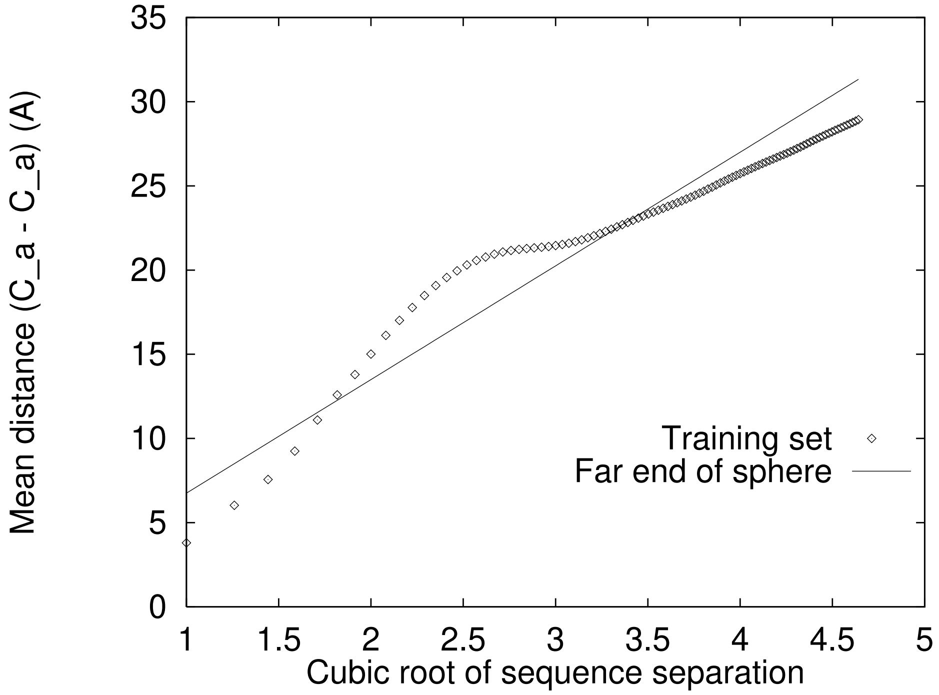
- Construction of non-redundant data set.
- Using entries from PDB, rel. 82; (Bernstein et al., 1977).
- X-ray resolution better than 2.5 Å.
- Threshold curve: $T = 290/\sqrt{L}$.
  (Sander & Schneider, 1991; Lund et al., 1997).
  (http://www.cbs.dtu.dk/services/RedHom/)
- Resulting set of 650 sequences.
- Use 10 cross validation sets.

Statistical analysis
Distance distributions

- Sequence separation and 3D distances
  (Reese et al., 1996).
- Distributions translated by their means.
- Bimodal for small sequence separations.
  Universal for large sequence separations.



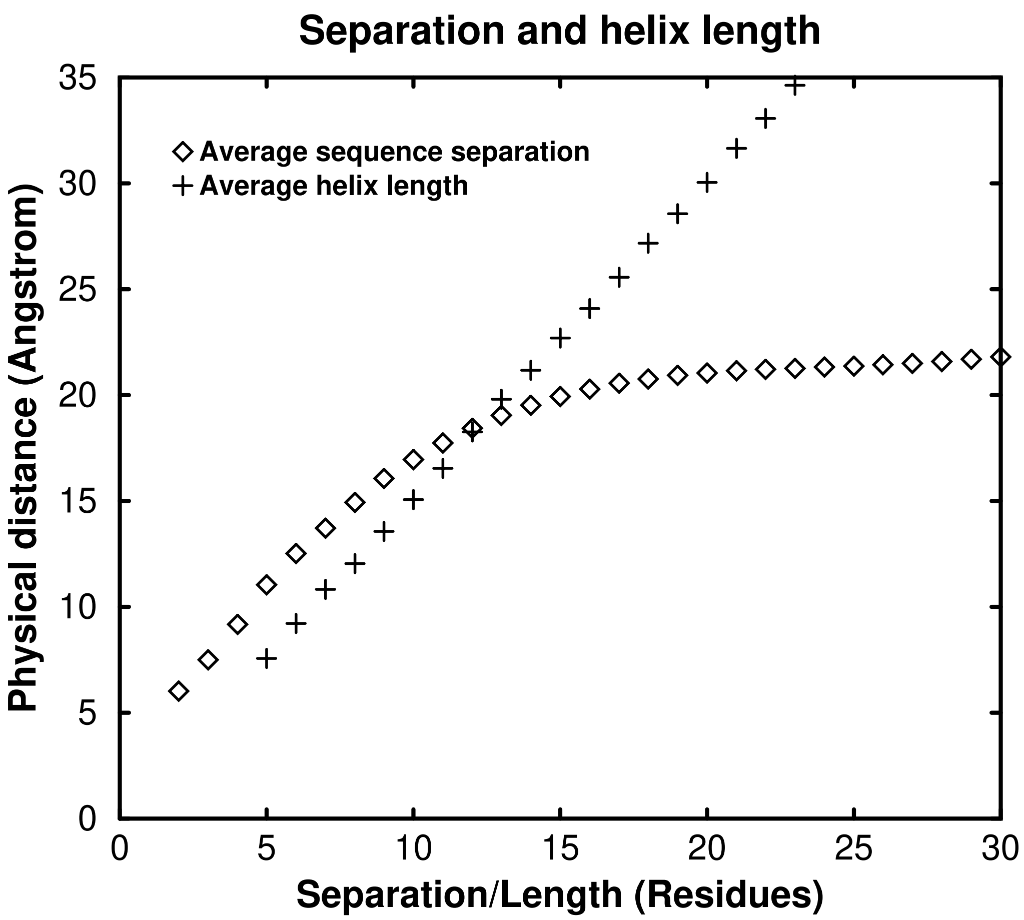




(Lund et al., 1997)

Helix lengths

- Computing the exact physical length in α-helices of increasing lengths.
- Compare to 3D distance at corresponding sequence separations.



Sequence motifs and information content

- Sequence windows corresponding to given separation.

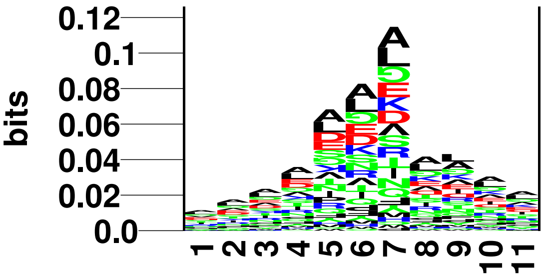
- Compare windows: $dist < t$ to $dist \geq t$.

- Sequence logos using the relative entropy: $I = \sum_k q_k log_2 \frac{q_k}{p_k}$

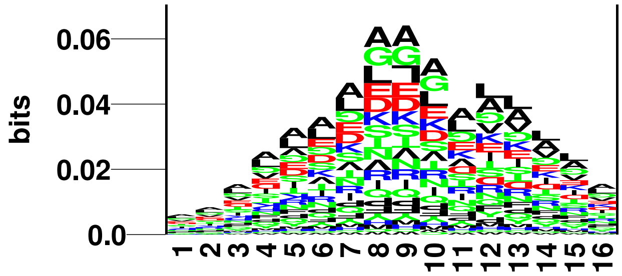
Sequence logos

(Schneider & Stephens, 1990)

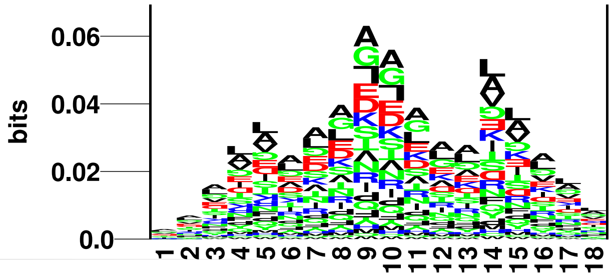
Separation: 2



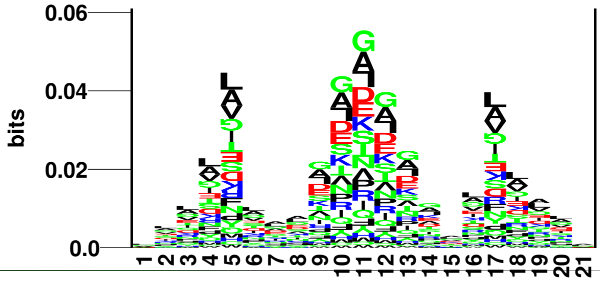
Separation: 7



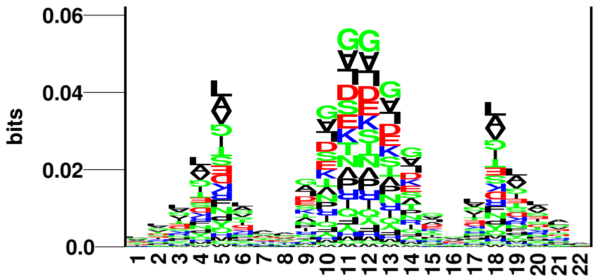
Separation: 9



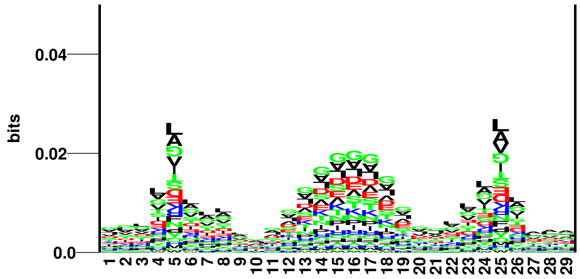
Separation: 12



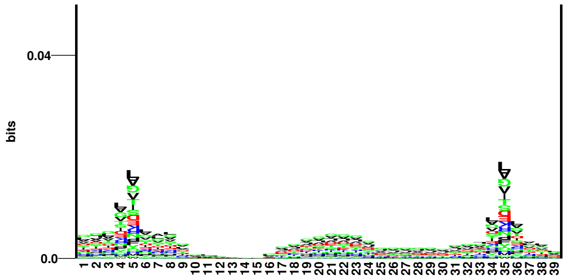
Separation: 13



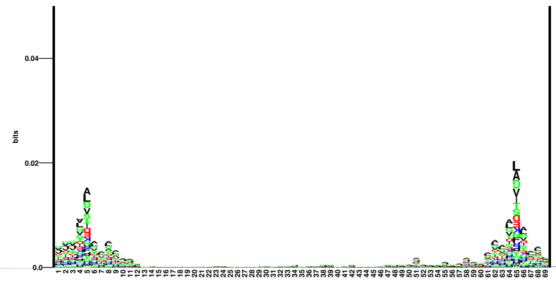
Separation: 20



Separation: 30

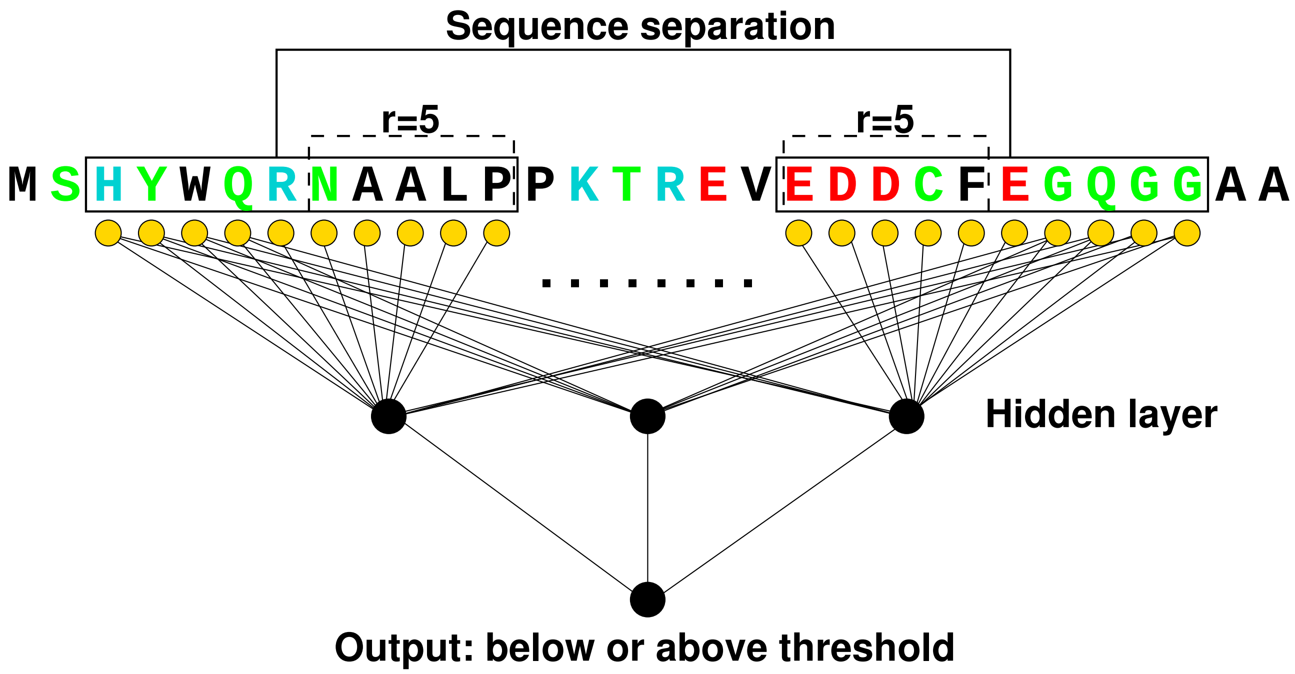


Separation: 60



Neural networks

Architectures:



Evaluation of performance:

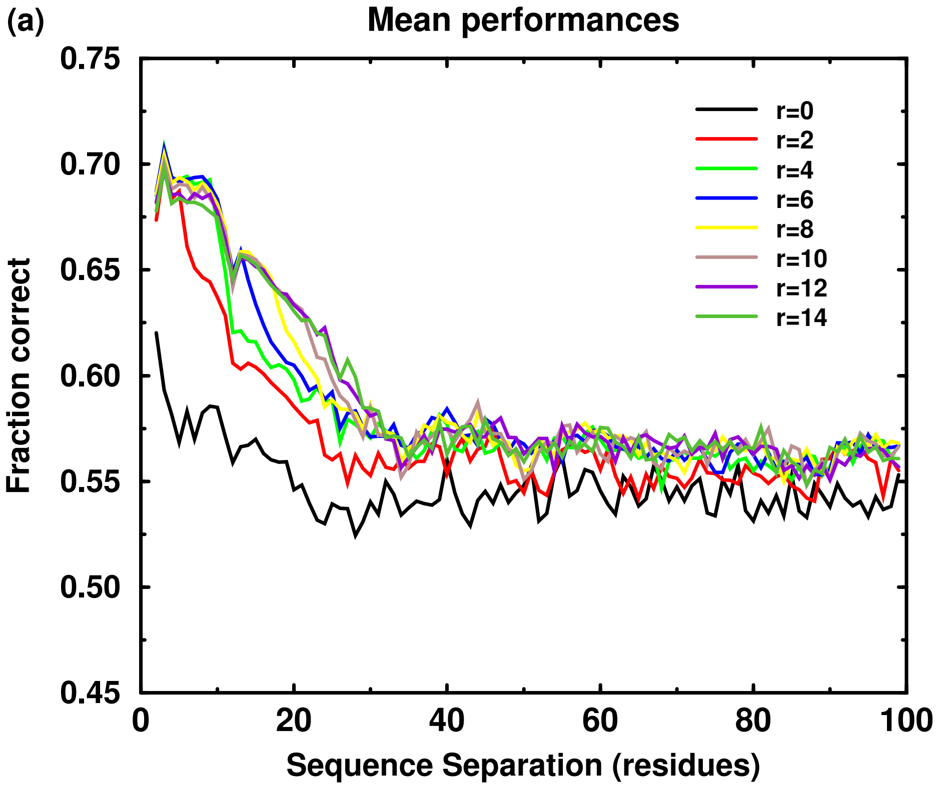
- average over 10 cross-validation sets.
- Correlation coefficient (Mathews, 1975)

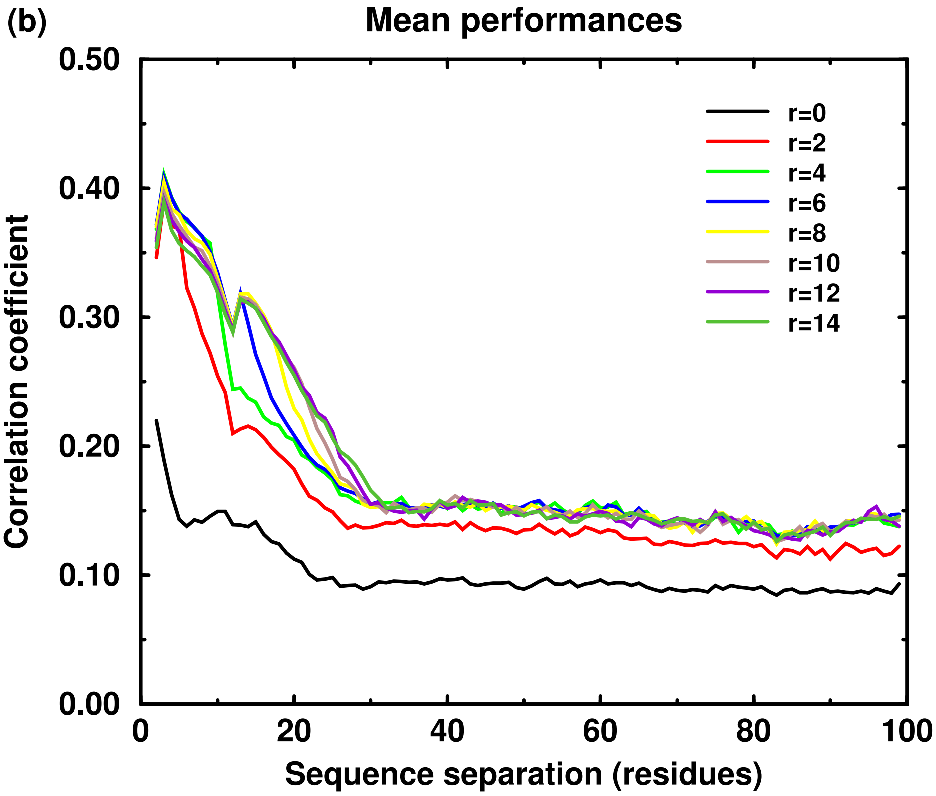
$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(N_t + N_f)(N_t + P_f)(P_t + N_f)(P_t + P_f)}}$$

P_t: true pos.; N_t: true neg.; P_f: false pos.; N_f: false neg.

- 8000 neural networks trained.

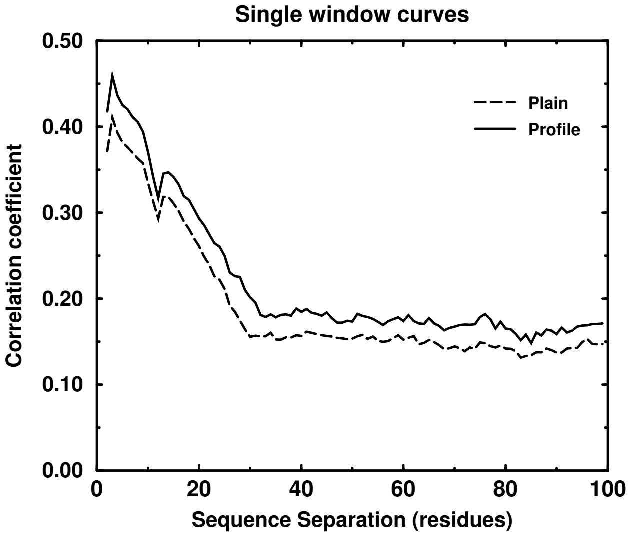
Network performances

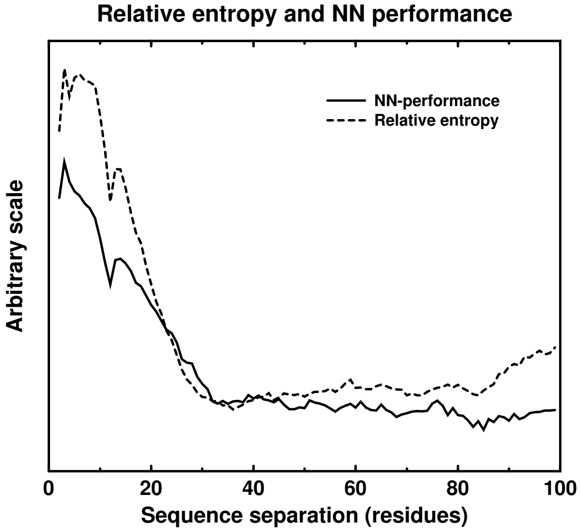




Network analysis

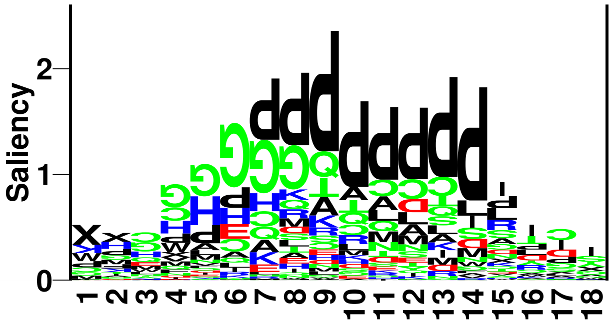
Performance correlated to statistical properties:

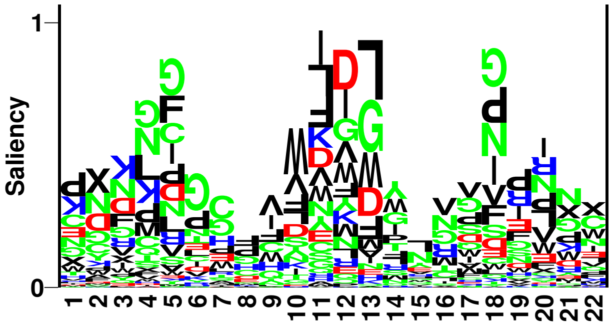




Weight-saliency logos:

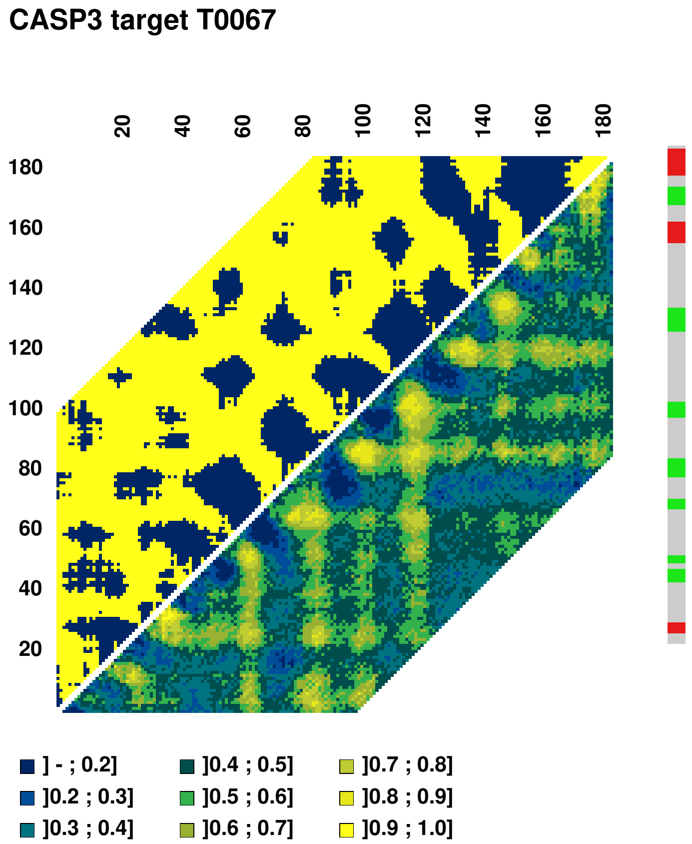
- Main features by the network
- Saliency of weight approx.: $s_{ji} = w_{ji}^2 W_j^2 K$
  (Gorodkin et al., 1997; http://www.wspc.com/journals/ijns/85_6/gorod.pdf)

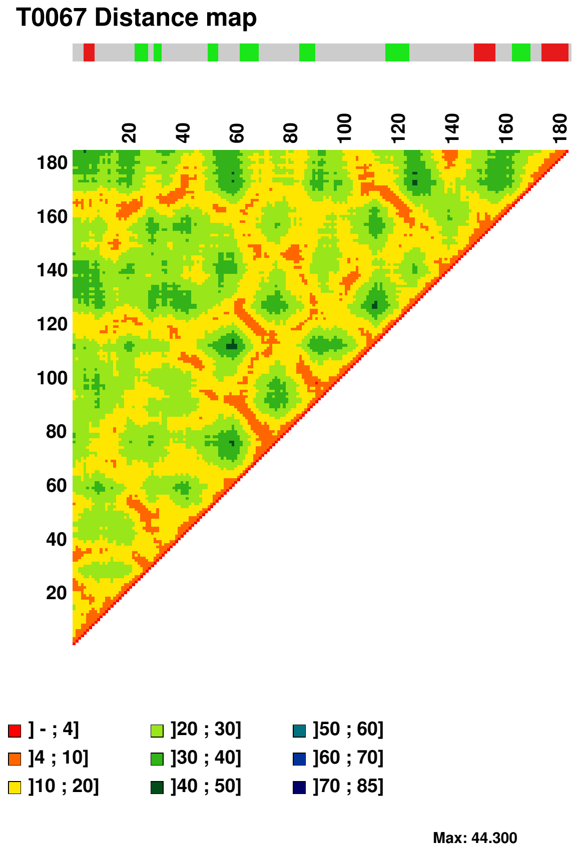




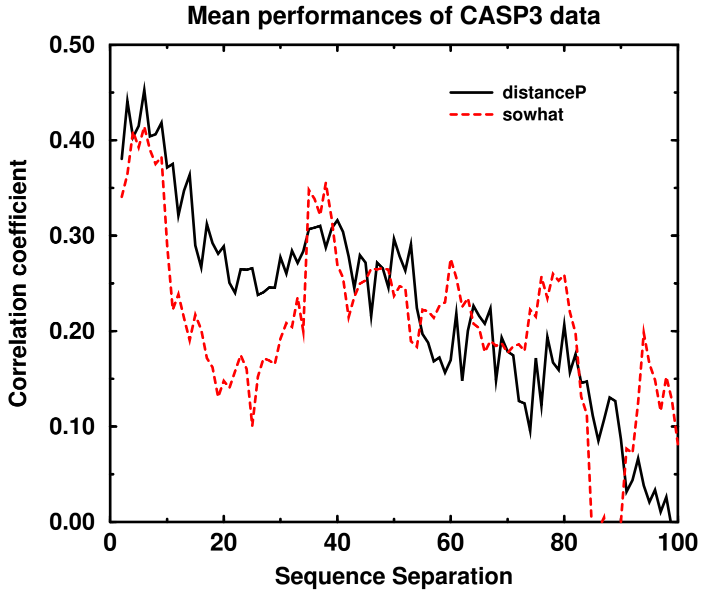
Evaluating on CASP3

Distance constraint prediction vs. true distance map





Performance for increasing sequence separation



Conclusion

- Statistical analysis of distance constraints:
  - Distance distributions and their universality.
  - Sequence motifs and information content.
- Explaining the prediction performance.
- Improved neural network performance.
- Future directions:
  - Include correlations between different sequence separations.
  - Automate the process of 3D determination.

Announcements

New servers

- Prediction of distance constraints:
  http://www.cbs.dtu.dk/services/distanceP/
- Server for graphical interface of matrix plots:
  http://www.cbs.dtu.dk/services/MatrixPlot/

Poster 53 today:

- Organization and Prediction of Structural Domains in Multidomain Proteins.

WWW demo tomorrow 17.30 (mail room):

- The distanceP server:
  predicting protein distance constraints.
- MatrixPlot: visualizing sequence constraints.