ISMB-99, 95-105

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

Jan Gorodkin[1,2], **Ole Lund**[1,*], **Claus A. Andersen**[1], and **Søren Brunak**[1]

[1]Center for Biological Sequence Analysis, Department of Biotechnology,
The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark
gorodkin@cbs.dtu.dk, olund@strubix.dk, ca2@cbs.dtu.dk, brunak@cbs.dtu.dk
Phone: +45 45 25 24 77, Fax: +45 45 93 15 85

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

Abstract

Correlations between sequence separation (in residues) and distance (in Angstrom) of any pair of amino acids in polypeptide chains are investigated. For each sequence separation we define a distance threshold. For pairs of amino acids where the distance between C^{α} atoms is smaller than the threshold, a characteristic sequence (logo) motif, is found. The motifs change as the sequence separation increases: for small separations they consist of one peak located in between the two residues, then additional peaks at these residues appear, and finally the center peak smears out for very large separations. We also find correlations between the residues in the center of the motif. This and other statistical analyses are used to design neural networks with enhanced performance compared to earlier work. Importantly, the statistical analysis explains why neural networks perform better than simple statistical data-driven approaches such as pair probability density functions. The statistical results also explain characteristics of the network performance for increasing sequence separation. The improvement of the new network design is significant in the sequence separation range 10–30 residues. Finally, we find that the performance curve for increasing sequence separation is directly correlated to the corresponding information content. A WWW server, *distanceP*, is available at http://www.cbs.dtu.dk/services/distanceP/.

Keywords: Distance prediction; sequence motifs; distance constraints; neural network; protein structure.

Introduction

Much work have over the years been put into approaches which either analyze or predict features of the three-dimensional structure using distributions of distances, correlated mutations, and more lately neural networks, or combinations of these *e.g.* (Tanaka & Scheraga 1976; Miyazawa & Jernigan 1985; Bohr *et al.* 1990; Sippl 1990; Maiorov & Crippen 1992; Göbel *et al.* 1994; Mirny & Shaknovich 1996; Thomas, Casari, & Sander 1996; Lund *et al.* 1997; Olmea & Valencia 1997; Skolnick, Kolinski, & Ortiz 1997; Fariselli & Casadio 1999). The ability to adopt structure from sequences depends on constructing an appropriate cost function for the native structure. In search of such a function we here concentrate on finding a method to predict distance constraints that correlate well with the observed distances in proteins. As the neural network approach is the only approach so far which includes sequence context for the considered pair of amino acids, these are expected not only to perform better, but also to capture more features relating distance constraints and sequence composition.

The analysis include investigation of the distances between amino acid as well as sequence motifs and correlations for separated residues. We construct a prediction scheme which significantly improve on an earlier approach (Lund *et al.* 1997).

For each particular sequence separation, the corresponding distance threshold is computed as the average of all physical distances in a large data set between any two amino acids separated by that amount of residues (Lund *et al.* 1997). Here, we include an analysis of the distance distributions relative to these thresholds and use them to explain qualitative behavior of the neural network prediction scheme, thus extending earlier studies (Reese *et al.* 1996). For the prediction scheme used here it is essential to relate the distributions to their means. Analysis of the network weight composition reveal intriguing properties of the distance constraints: the sequence motifs can be decomposed into sub-motifs associated with each of the hidden units in the neural network.

Further, as the sequence separation increases there is a clear correspondence in the change of the mean value, distance distributions, and the sequence motifs describing the distance constraints of the separated amino acids, respectively. The predicted distance constraints may be used as inputs to threading, *ab initio*, or loop modeling algorithms.

Materials and Method

Data extraction

The data set was extracted from the Brookhaven Protein Data Bank (Bernstein *et al.* 1977), release 82 containing 5762 proteins. In brief entries were excluded if: (1) the sec-

* Present address: Structural Bioinformatics Advanced Technologies A/S, Agern Alle 3, DK-2970 Hørsholm, Denmark

ondary structure of the proteins could not be assigned by the program DSSP (Kabsch & Sander 1983), as the DSSP assignment is used to quantify the secondary structure identity in the pairwise alignments, (2) the proteins had any physical chain breaks (defined as neighboring amino acids in the sequence having C^{α}-distances exceeding 4.0Å) or entries where the DSSP program detected chain breaks or incomplete backbones, (3) they had a resolution value greater than 2.5Å, since proteins with worse resolution, are less reliable as templates for homology modeling of the C^{α} trace (unpublished results).

Individual chains of entries were discarded if (1) they had a length of less than 30 amino acids, (2) they had less than 50% secondary structure assignment as defined by the program DSSP, (3) they had more than 85% non-amino acids (nucleotides) in the sequence, and (4) they had more than 10% of non-standard amino acids (B, X, Z) in the sequence.

A representative set with low pairwise sequence similarity was selected by running algorithm #1 of Hobohm *et al.* (1992) implemented in the program RedHom (Lund *et al.* 1997).

In brief the sequences were sorted according to resolution (all NMR structures were assigned resolution 100). The sequences with the same resolution were sorted so that higher priority was given to longer proteins.

The sequences were aligned utilizing the local alignment program, *ssearch* (Myers & Miller 1988; Pearson 1990) using the pam120 amino acid substitution matrix (Dayhoff & Orcutt 1978), with gap penalties -12, -4. As a cutoff for sequence similarity we applied the threshold $T = 290/\sqrt{L}$, T is the percentage of identity in the alignment and L the length of the alignment. Starting from the top of the list each sequence was used as a probe to exclude all sequence similar proteins further down the list.

By visual inspection seven proteins were removed from the list, since their structure is either 'sustained' by DNA or predominantly buried in the membrane. The resulting 744 protein chains are composed of the residues where the C^{α}-atom position is specified in the PDB entry.

Ten cross-validation sets were selected such that they all contain approximately the same number of residues, and all have the same length distribution of the chains. All the data are made publicly available through the world wide web page http://www.cbs.dtu.dk/services/distanceP/.

Information Content / Relative entropy measure

Here we use the relative entropy to measure the information content (Kullback & Leibler 1951) of aligned regions between sequence separated residues. The information content is obtained by summing for the respective position in the alignment, $I = \sum_{i=1}^{L} I_i$, where I_i is the information content of position i in the alignment. The information content at each position will sometimes be displayed as a sequence logo (Schneider & Stephens 1990). The position-dependent information content is given by

$$I_i = \sum_k q_{ik} \log_2 \frac{q_{ik}}{p_k} \quad , \tag{1}$$

where k refers to the symbols of the alphabet considered (here amino acids). The observed fraction of symbol k at position i is q_{ik}, and p_k is the background probability of finding symbol k by chance in the sequences. p_k will sometimes be replaced by a position-dependent background probability, that is the probability of observing letter k at some position in the alignment in another data set one wishes to compare to. Symbols in logos turned 180 degrees indicate that $q_{ik} < p_k$.

Neural networks

As in the previous work, we apply two-layer feed-forward neural networks, trained by standard back-propagation, see *e.g.* (Brunak, Engelbrecht, & Knudsen 1991; Bishop 1996; Baldi & Brunak 1998), to predict whether two residues are below or above a given distance threshold in space. The sequence input is sparsely encoded. In (Lund *et al.* 1997) the inputs were processed as two windows centered around each of the separated amino acids. However, here we extend that scheme by allowing the windows to grow towards each other, and even merge to a single large window covering the complete sequence between the separated amino acids. Even though such a scheme increases the computational requirements, it allow us to search for optimal covering between the separated amino acids.

As there can be a large difference in the number of positive (contact) and negative (no contact) sequence windows, for a given separation, we apply the balanced learning approach (Rost & Sander 1993). Training is done by a 10 set cross-validation approach (Bishop 1996), and the result is reported as the average performance over the partitions. The performance on each partition is evaluated by the Mathews correlation coefficient (Mathews 1975)

$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(N_t + N_f)(N_t + P_f)(P_t + N_f)(P_t + P_f)}} \quad , \tag{2}$$

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).

The analysis of the patterns stored in the weights of the networks is done through the *saliency* of the weights, that is the cost of removing a single weight while keeping the remaining ones. Due to the sparse encoding each weight connected to a hidden unit corresponds exactly to a particular amino acid at a given position in the sequence windows used as inputs. We can then obtain a ranking of symbols on each position in the input field. To compute the saliencies we use the approximation for two-layer one-output networks (Gorodkin *et al.* 1997), who showed that the saliencies for the weights between input and hidden layer can be written as

$$s_{ji}^k = s_{ji} = w_{ji}^2 W_j^2 K, \tag{3}$$

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. The kth symbol is implicitly given due to the

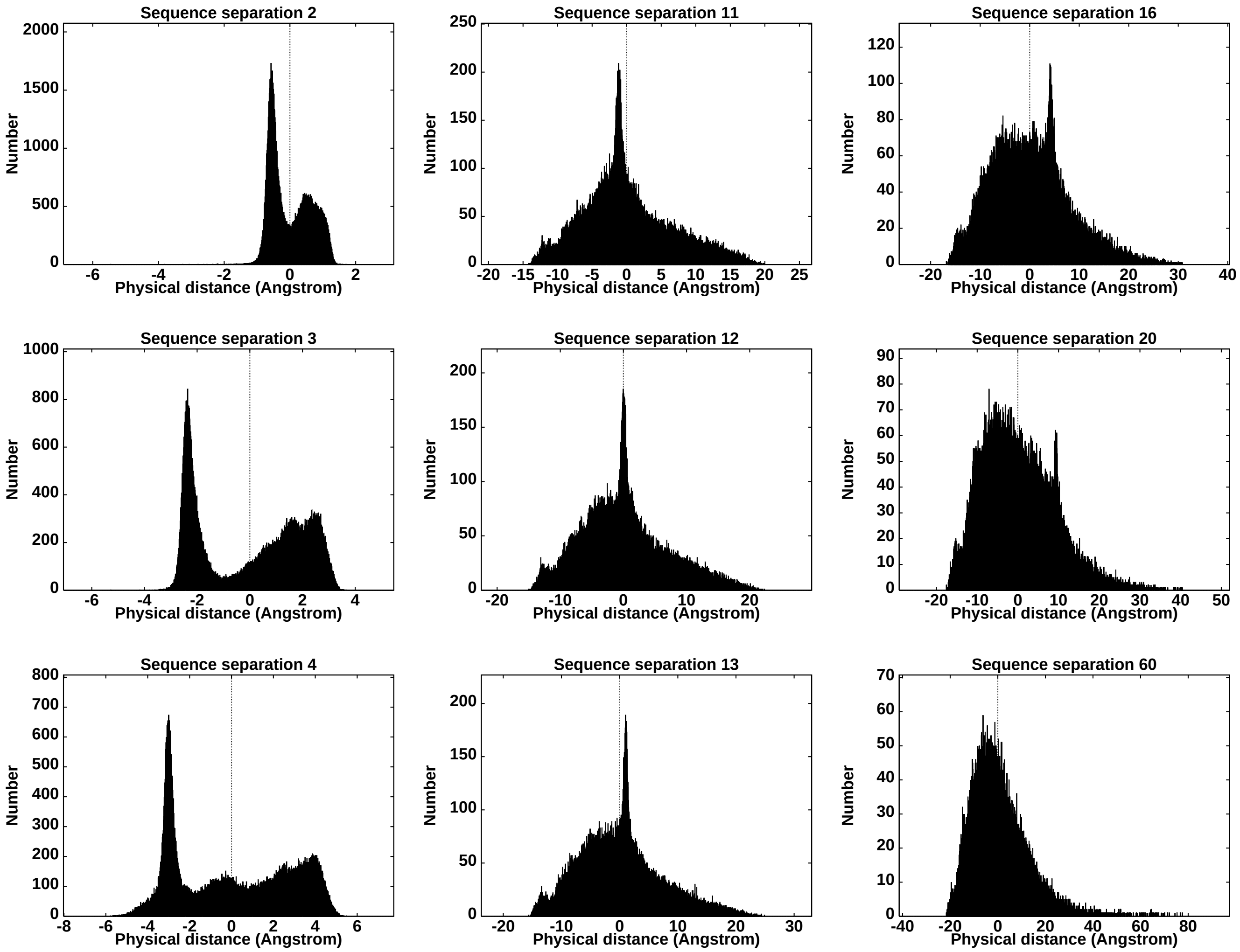

Figure 1: Length distributions of residue segments for corresponding sequence separations, relative the respective mean values. Sequence separations 2, 3, 4, 11, 12, 13, 16, 20, and 60 are shown. These show the representative shapes of the distance distributions. The vertical line through zero indicates the displacement with respect to the mean distance.

sparse encoding. If the signs of w_{ji} and W_j are opposite, we display the corresponding symbol upside down in the weight-saliency logo.

Results

We conduct statistical analysis of the data and the distance constraints between amino acids, and subsequent use the results to design and explain the behavior of a neural network prediction scheme with enhanced performance.

Statistical analysis

For each sequence separation (in residues), we derive the mean of all the distances between pairs of C^{α} atoms. We use these means as distance constraint *thresholds*, and in a prediction scheme we wish to predict whether the distance between any pair of amino acids is above or below the threshold corresponding to a particular separation in residues. To analyze which pairs are above and below the threshold, it is relevant to compare: (1) the distribution of distances between amino acid pairs below and above the threshold, and (2) the sequence composition of segments where the pairs of amino acids are below and above the threshold.

First we investigate the length distribution of the distances as function of the sequence separation. A complete investigation of physical distances for increasing sequence separation is given by (Reese *et al.* 1996). In particular it was found that the α-helices caused a distinct peak up to sequence separation 20, whereas β-strands are seen up to separations 5 only. However, when we perform a simple translation of these distributions relative to their respective means, the same plots provide essentially new qualitative information, which is anticipated to be strongly correlated to the performance of a predictor (above/below the threshold). In particular we focus on the distributions shown in Figure 1, but we also use the remaining distributions to make some important observations.

When inspecting the distance distributions relative to their mean, two main observations are made. First, the distance distribution for sequence separation 3, is the one distance distribution where the data is most bimodal. Thus sequence separation 3 provides the most distinct partition of the data points. Hence, in agreement with the results in (Lund *et al.* 1997) we anticipate that the best prediction of distance constraints can be obtained for sequence separation 3. Furthermore, we observe that the α-helix peak shifts relative to the mean when going from sequence separation 11 to 13. The length of the helices becomes longer than the mean distances. This shift interestingly involve the peak to be placed at the mean value itself for sequence separation 12. Due to this phenomenon, we anticipate, that, for an optimized predictor, it can be slightly harder to predict distance constraints for separation 12 than for separations 11 and 13.

The peak present at the mean value for sequence separation 12 does indeed reflect the length of helices as demonstrated clearly in Figure 2. Rather than using the simple rule that each residue in a helix increases the physical distance with 1.5 Angstrom (Branden & Tooze 1999), we computed the actual physical lengths for each size helix to obtain a more accurate picture. The physical length of the α-helices was calculated by finding the helical axis and measuring the translation per C^{α}-atom along this axis. The helical axis was determined by the mass center of four consecutive C^{α}-atoms. Helices of length four are therefore not included, since only one center of mass was present. We see that helices at 12 residues coincide with sequence separation 12. Again we use this as an indication that at sequence separation 12 it may be slightly harder to predict distance constraints than at separation 13.

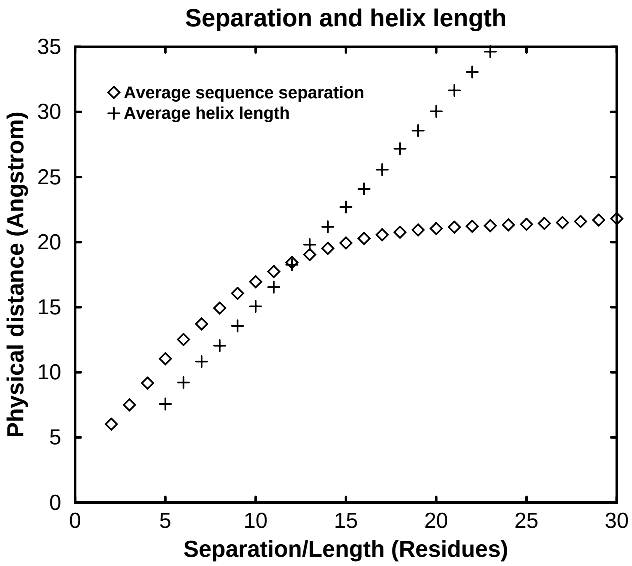

Figure 2: Mean distances for increasing sequence separation and computed average physical helix lengths for increasing number of residues.

As the sequence separation increases the distance distribution approaches a universal shape, presumably independent of structural elements, which are governed by more local distance constraints. The traits from the local elements do as mentioned by (Reese *et al.* 1996) (who merge large separations) vanish for separations 20–25. (Here we considered separations up to 100 residues.) In Figure 1 we see that the transition from bimodal distributions to unimodal distributions centered around their means, indicates that prediction of distance constraints must become harder with increasing sequence separation. In particular when the universal shape has been reached (sequence separation larger than 30), we should not expect a change in the prediction ability as the sequence separation increases. The universal distribution has its mode value approximately at -3.5 Angstrom, indicating that the most probable physical distance for large sequence separations corresponds to the distance between two C^{α} atoms.

Notice that the universality only appears when the distribution is displaced by its mean distance. This is interesting since the mean of the physical distances grows as the sequence separation increases. From a predictability perspective, it therefore makes good sense to use the mean value as a threshold to decide the distance constraint for an arbitrary sequence separation.

A useful prediction scheme must rely on the information available in the sequences. To investigate if there exists a detectable signal between the sequence separated residues, for each sequence separation, we constructed sequence logos (Schneider & Stephens 1990) as follows: The sequence segments for which the physical distance between the separated amino acids was above the threshold were used to generate a position-dependent background distribution of amino acids. The segments with corresponding physical distance below the threshold were all aligned and displayed in sequence logos using the computed background distribution of amino acids. The pure information content curves are shown for increasing sequence in Figure 3. The corresponding sequence logos are displayed in Figure 4. We used a "margin" of 4 residues to the left and right of the logos, *e.g.*, for sequence separation 2, the physical distance is measured between position 5 and 7 in the logo.

The change in the sequence patterns is consistent with the change in the distribution of physical distances. Up to separations 6–7, the distribution of distances (Figure 1) contains two peaks, with the β-strand peak vanishing completely for separation 7–8 residues. For the same separations, the sequence motif changes from containing one to three peaks. For larger sequence separations, the motif consists of three characteristic peaks, the two located at positions exactly corresponding to the separated amino acids. The third peak appear in the center. This peak tells us that for physical distances within the threshold, it indeed matters whether the residues in the center can bend or turn the chain. We see that exactly such amino acids are located in the center peak: glycine, apartic acid, glutamic acid, and lysine, all these are medium size residues (except glycine), being hydrophilic, thereby having affinity for the solvent, but not being on the outer most surface of the protein (see *e.g.*, (Creighton 1993)

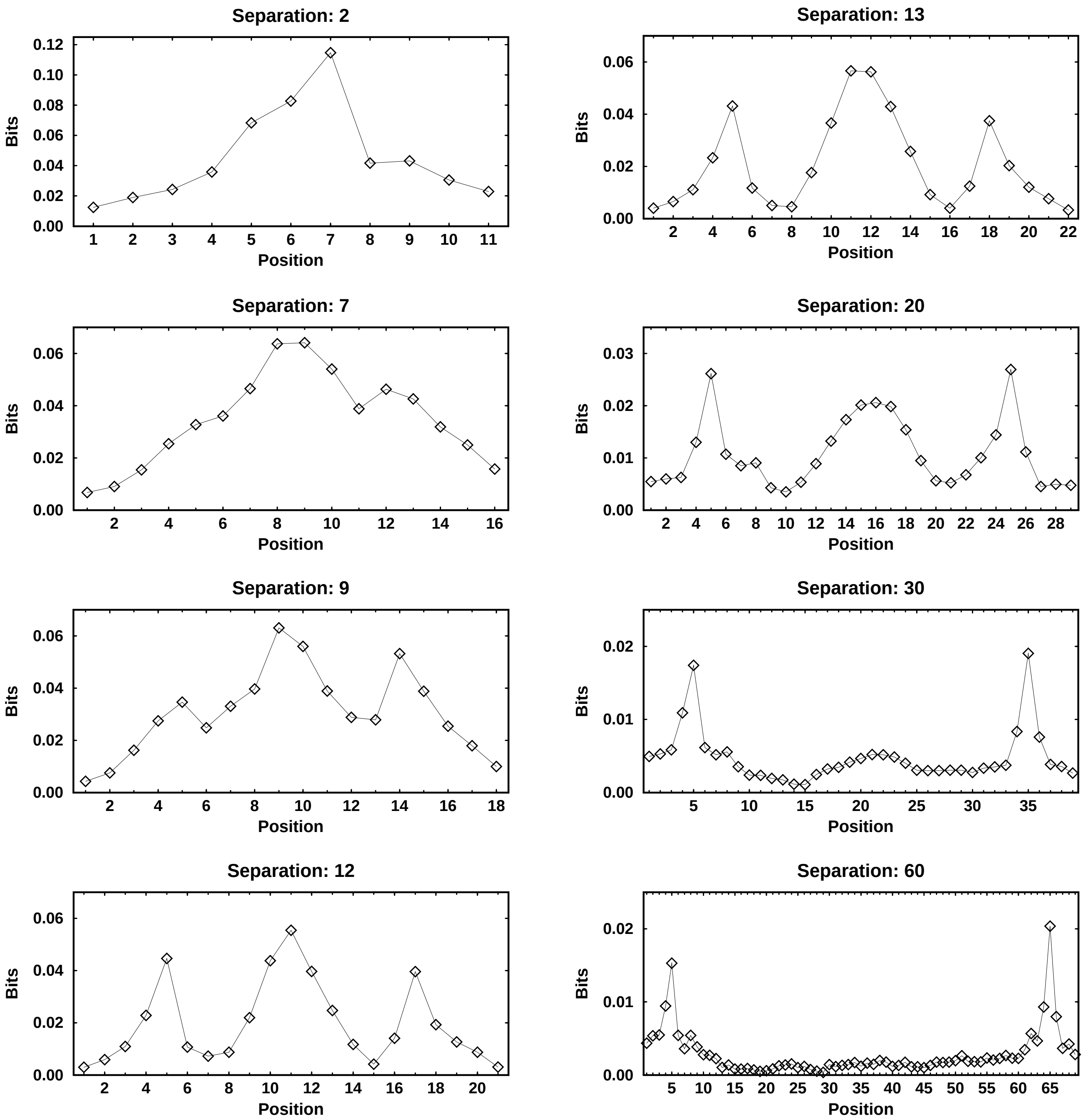

Figure 3: Sequence information profiles for increasing sequence separation.

for amino acid properties). As the sequence separation increases from about 20 to 30 the center peak smears out, in agreement with the transition of the distance distribution that shifts to the universal distribution in this range of sequence separation. The sequence motif likewise becomes "universal". Only the peaks located at the separated residues are left.

The composition of single peak motifs resemble to a large degree the composition of the center for motifs having three peaks. However, the slightly increased amount of leucine moves from the center of the one peak motifs to the separated amino acid positions in the three peak motifs. The reverse happens for glycine. Interestingly, as the sequence separation increases valine (and isoleucine) become present at the outermost peaks. In agreement with the helix length shift from below to above the threshold at separations 12

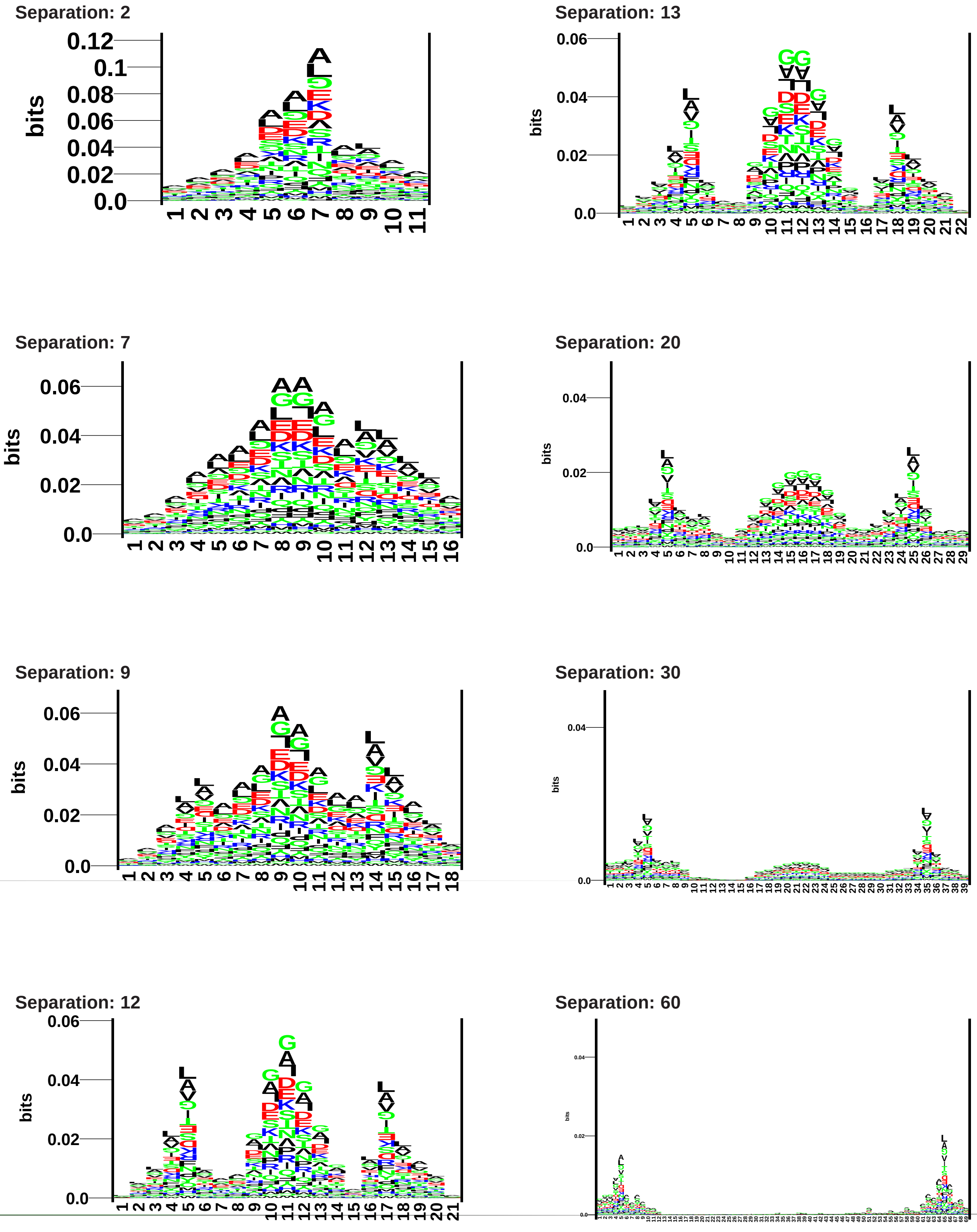

Figure 4: Sequence logos for increasing sequence separation. A margin of 4 residues on both ends of the logo is used. Symbols displayed upside down indicate that their amount is lower than the background amount. This figure is available in full color at http://www.cbs.dtu.dk/services/distanceP/.

and 13, the center peak shifts from slight over to slight underrepresentation of alanine: helices are no longer dominant for distances below the threshold.

The smearing out of the center peak for large sequence separations does not indicate lack of bending/turn properties, but reflects that such amino acids can be placed in a large region in between the two separated amino acids. What we see for small separations, is that the signal in addition must be located in a very specific position relative to the separated amino acids.

Neural networks: predictions and behavior

As found above, optimized prediction of distance constraints for varying sequence separation, should be performed by a non-linear predictor. Here we use neural networks. Clearly, far most of the information is found in the sequence logos, so we expect only a few hidden neurons (units) to be necessary in the design of the neural network architecture. As earlier, we only use two-layer networks with one output unit (above/below threshold). We fix the number of hidden units to 5, but the size of the input field may vary.

We start out by quantitatively investigating the relation between the sequence motifs in the logos above, and the amount of sequence context needed in the prediction scheme. First, we choose the amount of sequence context, by starting with local windows around the separated amino acids and the extending the sequence region, r, to be included. We only include additional residues in the direction towards the center. The number of residues used in the direction away from the center is fixed to 4, except in the cases of $r = 0$ and $r = 2$, where that number is zero and two respectively. At some point the two sequence regions may reach each other. Whether they overlap, or just connect does not affect the performance. For each $r = 0, 2, 4, 6, 8, 10, 12, 14$, we trained networks for sequence separations 2 to 99, resulting in the training of 8000 networks, since we used 10 cross-validation sets. The performances in "fraction correct" and "correlation coefficient" are shown in Figure 5.

Figure 5(b) shows the performance in terms of correlation coefficient. We see that each time the sequence separation becomes so large that the two context regions do not connect the performance drops when compared to contexts that merge into a single window. This behavior is generic, it happens consistently for increasing size of sequence context region, though the phenomenon is asymptotically decreasing. An effect can be found up to a region size of about 30, see Figure 6. Several features appear on the performance curve, and they can all be explained by the statistical observations made earlier. First, the curve with no context ($r = 0$) corresponds to the curve one obtains using probability pair density functions (Lund *et al.* 1997). The bad performance of such a predictor is to be expected due to the lack of motif that was not included in the model. The green curve ($r = 4$) is the curve that corresponds to the neural network predictor in (Lund *et al.* 1997). As observed therein a significant improvement is obtained when including just a little bit of sequence context. As the size of the context region is increased, so is the performance up to about sequence separation 30. Then the performance remains the same independent of the sequence separation. We even note the drop in performance for sequence separation 12, as predicted from the statistical analysis above. It is also characteristic that, as the distribution of physical distances approaches the universal shape, and as the center peak of the sequence logos vanishes, the performance drops and becomes constant with approximately 0.15 in correlation coefficient. The change in prediction performance takes place at sequence separations for which changes in the distribution of physical distance and sequence motifs change. Due to the not vanishing signal in the logos for large separations, it is possible to predict distance constraints significantly better than random, and better than with a no-context approach. The conclusion is not surprising: the best performing network is that which uses as much context as possible. However, for large sequence separations more than 30–35 residues, the same performance is obtained by just using a small amount of sequence context around the separated amino acids, as in (Lund *et al.* 1997). We found that the performance between the worst and best cross-validation set is about 0.1 for separations up to 12 residues, then about 0.07 for separations up to 30 residues, and then 0.1–0.2 for separations larger than 30. Hence, at sequence separation 31 the performance start to fluctuate indicating that the sequence motif becomes less defined throughout the sequences in the respective cross-validation sets. Hence we can use the networks as an indicator for when a sequence motif is well defined.

As an independent test, we predicted on nine CASP3 (http://predictioncenter.llnl.gov/casp3/) targets (and submitted the predictions). None of these targets (T0053 T0067 T0071 T0077 T0081 T0082 T0083 T0084 T0085) have sequence similarity to any sequence with known structure. The previous method (Lund *et al.* 1997) had on the average 64.5% correct predictions and an average correlation coefficient of 0.224. The method presented here has on average 70.3% correct predictions and an average correlation coefficient of 0.249. The performance of the original method corresponds to the performance previously published (Lund *et al.* 1997). However, the average measure is a less convenient way to report the performance, as the good performance on small separations are biased due to the many large separation networks that have essentially the same performance. Therefore we have also constructed performance curves similar to the one in Figure 5 (not shown). The main features of the prediction curve was present. In Figure 8 we show a prediction example of the distance constraints for T0067. The server (http://www.cbs.dtu.dk/services/distanceP/) provides predictions up a sequence separation of 100. Notice that predictions up to a sequence separation of 30 clearly capture the main part of the distance constraints.

We also investigated the qualitative relation between the network performance and information content in the sequence logos. Interestingly, and in agreement with the observations made earlier, we found that the two curves have the same qualitative behavior as the sequence separation increase (Figure 7). Both curves peak at separation 3, both curves drop at separation 12, and both curves reaches the plateau for sequence separation 30. We note that the relative entropy curve increases slightly as the separation increases.

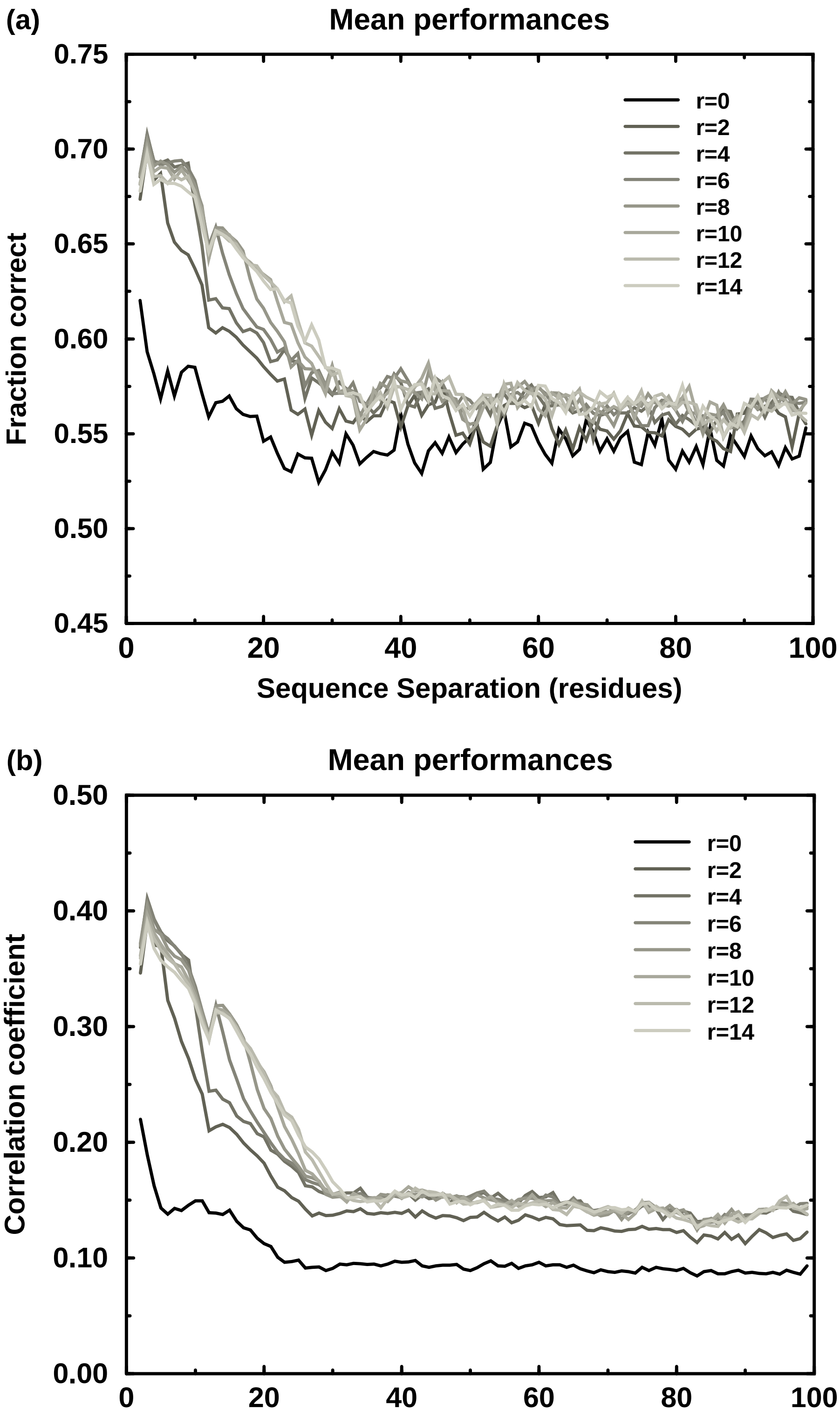

Figure 5: The performance as function of increasing sequence context when predicting distance constraints. The sequence context is the number of additional residues $r = 0, 2, 4, 6, 8, 10, 12, 14$ from the amino acid and toward the other amino acid. The performance is showed in percentage (a), and in correlation coefficient (b). This figure is available in full color at http://www.cbs.dtu.dk/services/distanceP/.

This is due to a well known artifact for decreasing sampling size, and entropy measures. The correlation between the curves also indicates that the most information used by the predictor stem from the motifs in the sequence logos.

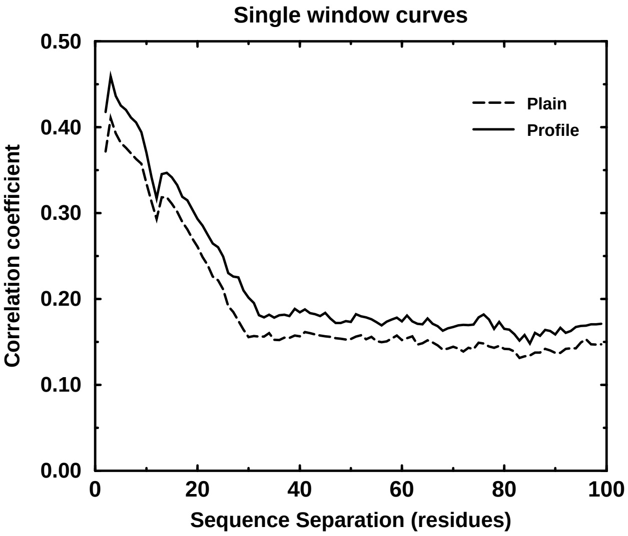

Figure 6: The performance curve using a single windows only. We also show the performance when including homologue in the respective tests. This complete a profile.

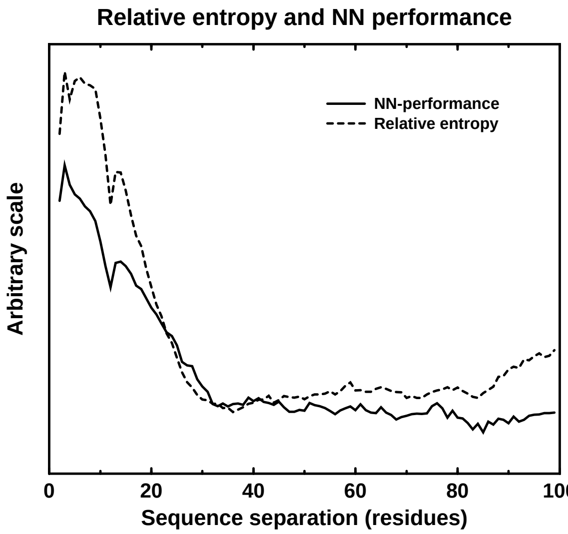

Figure 7: Information content and network performance for increasing sequence separation.

Finally, we conclude by an analysis of the neural network weight composition with the aim of revealing significant motifs used in the learning process. In general, we found the same motifs appearing for the each of the 10 cross-validation sets for a given sequence separation. As described in the Methods section, we compute saliency logos of the network parameters, that is we display the amino acids for which the corresponding weights, if removed, cause the largest increase in network error. We made saliency logos for each of the 5 hidden neurons. For the short sequence separations the network strongly punish having a proline in the center of the input field, that is a proline in between the separated amino acids. This makes sense as proline is not located within secondary structure elements, and the helix lengths for small separations are smaller than the mean distances. (The network has learnt that proline is not present in helices.) In contrast, the presence of aspartic acid can have a positive or negative effect depending on the position in the input field. For larger sequence separations, some of the neurons display a three peak motif resembling what was found in the sequence logos. The center of such logos can be very glycine rich, but also show a strong punishment for valine or tryptophan. Sometimes two neurons can share the motif of what is present for one neuron in another cross-validation set. Other details can be found as well. Here we just outlined the main features: the networks not only look for favorable amino acids, they also detects those that are not favorable at all. In Figure 9 we show an example of the motifs for 2 of the 5 hidden neurons for a network with sequence separation 13.

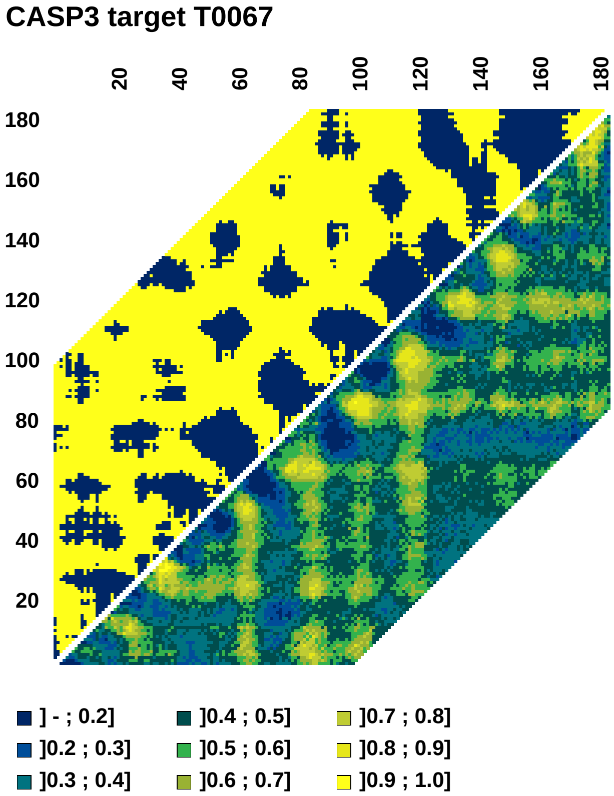

Figure 8: Prediction of distance constraints for the CASP3 target T0067. The upper triangle is the distance constraints of published coordinates, where dark points indicate distances above the thresholds (mean values) and light points distances below the threshold. The lower triangle shows the actual neural network predictions. The scale indicates the predicted probability for sequence separations having distances below the thresholds. Thus light points represent prediction for distances below the thresholds and dark points prediction for distances above the threshold. The numbers along the edges of the plot show the position in the sequence. This figure is available in full color at http://www.cbs.dtu.dk/services/distanceP/.

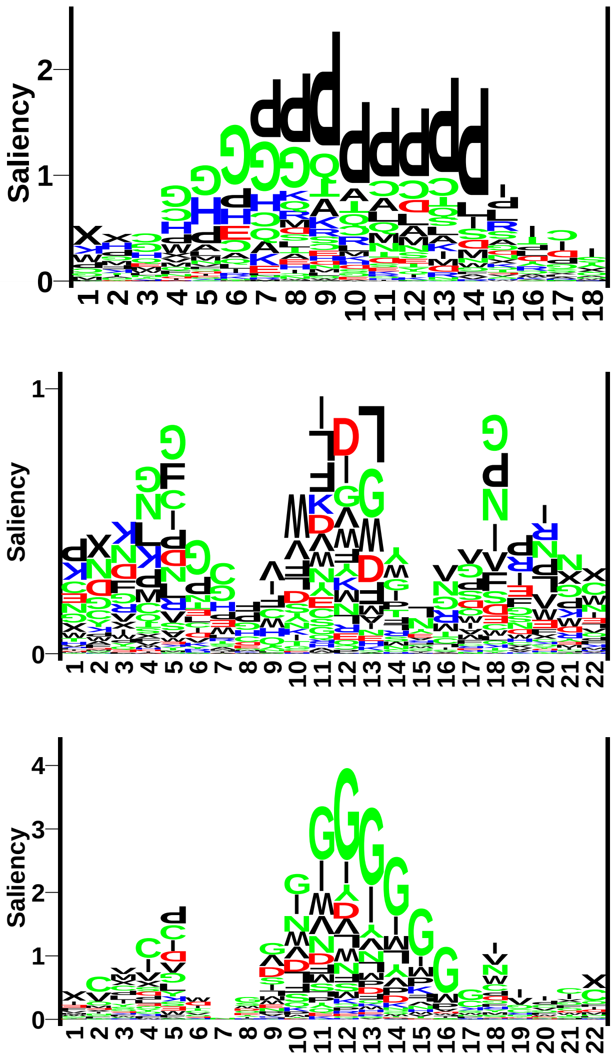

Figure 9: Three saliency-weight logos from the trained neural network at sequence separations 9 and 13. For each position of the input field the saliency of the respective amino acids is displayed. Letters turned upside down contribute negatively to the weighted sum of the network input. The top logo (sequence separation 9) illustrates strong punishment for proline and glycine upstreams. On the middle logo (sequence separation 13), N is turned upside down on the peaks near the edges. On the bottom logo the I's close to the edges contribute positively, and the N in the center peak also contribute positively. The I's in the center peak contributes negatively (they are upside down). This figure is available in full color at http://www.cbs.dtu.dk/services/distanceP/.

Final remarks

We studied prediction of distance constraints in proteins. We found that sequence motifs were related to the distribution of distances. Using the motifs we showed how to construct an optimal neural network predictor which improve an earlier approach. The behavior of the predictor could be completely explained from a statistical analysis of the data, in particular a drop in performance for sequence separation 12 residues was predicted from the statistical analysis. We also correctly predicted separation 3 as having optimal performance. We found that the information content of the logos has the same qualitative behavior as the network performance for increasing sequence separation. Finally, the weight composition of the network was analyzed and we found that the sequence motif appearing was in agreement with the sequence logos.

The perspectives are many: the network performance may be improved further by using three windows for sequence separations between 20 and 30 residues. The performance of the network can be used as inputs to a new collection of networks which can clean up certain types of false predictions. Combining the method with a secondary structure prediction should give a significant performance improvement on the short sequence separations. The relation between information content and network performance might be explained quantitatively through extensive algebraic considerations.

Acknowledgments

This work was supported by The Danish National Research Foundation. OL was supported by The Danish 1991 Pharmacy Foundation.

References

Baldi, P., and Brunak, S. 1998. *Bioinformatics — The machine learning approach*. Cambridge Mass.: MIT Press.

Bernstein, F. C.; Koetzle, T. G.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rogers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The protein data bank: A computer based archival file for macromolecular structures. *J. Mol. Biol.* 122:535–542. (http://www.pdb.bnl.gov/).

Bishop, C. M. 1996. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M. C.; Fredholm, H.; Lautrup, B.; and Petersen, S. B. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.* 261:43–46.

Branden, C., and Tooze, J. 1999. *Introduction to Protein Structure*. New York: Garland Publishing Inc., 2nd edition.

Brunak, S.; Engelbrecht, J.; and Knudsen, S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220:49–65.

Creighton, T. E. 1993. *Proteins*. New York: W. H. Freeman and Company, 2nd edition.

Dayhoff, M. O. ans Schwartz, R. M., and Orcutt, B. C. 1978. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* 5, Suppl. 3:345–352.

Fariselli, P., and Casadio, R. 1999. A neural network based predictor of residue contacts in proteins. *Prot. Eng.* 12:15–21.

Göbel, U.; Sander, C.; Schneider, R.; and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317.

Gorodkin, J.; Hansen, L. K.; Lautrup, B.; and Solla, S. A. 1997. Universal distribution of saliencies for pruning in layered neural networks. *Int. J. Neural Systems* 8:489–498. (http://www.wspc.com.sg/journals/ijns/85_6/gorod.pdf).

Hobohm, U.; Scharf, M.; Schneider, R.; and Sander, C. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Prot. Sci.* 1:409–417.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition and hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86.

Lund, O.; Frimand, K.; Gorodkin, J.; Bohr, H.; Bohr, J.; Hansen, J.; and Brunak, S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Prot. Eng.* 10:1241–1248. (http://www.cbs.dtu.dk/services/CPHmodels).

Maiorov, V. N., and Crippen, G. M. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888.

Mathews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta* 405:442–451.

Mirny, L. A., and Shaknovich, E. I. 1996. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* 264:1164–1179.

Miyazawa, S., and Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: cuasi-chemical approximation. *Macromolecules* 18:534–552.

Myers, E. W., and Miller, W. 1988. Optimal alignments in linear space. *CABIOS* 4:11–7.

Olmea, O., and Valencia, A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design* 2:S25–S32.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* 183:63–98.

Reese, M. G.; Lund, O.; Bohr, J.; Bohr, H.; Hansen, J. E.; and Brunak, S. 1996. Distance distributions in proteins: A six parameter representation. *Prot. Eng.* 9:733–740.

Rost, B., and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–599.

Schneider, T. D., and Stephens, R. M. 1990. Sequence logos: a new way to display concensus sequences. *Nucl. Acids Res.* 18:6097–6100.

Sippl, M. J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–83.

Skolnick, J.; Kolinski, A.; and Ortiz, A. R. 1997. MONSSTER: A method for folding globular proteins with a small number-comment of distance restraints. *J. Mol. Biol.* 265:217–241.

Tanaka, S., and Scheraga, H. A. 1976. Medium- and long range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 9:945–950.

Thomas, D. J.; Casari, C.; and Sander, C. 1996. The prediction of protein contacts from multiple sequence alignments. *Prot. Eng.* 9:941–948.