A New Neural Network Ranker to Evaluate Protein Structure Predictions

Davide Baú and Gianluca Pollastri*

University College Dublin, Technical Report UCD-CSI-2012-03, June 2012

Abstract

Assessing and ranking the quality of a predicted model represents an important and difficult problem in protein three-dimensional structure prediction. To accurately address this problem, we have developed a new machine learning based Model Quality Assessment program for protein structure prediction. The novelty of the approach relies on the integration of informative structural features represented by 7 Gaussian integrals (representing the curvature of the backbone) and other information representing solvent accessibility, hydrogen bonds and predicted geometrical constraints of the model.

Our tests indicate that the ranker can be very effective at discriminating good models from bad ones. A 5-fold cross-validation test on the whole CASP7 dataset yields a correlation of 0.831, which is very good in comparison to the performance of different scoring functions tested on the server models submitted to CASP7. We also perform a test using the CASP5 dataset as training set and the CASP7 dataset as testing set. The correlation obtained in this case is 0.821.

1 Introduction

In protein structure prediction, a large number of models are generated for every single query sequence in order to best search the three-dimensional conformational space. While this procedure improves the probability of producing native-like structures, it requires a robust and efficient method to spot, among all the models generated, the ones closest to the native structure. The ability of selecting the best model plays in fact a crucial role in protein structure prediction [6]. Model Quality Assessment Programs (MQAP) are computer programs developed to rank models generated by structure prediction algorithms. Based on the information used, they can be divided into three categories [23]: *consensus-based methods*, which rely on the similarity to other models, *structure-based methods*, which make use of features calculated from the three-dimensional

*School of Computer Science and Informatics and Complex and Adaptive Systems Laboratory, University College Dublin, Ireland

model, and *evolution-based methods* which take advantage of the similarity between the model and a template.

Consensus based methods rank models by trying to find mutual similarities within large ensembles of them, assuming that recurring structural patterns are more likely to be correct than patterns occurring only in few models [23].
Structure-based methods assess the quality of a model by looking at selected structural features of it, e.g. secondary structure[18, 17], solvent accessibility[15, 16, 11], structural motifs [12, 9], beta shee pairings [2], atomic interactions[14, 1, 22, 10, 24]. The existence of multiple domains[26] and disorder information[21, 25] may also be considered.
Evolution based methods evaluate models by comparing them to the homologous proteins used as templates during the model generation phase.

MQAPs programs can predict the overall quality of a model or local quality. The overall quality is an index that reflects how native-like a protein is as a whole, e.g. its TM-score to the native structure, and rank models according to it. This index does not account for regions predicted incorrectly, which are instead investigated by local quality measures and might be used to drive the reconstruction.

Here we present a new machine learning based Model Quality Assessment Program for protein structure prediction. The novelty of the approach relies on the integration of informative structural features embodied in 7 Gaussian Integrals (representing the curvature of the backbone) and other information representing solvent accessibility, hydrogen bonds and predicted geometrical constraints of the model. Our tests indicate that the ranker can be very effective at discriminating good models from bad ones.

The ranker included 27 features and was used in CASP8 as part of the Quality Measure used to rank our models and as a predictor in the Quality Assessment (QA) category.

2 Approach

The main disadvantage of using simplified representations (e.g Cα-only) for protein structures is the difficulty to derive a meaningful energy model to rank the structures generated. We try to overcome this problem by relying on geometrical constraints to discern native-like protein conformations from unfolded, or incorrectly folded ones. Often the pseudo-energy functions used in reconstruction methods to guide the search phase output numerical scores that do not accurately describe the quality of models, especially when the reconstructor is based on protein features that are predicted with a low accuracy (e.g. when relying on *ab initio* residue contact map predictions). This happens because the pseudo-function tries to encode constraints that are not real, ending up in a model whose three-dimensional structure is far from the native one. To solve this problem and correctly rank the models generated and to (ideally) select the best one, we implemented a new neural network ranker. The idea is to take a snapshot of the final model, by means of a vector of geometrical descriptors,

and to feed it to a neural network that will output an estimation of the model quality. The quality assessment is thus modelled as a regression problem solved by Machine Learning techniques.
A multi-layered feed-forward neural network is trained to map the set of input features, describing geometrical characteristics of the backbone, onto a numerical score between 0 and 1 (sigmoidal output neuron; the higher the score, the better the quality). In the training phase the exact score (target) is known because the native structure is known. Here we choose to adopt as target the TM-score, a scoring function which has no bias with respect to the target protein length and all the residues of the modelled proteins. For training, we use the classical back-propagation algorithm to minimise the sum of squared errors between the predicted output and the correct pseudo-energy value.

We chose a structure-based method over the consensus and evolution based ones because the number of structures generated in our reconstructing pipeline is not large enough to allow clustering of correct patterns over non-correct ones (which underpins consensus-based methods) and because we do not always have a template for comparison. What we need is an overall quality index that reflects how native-like a protein is, (e.g. that evaluates its TM-score) and ranks it according to such index.

3 Methods

A number of features in the input vector come from the one- and two-dimensional features predicted by the suite of predictors included in the pipeline of our method to model protein structures [4] (see Figure 1). The other features are derived from geometrical characteristics of the decoys. In this section a detailed description of the different features used to describe a model will be given (in parenthesis the number of features).

The 27 features used in our ranker as a set of descriptors for protein structures are:

- Contact density (4)[22]
- Distance between the contact map of the structure and a predicted one (5)
- Mutual sums (4)
- Relative position of Secondary Structure Elements(SSEs) (3)
- Half-Sphere Exposure (2)
- Non-local hydrogen bonds and hydrophobic contacts (2)
- Selected Gauss integrals (7) ($I_{(1,2)}$, $I_{(1,2)(3,4)}$, $I_{(1,2)(3,4)(5,6)}$, $I_{(1,2)(3,5)(4,6)}$, $I_{(1,2)(3,6)(4,5)}$, $I_{(1,4)(2,3)(5,6)}$, $I_{(1,6)(2,3)(4,5)}$)

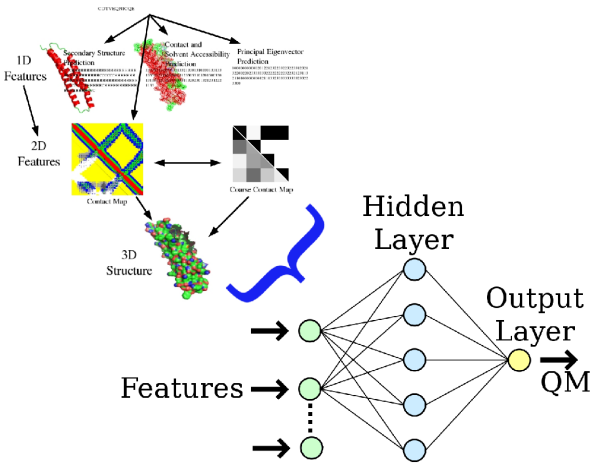

Figure 1: **NNetwork**: quality assessment is modelled as a regression problem solved by Machine Learning techniques. A multi-layered feed-forward neural network is trained to map the set of input features, describing geometrical characteristics of the backbone, onto a Quality Measure score (QM) between 0 and 1 (sigmoidal output neuron; the higher the score, the better the quality). Here we choose to adopt the TM-score, a scoring function which has no bias with respect to the target protein length. For training, we use the classical back-propagation algorithm to minimise the sum of squared errors between the predicted output and the correct pseudo-energy value.

Contact density. The number of an amino acid's neighbouring amino acids defines the degree of solvent exposure for it. Amino acids buried in the core of a globular protein, e.g., will be surrounded by a large number of other amino acids (neighbours), while amino acids staying at the interface with the solvent, will have a smaller number of residues around them. This number, or *contact density*, defines the solvent exposure for each amino acid in the protein [7]. We define the contact density in 4-classes as the Principal Eigenvector (PE) of a protein's residue contact map at 8 Å, multiplied by the principal eigenvalue [22]. The first class represents the least solvent exposed residues and the last class the most exposed ones. Each class, normalised by protein size, represent a feature.

Distance between the contact map of the structure and a predicted one. The 4-class contact map corresponding to a model is compared to the predicted one for that protein. The accuracy for each class, together with the total accuracy is then measured as the number of mismatches between the two maps (i.e. if the maps are identical, all the accuracies are equal to 1). In this way, 5 features are generated.

Mutual sums. These 4 features are the relative contact distribution of the four distance classes (in Å) $(3.8, 5.5]$, $(5.5, 7.5]$, $(7.5, 11]$ and $(11, 13]$. The thresholds have been derived from the mutual distance distribution of the Cα

atoms in our models. For every pair of Cα atoms, their mutual distance is calculated and the corresponding bin counter incremented. The four counters are then normalised by the total number of distances in the range (3.8, 13] Å and returned as features.

Relative position of Secondary Structure Elements (SSEs). The overall three-dimensional arrangement of a protein is defined by how its SSEs fold together. This means that the relative orientation of secondary structure elements is a measure of how correctly or incorrectly a protein folded. It is thus justified to take into consideration how SSEs are placed in the three-dimensional structure and in relation to each other.

For all the SSEs at least four amino acids long, the mutual distance is calculated between the middle amino acid, the previous amino acid and the next one (see Algorithm 1). If one of the three mutual distances is less then 12 Å the cosine of the angle between the involved SSEs is calculated as in Equation 1 and normalised between 0 and 1. For every type of SSE pairing (i.e. helix/helix, strand/strand, others), the total angle normalised by the total number of pairings for that angle is calculated and returned as a feature.

$$v_1 = \frac{\vec{a}_1 \cdot \vec{a}_2}{|\vec{a}_1||\vec{a}_2|} \tag{1}$$

Half-Sphere Exposure. Solvent exposure is an index measuring how exposed to the solvent an amino acid is. In globular proteins the difference in solvent exposure between residues at the surface and residues buried in the core of the protein is relevant. Solvent exposure can thus be interpreted as an index of protein globularity. Since most of the proteins carrying out biochemical reactions are globular, it is straightforward that solvent exposure is an important descriptor for this type of molecules.

A new measure of solvent exposure, the half-sphere exposure (HSE), has been defined as the number of Cα atoms in two half spheres around a residue's Cα atom. The two half spheres are defined by a plane that is perpendicular to the Cα − Cβ (or pseudo Cβ) vector and runs through the residue's Cα atom [7]. The HSE takes into account four different degrees of exposure, distinguishing between exposed, partly exposed, buried and deeply buried residues [7]. For this reason we preferred this measure of solvent exposure over traditional ones. Below is the description of how these two features (upper and lower semisphere) are calculated.

For each C$\alpha_{j,j\neq i} \in$ MODEL (see Algorithm 2), where MODEL is the set of Cα coordinates, and i is the center of the sphere, the vector $\vec{v} =$ Cα_i − Cα_j and the scalar product $\vec{v} \cdot \vec{b}$ are calculated, where $\vec{b}$ is the binormal (vector perpendicular to the plane cutting the sphere centered in i in a upper and a lower emisphere).The total number of residues in the upper emisphere and in the lower emisphere (both normalised by the size of the model) are returned.

Non-local hydrogen bonds and hydrophobic contacts. Hydrogen bonds are vital to maintain proteins in a folded state. The Energy difference between folded and unfolded proteins, in fact, is not high enough to disallow the

(theoretical) existence of unfolded states. A protein's conformation is largely stabilized by weak interactions like hydrogen bonds and hydrophobic interactions [13].

We define non-local hydrogen bonds as non-covalent interactions between the carbonyl group of the residue in position i and the the amide group of the residue in position $i + j$, with $j > i + 4$, and at a distance between 4.1 and 6.5 Å. In this case, if the distance is between 6.5 and 7.5 Å the interaction involved is accounted as an hydrophobic contact.
Each residue is constrained to form no more than two hydrogen bonds. Moreover, the binormal between the atoms involved in the bond must be within (-0.77,0.77) radians. These limitations are required because the strength of hydrogen bonds depends, among other factors, on the distance between the participating atoms, the number of bond atoms that are involved in the bond and their relative orientation (a relative angle of 0 degrees generates a strong bond) [13]. In our models, most of the interactions calculated as described above involve amino acids involved in α-helices. For this reason, and because α-helices are constrained to be right-handed, we enforce amino acids involved in non-covalent bonds to have a positive chirality [8]. We define chirality as the sign of $(\vec{r}_{i,i+1} \times \vec{r}_{i+1,i+2}) \cdot \vec{r}_{i+2,i+3}$.

Gauss integrals. The last 7 features are the most interesting ones and what differentiates this ranker from most of the other MQAPs. The novelty of the approach relies on the integration of informative structural features represented by 7 Gauss integrals. Since it is very hard, if not impossible, to derive a meaningful energy function able to accurately describe the goodness of a model represented only as a sequence of identical beads (the Cα), we decided to derive a quality measure from the geometric description of the shape of a protein. To achieve such goal, the use protein shape descriptors independent of translation and rotation that are able to distinguish among similar, but still different, morphologies is crucial. From this point of view, generalized Gauss integrals are an ideal choice [19]. Generalized Gauss integrals arise from Vassiliev knot invariants [3] and are based on crossings seen in planar projections of curves representing a protein (e.g. the planar projections of the Cα-traces). The crossings have a sign defined by the right-hand rule [20] (see Figure 2). The generalized Gauss integrals used here are based on the first-order Gauss integrals writhe and average crossing number.

The writhe is the total number of positive crossings minus the total number of negative crossings. Its natural definition for a polygonal space curve μ is

$$I_{(1,2)}(\mu) = Wr(\mu) = \sum_{0<i_1<i_2<N} W(i_1, i_2) \tag{2}$$

where $W(i_1, i_2)$ is the probability of seeing the i_1 and i_2^{th} segments cross when averaged over all directions in space multiplied by the sign of the crossing [19].

The average crossing number of a curve is the unsigned average number of crossings seen in the different planar projections of the curve and is defined by

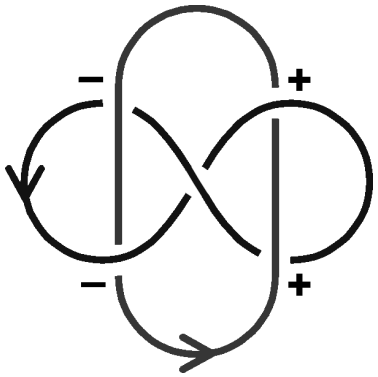

Figure 2: Example of inter-crossing curves.

$$I_{|1,2|}(\mu) = \sum_{0<i_1<i_2<N} |W(i_1, i_2)| \tag{3}$$

All the other second- and third-order Gauss integrals are derived from the writhe and the average crossing number [20].

4 Discussion

5 Dataset

The protein data set used to train and validate the ranker consists of a non redundant set of 258 protein structures (S258) showing no homology to the sequences employed to train the underlying predictive systems. This set includes proteins of moderate size (51 to 200 amino acids) and diverse topology as classified by SCOP (Structural Classification of Proteins database) (all-α, all-β, α/β, $\alpha+\beta$, surface, coiled-coil and small). For each protein in the dataset, we have generated 75 decoys of different quality, so that to have a balanced distribution of TM-score values (see Figure 3). The training set was composed of 19594 decoys and the validation set by 8355.

To test its applicability as an MQAP, the ranker was trained on the server models submitted to CASP5, while the server models submitted to CASP7 were used as test set (see Table 2). On the latter dataset, a k-fold (k=5) cross-validation test was then performed.

After preliminary experiments, training of the multi-layered feed-forward neural network has been carried out using a number of input units corresponding to the number of features.Each unit in one layer has directed connections to the units of the subsequent layer. For training, the classical back-propagation algorithm to minimise the sum of squared errors between the predicted output

Algorithm 1 SSE

Require: SSE, the set of Secondary Structure Elements of at least 4 amino acids.

for all $sse_{i,j} \in SSE$ **do**
  MID_AA($sse_{i,j}$) = $(AA^{last}_{sse_{i,j}} - AA^{first}_{sse_{i,j}})/2$
  $C\alpha^{i,j}_{mid} \leftarrow$ MID_AA($sse_{i,j}$)
  for all $C\alpha^{i,j}_{mid}$ **do**
    $d_1 = distance(C\alpha^{i,j}_{mid})$
    $d_2 = distance(C\alpha^{i-1,j-1}_{mid})$
    $d_3 = distance(C\alpha^{i+1,j+1}_{mid})$
    if d_1 or d_2 or $d_3 < 12$ **then**
      if $sse_{i,j}$ are helices **then**
        H_{count}++
        $H_{rot} \leftarrow$ ANGLE($C\alpha^{i,j}_{mid}$)
      else if $sse_{i,j}$ are strands **then**
        E_{count}++
        $E_{rot} \leftarrow$ ANGLE($C\alpha^{i,j}_{mid}$)
      else
        C_{count}++
        $C_{rot} \leftarrow$ ANGLE($C\alpha^{i,j}_{mid}$)
      end if
    end if
  end for
end for
return (H_{rot}/H_{count}++, E_{rot}/E_{count}++ C_{rot}/C_{count}++)

Algorithm 2 HSE

for all C$\alpha_i \in$ MODEL **do**
  d = DEFINE_SIGN(Cα_i);
  for all C$\alpha_{j,j\neq i} \in$ MODEL **do**
    dist = distance(C$\alpha_{i,j}$);
    if dist $<$ 12 and d $>=$ 0 **then**
      uprel++;
      upside++;
    else if dist $<$ 12 and d $<$ 0 **then**
      drel++;
      downside++;
    end if
    upcount += uprel/size;
    downcount += drel/size;
  end for
end for
return (upcount, downcount)

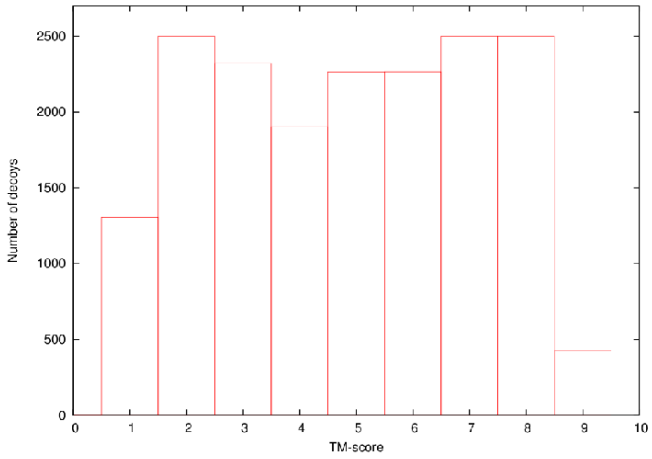

Figure 3: Training set TM-score distribution of the training set.

(modelled via a sigmoidal output neuron) and the correct pseudo-energy value was used. Since we are training the network to map the input vector onto a global quality measure index (i.e. the TM-score), the output unit is one. Both the momentum and the learning rate were set to 10^{-3} for all the experiments, while the number of hidden units and epochs vary for the different experiments and is reported in the tables below along with the correlation between the true TM-score and the predicted one calculated on the training and test sets. Here the correlation measures how accurately the predicted TM-score approaches the true one, as is defined as:

$$corr(x, y) = \frac{Cov(x, y)}{\sigma(x)\sigma(y)} \tag{4}$$

where x is the true TM-score, y the predicted one, Cov is the covariance of x and y (i.e. the measure of how much x and y change together), and σ the standard deviation. A correlation of 1 means that the predicted and the true TM-score value are identical across the set considered.

Correlations calculated for the the ranker on the S258 dataset (for both training and validation sets) are graphically reported in Figure 4. An ideal ranker with correlation 1 would produce a plot with all the points lying on the diagonal (i.e. same value for predicted and true TM-score). Thus the more "diagonal-like" a plot is, the more native-like the predicted TM-scores are. Keeping this consideration in mind, the good performance of our ranker is easily noticed by looking at these plots.

The ranker was trained on the server models submitted to CASP5 to test its applicability as an MQAP (see Table 2).

***Corr* on training set**	***Corr* on validation set**	**Hidden units**	**Epochs**	**Learning rate**
0.945	0.930	5	30000	10^{-3}

Table 1: Correlation values for the ranker trained and tested on two different subsets of models generated from the S258 dataset.

***Corr* on training set**	***Corr* on test set**	**Hidden units**	**Epochs**	**Learning rate**
0.926	0.821	3	10000	10^{-3}

Table 2: Correlation values for the ranker trained on the server models submitted to CASP5. The ranker was then tested tested on the server models submitted to CASP7.

Method	**Correlation**
Our Ranker	0.850
QMEAN5	-0.720
Modcheck	0.640
SSE PSIRED	-0.650

Table 3: Comparison between correlation values for different scoring function in predicting the quality of server models submitted to CASP7

We then performed a k-fold (k=5) cross-validation test on the server models submitted to CASP7. The correlation on this set was 0.831, which is very good in comparison to the performance of different scoring functions tested on the server models submitted to CASP7 [5] (see Table 3).

The average TM-score for CASP7 models ranked as first (i.e. for each target, the model to which the ranker assigned the best score) by the ranker was 0.471.

In table 4 the ranker correlations on CASP7 dataset are reported together with the correlation obtained training the network on models generated from our pipeline for CASP7 targets. The latter is part of our current reconstructing pipeline and is participating in the CASP8 MQAP category.

6 Conclusion

We have presented a new machine learning based Model Quality Assessment Program for protein structure prediction. The novelty of the approach relies on the integration of informative structural features represented by 7 Gaussian Integrals (representing the curvature of the backbone) and other information representing solvent accessibility, hydrogen bonds and predicted geometrical constraints of the model. Our tests indicate that the ranker can be very effective

k-fold test *Corr*	**Hidden units**	**Epochs**	**Learning rate**
0.831	3	30000	10^{-3}
Our CASP7 models			
0.853	5	50000	10^{-3}

Table 4: Correlation value for the k-fold (k=5) cross-validation test on server models submitted to CASP7. The last line in the Table is the correlation obtained training the network on models generated from our pipeline for CASP7 targets. The latter is part of our current reconstructing pipeline and is participating in CASP8 MQAP category

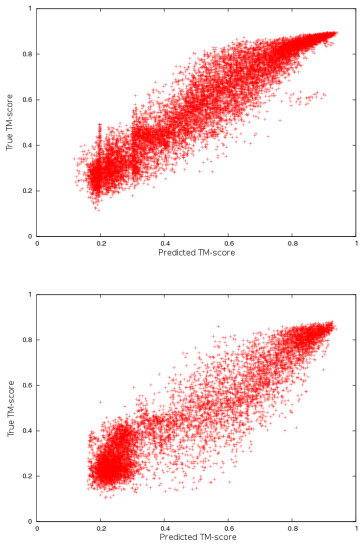

Figure 4: Predicted TM-score value versus true one for the 27 feature ranker using two different subset of the S258 dataset as training set (top) and as validation set (bottom).

at discriminating good models from bad ones.

The protein data set used to train and validate the ranker consisted of a non redundant set of 258 protein structures (S258) showing no homology to the sequences employed to train the underlying predictive systems. The correlation on this set was 0.945 (training set) and 0.930 (validation set).

The ranker was then tested on the server models submitted to CASP7. We first used the server models submitted to CASP5 as training set and those submitted to CASP7 as test set. The correlation on this set was 0.821. We then

performed a k-fold (k=5) cross-validation test on the server models submitted to CASP7. The correlation on this set was 0.831, which is very good in comparison to the performance of different scoring functions tested on the server models submitted to CASP7 (see Table 3).

The average TM-score for CASP7 models ranked as first (i.e. for each target, the model to which the ranker assigned the best score) by the ranker was 0.471.

Although this is a preliminary study, we plan on revisiting and expanding this research on newer, larger sets, including the CASP8 and CASP9 datasets.

Funding

This work is supported by Science Foundation Ireland grants 04/BR/CS0353, 05/RFP/CMS0029 and 10/RFP/GEN2749.

References

[1] P. Baldi and G. Pollastri. The principled design of large-scale recursive neural network architectures–dag-rnns and the protein structure prediction problem. *The Journal of Machine Learning Research*, 4:575–602, 2003.

[2] P. Baldi, G. Pollastri, C. A. F. Andersen, and S. Brunak. Matching protein β-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00)*, volume 8, pages 25–36, La Jolla, CA, 2000. AAAI Press.

[3] D. Bar-Natan. On the vassiliev knot invariants. *Topology*, 34:423–472, 1995.

[4] D. Baú, A.J.M. Martin, C. Mooney, A. Vullo, I. Walsh, and G. Pollastri. Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. *BMC Bioinformatics*, 7(402), 2006.

[5] P. Benkert, S.C.E. Tosatto, and D. Schomburg. Qmean: A comprehensive scoring function for model quality assesment. *PROTEINS: Structure, Function, and Bioinformatics*, 2007.

[6] D. Fischer. Servers for protein structure prediction. *Current opinion in structural biology*, 16:178–182, 2006.

[7] T. Hamelryck. An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *PROTEINS: Structure, Function, and Bioinformatics*, 59:38–48, 2005.

[8] T.X. Hoang, A. Trovato, F. Seno, J.R. Banavar, and A. Maritan. Geometry and symmetry presculpt the free-energy landscape of proteins. *PNAS*, 101(21):7960–7964, 2004.

[9] Q. Le, G. Pollastri, and P. Koehl. Structural alphabets for protein structure classification: a comparison study. *Journal of molecular biology*, 387(2):431–450, 2009.

[10] A.J.M. Martin, D. Baù, A. Vullo, I. Walsh, and G. Pollastri. Long-range information and physicality constraints improve predicted protein contact maps. *Journal of Bioinformatics and Computational Biology*, 6(5):1001, 2008.

[11] C. Mooney and Pollastri. Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, 77(1):181–90, 2009.

[12] C. Mooney, A. Vullo, and G. Pollastri. Protein structural motif prediction in multidimensional ϕ-ψ space leads to improved secondary structure prediction. *Journal of Computational Biology*, 13(8):1489–1502, 2006.

[13] D.L. Nelson and M.M. Cox. *Lehninger principles of biochemistry*. Worth Publishers, third edition, 2000.

[14] G. Pollastri and P. Baldi. Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18(suppl 1):S62–S70, 2002.

[15] G. Pollastri and P. Baldi. Prediction of contact maps by recurrent neural network architectures and hidden context propagation from all four cardinal corners. *Bioinformatics*, 18, Suppl.1:S62–S70, 2002.

[16] G. Pollastri, A. J. M. Martin, C. Mooney, and A. Vullo. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, 8(201), 2007.

[17] G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–20, 2005.

[18] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47:228–235, 2002.

[19] P. Rogen. Evaluating protein structure descriptors and tuning gauss integrals based descriptors. *Journal of Physics: Condensed Matter*, 17:1523–1538, 2005.

[20] P. Rogen and H. Bohr. A new family of protein shape descriptors. *Mathematical Biosciences*, 182:167–181, 2003.

[21] A. Vullo, O. Bortolami, G. Pollastri, and S. Tosatto. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, 34:W164–168, 2006. Web Server Issue.

[22] A. Vullo, I. Walsh, and G. Pollastri. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, 7(180), 2006.

[23] B. Wallner and A. Elofsson. Prediction of global and local model quality in casp7 using pcons and proq. *Proteins*, 69(8):184–193, 2007.

[24] I. Walsh, D. Baù, A.J.M. Martin, C. Mooney, A. Vullo, and G. Pollastri. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC structural biology*, 9(1):5, 2009.

[25] I. Walsh, A.J.M. Martin, T. Di Domenico, A. Vullo, G. Pollastri, and S.C.E. Tosatto. Cspritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic acids research*, 39(suppl 2):W190–W196, 2011.

[26] I. Walsh, A.J.M. Martin, C. Mooney, E. Rubagotti, A. Vullo, and G. Pollastri. Ab initio and homology based prediction of protein domains by recursive neural networks. *BMC bioinformatics*, 10(1):195, 2009.