Long-range information and physicality constraints improve predicted protein contact maps

Alberto J. M. Martin

School of Computer Science and Informatics and
Complex and Adaptive Systems Labs
University College Dublin
Belfield, Dublin 4, Ireland

albertoj@ucd.ie

Alessandro Vullo

School of Computer Science and Informatics and
Complex and Adaptive Systems Labs
University College Dublin
Belfield, Dublin 4, Ireland

alessandro.vullo@ucd.ie

Gianluca Pollastri

School of Computer Science and Informatics and
Complex and Adaptive Systems Labs
University College Dublin
Belfield, Dublin 4, Ireland

gianluca.pollastri@ucd.ie

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Protein topology representations such as residue contact maps are an important intermediate step towards *ab initio* prediction of protein structure, but the problem of predicting reliable contact maps is far from solved. One of the main pitfalls of existing contact map predictors is that they generally predict unphysical maps, i.e. maps that cannot be embedded into three-dimensional structures, or, at best, violate a number of basic constraints observed in real protein structures, such as the maximum number of contacts for a residue.

Here we focus on the problem of learning to predict more "physical" contact maps. We do so by first predicting contact maps through a traditional system (XXStout), then filtering these maps by an ensemble of artificial neural networks. The filter is provided as input not only the bare predicted map, but also a number of global, or long-range features extracted from it.

In a rigorous cross-validation test, we show that the filter greatly improves the predicted maps it is input. CASP7 results, on which we report here, corroborate this finding. Importantly, since the approach we present here is fully modular, it may be beneficial to any other *ab initio* contact map predictor.

Keywords: protein structure prediction; contact maps; neural networks.

1. Introduction

Protein topology representations such as residue contact maps are an important intermediate step towards *ab initio* prediction of protein structure. In fact, it has been argued that, for an adequate definition of contact, a contact map is roughly equivalent to the structure itself [1].

Not surprisingly, the problem of accurately predicting residue contact maps from primary sequences is still largely unsolved. Among the reasons for this are the unbalanced nature of the problem (with far fewer examples of contacts than non-contacts, although this depends on the definition of contact) and, especially, the formidable challenge of capturing long-range interactions in the maps.

In order to mitigate the intrinsic difficulty of mapping one-dimensional input sequences into two-dimensional outputs, in virtually all existing predictive systems protein contact maps are inferred by modelling a set of independent tasks, each task being the prediction of whether two residues are in contact[2,3,4,5,6].

Although the resulting problem turns out to be simplified, this approach does not take into account the global nature of the task. For this reason, joining the predictions for all residue pairs generally corresponds to a map that is not physically realisable. This means that the structure derived from this contact map violates basic constraints observed in real protein structures (e.g. the maximum number of contacts per amino acid) or, in the worst case, that the residues cannot be embedded into the three dimensional (3D) euclidean space.

Here we focus on the problem of learning to predict physically realisable residue contact maps. How to directly incorporate rules encoding protein structural principles into a learning framework remains an open issue. In this article we present the results of a straightforward approach: we predict contact maps by a traditional algorithm[6] and then train an ensemble of second stage artificial neural networks (NN-Filter) to filter these maps, with the aim of correcting some of the errors they contain and of increasing their physicality. This NN-Filter is explicitly provided as input not only the bare predicted map, but also information about its physical realisability and compressed long-range contact information extracted from it.

Predicted contact maps are obtained from XXStout[6] that uses as input evolutionary information from alignments of multiple homologous sequences, predicted secondary structure[7], solvent accessibility and contact density [6,8]. XXStout's predictions contain errors clearly due to to the fact that each pair of amino acids is modelled as a separate instance of the problem. For instance patterns of contacts between secondary structure elements, especially for large sequence separations, are shaped differently from those found in real contact maps (they tend to come in squares - see figures XXX for examples). Moreover amino acids are often predicted to be in contact with too many or with too few other amino acids, yielding unphysical overall maps.

We implement the NN-Filter stage by training two different ensembles of artificial neural networks to process different regions of the contact map: the region close

to the diagonal of the map, corresponding to small sequence separations; the region further away from the diagonal. We do so because the rules governing contact probability are likely to be different for positions near the main diagonal (mainly made by backbone atoms, reflecting secondary structure contacts) than for positions away from it (where contacts mainly occur between the side chains of amino acids placed in different secondary structure elements).

The NN-Filter stage is input a combination of local information about the predicted map -such as estimated probabilities of contact for pairs of residues, or secondary structure predictions for individual residues[7]- and global information, such as total number of contacts predicted for a residue or pair of residues, residue contact order, etc.

We perform experiments to validate the ability of NN-Filter to recover correct contact information and to model maps with higher chance of physical realisability. To measure this, beside standard performance measures like Precision, Recall and their harmonic mean F_1, we compare a number of distributions in real, predicted and filtered contact maps. The distributions considered are that of the number of contacts per residue, of the number of contacts with both residues in each pair in contact, and of the contact order per residue.

In a rigorous cross-validation experiment we show that filtered maps are a significant improvement over first stage predictions. We submitted both XXStout and NN-Fiter predictions to the CASP7 competition. Here we report our CASP7 results, which confirm that NN-Filter greatly improves on "raw" XXstout predictions.

Since the NN-Filter stage is entirely modular, we believe that our approach is likely to be beneficial to any contact map predictor, provided that the filter is trained on its error distribution.

2. Methods

2.1. *Data*

2.1.1. *Datasets*

The training dataset used in this study was extracted from the December 2003 25% PDBSELECT list [9]. First, sequences longer than 200 amino acids are excluded, leaving a total of 1601 sequences with 163376 residues (set $PDB1601$). One every five sequences is then included into the validation set, which we use to select the hyperparameters of the systems, such as the architectural details of the neural networks, the number of training epochs to be performed, etc. The validation set contains 321 sequences and 32725 residues and we will refer to it as $VS321$. The remaining sequences (1280 in total - we will refer to the overall set as $TRS1280$) are adopted as a training set.

We create a further set ($TES481$) to assess the quality of filtered maps vs "raw" maps in an unbiased manner. From the July 2005 25% PDBSELECT list we remove all those sequences with sequence identity greater than 25% with any of the

sequences found in $VS321$, $TRS1280$ and $TES258$. Of the remaining sequences we remove those that have amino acids with no 3D coordinates, are longer than 200 residues. This set contains 481 sequences (49159 amino acids).

Finally, we adopt a further set of 258 proteins (33667 residues - $TES258$) of maximal length 200 to test if NN-Filter maps yield better 3D reconstructions than "raw" XXStout ones. The set, also described in [11,12], is composed of sequences with no sequence identity greater than 25% among them, and no sequence identity greater than 25% against any sequence in the 2003 25% PDBSELECT list, hence no sequence identity greater than 25% against any sequence in $VS321$ or $TRS1280$.

There are two main reasons for using only chains up to 200 amino acids in length: to restrict our analysis mostly to proteins containing single domains; because XXstout[6], which is the source of "raw" maps, is trained on proteins of 200 residues at most.

It is worth noting that XXStout is trained on the $PDB1601$ set, which includes both training and validation sets for NN-Filter (respectively $TRS1280$ and $VS321$). For this reason, any improvement by NN-Filter over XXStout cannot be attributed to differences in training set size or coverage.

2.1.2. *Multiple sequence alignments*

For all proteins in all sets we created multiple sequence alignments from NR database as available on March 2004, redundancy reduced at a 98% threshold (1.05 million sequences). The alignments are generated by three runs of PSI-BLAST [13] with parameters b = 3000, e = 10^{-3} and h = 10^{-10}.

2.1.3. *Contact Maps*

The contact map of a protein with N amino acids is a symmetric $N \times N$ matrix C, with elements C_{ij} defined as:

$$C_{ij} = \begin{cases} 1 & \text{if amino acid i and j are in contact} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We define two amino acids as being in contact if the euclidean distance between their C_α is less than a given threshold. Here we use 12Å and 8Å thresholds.

Predicted contacts are obtained from contact map prediction system XXStout [6] that uses as inputs evolutionary information from multiple sequence alignments, predicted secondary structure, solvent accessibility and contact density. Predicted contact maps used in this work are at 12Å and 8Å thresholds.

Positions beside the diagonal (the diagonal being all C_{ij} for which $i = j$) likely follow different rules than those away from it, those contacts close to the diagonal reflecting secondary structure/local interactions, with those further away reflecting long range interactions mainly between different secondary structure elements (see [14]). For this reason we train two different sets of neural networks, one to filter map positions beside the diagonal and one for positions away from it. The cutoff between

"beside" and "away" contacts is chosen as follows: we check sequence separations for all contacts between amino acids that are part of the same secondary structure element in all the proteins belonging to December 2003 25% PDBSELECT list; we increase the highest margin found (10 for 12Å and 6 for 8Å) by 2 to allow for border effects, yielding cutoffs of 12 and 8 for 12Å and 8Å respectively. As the C_α of an amino acid is always closer than 8Å to the C_α of the neighbouring ones in the sequence, map postions just beside diagonal are set to contact by default.

2.2. *Artificial Neural Networks*

We use fully connected, feed forward neural networks trained via the backpropagation algorithm by minimising the cross-entropy error between the output and target probability distributions. We use a hybrid between batch and online training, in which network weights are updated more than once during a training epoch. Training examples are randomly shuffled before each training epoch. We set the number of hidden neurons to 20 after preliminary experiments and do not change it thereafter. Training is always stopped after 2000 epochs.

Due to the large number of examples and the long time needed to train a single network on all training instances, the training set is further divided into 20 sets (away from diagonal) and 5 sets (beside diagonal), each containing approximately the same number of examples. A different network is trained on each of these subsets and then the networks are joint in an ensemble of 20 for map positions away from the diagonal, 5 for map positions beside the diagonal.

2.3. *Input representation*

2.3.1. *Local inputs*

The main input to the filtering neural networks is the "raw" map predicted by XXstout. We input a whole patch of 11x11 contact predictions. Thus a network that predicts the probability of contact between residues in positions i and j along the sequence is shown contact predictions (in the form of predicted probabilities of contact) for all pairs of amino acids (k, l) with $i-5 \leq k \leq i+5$ and $j-5 \leq l \leq j+5$, for a total of 121 inputs. We also input Porter[7] secondary structure predictions for all 22 residues in the windows above, in the form of predicted probabilities of being in one of the 3 standard secondary structure classes (Helix, Strand, Coil). It should be noted that Porter is trained on the December 2003 25% PDBSELECT list, no sequence of which shows more than 25% sequence identity to any sequences in the sets we use for testing here ($TES258$ and $TES481$), hence the use of Porter does not violate the separation between training, validation and test sets. Sequence and evolutionary information are also provided to the network in the form of: identities of residues in positions $i-1$, i, $i+1$ and $j-1$, j, $j+1$; frequency profiles for residues in positions $i-1$, i, $i+1$ and $j-1$, j, $j+1$, extracted from multiple sequences alignments.

2.3.2. *Global inputs*

The groups of inputs above provide local information about the map. We also adopt a number of input features that contain global information about a map:

- Binary 8Å and 12Å map values for (i, j) - probability maps predicted by XXStout are transformed into binary maps by summing all the probabilities of contact for pairs of amino acids at the same sequence separation ($|i-j|$), then classifying the L top-scoring pairs at the given separation as contacts, where L is the integer part of the sum of probabilities of contact. Notice that because of its definition, each element of these maps implicitly contains information about all those elements on the map with sequence separation $|i - j|$. This definition of contact is also used to determine the following features.
- Number of contacts per amino acid (NC) of amino acids i and j.
- Number of amino acids in contact with both i and j (NCIJ).
- Residue Contact Order (CO): for each amino acid i its CO (CO_i) is computed using equation 2, where d_{ij} is the (euclidean) distance between amino acids i and j, t is the contact threshold, L is the protein length and NC_i is the total number of contacts for i.

$$CO_i = \frac{\sum_{\forall j:d_{ij}<t} |i-j|}{L \cdot NC_i} \tag{2}$$

- Sequence separation between amino acids i and j: $|i - j|$.
- Protein length.

2.4. *Quality measures*

2.4.1. *Standard measures*

Standard methods used to measure predicted map quality take into account correctly or incorrectly classified map positions (true positives, TP; false positives, FP; true negatives, TN; and false negatives, FN), that are used to compute Recall (R, equation 3), Precision (P, equation 4) and their harmonic average, F1 (equation 5), as follows:

$$R = \frac{TP}{(TP + FN)} \tag{3}$$

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

Both XXstout and NN-Filter output, for each pair of residues (i, j), a number between 0 and 1 representing the estimated probability of contact, given the training examples observed. To match these values into "hard" predictions, a threshold needs to be established above which a network output considered a contact and below

which it is not. A ROC curve is obtained by testing precision P and recall R of a predictor for a sample of threshold values and plotting R against $1-P$. We present ROC curves for all systems tested, and compute performances at breakeven points, i.e. those points on the ROC curve at which $R = P$. A further measure we report is the total number of contacts predicted and compare it in filtered and "raw" maps vs. native maps.

2.4.2. *Physicality measures*

Maps physicality (if the map represents a truly realisable 3D structure) is not taken into account by the quality measures above. To measure this, we run Kolmogorov-Smirnov tests on the distributions of a number of features related to physicality, against the distribution in native maps. We run the tests on filtered maps vs. native maps, and "raw" maps vs. native to test wheter filtering improves the overall physicality of the maps. Features whose distribution we compare are:

- Number of contacts per amino acid (NC)
- Number of amino acids in contact with 2 amino acids already in contact (NCIJ)
- Residue Contact Order (CO)

3. Results

3.1. *ROC curves*

In figures 1, 2, 3 and 4, we present ROC curves on the $TES481$ set for XXstout and NN-Filter predictions, both for near-diagonal and long-range contacts, and for both contact definitions adopted (8Å and 12Å). In all cases the ROC curve of the filter is above that of XXstout, indicating an improvement.

3.2. *Standard quality measures*

Based on ROC curves on the validation set $VS321$ (not shown) we determine breakeven points for all networks, then compare performances of all systems at breakeven. Breakeven points for 8Å maps are 0.34 for the region beside the diagonal, and 0.14 for long-range contacts. For 12Å maps breakeven points are, respectively, 0.42 and 0.08.

We measure all performances on the $TES481$ data set. The $F1$ measure for NN-Filter is an improvement over XXstout's both for 12Å and 8Å maps, and for all bands of sequence separation considered. For long range contacts (sequence separation of 24 or greater) $F1$ grows from 0.286 for XXstout to 0.316 for NN-Filter for 12Å maps and from 0.097 to 0.168 for 8Å maps.

In all cases the improvements come from a small reduction in precision P, and a much greater increase in recall R. In the case of contacts between residues with a sequence separation of 24 or greater and 8Å maps P decreases only from 23%

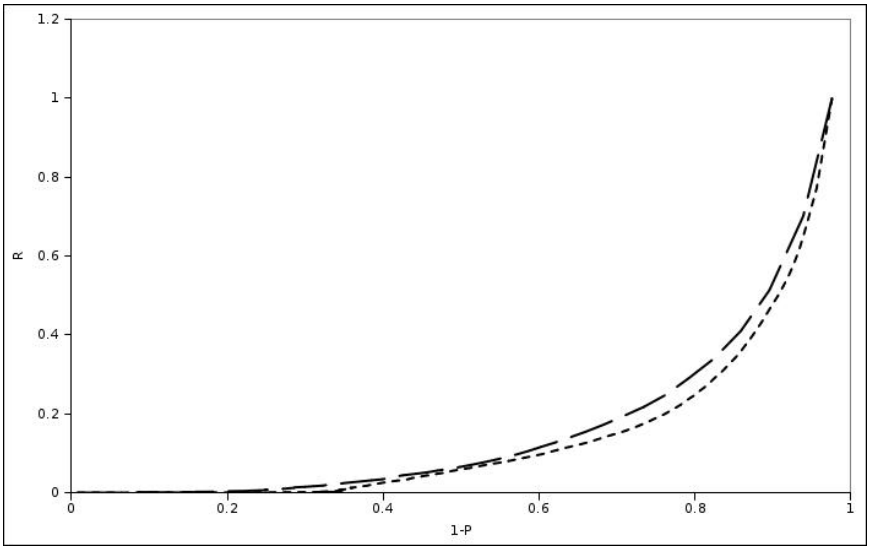

Fig. 1. ROC Curve for 8Å maps, positions away from the diagonal ($|i-j|>8$). Long dashes=NN-Fiter. Short-dashes=XXStout.

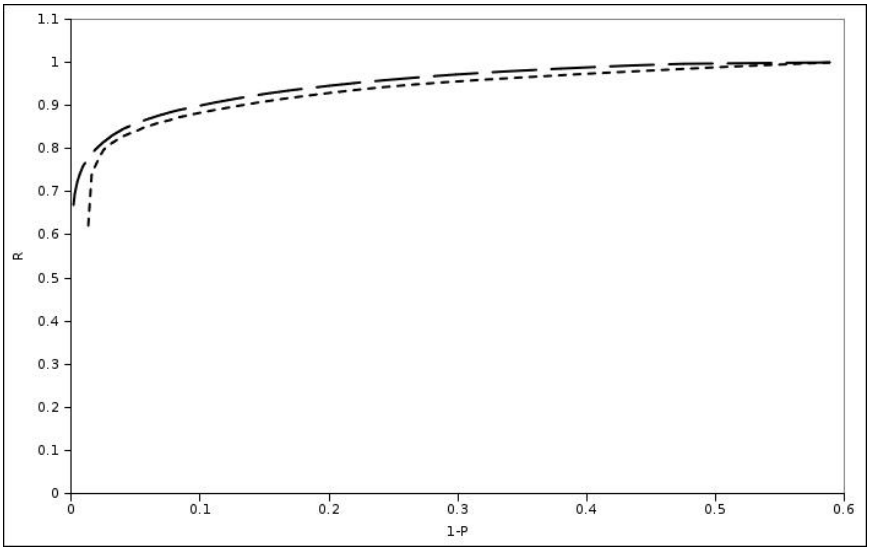

Fig. 2. ROC Curve for 8Å maps, positions beside the diagonal ($|i-j|\leq 8$). Long dashes=NN-Fiter. Short-dashes=XXStout.

to 21.9%, while R more than doubles from 5.7% to 13.7%. Thus NN-Filter is able to guess more than twice as many long range contacts than XXstout, with only a minor loss of accuracy.

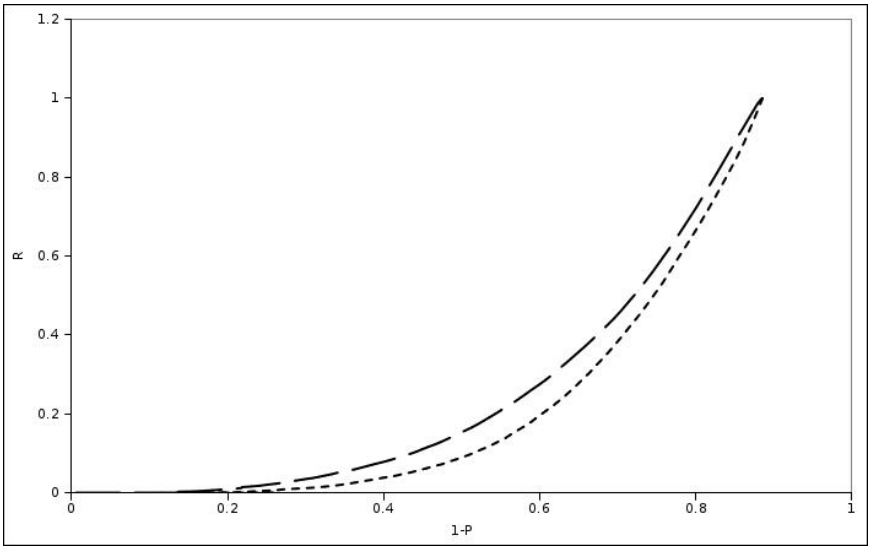

Fig. 3. ROC Curve for 12Å maps, positions away from the diagonal ($|i-j| > 12$). Long dashes=NN-Fiter. Short-dashes=XXStout.

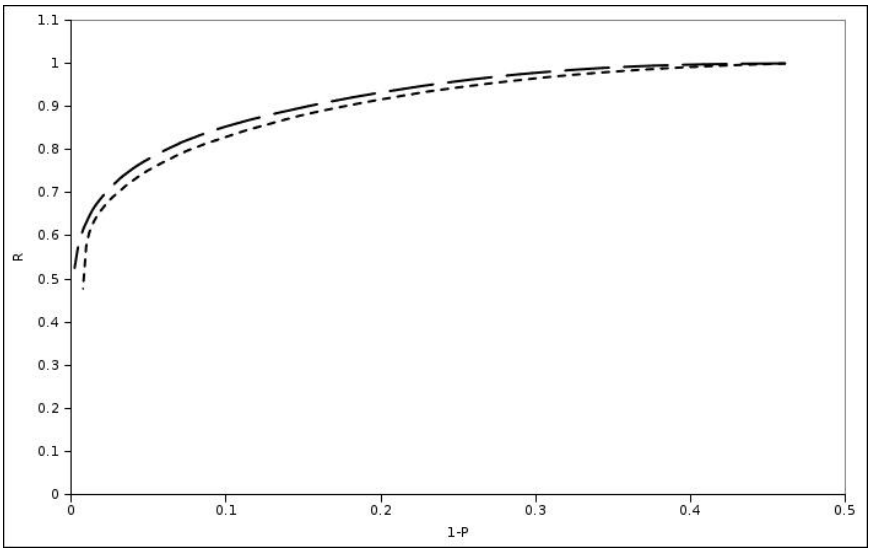

Fig. 4. ROC Curve for 12Å maps, positions away from the diagonal ($|i-j| \leq 12$). Long dashes=NN-Fiter. Short-dashes=XXStout.

3.3. *Physicality measures*

Figures 5, 6, 7, 8, 9, 10 show the distributions of the various physicality measures (NC, NCIJ and CO) for both 8Å and 12Å maps.

	XXStout			NN-Filter		
	$F1$	P	R	$F1$	P	R
$\|i-j\| \geq 6$	0.396	0.386	0.407	0.411	0.336	0.527
$\|i-j\| \geq 12$	0.323	0.313	0.334	0.353	0.278	0.484
$\|i-j\| \geq 24$	0.286	0.275	0.298	0.316	0.254	0.417
$\|i-j\| > 12$	0.319	0.309	0.330	0.352	0.275	0.488
$\|i-j\| \leq 12$	0.858	0.857	0.858	0.876	0.875	0.876

Table 1. Performances for XXstout and NN-Filter on the $TES481$ data set for different sequence separations. 12Å maps.

	XXStout			NN-Filter		
	$F1$	P	R	$F1$	P	R
$\|i-j\| \geq 6$	0.213	0.354	0.153	0.271	0.277	0.265
$\|i-j\| \geq 12$	0.159	0.315	0.107	0.235	0.247	0.224
$\|i-j\| \geq 24$	0.097	0.230	0.057	0.168	0.219	0.137
$\|i-j\| > 8$	0.186	0.335	0.129	0.254	0.256	0.252
$\|i-j\| \leq 8$	0.889	0.888	0.890	0.899	0.891	0.906

Table 2. Performances for XXstout and NN-Filter on the $TES481$ data set for different sequence separations. 8Å maps.

	NC	NCIJ	CO
XXStout	0.3315	0.2528	0.1849
NN-Filter	0.3233	0.3111	0.0933

Table 3. Comparison of distributions of XXStout vs. native maps and NN-Filter vs. native maps, 12Å maps. Kolmogorov-Smirnov test.

To gauge the similarity between distributions, we ran Kolmogorov-Smirnov tests for XXStout vs. Native and NN-Filter vs. Native for NC, NCIJ and CO. The results, reported in Tables 3 and 4, suggest that NN-Filter is closer to the native distribution for the case of Contact Order, slightly further for NCIJ and NC at 8Å, and roughly the same for the other quantities.

3.3.1. *Total number of contacts*

The total number of contacts predicted by XXStout and NN-Filter, vs. the number of contacts in the native maps, are reported in Table 5. While for 8Å maps NN-Filter predicts a number of contacts that is much closer to the correct one than XXStout's, this is actually reversed for 12Å maps, where NN-Filter over-predicts

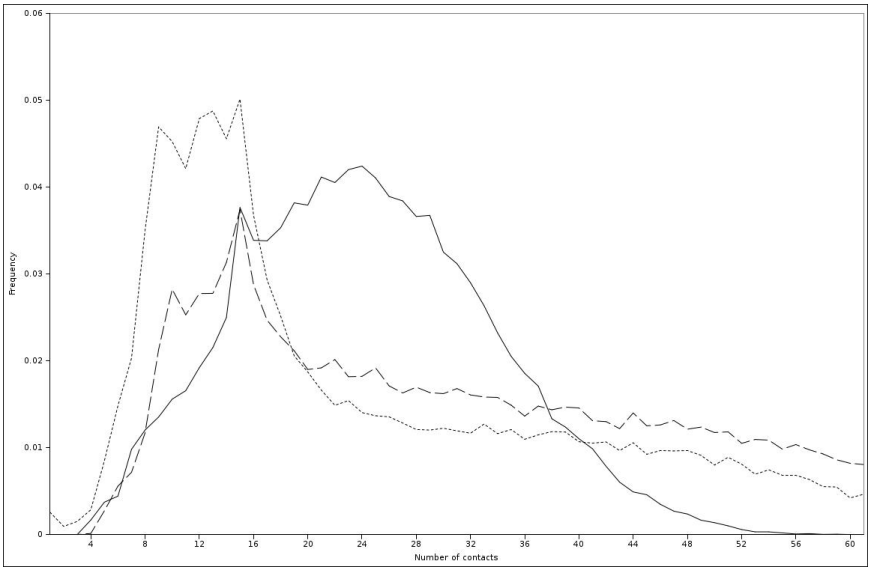

Fig. 5. Number of contacts per amino acid (NC) in the 481 dataset, 12Å maps, before and after the filter. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

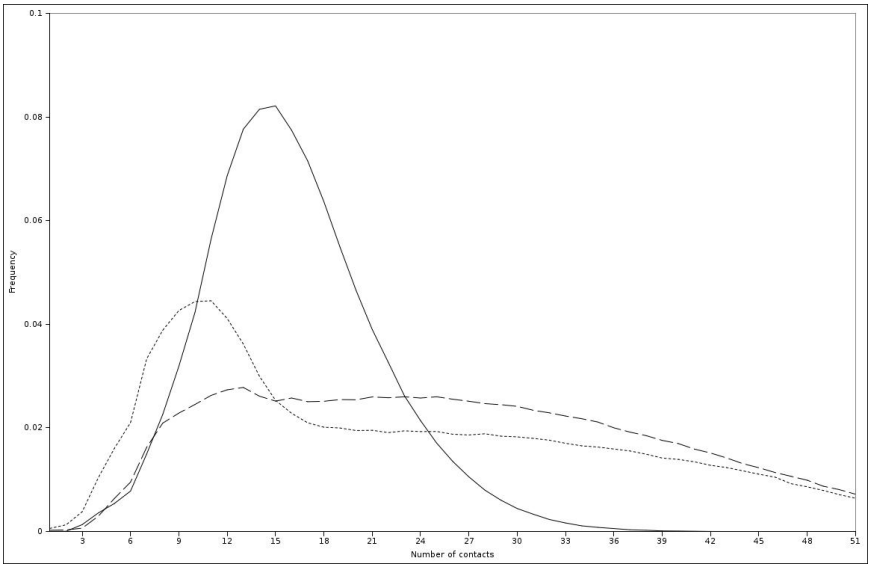

Fig. 6. Number of amino acids Vs Number of amino acids in contact with both AAi and AAj if they are in contact (NCIJ) in the 481 dataset, 12Å maps, before and after the filter. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

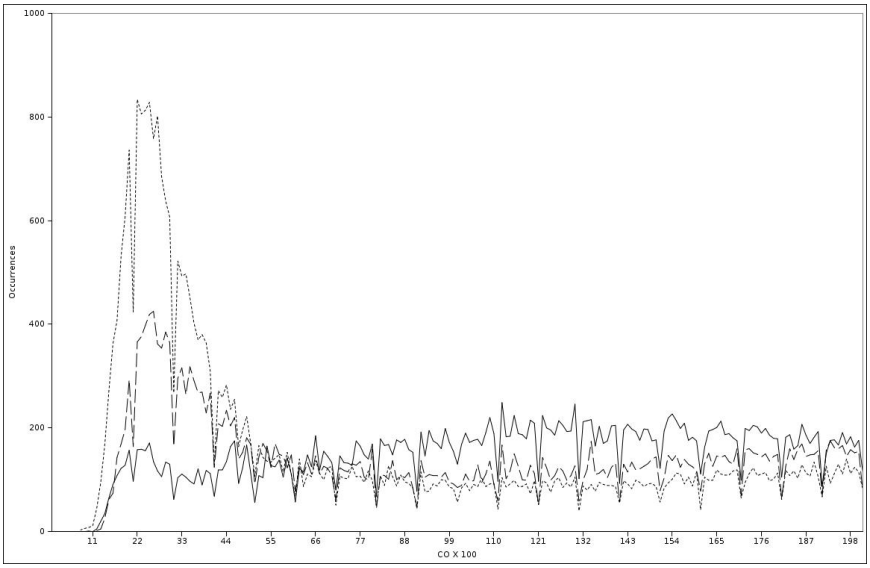

Fig. 7. Comparison of Residue CO distributions before and after filtering with true distributions for 12Å maps. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

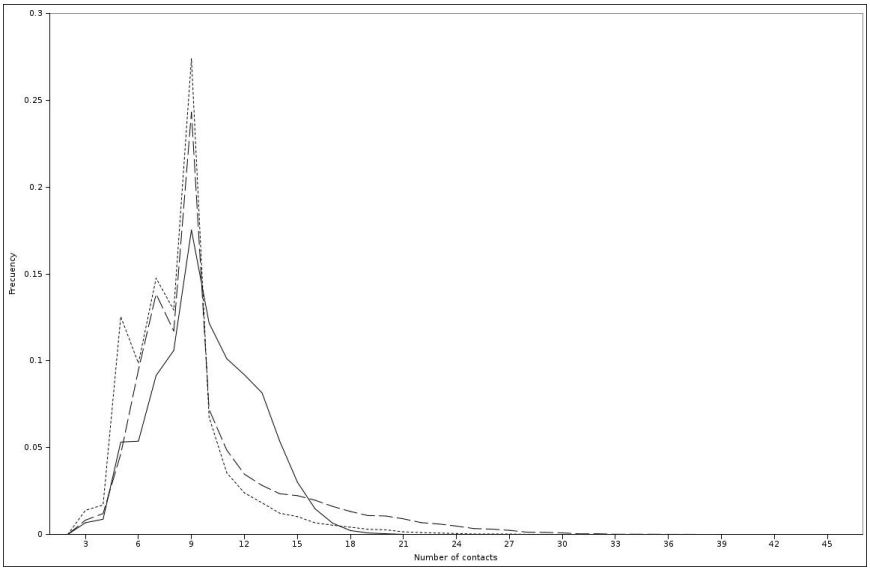

Fig. 8. Number of contacts per amino acid (NC) in the 481 dataset, 8Å maps, before and after the filter. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

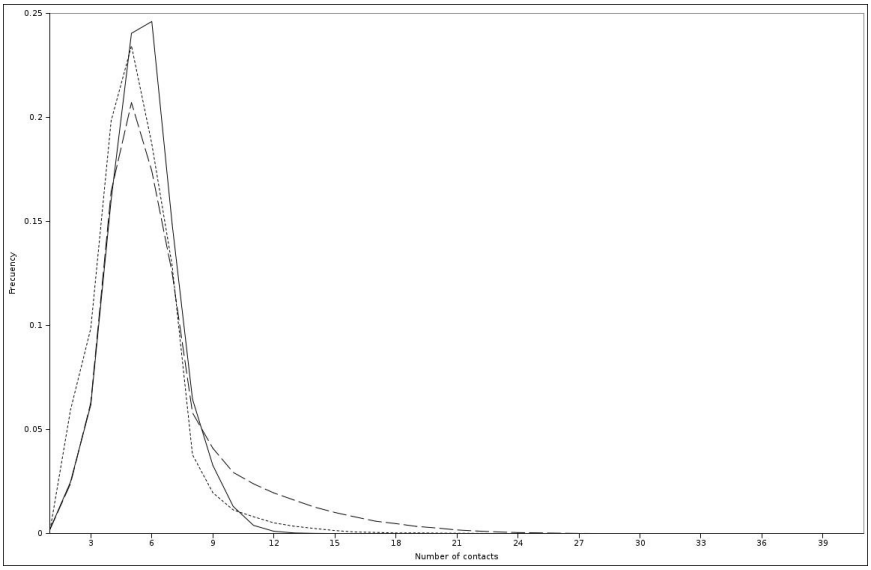

Fig. 9. Number of amino acids Vs Number of amino acids in contact with both AAi and AAj if they are in contact (NCIJ) in the 481 dataset, 8Å maps, before and after the filter. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

	NC	NCIJ	CO
XXStout	0.2905	0.2292	0.218
NN-Filter	0.3373	0.2576	0.1669

Table 4. Comparison of distributions of XXStout vs. native maps and NN-Filter vs. native maps, 8Å maps. Kolmogorov-Smirnov test.

	12Å	8Å
Native	573126	216719
XXStout	591856	178749
NN-Filter	786035	218260

Table 5. Number of true (Native) and predicted contacts by XXStout and NN-Filter, 12Å and 8Å.

contacts compared to XXStout.

3.4. *3D reconstructions*

We also test whether improved, filtered maps yield more accurate reconstructions of protein structures. The reconstruction procedure we adopt is similar to that in [1], and is described in detail in [11] and [12]. Briefly, proteins are represented simply

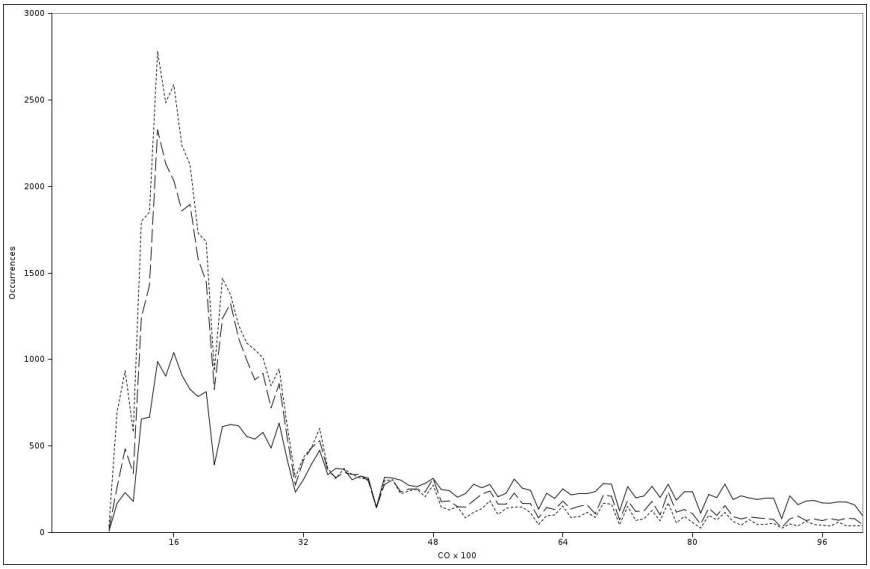

Fig. 10. Comparison of residue CO distributions before and after filtering with true distributions for 8Å maps. The solid line represents the native distribution, the dashed line NN-Filter, and the dotted line XXStout.

	TM	GDT	RMSD
XXStout	0.261	0.236	14.3
NN-Filter	0.260	0.232	13.8

Table 6. Reconstruction quality on $TES258$ for XXstout vs. NN-Filter 12Å maps.

as C_α traces and a stochastic search (simulated annealing) of the space of protein conformations is operated, starting from a random initial structure and trying to minimise a simple cost that favours the implementation of a contact map. We reconstruct all proteins in the $TES258$ set, running 10 reconstructions for each protein, and adopting as contact maps 12Å maps from XXstout and NN-Filter. Table 6 reports root mean square distances (RMSD), GDT and TM score[15], averaged over all reconstructions, between reconstructed and native structures.

No clear conclusion can be reached from the results. Although filtered maps are an improvement over XXstout predictions, they still do not seem to contain enough information to reliably predict the structure, at least through the simple reconstruction procedure we adopt here.

	Distill			Filter		
$\|i-j\|$	mF1	mP	mR	mF1	mP	mR
$\|i-j\| \geq 6$	0.0992	0.1715	0.0698	0.1320	0.1703	0.1077
$\|i-j\| \geq 12$	0.0812	0.1469	0.0561	0.1182	0.1584	0.0942
$\|i-j\| \geq 24$	0.0536	0.1039	0.0361	0.0831	0.1316	0.0608
$\|i-j\| \geq 8$	0.0883	0.1563	0.0615	0.1244	0.1609	0.1014
$\|i-j\| \leq 8$	0.8364	0.8236	0.8496	0.8339	0.8137	0.8551

Table 7. CASP7 results for XXstout (Distill) and NN-Filter. 8Å maps. The results are micro-averages, i.e. averaged over all pairs of residues in the set.

	Distill			Filter		
$\|i-j\|$	F1	P	R	F1	P	R
$\|i-j\| \geq 6$	0.0990	0.1712	0.0763	0.1229	0.1503	0.1187
$\|i-j\| \geq 12$	0.0773	0.1354	0.0579	0.1098	0.1382	0.1057
$\|i-j\| \geq 24$	0.0446	0.0891	0.0332	0.0640	0.0858	0.0599
$\|i-j\| \geq 8$	0.0871	0.1463	0.0665	0.1159	0.1410	0.1148
$\|i-j\| \leq 8$	0.8272	0.8083	0.8542	0.8252	0.8037	0.8558

Table 8. CASP7 results for XXstout (Distill) and NN-Filter. 8Å maps. The results are macro-averages, i.e. averaged over all proteins in the set.

3.5. *CASP7 results*

We tested both XXstout (as part of the Distill predictor) and NN-Filter at the CASP7 competition. As NN-Filter became available during the prediction season and after some of the targets has expired (i.e. no further predictions were accepted on them) we submitted predictions only for 66 of the 95 targets. The results in tables 7 and 8 refer to these 66 targets. The results are for a contact threshold of 8Å, as per CASP7 regulations. The first table reports micro-averaged results, meaning that precision and recall are averaged over the whole set (each pair of residues counts the same), while the second table reports macro-averages, i.e. precision and recall are computed for each protein and then averaged over all proteins (in this case each protein counts the same, but pairs of residues in long proteins count less than pairs in short proteins).

The tables broadly confirm the results on $TES481$, with NN-Filter improving over "raw" XXstout predictions on all sequence separation bands, with the exception of the area just beside the diagonal (sequence separation of at most 8), on which the two predictors are substantially undistinguishable, and quite accurate (F1 over 82%). For contacts with a sequence separation greater than 23 residues F1 grows from 5.4% to 8.3% (4.5% to 6.4% for macro-averages), and from 8.1% to 11.8% (7.7% to 11% for macro-averages) for sequence separation greater than 11.

4. Conclusions

One of the main problems with residue contact map predictors is the fact that they generally fail to incorporate global information about a protein. As a result, often, predicted contact maps are inherently local, in that their predictions do not respect some basic rules about the number of residue contacts allowed for an amino acid, its contact order, etc. In this paper we proposed a simple attempt to improve predicted contact maps by building a filter that has access not only to local properties of the maps, but also to a number of relevant global quantities. Although, somewhat surpisingly, it is unclear whether the filter improves the overall "physicality" of the maps (with only residue contact orders clearly becoming closer to native), our results show that the filter greatly enhances the quality of the map, especially for long-range contacts, that are both harder to predict, and more informative to determine the overall topology of a protein structure. These results are confirmed by the latest CASP7 competition, in which the filtered architecture clearly outperformed the first-stage predictor. Importantly, since the filtering stage is entirely modular, we believe that our approach is likely to be beneficial to any contact map predictor.

We are currently extending this work to deal with distance maps, and multi-class distance maps, which are more informative than binary contact maps, and with maps incorporating homology to structures in the PDB, which are greatly more accurate than *ab initio* predicted ones. In both cases, it is likely that improvements in map quality will also translate into improvements in 3D structure prediction.

References

1. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, (2):295–306, 1997.
2. P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12(1):15–21, 1999.
3. 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.
4. R.M. McCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20, Suppl. 1:224–231, 2004.
5. M. Punta and B. Rost. Profcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–8, 2005.
6. A Vullo, I Walsh, and G Pollastri. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, 7:180, 2006.
7. G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–20, 2005.
8. G Pollastri, AJ 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.
9. U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–24, 1994.
10. W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637,

1983.
11. G Pollastri, D Baú, and A Vullo. Distill: A machine learning approach to ab initio protein structure prediction. In S Bandyopadhyay, U Maulik, and JTL Wang, editors, *Analysis of Biological Data: A Soft Computing Approach*. World Scientific. in press.
12. D Bau, AJ 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.
13. S.F. Altschul, T.L. Madden, and A.A. Schaffer. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.*, 25:3389–3402, 1997.
14. M.M. Gromiha and S. Selvaraj. Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, (86):235–277, 2004.
15. Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, (57):702–710, 2004.
16. A. G. Murzin, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, (247):536–540, 1995.
17. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.
18. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000.