**JOURNAL OF COMPUTATIONAL BIOLOGY**
**Volume 13, Number 8, 2006**

**Pp. 1489–1502**

# Protein Structural Motif Prediction in Multidimensional $\phi$-$\psi$ Space Leads to Improved Secondary Structure Prediction

CATHERINE MOONEY, ALESSANDRO VULLO, and GIANLUCA POLLASTRI

## ABSTRACT

**A significant step towards establishing the structure and function of a protein is the prediction of the local conformation of the polypeptide chain. In this article, we present systems for the prediction of three new alphabets of local structural motifs. The motifs are built by applying multidimensional scaling (MDS) and clustering to pair-wise angular distances for multiple $\phi$-$\psi$ angle values collected from high-resolution protein structures. The predictive systems, based on ensembles of bidirectional recurrent neural network architectures, and trained on a large non-redundant set of protein structures, achieve 72%, 66%, and 60% correct motif prediction on an independent test set for di-peptides (six classes), tri-peptides (eight classes) and tetra-peptides (14 classes), respectively, 28–30% above baseline statistical predictors. We then build a further system, based on ensembles of two-layered bidirectional recurrent neural networks, to map structural motif predictions into a traditional 3-class (helix, strand, coil) secondary structure. This system achieves 79.5% correct prediction using the "hard" CASP 3-class assignment, and 81.4% with a more lenient assignment, outperforming a sophisticated state-of-the-art predictor (Porter) trained in the same experimental conditions. The structural motif predictor is publicly available at: *http://distill.ucd.ie/porter+/*.**

**Key words:** protein structure prediction, secondary structure, structural motifs, neural networks.

## 1. INTRODUCTION

A SIGNIFICANT STEP towards establishing the structure and function of a protein is the prediction of the local conformation of the polypeptide chain. Secondary structure, consisting of folding regularities maintained by hydrogen bonds, is a widely used representation. A large number of secondary structure predictors have been developed, with the most accurate ones classifying correcly up to 76–80% of residues (Jones, 1999; Pedersen et al., 2000; Baldi et al., 1999; Pollastri et al., 2002; Pollastri and McLysaght, 2005; Rost and Eyrich, 2001). Many other representations of local conformations are possible, based on patterns of hydrogen bonds (e.g., different secondary structure assignments [Pollastri et al., 2002]), or on a protein's backbone angles. Protein dihedral backbone angles $\phi$ and $\psi$ are a compact, vectorial representation of a protein's structure. The Ramachandran plot, which plots $\phi$ against $\psi$, has proved to be

School of Computer Science and Informatics, University College Dublin, Belfield, Dublin, Ireland.

a valuable tool to identify and visualize the $\phi$-$\psi$ conformational space allowed. However it is harder to deal with, or visualize, the conformations of $n$-peptides—protein fragments which can be described by $n$ pairs of dihedral angles.

The identification, and classification of conformations of $n$-peptides has been attempted in many different forms and through an array of different algorithms. A simple approach is building libraries of relevant fragments (Bystroff and Baker, 1998; Yang and Wang, 2003), for instance by identifying structural motifs by some clustering method (e.g., k-means [Bystroff et al., 2000], Self-Organizing Maps [de Brevern et al., 2000, 2004]). After relevant structural patterns are identified, the prediction of torsional angles can be modelled as a classification problem in the space induced by these patterns. This means mapping the string representing the primary sequence into a string from an alphabet of $p$ letters, each representing a structural motif. Numerous machine learning or, more in general, statistical tools are available to solve this problem, and have been adopted to try to classify protein local structural motifs (Bystroff et al., 2000; de Brevern et al., 2004). In Karchin et al. (2003), nine alphabets of local structure descriptions were examined to establish which descriptions are most useful for improving fold recognition and alignment quality. This study showed that detailed alphabets have greater potential for fold recognition and that the best results can be achieved by combining several alphabets. Developing new, informative alphabets of structural motifs, and efficient methods to predict them from the primary sequence is thus of great interest: these alphabets may help improving the performances of current algorithms for protein folding, and for *ab initio* protein structure prediction; richer, subtler predicted information about torsional angles may feed back into secondary structure prediction algorithms, boosting their performance, as shown in Wood and Hirst (2005).

Recently, multidimensional scaling (MDS) was applied to pair-wise angular distances for multiple $\phi$-$\psi$ values collected from high-resolution protein structures (Sims et al., 2005). This principled method allowed the visualization of protein backbone fragments in a reduced 3D conformational space, and led to the identification of a small number of conformational clusters that are populated by real backbones. Here we map the clusters identified in Sims et al. (2005) into three conformational alphabets of 6, 8, and 14 letters, for di-, tri-, and tetra-peptides respectively. Based on this novel representation we develop architectures composed of ensembles of bidirectional recurrent neural networks to predict the structural motif of protein backbone fragments from a protein's primary sequence. These architectures achieve 72%, 66%, and 60% correct prediction in the 6, 8, and 14 class problem, respectively, 28–30% above base-line statistical predictors.

Furthermore, we adopt the three systems as the first stage of a pipeline for the prediction of traditional 3-class secondary structure. To do so, we build a further system to map structural motif predictions into secondary structure. This system achieves 79.5% correct classification using the "hard" CASP 3-class assignment, and 81.4% using the "easier" assignment in Pedersen et al. (2000), outperforming the state-of-the-art predictor Porter (Pollastri and McLysaght, 2005) trained in the same experimental conditions.

## 2. METHODS

### *2.1. Dataset*

The data set used in our simulations is extracted from the December 2003 25% pdb_select list (Hobohm et al., 1992). We use the DSSP program (Kabsch and Sander, 1983) to assign secondary structure and $\phi$ and $\psi$ angles, and remove sequences for which DSSP does not produce an output due to missing entries or format errors. After processing by DSSP, the set contains 2171 proteins and 344,653 amino acids. For our experiments we split the data into a training set containing 1736 sequences (S1736) and a test set of 435 (S435), or 1/5 of the total. The test set sequences are selected in an interleaved fashion (i.e., every fifth sequence is picked) from the whole set sorted alphabetically by PDB code.

Prediction from a multiple alignment of protein sequences rather than a single sequence has long been recognized as a way to improve prediction accuracy for virtually all protein structural features, including secondary structure (Rost and Sander, 1994; Riis and Krogh, 1996; Jones, 1999; Baldi et al., 1999; Pollastri et al., 2002; Pollastri and McLysaght, 2005), solvent accessibility (Pollastri et al., 2002), beta-sheet pairing (Baldi et al., 2000), and contact maps (Pollastri and Baldi, 2002; Baldi and Pollastri, 2003). Here we exploit

evolutionary information in the form of frequency profiles compiled from multiple sequence alignments. Alignments for the 2171 proteins in S1736 abd S435 are extracted from the NR database as available on March 3, 2004, containing over 1.4 million sequences. The database is first redundancy reduced at a 98% threshold, leading to a final 1.05 million sequences. The alignments are generated by three runs of PSI-BLAST (Altschul et al., 1997) with parameters $b = 3000$, $e = 10^{-3}$ and $h = 10^{-10}$. Data sets, multiple alignments, and training/test set splitting are identical to those in Pollastri and McLysaght (2005) and Vullo et al. (2006).

### *2.2. Data clustering*

To cluster sequences of $\phi$ and $\psi$ angles in the sets, we follow the scheme devised in Sims et al. (2005). In this study, protein conformational space from two to seven residue lengths is mapped into a three-dimensional space employing multidimensional scaling (MDS). MDS is run on a data matrix composed of pair-wise angular distances for multiple $\phi$-$\psi$ values collected from high-resolution protein structures. The resulting data points are then clustered. The analysis in Sims et al. (2005) identifies 6, 8, and 14 clusters for the case of di-, tri-, and tetra-peptides respectively. We adopt the centroids of these clusters as structural motifs. We map each $n$-peptide $s$ in the S1736 and S435 datasets into the motif corresponding to the cluster $i$ that minimizes the following distance:

$$d(c_i, s) = \left( \sum_{k=1}^{n} [(180^\circ - |180^\circ - |\phi_k^{c_i} - \phi_k^s||)^2 + (180^\circ - |180^\circ - |\psi_k^{c_i} - \psi_k^s||)^2] \right)^{\frac{1}{2}}$$

where $c_i$ is the centroid of the $i$-th cluster, $x_k^{c_i}$ $x_k^s$ are the values of torsional angle $x$ for the $k$-th residue of centroid $c_i$ and of n-peptide $s$ respectively. All the angles are transformed to fall between 0° and 360°. The distance in equation 1 is the same adopted in Sims et al. (2005), and takes into account the circular quality of angles.

The compositions of all clusters in terms of DSSP secondary structures are reported in Tables 1–3. DSSP classes are defined as: H = $\alpha$-helix; G = 3-10-helix; I = $\pi$-helix; E = extended strand; B = $\beta$-bridge; T = turn; S = bend; C = the rest.

In Figures 1–3, we show sequence logos (Crooks et al., 2004) for the three clustering schemes. In these, each cluster is represented by a stack of letters corresponding to the 8 DSSP classes. The height of each stack is proportional to the difference, measured in bits, between the maximum possible entropy for that cluster (that one would obtain if DSSP classes were equally probable) and the observed entropy:

$$log_2 8 - \left( - \sum_{i=1}^{8} p(s_i) log_2 p(s_i) \right) \tag{1}$$

where $s_i$ is the $i$-th DSSP class, and $p(s_i)$ is its frequency within a given cluster. The height of each letter is proportional to its relative frequency within the stack.

TABLE 1. DSSP 8-CLASS SECONDARY STRUCTURE COMPOSITION (%) OF 6 CLUSTERS (DI-PEPTIDES)

| | *H* | *G* | *I* | *E* | *B* | *T* | *S* | *C* |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.67 | 1.99 | 0.05 | 12.59 | 1.49 | 45.72 | 21.96 | 15.53 |
| 2 | 7.19 | 3.66 | 0.04 | 1.15 | 0.24 | 41.92 | 16.72 | 29.08 |
| 3 | 75.45 | 6.44 | 0.05 | 0.07 | 0.02 | 12.61 | 3.12 | 2.23 |
| 4 | 9.42 | 3.69 | 0.02 | 12.46 | 1.64 | 11.73 | 18.21 | 42.83 |
| 5 | 0.01 | 0.04 | 0.00 | 57.50 | 2.42 | 1.32 | 6.16 | 32.55 |
| 6 | 3.15 | 2.93 | 0.02 | 12.88 | 1.63 | 17.92 | 28.86 | 32.62 |

Cluster number in the first column. Rows add up to 100.

TABLE 2. DSSP 8-CLASS SECONDARY STRUCTURE COMPOSITION (%) OF 8 CLUSTERS (TRI-PEPTIDES)

| | *H* | *G* | *I* | *E* | *B* | *T* | *S* | *C* |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.05 | 0.06 | 0.01 | 21.57 | 1.90 | 11.32 | 28.35 | 36.73 |
| 2 | 0.01 | 0.03 | 0.00 | 64.79 | 1.94 | 0.89 | 4.65 | 27.69 |
| 3 | 1.35 | 1.71 | 0.01 | 25.17 | 2.46 | 11.80 | 20.73 | 36.77 |
| 4 | 10.80 | 7.37 | 0.04 | 4.90 | 1.21 | 31.83 | 19.86 | 23.98 |
| 5 | 78.59 | 5.34 | 0.05 | 0.03 | 0.01 | 10.44 | 2.39 | 3.15 |
| 6 | 19.42 | 7.66 | 0.03 | 5.76 | 1.23 | 25.01 | 12.56 | 28.34 |
| 7 | 5.06 | 2.18 | 0.01 | 25.18 | 2.53 | 13.06 | 13.94 | 38.04 |
| 8 | 9.33 | 4.28 | 0.06 | 2.60 | 1.06 | 21.04 | 27.94 | 33.69 |

Cluster number in the first column. Rows add up to 100.

TABLE 3. DSSP 8-CLASS SECONDARY STRUCTURE COMPOSITION (%) OF 14 CLUSTERS (TETRA-PEPTIDES)

| | *H* | *G* | *I* | *E* | *B* | *T* | *S* | *C* |
|---|---|---|---|---|---|---|---|---|
| 1 | 13.43 | 5.61 | 0.01 | 15.31 | 2.12 | 19.64 | 10.96 | 32.92 |
| 2 | 31.33 | 11.25 | 0.05 | 3.67 | 0.91 | 17.03 | 11.17 | 24.59 |
| 3 | 84.20 | 3.69 | 0.05 | 0.01 | 0.01 | 7.98 | 1.51 | 2.56 |
| 4 | 14.53 | 6.66 | 0.09 | 2.32 | 0.85 | 28.18 | 19.98 | 27.40 |
| 5 | 8.10 | 3.83 | 0.01 | 7.61 | 2.50 | 16.98 | 22.60 | 38.37 |
| 6 | 3.11 | 1.48 | 0.01 | 35.62 | 2.34 | 10.38 | 12.12 | 34.95 |
| 7 | 0.01 | 0.02 | 0.00 | 68.59 | 1.56 | 0.72 | 4.05 | 25.05 |
| 8 | 0.14 | 0.39 | 0.01 | 28.89 | 2.09 | 9.70 | 23.73 | 35.05 |
| 9 | 0.11 | 1.45 | 0.01 | 12.32 | 2.10 | 30.94 | 22.45 | 30.62 |
| 10 | 21.11 | 11.14 | 0.05 | 3.00 | 0.89 | 27.20 | 15.35 | 21.28 |
| 11 | 6.12 | 5.45 | 0.01 | 13.32 | 2.20 | 24.02 | 17.51 | 31.36 |
| 12 | 0.84 | 1.17 | 0.00 | 36.61 | 2.27 | 8.98 | 16.95 | 33.18 |
| 13 | 0.10 | 0.25 | 0.01 | 30.37 | 2.14 | 12.25 | 21.62 | 33.27 |
| 14 | 18.10 | 6.59 | 0.08 | 1.23 | 0.95 | 22.96 | 24.49 | 25.59 |

Cluster number in the first column. Rows add up to 100.

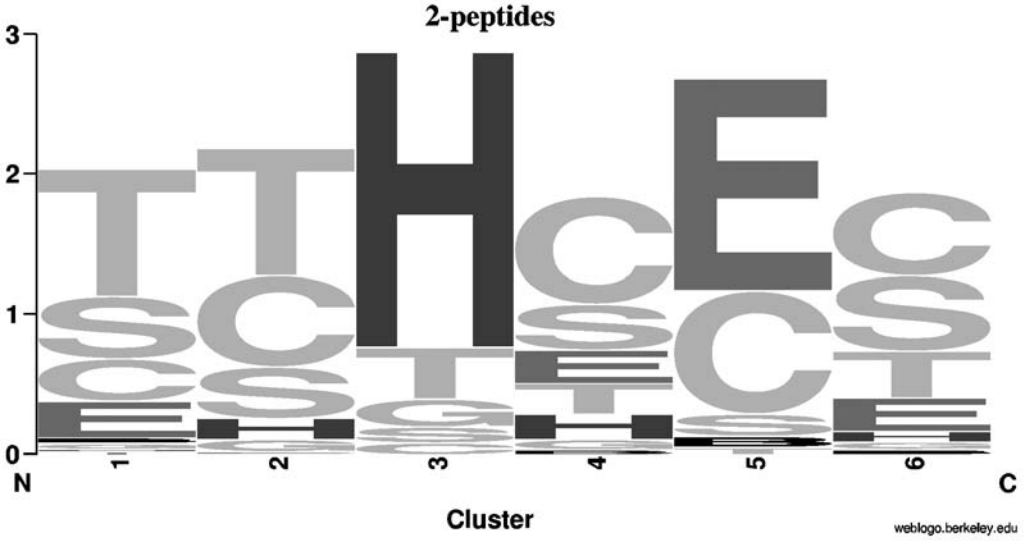


**FIG. 1.** Sequence logos of DSSP 8-class secondary structures in 6 clusters (di-peptides).

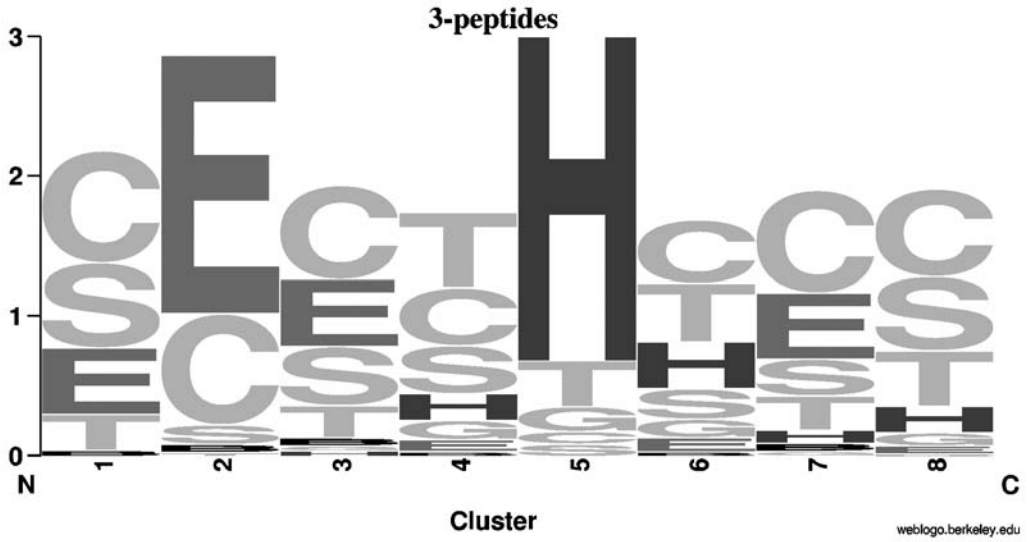


**FIG. 2.** Sequence logos of DSSP 8-class secondary structures in 8 clusters (tri-peptides).

### *2.3. Predictive algorithms and implementation*

*2.3.1. From primary sequence to structural motifs.* We model the prediction of a residue's closest structural motif as a classification task with multiple classes. Formally, this consists in learning a mapping $f(\cdot) : \mathcal{I} \rightarrow \mathcal{O}$ from the space $\mathcal{I}$ of labeled input sequences to the space $\mathcal{O}$ of labeled output sequences. In practice, we want to predict a sequence of labels $\mathcal{O} = \mathcal{O}^{(n)} = (o_1^{(n)}, \ldots, o_N^{(n)})$, for a given sequence of inputs $I = (i_1, \ldots, i_N)$, where each $i_j \in I$ is the input coding of residue $r_j$ in position $j$, and output $o_k^{(n)}$ represents the $n$-peptide structural motif the $k$-th residue belongs to. In our case the problem is modelled for $n = 2$, $n = 3$, and $n = 4$, corresponding to 6-, 8-, and 14-class classifications as in Sims et al. (2005).

To learn the mapping between inputs $\mathcal{I}$ and outputs $\mathcal{O}^{(n)}$ (sequence to structural motif) we use an architecture composed of Bidirectional Recurrent Neural Networks (BRNN) (Baldi et al., 1999; Baldi and Pollastri, 2003) of the same length $N$ as the amino acid sequence. Similarly to Pollastri and McLysaght (2005), we use BRNNs with shortcut connections. In these BRNNs, connections along the forward and

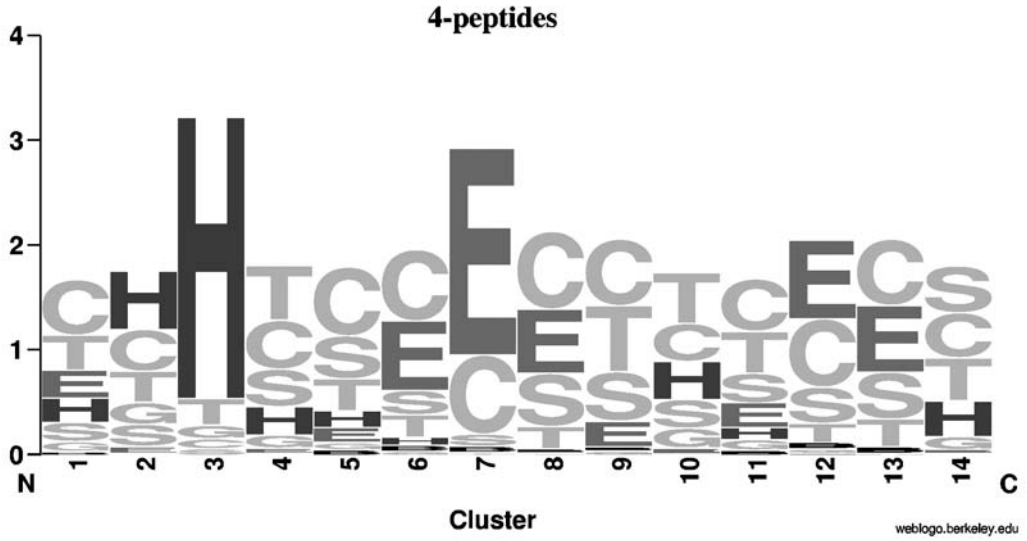


**FIG. 3.** Sequence logos of DSSP 8-class secondary structures in 14 clusters (tetra-peptides).

backward hidden chains span more than 1-residue intervals, creating shorter paths between inputs and outputs. These networks take the form:

$$o_j = \mathcal{N}^{(O)}\left(i_j, h_j^{(F)}, h_j^{(B)}\right)$$

$$h_j^{(F)} = \mathcal{N}^{(F)}\left(i_j, h_{j-1}^{(F)}, \ldots, h_{j-S}^{(F)}\right)$$

$$h_j^{(B)} = \mathcal{N}^{(B)}\left(i_j, h_{j+1}^{(B)}, \ldots, h_{j+S}^{(B)}\right)$$

$$j = 1, \ldots, N$$

where $h_j^{(F)}$ and $h_j^{(B)}$ are forward and backward chains of hidden vectors with $h_0^{(F)} = h_{N+1}^{(B)} = 0$, and $S$ is the longest shortcut length. We parametrize the output update, forward update, and backward update functions (respectively, $\mathcal{N}^{(O)}$, $\mathcal{N}^{(F)}$, and $\mathcal{N}^{(B)}$) using three two-layered feed-forward neural networks.

In the tests presented in this work, the input associated with the $j$-th residue $i_j$ contains amino acid information obtained from multiple sequence alignments of the protein sequence to its homologues, to leverage evolutionary information. Amino acids are coded as letters out of an alphabet of 25 (Pollastri and McLysaght, 2005). Beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine), and . (gap) are considered. The input presented to the networks is the frequency of each of the 24 non-gap symbols, plus the overall frequency of gaps in each column of the alignment. That is, if $n_{sk}$ is the total number of occurrences of symbol $s$ in column $k$, and $g_k$ the number of gaps in the same column, the $s^{th}$ input to the networks in position $k$ is:

$$\frac{n_{sk}}{\sum_{v=1}^{24} n_{vk}} \tag{2}$$

for $j = 1 \ldots 24$, while the $25^{th}$ input is:

$$\frac{g_k}{g_k + \sum_{v=1}^{24} n_{vk}} \tag{3}$$

This input coding scheme is richer than simple 20-letter schemes and has proven effective in Pollastri and McLysaght (2005).

Obviously, when encoding the output for the $n$-peptide case, one needs to consider that each residue belongs to $n$ distinct $n$-peptides, which in turn may belong to different structural motifs. In the experiments presented in this work, the output label $o_j^{(n)}$ for the $j$-th residue $r_j$ is the structural motif of the $n$-peptide formed by residues $r_j \ldots r_{j+n-1}$. Different offsets between inputs and output labels do not significantly affect classification performances (not shown).

To predict structural motifs we train seven BRNNs of different sizes and with different architectural details for each of the three classification problems. The ranges for the number of free parameters per network are shown in Table 4.

Table 4. The Number of Free Parameters per Network Type

| *No. of clusters* | *Minimum* | *Maximum* |
|---|---|---|
| 6 | 6438 | 12834 |
| 8 | 7136 | 13946 |
| 14 | 9230 | 17282 |

*2.3.2. From structural motifs to 3-class secondary structure.* We model the prediction of a residue's secondary structure from predicted structural motifs as a 3-class classification task, formally: $g(\cdot) : \mathcal{O} \rightarrow \mathcal{O}'$ from the space $\mathcal{O}$ of structural motifs to the space $\mathcal{O}'$ of labeled secondary structure sequences. In this case the input for residue $j$ is $o_j = (o_j^{(2)}, o_j^{(3)}, o_j^{(4)})$, or the predicted structural motifs for all three clustering schemes, and the corresponding output $o'_j$ is the secondary structure label (Helix, Strand, Coil) for the same residue.

To model this mapping ($\mathcal{O} \rightarrow \mathcal{O}'$) we train the same architecture adopted in Pollastri and McLysaght (2005) and Vullo et al. (2006), composed of an ensemble of 5 two-layered BRNNs. The first layer is similar to the BRNN adopted to predict structural motifs, except that the input associated with the $j$-th residue now contains 28 numbers representing the probabilities of each of the 6, 8, and 14 clusters, as estimated by the structural motif predictor. The output (target) is the 3-class secondary structure. We also adopt a second filtering BRNN, similarly to Pollastri and McLysaght (2005) and Vullo et al. (2006). The network is trained to predict secondary structure given the first-layer secondary structure predictions. The $i$-th input to this second network includes the first-layer predictions in position $i$ augmented by first stage predictions averaged over multiple contiguous windows. In other words, if $o'_{j1}, \ldots o'_{jm}$ are the outputs in position $j$ of the first stage network corresponding to estimated probability of secondary structure $j$ being in class $m$, the input to the second stage network in position $j$ is the array $O'_j$:

$$O'_j = \left( o'_{j1}, \ldots, o'_{jm}, \sum_{h=k_{-p}-w}^{k_{-p}+w} o'_{h1}, \ldots, \sum_{h=k_{-p}-w}^{k_{-p}+w} o'_{hm}, \ldots, \sum_{h=k_p-w}^{k_p+w} o'_{h1}, \ldots, \sum_{h=k_p-w}^{k_p+w} o'_{hm} \right) \tag{4}$$

where $k_f = j + f(2w+1)$, $2w+1$ is the size of the window over which first-stage predictions are averaged and $2p+1$ is the number of windows considered. In the tests we use $w = 7$ and $p = 7$. This means that 15 contiguous, non-overlapping windows of 15 residues each are considered; i.e., first-stage outputs between position $j - 112$ and $j + 112$, for a total of 225 contiguous residues, are taken into account to generate the input to the filtering network in position $j$. This input contains a total of $16 \times 3$ real numbers: 3 representing the first-stage secondary structure prediction in position $j$; $15 \times 3$ representing the 3-class first-stage secondary structure predictions averaged over each of the 15 windows.

*2.3.3. Training procedure.* In all the BRNNs, we adopt softmax outputs and train the networks by minimizing the cross-entropy error between the output and target probability distributions, using gradient descent with no momentum term or weight decay. The gradient is computed using the Back-Propagation Through Structure (BPTS) algorithm (Frasconi et al., 1998). We use a hybrid between online and batch training, with 580 batch blocks (roughly 3 proteins each) per training set; i.e., the weights are updated 580 times per epoch. Training examples are presented to the networks in a different random order every epoch. For this reason the error does not, in general, decrease monotonically. When the error does not decrease for 50 consecutive epochs, the learning rate is halved. Training stops after 1000 epochs. Typically, by the end of training, the learning rate is between $1/8$ and $1/1024$ of the initial one. In the system mapping structural motif predictions into 3-class secondary structure first-layer and filtering BRNNs are trained simultaneously, but supervised independently.

## 3. RESULTS AND DISCUSSION

On the S435 set (test set), our systems for structural motif prediction for di-, tri-, and tetra-peptides into 6, 8, and 14 clusters achieve 72.0%, 65.9%, and 59.8% correct classifications, respectively. A base-line statistical predictor assigning each type of residue to the class it most frequently belongs to (Richardson and Barlow, 1999) achieves, respectively, 41.9%, 37.0%, and 31.9%, or 28–30% below our architectures. Figures 4–6 show the percentage of times the true structural motif is in the $k$ top ranking classes according to the predictors. For di-peptides in about 87% of the cases the true class is either the first or the second estimated as most probable. This is also true for over 80% of residues in the tri-peptide case, and nearly 75% of the times for tetra-peptides. In the case of tetra-peptides (14 classes), the true class is ranked within the 3 most probable for well over 80% of residues. The predictors are also significantly more confident

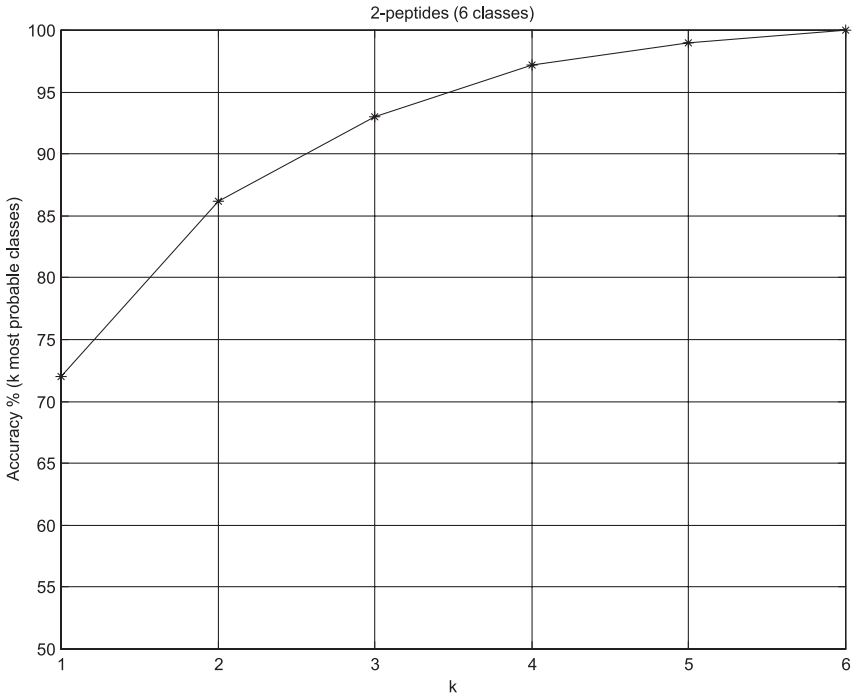


**FIG. 4.** Percentage of residues for which the correct structural motif is in the *k* top ranking classes according to the predictors: di-peptides.

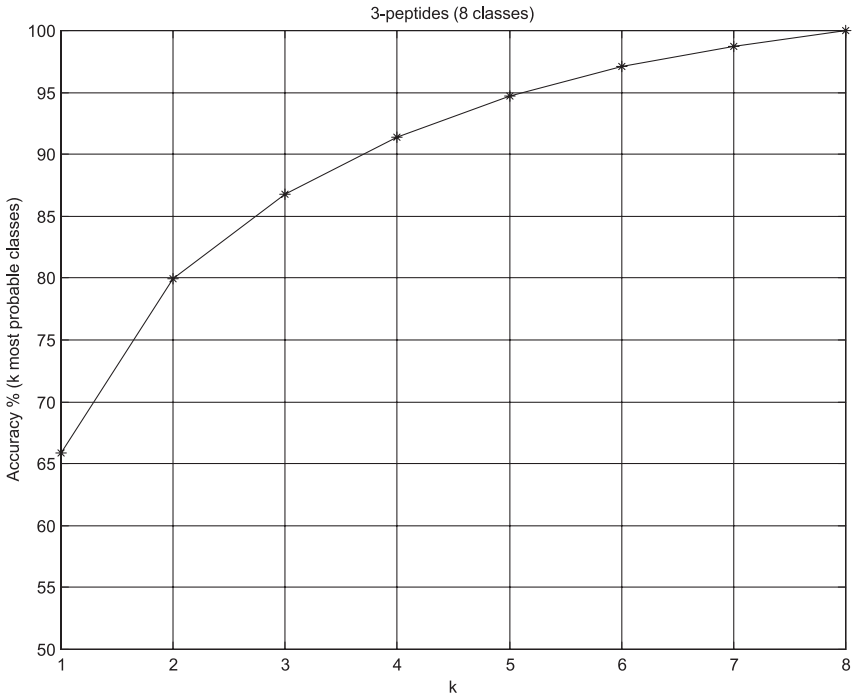


**FIG. 5.** Percentage of residues for which the correct structural motif is in the *k* top ranking classes according to the predictors: tri-peptides.

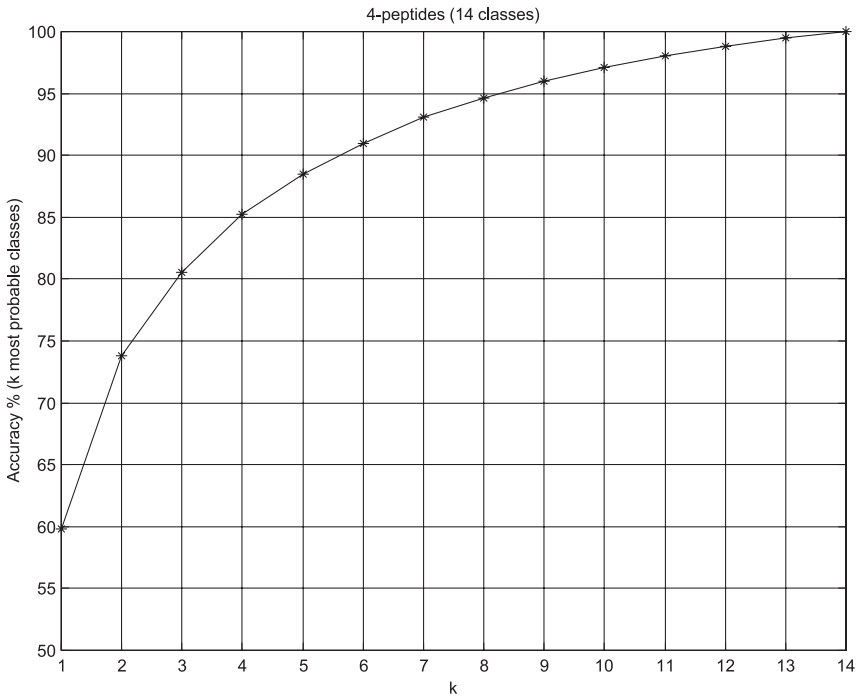


**FIG. 6.** Percentage of residues for which the correct structural motif is in the $k$ top ranking classes according to the predictors: tetra-peptides.

(Table 5) on their correct predictions than on their incorrect ones. In the former case the difference between the outputs corresponding to the class ranked first and the class ranked second is on average 0.57–0.64, in the latter 0.2–0.27. This suggests that the full output of the predictors (i.e., the 6, 8, or 14 estimated probabilities of the structural motifs for each residue) contains substantially more information than the simple identity of the class ranked first.

When the results are examined in detail we observe that: not surprisingly, the largest clusters tend to be predicted most accurately; these clusters are populated mainly by helical structures (the largest cluster), and by strands (the second largest; Table 6). In the di-peptide clustering scheme nearly 40% of all residues fall into cluster 3. This cluster contains 94.9% of all $\alpha$-helical residues and 73.6% of all 3-10-helices, with these two categories making up 82% of all the residues contained in the cluster (Table 1). The prediction accuracy for this cluster is the highest of all classes, at 87%. In the tri-peptide clustering scheme, cluster 5 contains 37% of all residues, including 90.5% of those labeled as $\alpha$-helix, and 55.9% labeled as 3-10-helix, and is predicted with 89.7% accuracy (Table 2). In the tetra-peptide case (14 clusters) cluster 3 holds 32% of all residues, including 84.5% of all $\alpha$-helical residues and 33.7% of 3-10-helices, and is predicted at 90.3% accuracy (Table 3). These clusters also contain a sizeable fraction (44.1%, 33.4%, and 22.3%, respectively) of DSSP turns (T).

TABLE 5. AVERAGE NETWORK CONFIDENCE FOR CORRECTLY AND INCORRECTLY PREDICTED RESIDUES

| | *Correct* | *Incorrect* |
|---|---|---|
| Di-peptides | 0.639 | 0.273 |
| Tri-peptides | 0.604 | 0.225 |
| Tetra-peptides | 0.573 | 0.195 |

Network confidence: the difference between the highest and second highest prediction per residue.

Table 6. Percentages of Correct Predictions for Di-, Tri-, and Tetra-Peptides and 6, 8, and 14 Clusters, with Cluster Means Determined Using MDS from Sims et al. (2005)

| *Cluster* | *Di-peptides* | *Tri-peptides* | *Tetra-peptides* |
|---|---|---|---|
| 1 | 25.5 | 16.7 | 46.1 |
| 2 | 52.2 | 80.9 | 49.0 |
| 3 | 87.0 | 34.6 | 90.3 |
| 4 | 38.7 | 37.1 | 17.7 |
| 5 | 83.2 | 89.8 | 17.9 |
| 6 | 35.8 | 47.5 | 38.2 |
| 7 | | 38.9 | 80.8 |
| 8 | | 18.3 | 18.1 |
| 9 | | | 17.9 |
| 10 | | | 36.5 |
| 11 | | | 31.3 |
| 12 | | | 35.9 |
| 13 | | | 22.0 |
| 14 | | | 20.5 |
| $Q_{tot}$ | 72.0 | 65.9 | 59.8 |

The second most accurately predicted class is the second largest cluster in each case, and this time the clusters are dominated by strands. In the first case (di-peptides) cluster 5 contains 85.2% of all strands and 65.1% of all $\beta$-bridges, and is predicted with an accuracy of 83.2%. In the second case (tri-peptides), 67% of the residues in the second most accurately predicted cluster are strands (71.8% of all strands and 39.1% of all $\beta$-bridges), and the prediction accuracy is 80.9%. For tetra-peptides, cluster 7 is labeled 70% strand (57.8% of all strands and 24% of $\beta$-bridges) and is classified with a prediction accuracy of 80.8%.

All the other clusters are substantially smaller (most containing less than 10% of the residues), which makes them harder to predict. Prediction accuracy per cluster drops to between 52.2% to 25.5% for 6 clusters, 47.5% to 16.7% for 8 clusters, and 40% to 17.7% for 14 clusters. However small, some of these classes seem to represent genuine splits within DSSP-assigned secondary structures, and in some cases possibly "clean up" the ambiguity of some DSSP choices. For instance, clusters 2 and 10 in the 14-cluster problem cover 33.1% of 3-10-helices (roughly the same fraction contained in the mainly-helical cluster 3), but a much smaller fraction of $\alpha$-helices: at a local backbone conformation level, about a third of DSSP-assigned 3-10-helices are more similar to $\alpha$-helices while a further third is clustered separately. Cluster 2 is the third most accurately predicted (49.0%), while cluster 10 is predicted with a 36.5% accuracy. Further examples are clusters 6 and 12, which together contain roughly 21% of all strands, mainly parallel (while strands in cluster 7 are mainly anti-parallel) and nearly a quarter of all $\beta$-bridges (roughly the same amount as cluster 7). Both clusters are predicted at around 35–40% accuracy.

### *3.1. Secondary structure prediction*

We define the 3 secondary structural classes by mapping the 8 DSSP classes in two different ways: H, G, I $\rightarrow$ Helix; E, B $\rightarrow$ Strand; S, T, C $\rightarrow$ Coil (HARD); this assignment is known to be "hard" and has been adopted at CASP (Lesk et al., 2001; Moult et al., 2003); H $\rightarrow$ Helix; E $\rightarrow$ Strand; G, I, B, S, T, C $\rightarrow$ Coil (EASY)—this assignment generally leads to an "easier" classification task and is adopted for instance in Pedersen et al. (2000).

The system achieves 79.5% correct classification on the HARD assignment, and 81.4% on the EASY one. In the HARD case 83.6% of all actual helices and 71.1% of all actual strands are predicted correctly, while 86% of predicted helices and 77.7% of predicted strands are in fact helices and strands, respectively. Matthews' correlation coefficients are, respectively, 76.8% and 67.1%. Predictive accuracies and confusion matrices for HARD and EASY are reported in Tables 7–9.

We compare secondary structure predictions based on predicted structural motifs with the state-of-the-art predictor Porter (Pollastri and McLysaght, 2005). In tests reported in Pollastri and McLysaght (2005),

TABLE 7. PREDICTION OF SECONDARY STRUCTURE USING HARD AND EASY 3-CLASS ASSIGMENTS

| | *HARD* | | | *EASY* | | |
|---|---|---|---|---|---|---|
| | $Q_{class}$ | | $C_{class}$ | $Q_{class}$ | | $C_{class}$ |
| | *Obs* | *Pred* | | *Obs* | *Pred* | |
| Helix | 83.6 | 86.0 | 76.8 | 85.1 | 86.6 | 79.4 |
| Strand | 71.1 | 77.7 | 67.1 | 71.3 | 77.9 | 67.8 |
| Coil | 80.7 | 75.5 | 61.2 | 83.6 | 79.5 | 64.6 |
| $Q_3$ | 79.5 | — | — | 81.4 | — | — |

Ensembles of 45 BRNNs, same architecture as in Pollastri and McLysaght (2005).

$Q_{class}obs$, the percentage of residues in a given class that are correctly predicted; $Q_{class}pred$, the percentage of residues *predicted* in a given class that are correctly predicted.

TABLE 8. CONFUSION MATRIX FOR SECONDARY STRUCTURE PREDICTION INTO 3 CLASSES, HARD ASSIGNMENT

| | *Hobs* | *Eobs* | *Cobs* | *Tot pred* |
|---|---|---|---|---|
| Hpred | 20,216 | 420 | 2,872 | 23,508 |
| Epred | 498 | 11,309 | 2,744 | 14,551 |
| Cpred | 3,463 | 4,172 | 23,480 | 31,115 |
| Tot obs | 24,177 | 15,901 | 29,096 | 69,174 |

Xpred, structure X is predicted. Yobs, structure Y is observed.

The number in row Xpred and column Yobs represents the number of residues for which structure Y is observed and structure X is predicted.

TABLE 9. CONFUSION MATRIX FOR SECONDARY STRUCTURE PREDICTION INTO 3 CLASSES, EASY ASSIGNMENT IN PEDERSEN ET AL. (2000)

| | *Hobs* | *Eobs* | *Cobs* | *Tot pred* |
|---|---|---|---|---|
| Hpred | 18,540 | 277 | 2,593 | 21,410 |
| Epred | 334 | 10,732 | 2,713 | 13,779 |
| Cpred | 2,913 | 4,048 | 27,024 | 33,985 |
| Tot obs | 21,787 | 15,057 | 32,330 | 69,174 |

Xpred, structure X is predicted. Yobs, structure Y is observed.

The number in row Xpred and column Yobs represents the number of residues for which structure Y is observed and structure X is predicted.

Porter outperforms the main other state-of-the-art public servers on an independent set. Porter currently has the highest performance of all servers evaluated by EVA (Rost and Eyrich, 2001). Porter is trained in 5-fold cross validation, with the first fold's training and test sets being identical to the training and test set used in this work. Porter's architecture is identical to the architecture we present here. These two elements ensure that the comparison between the system presented here and Porter's first fold is fair, and that every gain that we obtain originates from the adoption of a different predictive pipeline.

The results of the comparison for the CASP assignment are reported in Table 10. The overall correct prediction of the structural motif $\rightarrow$ secondary structure pipeline exceeds Porter's by 0.4%. To estimate the statistical significance of this result, we measure the standard deviation of the error distribution by sampling

TABLE 10. PREDICTION OF SECONDARY STRUCTURE

| | $(\phi - \psi \rightarrow SS)$ | | | *Porter's fold 1* | | |
|---|---|---|---|---|---|---|
| | $Q_{class}$ | | | $Q_{class}$ | | |
| | *Obs* | *Pred* | $C_{class}$ | *Obs* | *Pred* | $C_{class}$ |
| Helix | 83.6 | 86.0 | 76.8 | 82.8 | 86.0 | 76.2 |
| Strand | 71.1 | 77.7 | 67.1 | 70.0 | 78.0 | 66.7 |
| Coil | 80.7 | 75.5 | 61.2 | 81.3 | 74.8 | 60.9 |
| $Q_3$ | 79.5 | — | — | 79.1 | — | — |

Comparison between the $\phi - \psi \rightarrow$ secondary structure pipeline and Porter's fold 1 (Pollastri and McLysaght, 2005), HARD assignment.

$Q_{class}obs$, the percentage of residues in a given class that are correctly predicted. $Q_{class}pred$, the percentage of residues *predicted* in a given class that are correctly predicted.

with replacement $N$ residues from the S435 set $M$ times. In our case $M = 1000$ and $N = 69,174$ (the size of the set). We obtain nearly identical standard deviations of 0.15% for the error of both predictors. Given these deviations, the observed difference of 0.4% is significant at $p = 0.05$.

## 4. CONCLUSION

One of the conclusions of the now six CASP competitions is that our progress towards protein structure prediction so far was mainly based on incremental improvements, rather than large leaps ahead. Another conclusion seems to be that, in order to maximize our chance of success, predictive systems need to be built that integrate a vast number of sources of information, be these derived from theoretical laws, or based on higher levels of abstraction and inferred from the available data sets of examples through learning or statistical techniques. In particular, in the latter case, the trend over the years has been towards increasing the number of representations adopted (e.g., secondary structure, solvent accessibility, coordination number, disordered regions, residue contact maps, coarse contact maps), and increasing the resolution of these predictions (e.g., from 3-class to $n$-class secondary structure, from 2- or 3-class solvent accessibility to $n$-class or even direct prediction of surface area).

In this work we followed both directions: we adopted a new set of alphabets of structural motifs describing a protein's backbone configuration based on $\phi$-$\psi$ angles; these alphabets represent subtle, high-resolution classification schemes. We developed systems for the prediction of structural motifs from a protein's primary sequence. These architectures achieve large gains (28–30%) over base-line statistical predictors. Adopting their predictions as the first stage towards prediction of 3-class secondary structure yields a state-of-the-art system, classifying correctly 79.5–81.4% residues, and outperforming a copy of Porter, a state-of-the-art predictor (Pollastri and McLysaght, 2005), trained in identical, rigorous experimental conditions.

Our structural motif predictors may feed in a number of other stages of *ab initio* protein structure prediction systems in at least three ways:

- They may be used as an additional input to predictors of other structural features (e.g., solvent accessibility, residue contact maps). This may lead to improved prediction of these features, and ultimately translate into improved structure predictions.
- They may be directly adopted, in combination with other features, to guide the *ab initio* reconstruction of protein backbones.
- They may help selecting and ranking "decoys" obtained by any *ab initio* structure prediction method.

The predictor of structural motifs for tetra-peptides described in this work is available as a public web server accessible at *http://distill.ucd.ie/*.

## ACKNOWLEDGMENTS

We wish to thank Brett Becker and Quan Le for useful suggestions. This work is supported by Science Foundation Ireland grants 04/BR/CS0353 and 05/RFP/CMS0029, grant RP/2005/219 from the Health Research Board of Ireland, a UCD President's Award 2004, and an Embark Fellowship from the Irish Research Council for Science, Engineering and Technology to A.V.

## REFERENCES

Altschul, S.F., Madden, T.L., and Schaffer, A.A. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.

Baldi, P., Brunak, S., Frasconi, P., et al. 1999. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15, 937–946.

Baldi, P., and Pollastri, G. 2003. The principled design of large-scale recursive neural network architectures—dag-rnns and the protein structure prediction problem. *J. Mach. Learn. Res.* 4, 575–602.

Baldi, P., Pollastri, G., Andersen, C.A.F., et al. 2000. Matching protein $\beta$-sheet partners by feedforward and recurrent neural networks. *Proc. 2000 Conf. Intell. Syst. Mol. Biol. (ISMB00)* 25–36.

Bystroff, C., and Baker, D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281, 565–577.

Bystroff, C., Thorsson, V., and Baker, D. 2000. Hmmstr: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* 301, 173–190.

Crooks, G.E., Hon, G., Chandonia, J.M., et al. 2004. Weblogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.

de Brevern, A.G., Etchebest, C., and Hazout, S. 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 41, 271–287.

de Brevern, A.G., Etchebest, C., and Hazout, S. 2004. Local backbone structure prediction of proteins. *In Silico Biol.* 4, 31.

Frasconi, P., Gori, M., and Sperduti, A. 1998. A general framework for adaptive processing of data structures. *IEEE Trans. Neural Networks* 9, 768–786.

Hobohm, U., Schard, M., Schneider, R., et al. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science* 1, 409–417.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrix. *J. Mol. Biol.* 292, 195–202.

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

Karchin, R., Cline, M., Mandel-Gutfreund, Y., et al. 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 51, 504–514.

Lesk, A.M., Lo Conte, L., and Hubbard, T.J.P. 2001. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. *Proteins* S5, 98–118.

Moult, J., Fidelis, K., Zemla, A., et al. 2003. Critical assessment of methods of protein structure prediction (casp)-round v. *Proteins* 53, 334–339.

Pedersen, T.N., Lundegaard, C., Nielsen, M., et al. 2000. Prediction of protein secondary structure at 80% accuracy. *Proteins* 41, 17–20.

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

Pollastri, G., Baldi, P., Fariselli, P., et al. 2002. Improved prediction of solvent accessibility and number of residue contacts in proteins. *Proteins* 47, 142–153.

Pollastri, G., and McLysaght, A. 2005. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21, 1719–1720.

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

Richardson, C.J., and Barlow, D.J. 1999. The bottom line for prediction of residue solvent accessibility. *Protein Eng.* 12, 1051–1054.

Riis, S.K., and Krogh, A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.* 3, 163–183.

Rost, B., and Eyrich, V.A. 2001. EVA: large-scale analysis of secondary structure prediction. *Proteins* S5, 192–199.

Rost, B., and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19, 55–72.

Sims, G.E., Choi, I., and Kim, S. 2005. Protein conformational space in higher order $\psi$-$\phi$ maps. *Proc. Natl. Acad. Sci. USA* 18, 618–621.

Vullo, A., Walsh, I., and Pollastri, G. 2006. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinform.* 7, 180.

Wood, M.J., and Hirst, J.D. 2005. Protein secondary structure prediction with dihedral angles. *Proteins* 59, 476–481.

Yang, A., and Wang, L. 2003. Local structure prediction with local structure-based sequence profiles. *Bioinformatics* 19, 1267–1274.

Address reprint requests to:
*Dr. Gianluca Pollastri*
*School of Computer Science and Informatics*
*University College Dublin*
*Belfield, Dublin 4, Ireland*

*E-mail:* gianluca.pollastri@ucd.ie