PROFcon: novel prediction of long-range contacts
Marco Punta\textsuperscript{1,3,*} and Burkhard Rost\textsuperscript{1,2,3}

\textsuperscript{1}CUBIC and \textsuperscript{2}NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA and \textsuperscript{3}Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 Street Nicholas Avenue, New York, NY 10032, USA

Received on January 4, 2005; revised on April 8, 2005; accepted on April 14, 2005
Advance Access publication May 12, 2005

ABSTRACT

Motivation: Despite the continuing advance in the experimental determination of protein structures, the gap between the number of known protein sequences and structures continues to increase. Prediction methods can bridge this sequence–structure gap only partially. Better predictions of non-local contacts between residues could improve comparative modeling, fold recognition and could assist in the experimental structure determination.

Results: Here, we introduced PROFcon, a novel contact prediction method that combines information from alignments, from predictions of secondary structure and solvent accessibility, from the region between two residues and from the average properties of the entire protein. In contrast to some other methods, PROFcon predicted short and long proteins at similar levels of accuracy. As expected, PROFcon was clearly less accurate when tested on sparse evolutionary profiles, that is, on families with few homologs. Prediction accuracy was highest for proteins belonging to the SCOP alpha/beta class. PROFcon compared favorably with state-of-the-art prediction methods at the CASP6 meeting. While the performance may still be perceived as low, our method clearly pushed the mark higher. Furthermore, predictions are already accurate enough to seed predictions of global features of protein structure.

Availability: http://www.predictprotein.org/submit_profcon.html
Contact: punta@cubic.bioc.columbia.edu
Supplementary information: http://www.rostlab.org/results/2005/profcon

INTRODUCTION

Protein three-dimensional (3D) structure is one key to understanding biological function. Structures unravel details needed to engineer residue mutations and to design protein-specific ligands. Thus, the knowledge of protein structure can impact medical and clinical research. Structural genomics is a large-scale effort that has the experimental determination of most unknown folds as one aim (Friedberg et al., 2004; Liu and Rost, 2002; Portugaly et al., 2002; Rost, 1998; Shapiro and Lima, 1998). Structural genomics consortia are progressing rapidly and already contribute every third (J. Liu and B. Rost, unpublished data) of the sequence-unique structures added to the Protein Data Bank (PDB) (Berman et al., 2002) of 3D structures. Nevertheless, the sequence–structure gap, that is, the difference between the number of proteins with known sequence and those with known structure is increasing much more rapidly. Methods that predict aspects of protein structure continue to be a crucial means of obtaining structural information that helps in the unraveling of protein function (Goldsmith-Fischman and Honig, 2003; Skolnick and Fretow, 2000; Thornton, 2001; Zhang et al., 1999).

Despite significant advances over the last years, computational biology can still not reliably generate biologically meaningful 3D models for proteins that have no detectable homology to proteins of known structure (Moule et al., 2003). In the absence of a reliable solution to the protein structure prediction problem, developers have addressed simplified problems, such as the prediction of protein secondary structure and solvent accessibility; such methods have evolved into successful, automatic tools that continue to significantly impact experimental and computational biology (Rost, 2001; Rost et al., 2003). One problem with these methods is that they only predict local features of 3D structure; these local features are further simplified by projecting 3D information onto a 1D representation of protein structure (e.g. strings of secondary structure) that neither captures the 2D information of contacts between residues, nor local ‘irregularities’, such as bends of helices. In contrast, 2D maps of distances between residues reduce the dimensionality of the problem in a way that, in principle, allows the reconstruction of the full 3D structure (Galaktionov and Rodionov, 1980; Havel et al., 1983; Nilges, 1995). NMR spectroscopy exploits this fact in determining structures from distance constraints. Most of the available 2D prediction methods simplify the description from real-valued distances to binary-valued contacts (below an assigned distance threshold; here we defined residues as in contact when their C beta atoms were closer than 8 Å, i.e. 0.8 nm, Methods). Contact prediction methods extract information from correlated mutations (Goebel et al., 1994; Olmea et al., 1999; Olmea and Valencia, 1997), use neural networks with (Fariselli et al., 2001) or without correlated mutations (Pollaustri and Baldi, 2002), hidden Markov models (Bystroff and Shao, 2002; Shao and Bystroff, 2003), Support Vector Machines (Zhao and Karypis, 2003) and genetic programming (MacCallum, 2004). The prediction of 2D maps continues to be a very difficult problem. Consequently, performance is rather limited. Nevertheless, automatic 2D predictions have been used successfully for the prediction of protein structure (Olmea et al., 1999; Ortiz et al., 1999; Skolnick et al., 2003). Furthermore, no matter how inaccurate 2D predictions, they are still better than constraints from the best de novo 3D prediction methods (Eyrich et al., 2003).
Here, we introduced PROFcon, a new method for predicting inter-residue contacts through a simple neural network. For the network input we mixed different sources of information most of which had been used separately in some way before (Methods). We considered information from two "windows" around two residues \(i\) and \(j\) for which the probability of a spatial contact was predicted. Each sequence position \(k\) within the two windows \((k \in \{i-n, \ldots, i, i+n, j-n, \ldots, j+n\})\) was characterized by evolutionary substitution profiles from multiple sequence alignments, conservation weights, predicted secondary structure and predicted solvent accessibility. Additionally, we used the complexity of residues \(i\) and \(j\) and classified the pair \(ij\) into one of seven classes based on their physico-chemical properties. Information from the sequence segment that connects a pair \(ij\) has been shown to correlate with the probability of contact formation (Gorodkin et al., 1999). The segment length—usually referred to as sequence separation—has been used successfully to predict contacts (Fariselli et al., 2001a; Zhao and Karypis, 2003). However, characterizing the connecting segment in more detail is likely to further improve predictions for residues that are not too far apart (Gorodkin et al., 1999). Therefore, we added a third window describing the region at the center of a segment between \(i\) and \(j\) (analogous to the windows around \(i\) and \(j\)). Finally, we introduced global features, such as the overall compositions of amino acids, predicted secondary structure composition and protein length. We evaluated the sustained level of performance of our method on a dataset of experimentally determined X-ray structures from the PDB (Berman et al., 2002; Bernstein et al., 1977). We benchmarked proteins of different lengths, different number of aligned homologous sequences and different structural classes assigned according to SCOP (Murzin et al., 1995). PROFcon performed favorably compared to other state-of-the-art contact prediction methods in the recent CASP6 (December 2004) assessment of blind predictions (O. Grana and A. Valencia, manuscript in preparation). The method is available as an Internet prediction server at http://www.predictprotein.org/submit_PROFcon.html; all datasets used for this work are at http://www.rostlab.org/results/2005/profcon/.

**SYSTEMS AND METHODS**

Datasets and cross-validation

All proteins used for the development of our methods were taken from the PDB (Berman et al., 2002), i.e. have known structure. To avoid biasing methods on account of the accidental composition of proteins in the PDB, we extracted a subset of proteins that are not clearly related in sequence. The EVA server evaluates structure prediction methods (Koh et al., 2003) and maintains a continuously updated subset of sequence-unique PDB chains [no pair of proteins in this set has HSSP-value above 0 (Rost, 1999; Sander and Schneider, 1991)]. In particular, we used the December 2003 EVA release, a set of 3201 protein chains of known structure. We removed all non-X-ray structures, all membrane and coiled-coil proteins and proteins with physical chain breaks (Gorodkin et al., 1999). We divided the proteins into three datasets: (1) for training, we selected structures with high resolution (\(\leq 2.0\)Å), (2) for cross-training, i.e. the optimization of all free neural network parameters, such as ‘stop training’ structures with low resolution in the interval (2.5–3.0 Å) and (3) for testing structures with medium resolution (2.0–2.5 Å). Owing to CPU limitations, we had to reduce the test set further by excluding all proteins longer than 400 residues. Training, cross-training and test set contained 748, 466 and 633 proteins, respectively.

**Definition of contact**

More for the sake of enabling the direct comparison with other methods than for any other reason, we choose the standard threshold for considering a pair of residues to be in contact (Galaktionov and Marshall, 1994; Goebel et al., 1994; Hubbard, 1994; Lund et al., 1997; Miyazawa and Jernigan, 1985; Sippl, 1990; Taylor and Hatrick, 1994), namely a maximal distance of 8 Å between their C-beta atoms (C-alpha for glycines).

**Neural network architecture**

We trained standard feed-forward neural networks with back-propagation and momentum term (Rost and Sander, 1993). We addressed the extremely unequal distribution of true (contact) and false (non-contact) samples by balanced training (Rost and Sander, 1993). Symmetry between the contact probabilities for the prediction between \(ij\) and \(ji\) was enforced through a simple post-processing by averaging over both raw output values (Pollastri and Baldi, 2002). In total, we used 738 input, 100 hidden and 2 output units (contact, non-contact). The input features corresponded to three different ways of describing each pair of residues (Table 1). Supplement in Supplementary Materials), we used: (1) information from the local environment of both residues, (2) information from the segment connecting \(i\) and \(j\) and (3) global information from the entire protein.

**Local information from immediate residue environment.** For each residue pair \(ij\) in a protein, the network incorporates information from all residues in two windows of size 9 centered around \(i\) and \(j\) (corresponding to the intervals \((i-4; i+4)\) and \((j-4; j+4)\)). Each residue position within the two windows was characterized by 29 input units: \(20\) for the evolutionary profile [i.e. frequency of occurrence of the 20 amino acid types at that position, as obtained from multiple sequence alignments (Przybylski and Rost, 2002; Rost, 1996)], \(1\) additional unit served as a spacer accounting for the \(N\)- and \(C\)-terminal residues (Bohr et al., 1988; Qian and Sejnowski, 1988), \(4\) units coded for the predicted secondary structure (3 for helix/strand/other and \(1\) for the reliability of the secondary structure prediction at that residue), \(3\) units for the predicted solvent accessibility (2 units for buried/exposed and \(1\) unit for prediction reliability) and, finally, \(1\) for the conservation weight (Rost, 1996). Alignments were obtained through PSI-BLAST (Altschul et al., 1997) using our standard protocol of three automatic iterations (Przybylski and Rost, 2002) and then filtering the aligned sequences at \(80\%\) sequence identity, i.e. any two sequences with \( \geq 80\%\) percentage pairwise sequence identity were removed at the end. We used PROPhd (Rost, 2001, 2005; Rost and Liu, 2003) to predict secondary structure and solvent accessibility. Note that we trained and tested on predicted rather than observed values for 1D structure to account for the fact that secondary structure predictions are more correlated to each other than they are to observed secondary structure (Przybylski and Rost, 2004). As a consequence, using predictions both in training and testing can result in a more coherent input to the networks and hence can help to ease the classification task [note that similar advantages hold for the prediction of secondary structure itself (Rost, 1996, 2005; Rost and Sander, 1993)]. As the two local windows together accounted for 18 residue positions, we needed a total of 522 input units for their description (18x29). We also introduced additional features to better characterize the central residues \(i\) and \(j\), namely a coarse-grained bio-physical classification (Creighton, 1992) (7 input units: hydrophobic–hydrophobic, polar–polar, charged–charged, opposite charges, same charges, aromatic–aromatic, other) and we specified whether or not \(i\) and \(j\) were in low-complexity regions [according to SEG program (Wootton and Federhen, 1996), 2 input units].

**Local information from connecting segment.** Since the centre of segments that connect two residues \(i\) and \(j\) has been shown to be most informative for the contact formation between these two residues (Gorodkin et al., 1999), we introduced another window of five consecutive residues that spanned the interval \((i+(j-1)/2) - 2 \text{; }\text{int}(j+(j-1)/2)\). Each residue in this window was characterized by the same information as used for the windows around \(i\) and \(j\), i.e. we used 29 input units for each residue. Also, it has been shown that the probability for contact formation decreases as sequence
separation increases (Fariselli and Casadio, 1999; Galaktionov and Marshall, 1994; Hubbard, 1994; Lifson and Sander, 1979). Therefore, we also had to encode the length of the segment that connects $i$ and $j$; for this, we used 11 input units that corresponded to sequence separations of 6, 7, 8, 9, 10–14, 15–19, 20–24, 25–29, 30–39, 40–49 and >49 (values chosen by intuition instead of by optimization). Finally, we added features that described the entire sequence, namely its amino acid composition (20 units), its secondary structure composition (3 units) and the fraction of SEG-low-complexity (Wootton and Federhen, 1996) residues in that segment (1 unit). Overall, we used 180 input units to describe the connecting segment.

**Global information.** The use of global information can help the network to ‘decide’ whether or not two residues are in contact. For example, knowing that the protein is very short should increase the probability of having a contact between two cysteine residues (disulfide bridges); knowing that the protein is longer than the average domain length (~100 residues) should decrease the probability of long-range contacts (fewer inter-domains contacts). Here, we used only very coarse-grained global features, namely 20 + 3 units to describe the composition of amino acids and secondary structure of the entire protein, and 4 units to describe the protein length [intervals 1–61, 61–120, 121–240 and >241; again, values were not optimized but chosen identically to our PHD methods (Rost, 1996; Rost and Sander, 1993, 1994)].

**Measuring performance**

Many measures have been introduced to evaluate the performance of 2D predictions. We applied criteria that were basically identical to those used at CASP/CAFAASP (Eyrich et al., 2003; Fariselli et al., 2001b; Fischer et al., 1999, 2001, 2003). Here, we only brieﬂy sketched these scores that are described in detail elsewhere (Eyrich et al., 2003; Fariselli et al., 2001b; Goebel et al., 1994; Olmea et al., 1999). Accuracy (also referred to as ‘speciﬁcity’) was deﬁned by:

$$\text{Acc} = \frac{TP}{TP + FP} = \frac{NC_{\text{ok}}}{NC_{\text{pred}}} \quad (1)$$

where $NC_{\text{ok}}$ is the number of correctly predicted contacts, i.e. the true positives (TP), $NC_{\text{pred}}$ is the number of predicted contacts, which corresponds to the sum of true positives (TP) and false positives (FP). We also followed the tradition to evaluate performance on a number of predicted contacts that is proportional to a fraction of the protein length $L$. The rationale behind this choice is that the overall number of contacts in a protein is linearly correlated to the protein length. The advantage is that accuracy estimates are related to a quantity that can be evaluated from sequence alone (hence, it is known a priori). However, in isolation accuracy alone does not suffice, instead we need to contrast it with the coverage (also referred to as ‘sensitivity’), defined as:

$$\text{Cov} = \frac{TP}{TP + FN} = \frac{NC_{\text{ok}}}{NC_{\text{obs}}} \quad (2)$$

where $NC_{\text{obs}}$ is the number of correctly predicted contacts, i.e. the true positives (TP), $NC_{\text{obs}}$ is the number of observed contacts, which corresponds to the sum of true positives (TP) and false negatives (FN). The following example illustrates the importance of considering coverage in contact predictions. If we fix the number of predictions to a fraction of the length of a protein (e.g. $L/2$), and if we assume a linear relation between the number of contacts ($NC_{\text{obs}}$) and the protein length ($L$), i.e. $NC_{\text{obs}} = \alpha + \beta L$, we can write the coverage as:

$$\text{Cov} = \frac{NC_{\text{pred}}}{NC_{\text{obs}}} \text{Acc} = \frac{0.5L}{\alpha + \beta L} \text{Acc}, \quad (3)$$

where $NC_{\text{pred}}$ is the number of predicted contacts. A simple regression on our test set (633 proteins with $L \leq 400$ and considering only sequence separations $s \geq 6$) estimated the two free parameters to be: $\alpha = -220$ and $\beta = 5$ [valid only for $L \geq 45$, see denominator in Equation (3)]. For a protein of 100 residues, we would obtain Cov $\approx 0.18 \times $ Acc, while a protein with 400 residues would yield Cov $\approx 0.11 \times $ Acc. In other words, the same level of accuracy corresponds to different coverage for proteins of different length.

**Table 1. Benefit from using connecting segments**

<table>
<thead>
<tr>
<th>Method $^a$</th>
<th>Sequence separation $^b$</th>
<th>Nprot $^b$</th>
<th>Acc $^b$</th>
<th>Err$^2$ Acc $^d$</th>
<th>Cov $^b$</th>
<th>Err$^2$ Cov $^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local only</td>
<td>6</td>
<td>633</td>
<td>29.7</td>
<td>0.5</td>
<td>8.6</td>
<td>0.1</td>
</tr>
<tr>
<td>PROFCon</td>
<td>6</td>
<td>633</td>
<td>32.4</td>
<td>0.5</td>
<td>9.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Local only</td>
<td>24</td>
<td>621</td>
<td>19.5</td>
<td>0.5</td>
<td>8.8</td>
<td>0.3</td>
</tr>
<tr>
<td>PROFCon</td>
<td>24</td>
<td>621</td>
<td>20.0</td>
<td>0.5</td>
<td>9.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$^a$Two different methods are compared: networks that use only information from the local sequence environment of two contacting residues (local only) and our system that also uses information from the connecting segment (PROFCon).

$^b$Minimal sequence separation: $s$ (=6 or 24) means that only contacts between pairs $i$, $j$ minimally $r$ residues apart are considered, i.e. $|i - j| \geq s$.

$^d$Number of proteins in respective test set.

$^c$Acc: Accuracy [aka Specificity, Equation (1)] and Cov: coverage [aka Selectivity, Equation (2)]. Note that we report numbers valid for the $L/2$ strongest predictions.

$^e$Standard error from bootstrapping.

**Table 2. Improvement through evolutionary information**

<table>
<thead>
<tr>
<th>Nali $^a$</th>
<th>Sequence separation $^b$</th>
<th>Nprot $^b$</th>
<th>Acc $^b$</th>
<th>Err$^2$ Acc $^d$</th>
<th>Cov $^b$</th>
<th>Err$^2$ Cov $^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–14</td>
<td>6</td>
<td>138</td>
<td>23.0</td>
<td>1.0</td>
<td>9.5</td>
<td>0.8</td>
</tr>
<tr>
<td>15–49</td>
<td>123</td>
<td>31.0</td>
<td>1.0</td>
<td>9.8</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>50–199</td>
<td>187</td>
<td>35.6</td>
<td>0.9</td>
<td>9.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>&gt;200</td>
<td>185</td>
<td>37.0</td>
<td>0.9</td>
<td>10.2</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>0–14</td>
<td>24</td>
<td>132</td>
<td>13.2</td>
<td>0.7</td>
<td>10.0</td>
<td>1.0</td>
</tr>
<tr>
<td>15–49</td>
<td>120</td>
<td>19.0</td>
<td>1.0</td>
<td>9.1</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>50–199</td>
<td>185</td>
<td>21.5</td>
<td>0.9</td>
<td>9.0</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>&gt;200</td>
<td>184</td>
<td>24.2</td>
<td>0.9</td>
<td>9.5</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Number of proteins in multiple sequence alignment used to extract evolutionary profiles.

$^b$As in Table 1; note all values compiled for the first $L/2$ predictions (Methods).

**Table 3. Performance versus protein length**

<table>
<thead>
<tr>
<th>$L^a$</th>
<th>Sequence separation $^b$</th>
<th>Nprot $^b$</th>
<th>Acc $^b$</th>
<th>Err$^2$ Acc $^d$</th>
<th>Cov $^b$</th>
<th>Err$^2$ Cov $^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>6</td>
<td>78</td>
<td>31.0</td>
<td>2.0</td>
<td>18.0</td>
<td>1.0</td>
</tr>
<tr>
<td>101–200</td>
<td>230</td>
<td>32.5</td>
<td>0.9</td>
<td>9.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>201–300</td>
<td>191</td>
<td>33.9</td>
<td>0.9</td>
<td>8.6</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>301–400</td>
<td>134</td>
<td>31.0</td>
<td>1.0</td>
<td>7.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>6</td>
<td>633</td>
<td>32.4</td>
<td>0.5</td>
<td>9.8</td>
<td>0.2</td>
</tr>
<tr>
<td>120</td>
<td>24</td>
<td>66</td>
<td>19.0</td>
<td>1.0</td>
<td>22.0</td>
<td>1.0</td>
</tr>
<tr>
<td>101–200</td>
<td>230</td>
<td>18.5</td>
<td>0.7</td>
<td>8.8</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>201–300</td>
<td>191</td>
<td>22.0</td>
<td>1.0</td>
<td>8.2</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>301–400</td>
<td>134</td>
<td>19.4</td>
<td>0.8</td>
<td>6.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>24</td>
<td>621</td>
<td>20.0</td>
<td>0.5</td>
<td>9.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$^a$Protein length.

$^b$As in Table 1; note all values compiled for the first $L/2$ predictions (Methods).

Therefore, even if we predict a number of contacts proportional to the length of the protein $L$, we need to report Cov along with Acc in order to capture the performance of a method.

For each score we reported the average over all proteins in the test set and the associated estimates for standard deviations of the averages obtained from bootstrapping (Efros and Tishbyarni, 1993). In Tables 1–4 we reported performance for two particular values in the minimal sequence separation ($s \geq 6$ and $s \geq 24$), i.e. two different definitions of ‘non-local’ contacts, and for a number $L/2$ of predictions. In Figure 1, we reported accuracies for different sequence separation values, and in Figure 1 Supplement
Table 4. Performance differs between structural classes

<table>
<thead>
<tr>
<th>SCOP class</th>
<th>Sequence separation</th>
<th>Nprot</th>
<th>Acc</th>
<th>Err</th>
<th>Imp</th>
<th>Cov</th>
<th>Err(Cov)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-alpha</td>
<td>6</td>
<td>131</td>
<td>24.0</td>
<td>1.0</td>
<td>14.3</td>
<td>0.8</td>
<td>11.0</td>
</tr>
<tr>
<td>All-beta</td>
<td>103</td>
<td>35.0</td>
<td>35.0</td>
<td>1.0</td>
<td>13.0</td>
<td>0.7</td>
<td>7.8</td>
</tr>
<tr>
<td>Alpha + beta</td>
<td>119</td>
<td>36.0</td>
<td>1.0</td>
<td>14.8</td>
<td>0.6</td>
<td>10.0</td>
<td>0.4</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td>169</td>
<td>35.9</td>
<td>0.8</td>
<td>19.4</td>
<td>0.5</td>
<td>8.7</td>
<td>0.2</td>
</tr>
<tr>
<td>All-alpha</td>
<td>24</td>
<td>128</td>
<td>13.5</td>
<td>0.6</td>
<td>9.4</td>
<td>0.6</td>
<td>9.9</td>
</tr>
<tr>
<td>Alpha + beta</td>
<td>103</td>
<td>17.0</td>
<td>1.0</td>
<td>8.3</td>
<td>0.6</td>
<td>6.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td>118</td>
<td>17.0</td>
<td>1.0</td>
<td>8.8</td>
<td>0.5</td>
<td>7.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td>169</td>
<td>29.0</td>
<td>1.0</td>
<td>18.5</td>
<td>0.6</td>
<td>10.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

*As in Table 1; note all values compiled for the first L/2 predictions (Methods).

Fig. 1. PROFcon accuracy versus sequence separation. Accuracy [aka specificity, Equation (1)] for three particular values of the number of predicted contacts that were considered, as common practice, these values were chosen depending on protein length L (L/20, L/10, L/5; black symbols) and the random baseline (gray circles); the random accuracy is defined as Rnd(s) = N_{\text{obs}}(s)/(L - s), where N_{\text{obs}}(s) is the number of contacts separated by a sequence separation of exactly s residues; (L - s) is the total number of residue pairs in a protein of length L, which are separated by s residues. Trivially, if fewer contacts are predicted (L/20 < L/10 < L/5), accuracy is higher. However, this increased accuracy comes at the expense of a low coverage. Note that all thresholds shown implied that the number of predicted contacts was smaller than the number of observed contacts.

We considered nine amino acids as hydrophobic (alanine, leucine, isoleucine, valine, tryptophane, phenylalanine, proline, cysteine and methionine).

RESULTS AND DISCUSSION
Connecting segment very informative for contact formation
First, we confirmed (Gorodkin et al., 1999) that the information from the segment connecting two residues i and j improves the prediction of contacts (Fig. 1, Table 1). For sequence separations > 20, accuracy was always lower than 20% (Fig. 1), while it reached almost 40% (s = 6 and L/20, Fig. 1) for ‘less global’ residues. This difference could partially be explained by the background probability (more contacts at shorter separations as illustrated by the curve for random in Fig. 1). On the other hand, it may be easier for the neural networks to learn what determines the formation of more local contacts (more samples, stronger sequence signatures, e.g. for beta turns and beta hairpins). In order to take this strong dependency of performance on sequence separation into account, we always reported values for performance for two different values of minimal sequence separation (s ≥ 6 and s ≥ 24). Note that we provided more details about the
explicit dependency of performance in the supplementary material (in particular Fig. 1_Supplement).

**Evolutionary profiles were crucial for performance**

Due to the large size of our datasets and our limited resources we could not systematically test the relevance of all the input features that we used through a leave-one-out type of test. One aspect that we investigated by training separate networks was the contribution of non-local input information, in general: networks using only local features were less accurate than our final system (Table 1). Detailed analyses of our results and preliminary work on smaller datasets suggested which of the remaining input features were most important for performance. Evolutionary information was clearly most relevant, as had been noted before (Fariselli and Casadio, 1999). Even the simplest measure for the information in a multiple sequence alignment, namely the number of proteins aligned, clearly correlated with performance, e.g. the accuracy dropped from 37% for alignments with >200 proteins to 23% for alignments with <15 proteins at sequence separations ≥6 and from 24% to 13% at separations ≥24 (Column Acc in Table 2). Note that, although accuracy correlated dramatically with the number of aligned sequences, differences in coverage (Column Cov in Table 2) were not statistically significant. This was probably related to the different protein composition of the reported subsets, in terms of length and structural class.

**Contact density dependent on type of protein**

It is well known that the contact density decreases with increasing protein length. Thus, contact predictions are more difficult for longer proteins (Fariselli et al., 2001b; Pollastri and Baldi, 2002). We observed that the contact density [Equation (4)] also depends on the structural class: all-beta proteins had the highest density \( D(6) = 0.040 \), while all-alpha proteins had the lowest contact density \( D(6) = 0.022 \). Since it is more difficult to predict low-density than high-density contacts, most existing methods for contact prediction strongly depend on protein length (Fariselli et al., 2001a; Pollastri and Baldi, 2002) and structural class (MacCallum, 2004; Zhao and Karypis, 2003).

**Similar accuracy but better performance for short proteins**

PROFcon reached surprisingly similar levels of accuracy for proteins of very different lengths (Table 3, column Acc). However, when also considering the coverage/selectivity of our predictions (Table 3, column Cov), we noted the length-dependency of our method: short proteins (especially of length <100) had, by far, the highest coverage (1.5–2 times higher than for long proteins).

**All-alpha worst and alpha/beta best**

Our test set was large enough to distinguish between the four major structural classes in SCOP (Murzin et al., 1995), namely all-alpha, all-beta, alpha/beta and alpha + beta (at least 100 proteins in each class). We found that PROFcon performed clearly worst on all-alpha, especially for shorter sequence separations (Table 4); we verified that this effect was true at levels of identical coverage (data not shown). Performance was largely similar for the other three classes with the exception of very long-range contacts (\( s ≥ 24 \)) that were predicted best in alpha/beta proteins (Table 4). A similar trend has previously been reported on a much smaller dataset (MacCallum, 2004). This trend might originate from particular strand-helix-strand modules that are abundant in alpha/beta proteins, namely those with two flanking strands that contact each other. This type of structural motif may be easy to predict, especially for a system that relies on information about the connecting segments.

**Of the predicted contacts 50% within two residues of an observed contact**

‘\( δ \) analysis’ (Methods) shows that many predicted contacts fall very close to observed contacts (Table 2_Supplement in the Supplementary Material). For example, for sequence separation \( s ≥ 6 \), 50% of all predicted contacts (L/2 predictions) are within 2 residues of an experimental contact (Table 2_Supplement).

**Correct for core, hydrophobic and regular secondary structure**

At least half of the contacts correctly predicted by PROFcon were between residues in identical regular secondary structures (helix–helix and strand–strand, Table 3_Supplement in the Supplementary Material); this was independent of the structural class and of sequence separation (Table 3_Supplement). Although most correctly predicted contacts were between regular secondary structures, all residues contacting between helices were, on average, predicted the least accurately (slightly worse than mixed). Strand–strand contacts were by far the most accurately predicted (\( >40\% \) for \( s ≥ 6 \) and \( >20\% \) for \( s ≥ 24 \), in all classes, Table 3_Supplement). In alpha/beta proteins long-range strand–strand contacts (\( s ≥ 24 \)) were predicted at levels of accuracy as high as 42% compared with 20 and 24% in all-beta and alpha + beta, respectively. This indicated that PROFcon captured a strong signal from long-range preferences determining the formation of sheets in alpha/beta proteins that may be related to strand–helix–strand modules. In analogy to contacts between regular secondary structures, contacts between hydrophobic residues also constitute most of the correctly predicted contacts (Table 4_Supplement in the Supplementary Material). Unlike for secondary structure, the contribution of hydrophobic pairs is slightly increasing for higher sequence separations. As expected, predicted contacts are on the average more buried (core residues) and less distant in sequence than observed contacts (solvent accessibility averages between 11 and 22 Å² for correctly predicted contacts, Table 4_Supplement; L/2 predictions).

**CASP6 and comparisons with other methods**

Comparing the performance of PROFcon with that of other contact prediction methods is not an easy task; different groups use different datasets and, as shown (Tables 2–4), the dataset composition (number of sequences in alignments, protein length, structural class composition) significantly alters average scores. Furthermore, different groups use different scores, often even different definitions for what is considered a long-range inter-residue contact. The only reasonable comparison of the performance of methods is based on the same scores and the same dataset. Such a dataset must be sufficiently large to contain a representative sequence-unique subset of proteins (Eyrich et al., 2001; Rost et al., 2003; Rost and O’Donoghue, 1997; Rost and Sander, 1993, 1994). Furthermore, the set should not overlap with any of the proteins used for method development. While cross-validation provides some clues, it does not suffice for rigorous comparisons. At the moment, there is no set available that meets all conditions for a comprehensive, meaningful comparison of our method with others. The best approximation might be the data from
Fig. 2. Predicted and observed contacts for two CASP6 targets: T230 (A) and T216_2 (B, second domain of target T216). Right side: simple sketch of 3D structure using VMD (Humphrey et al., 1996), helices shown as red cylinders, strands as yellow arrows; other residues are marked in blue. Left side: contact maps; the upper triangles contain the best 2 × L predictions of PROFcon (red dots); the lower triangles contain the experimental contact maps (black dots). All contacts between residues closer than 6 positions (sequence separation < 6) are removed, and the lines parallel to the diagonals indicate sequence separations of 24. The blue boxes highlight specific clusters of interactions between selected secondary structure elements in the observed structure; gray boxes highlight over-predicted (predicted, not observed) clusters of contacts between selected secondary structure elements. The labels on the boxes are taken from the 3D sketches on the right. For example, on the top panel (A) the blue box label a-b means that the highlighted area corresponds to contacts occurring between strands a and b of target T0230.

CASP6 (December 2004), with the caveat that this set was far too small to draw definite conclusions. PROFcon appeared to be one of the top three contact prediction methods at CASP6, as judged by the assessor (A. Valencia, CASP6 website; note that the limitation of the dataset did not allow any distinction in performance between the top three methods).

Contact predictions capture relevant information and are useful!
More than many other structure prediction methods, contact predictions continue to suffer from the fact that performance appears to be so low. Are 2D predictions of any use? Our predictions clearly capture important globular information as demonstrated by a method that
succeeds in predicting folding rates exclusively based on PROFcon predictions (Punta and Rost, 2005). Ortiz, Skolnick and colleagues have shown that even more noisy contact predictions provide important constraints for the prediction of 3D structures (Ortiz et al., 1999), and used by experts, contact predictions have also been shown to aid in fold recognition (Olmea et al., 1999; Pazos et al., 1999). Two particular examples from CASP6 may help elucidate what exactly is captured by 2D predictions.

Example 1. CASP6 target T0230 (Fig. 2A) is a small protein (~100 residues) characterized by the following topology: two helices, two strands (labeled A and B in Fig. 2A), one helix (labeled 1), another strand (labeled C) and one last helix. At CASP6, T0230 was classified as fold recognition analogous target (FR/A), i.e. a protein for which a structure was known in PDB, however, the similarity between the template and the known structure could not be identified through sequence homology. PROFcon strongly predicted the interaction between the two anti-parallel strands A and B that are separated by a short loop (Fig. 2). It also correctly identified the sparse cluster of interactions between helix 1 and strand C. However, it wrongly predicted the main interaction of strand C. In fact, while predicting only a few interactions between parallel strands C and B (that are in contact in the structure), it suggested a strong contact between C and A that is not observed.

Example 2. The substantially longer (~210 residues) domain labeled T0216_2 in CASP6 (Fig. 2B) is part of a two-domain protein. CASP6 assessors classified this target as a novel fold (NF), i.e. a domain for which no other domain from PDB had a structural similarity. We focused on some of the strands because this domain has a rather complicated architecture. Specifically, we considered two groups of strands labeled A, B, C and 1, 2, 3, 4 in Fig. 2B. PROFcon correctly identified contacts between pairs of strands A–B, 2–3 and 3–4, while it completely missed the interactions between A–C and 1–2. In both examples (Fig. 2), incorrect contact predictions between regular secondary structure segments often occurred far from the main diagonal of the map (i.e. at large sequence separations, see lines marking s = 24). The overall accuracy for 2 ≤ L predictions and s ≥ 6 was close to average for both T0230 (26%) and T0216_2 (21%).

These two examples seem to suggest that even when the overall prediction accuracy is rather low, PROFcon still correctly identifies contacts between regular secondary structures that are separated by <20–30 residues. This ability could be exploited by integrating the predictions into fold recognition and/or de novo prediction methods.

Unique combination of information makes the difference
Our approach to contact prediction was by no means ‘radically new’. We simply combined sources of information slightly differently from what other groups did. In doing so, we ended up with a simple neural network the size of which is not unusual for the predictions of 1D structure (Rost, 2001; Rost et al., 2003), but is slightly larger than what has previously been used to predict 2D structure. As a result we needed a very large training set with almost 400,000 positive (contact) samples. Preliminary tests demonstrated that we needed at least these many samples to fully profit from all the input features that we combined. The downside was the increase in computational complexity: the development required several terabytes and many CPU years. These constraints also impacted our ability to separately test—and possibly optimize—the various input features that we considered. One outstanding feature of PROFcon is its consistent performance across a wide range of protein lengths and for both less and more long-range contacts. This consistency was clearly related to the introduction of information from the segment connecting two contacting residues (Table 1).

Future
Major improvements from here may have to introduce ways of post-processing raw predictions. Our method did not exploit any of the constraints imposed on a contact map of an entire protein, e.g. that residues can form only a limited number of contacts. In the past, an intricate post-processing has been proposed for beta-strand pairing (Asogawa, 1997). Although the concepts embedded in HMMSTR also address this task (Shao and Bystroff, 2003), no solution exists that comprehensively refines inter-residue contact predictions.

CONCLUSIONS
Better predictions of 2D information captured by inter-residue contacts from sequence could help predict important aspects of protein structures. However, the difficulty of the task and the perceived lack of acceptable performance have so far hampered progress. We presented a new method that exploits evolutionary information in the form of multiple sequence alignments and other sequence information relevant for predicting contacts through simple neural networks. While none of our ideas revolutionized the field, the particular combination of information chosen made a significant difference in sustained prediction performance; the major novelty was the particular way in which we successfully used information other than from the sequence environment of the two contacting residues (Table 1). Our method, PROFcon, was particularly successful in its consistent predictions of contacts across a wide range of protein lengths, as well as for residues closer in sequence (separated by at least 6 residues) as for residues very far apart in sequence (separated by at least 24 residues). Nevertheless, PROFcon performed better for short proteins (Table 3), for proteins for which sequence alignment methods detected many homolog (Table 2) and for proteins with beta-strands (Table 4); it was particularly successful for alpha/beta proteins (Table 4). Overall, visual inspections for individual contact maps suggest that the predictions contain more useful information than might be expected from the low levels of accuracy and coverage.

ACKNOWLEDGEMENTS
Thanks to Jinfeng Liu and Megan Restuccia (Columbia) for computer assistance; to the EVA team, in particular, to Dariusz Przybylski, Ingrid Koh and Volker Eyrich (all Columbia), Osvaldo Grana and Alfonso Valencia (CNB Madrid). Many thanks for insightful discussions to Yanay Ofran (Columbia), Soren Brunak (CBS Copenhagen), Piero Fariselli and Rita Casadio (both Bologna University), Alfonso Valencia (CNB Madrid), Reinhard Schneider (LJON) and Chris Sander (Sloan Kettering, NYC). This work was supported by the grant R01-GM64633-01 from the NIH. Last, not least, thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego University) and their crews for maintaining excellent databases, and to all experimentalists who enabled this analysis by making their data publicly available.