# Fast Modelling of Protein Structures Through Multi-level Contact Maps.

Davide Baù[*] Ian Walsh[*] Gianluca Pollastri, Alessandro Vullo

School of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland

**Abstract**

We present an algorithm to reconstruct protein C$_\alpha$ traces from 4-class distance maps, and benchmark it on a non-redundant set of 258 proteins of length between 51 and 200 residues. We first represent proteins as contact maps, and show that even when exact maps are available, only low-quality models can often be obtained. We then adopt a more powerful simplification of distance maps: multi-class contact maps. We show that the reconstructions based on 4-class native maps are significantly better than those from binary maps. Furthermore, we build two predictors of 4-class maps based on recursive neural networks: one ab initio, or relying on the sequence and on evolutionary information; one in which homology information is provided as a further input, showing that even very low sequence similarity to PDB templates yields more accurate maps than the ab initio predictor. We reconstruct C$_\alpha$ traces based on both ab initio and homology-based 4-class map predictions. We show that homology-based predictions are generally more accurate than ab initio ones even when homology is dubious.

## 1 Introduction

Although a protein can be first characterised by its amino acid sequence, most proteins fold into three-dimensional structures that encode their function. Genomics projects leave us with millions of protein sequences, currently $\approx 6 \times 10^6$, of which only a small fraction ($\approx 2\%$) have their 3D structure experimentally determined. In the near future, we will probably devolve upon structural genomics projects in order to bridge the huge gap between sequence and structure. The current high throughput pipelines have to deal with serious bottlenecks, e.g. a large fraction of targets are found to be unsuitable for structural determination with available methods [1]. Therefore, computational protein structure prediction

[*] Contributed equally to this work

remains an irreplaceable instrument for the exploration of sequence-structure-function relationships. This is especially important for analyses at the genome or inter-genome level, where informative structural models need to be generated for thousands of gene products (or a portion of them) in reasonable amounts of time.

The faster and more reliable methods for structure prediction rely on the transfer of knowledge between closely related proteins accumulated in sequence and structure databases – the field known as template-based modelling. The algorithms employed typically adopt heuristics based on sequence and/or structural similarity to model the unknown target structure based on known structures that are fathomed to be homologous to it. Automating the modelling process is difficult: there are several stages and critical points in the design (choice of templates, the creation of a correct structural alignement etc.) and for some of them manual intervention is at least helpful. The accuracy of template-based techniques strongly depends on the amount of detectable similarity, thus preventing the reliable application of these methods for a significant fraction of unannotated proteins. This is the realm of the so called *ab initio* or de novo protein structure prediction, where models are predicted not relying on similarity to proteins with known structure. Ab initio techniques are obviously not as accurate as those based on templates, but the design in this case is generally much simpler. Moreover, improvements can be obtained by relying on fragment-based algorithms [2], that use fragments of proteins of known structure to reconstruct the complete structure of the target protein. A system for the prediction of protein structures ab initio is generally composed of two elements: an algorithm to search the space of possible protein configurations to minimise some cost function; the cost function itself, composed of various constraints being either derived from physico-chemical laws, experimental results, or being structural features (e.g. secondary structure or solvent accessibility) predicted by machine learning or other kinds of statistical systems [3].

We describe and benchmark all the components of a fully-automated system for protein structure modelling which is fast and simple in the design (modular, few stages). The same protocol is applied whether or not the unknown input protein shares significant levels of similarity to other proteins with known structure, and is based on two steps, solved efficiently. Given the input protein, we first encode information about the family of homologous sequences and possibly structures. Sequence information has the form of profiles extracted from multiple sequence alignemnts. Unlike the usual template-based methodology, there is no a priori choice of the best available templates used to model the unknown structure. For each position of the input sequence, structural information from putative templates (if present) is carefully weighed according to the quality of their respective structures and the amount of similarity. Based on sequence and structural information, we make inferences about the geometry of the unknown structure by predicting a set of soft constraints by machine learning. The unknown structure is found in the second and final stage. Here, the system features the typical structure of ab initio methods, where modelling occurs as a result of searching the configurational space of 3D structures with a suitable potential or pseudo-energy function. At the current stage of development, the potential is a non linear function of the soft constraints found in the first stage, with few parameters and simple enough to be globally optimised by quick Monte Carlo searches using a linear schedule. In order to keep the simulations within manageable times, protein structures are represented by the trace of their backbone $C_\alpha$ atoms, bearing in mind it is hard to derive a meaningful energy model for such stripped-down representation of a protein. We overcome this problem by relying on informative geometrical constraints to discern native-like protein conformations from unfolded, or incorrectly folded ones. The constraints predicted in the first stage have the form of residue based pairwise distance attributes labelled into two or more classes.

In the past, research has focussed on studying binary contact maps (i.e. two classes, contact or not). It is generally believed that binary maps provide sufficient information to unambiguosly reconstruct native or near-native models [4]. Efforts have therefore been put on the prediction of this kind of distance restraints. Unfortunately, the expected success rates of the most promising techniques developed for this problem have not improved to satisfactory levels, despite years of attempts [5]. The reason for this is at least twofold. First, contact map prediction is an unbalanced problem, with far fewer contacts than non-contacts. Especially for long-range contacts (i.e. those between amino acids that are tens or hundreds of positions apart in the sequence) the ratio between negative and positive examples can exceed 100. Second, contact map predictors are generally ab initio, i.e. do not exploit all available information. Another problem with binary contact maps is that, although it has long been stated that native maps yield correct structures, this is true only at a relatively low resolution (3-4Å on average, in the best case).

In this chapter, we introduce a representation of protein structures based on a generalisation of binary contact maps, multi-class distance maps, and show that it is powerful and predictable with some success. Our tests suggest that multi-class maps, when using experimental restraints, are informative enough to quickly guide simple optimisation searches to nearly correct models - significantly better than with binary contact maps. We compare reconstructions based on binary and multi-class maps on a non-redundant set of 258 proteins of length between 51 and 200 residues. The reconstructions based on multi-class maps have an average Root Mean Square Deviation (RMSD) of roughly 2 Å and a TM-score of 0.83 to the native structure (4 Å and 0.65 for binary maps).

We then develop high-throughput systems for the prediction of multi-class contact maps, which exploit similarity to proteins of known structure, where available, in the form of simple structural frequency profiles from sets of PDB templates. We build two predictors of multi-class maps based on recursive neural networks: one ab initio, or relying on the sequence and on evolutionary information; one in which homology information is provided as a further input. We show that even very low sequence similarity to PDB templates (PSI-BLAST e-value up to 10) yields more accurate maps than the ab initio predictor. Furthermore, the predicted map is generally more accurate than the maps of the templates, suggesting that the combination of sequence and template information is more informative than templates alone. Finally, the optimisation search protocol we developed is benchmarked using both ab initio and homology-based multi-class map predictions. We show that homology-based predictions are generally more accurate than ab initio ones even when homology is dubious, and that fair to accurate protein structure predictions can be generated

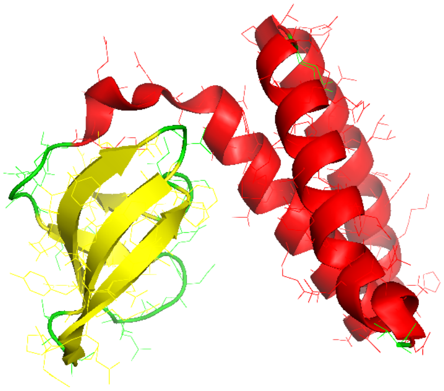

Figure 1: Full atoms three dimensional structure for protein 1SF9. While computationally intense, the full atoms representation allows to display all structural information, like side chains orientation and secondary structure elements.

for a broad range of homology to structures in the PDB.

Using the current reconstruction protocol, hundreds of reconstructions for the same protein can be performed in few minutes on current machines. On small clusters of machines it is possible to perform predictions on a genomic scale in few hours for simpler organisms, or few days for the most complex ones.

## 2 Representing Protein Structures

Protein three-dimensional (3D) structures are fully represented by the coordinates of their atoms. For a protein with $N$ atoms, $3N$ coordinates are then needed to describe its 3D structure (Figure 1). Although this is the ideal representation, it has the drawback of yielding a computationally intense model. Simplified representations have been proposed before, in which an amino acid is typically described by fewer points than the atoms it contains, thus reducing the degrees of freedom of the model. Typical simplified representations include backbone only models, where all the side chain atoms are excluded, and virtual atom models, where each residue in the sequence is assigned a virtual (i.e. geometrical, not physical) point, to represent a subset of its atoms [6]. At the extreme of the above cases are representations with only one point per amino acid, typically the $C_\alpha$ atom (see Figure 2), or $C_\beta$ atoms. This way the degrees of freedom needed to represent a protein of $N$ atoms and $M$ amino acids are reduced to $3M$, with $M << N$.

The prediction algorithm described in this work represents protein structures as the trace of their backbone $C_\alpha$ atoms, one for each amino acid of the sequence. An obvious advantage of this choice is its extreme simplicity, given that one order of magnitude fewer points are

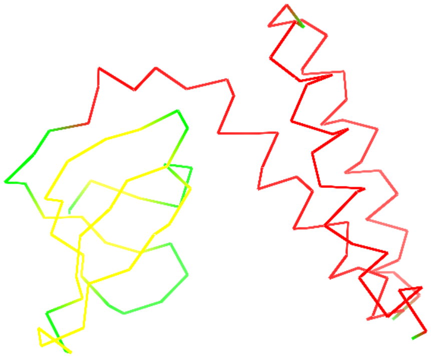

Figure 2: C$\alpha$ trace for protein 1SF9. Although computationally less intense than a full atom model (it uses one order of magnitude fewer points, and yields two orders of magnitude fewer interactions), using a representaion from which most of the native protein details have been removed has the drawback of making it hard to derive a meaningful energy model.

used, which yields two orders of magnitude fewer interactions than a full atom model. It is worth noting that reliable full-atom model can be generally derived from $C_\alpha$ traces close to the native ones, for instance by refining them using molecular dynamics simulations, or optimisations of detailed energy functions applied to full atom models predicted from the backbone [7]. The real difficulty is to derive a meaningful energy model for a protein from which most of the details have been removed, to effectively explore the search space from random initial configurations. Here we overcome the problem by relying entirely on non-physical (i.e. geometrical) constraints to discern good (native-like) conformations from bad (unfolded, or incorrectly folded) ones. Although this is a simplified goal, we show that success in this task generally yields informative predictions.

The potentials we develop in the next section are based on terms measured using another simplified representation of protein structures, the contact map. A protein's contact map belongs to a class of two-dimensional (2D) projections of 3D representations of geometrical objects. In the next subsection we give a brief description of contact maps, particularly focussing on multi-class ones, which are the representations our prediction algorithm relies on.

## Distance matrix and contact map

Using two-dimensional projections of 3D objects is an attractive way of encoding geometrical information of protein structures, as these are scale and rotation invariant and do not depend on the coordinate frame. Therefore, 2D projections can be modelled as the output variable of learning or statistical systems trained in a supervised fashion, i.e. using sam-

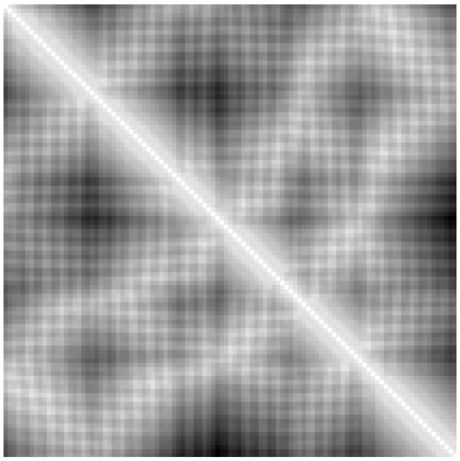

Figure 3: Distance matrix in greyscale image format. White is 0 Å and black is the maximum distance in the protein.

ples of (input, target) pairs collected from structure databases. We call this encoding of a structure 2D because it can be graphically represented as a two-dimensional matrix, where the cells denote properties of pairs of objects in the 3D space. In the case of proteins, the geometrical relationship may involve fragments of the structure at different scales, using for instance amino acid [8] or secondary structure segment pairs [9], the former being much more common than the latter. Contact maps at 8Å have been assessed as a special category at CASP for several years [5].

Geometrical relationships between amino acids can be expressed as a set of distance restraints, e.g. in the form $L \leq d(i,j) \leq U$, where $d(i,j)$ is the distance between residues in positions $i$ and $j$ and $L$ (resp. $U$) is lower (resp. upper) bound on the distance. Restraints such as the above ones can be experimentally determined, e.g. from NMR experiments. Indeed, algorithms for modelling protein structures from distance restraints are borrowed from the NMR literature and use for instance stochastic optimisation methods [4, 10], distance geometry [11, 12], and variants of them [13–15].

There is a trade-off between the resolution of the input restraints, e.g. the uncertainty with which they specify the property of the pairs, and the ability of the reconstruction algorithm to recover the correct model from these inputs. In the best case, the complete noise-free distance matrix is available, and the optimisation problem can be solved analyt-

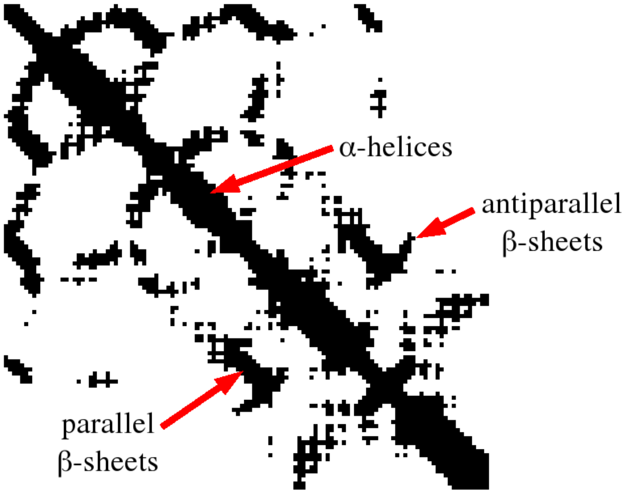

Figure 4: Different secondary structure elements like helices (thick bands along the main diagonal) and parallel – or anti-parallel – $\beta$-sheets (thin bands parallel – or anti-parallel – to the main diagonal) are easily detected from the contact map.

ically by finding a 3D embedding of the 2D restraints. A distance matrix consists in the set $\{d(i,j)\}_{i>j}$ of $N(N-1)/2$ distances between any two points in positions $i$ and $j$ of a protein with $N$ amino acids. Note how the distance matrix corresponds to the above form of constraints with lower distance bound equal to the upper one. Figure 3 shows a greyscale picture of the distance matrix of the protein with PDB code 1ABV, where the distances are calculated between the $C_\alpha$ atoms.

The distance matrix or even detailed distance restraints cannot be reliably determined by means of computational techniques, unless experimental data is available or when there is strong homology to proteins with known structure. This is why in the past research has focussed on predicting representations of the distance matrix which are at the same time simpler to learn and able to retain substantial structural information. The contact map of a protein is a labelled matrix derived by thresholding the distance matrix and assigning labels to the resulting discrete intervals. The simplest alternative is the binary contact map, where one distance threshold $t$ is chosen to describe pairs of residues that are in contact ($d(i,j) < t$) or not ($d(i,j) \geq t$). The binary contact map can also be seen as the adjacency matrix of the graph with $N$ nodes corresponding to the residues. Binary contact maps are popular as noise-tolerant alternatives to the distance map, and algorithms exist to recover protein structures from these representations [4,16,17]. Unfortunately, our studies and other empirical evidence indicates that recovering good-quality models even from the binary map of the native fold is difficult [17] The definition of contact among amino acid is based on a single atom (normally $C_\alpha$ or $C_\beta$) and depends on a geometrical threshold. This may be ambigous in situations where other knowledge must be taken into account, for instance when the orientation of the side-chain is important.

Although numerous methods have been developed for binary contact map prediction [18–24], improvements are only slowly occurring (e.g. in [21], as shown by the CASP6 experiment [25]). Accurate prediction is far from being achieved and limitations of existing prediction methods have again emerged at CASP7 and from automatic evaluation of structure prediction servers such as EVA [26]. There are various reasons for this: the number of positive and negative examples (contacts vs. non contacts) is strongly unbalanced; the number of examples grows with the squared length of the protein making this a tough computational challenge; capturing long ranged interactions in the primary sequence is difficult, hence grasping an adequate global picture of the map is a formidable problem.

Based on the above considerations, we believe that alternative representations of protein topologies are particularly appealing, provided that they are informative and, especially, predictable. Here we focus on a representation of the distance matrix called multi-class contact map and based on a set of categorical attributes or classes. Each class corresponds to an interval of distances (in Å) where a given pair of residues may fall into. Formally, given a set of distance thresholds $\{t_k\}_{k=0\ldots T}$ (where $t_0 = 0$ and $t_T = \infty$), a multi-class contact map of a protein with $N$ amino acids is a symmetrix $N \times N$ matrix $\mathbf{C}$ where the element corresponding to the amino acids in positions $i$ and $j$ is defined as $\mathrm{C}_{ij} = k$ if $d(i,j) \in [t_k, t_{k+1})$. Obviously, this class of projections contains richer information than binary contact maps (so long as $T > 3$). Therefore, using multi-class contact maps is expected to improve the resolution of reconstruction algorithms on geometrical constraints. Moreover, if a suitable set of distance thresholds is chosen, the number of instances in each class may be kept approximately balanced, which in turn may improve generalisation performances of learning algorithms over the (normally unbalanced) binary prediction case.

For our experiments, we derived a set of five distance thresholds to define multi-class contact maps based on four distance intervals. As shown in Figure 5, the four classes are empirically chosen from the distribution of distances among amino acids in the training set, ignoring trivial pairs $|i - j| \leq 3$ and by trying to keep informative distance constraints and the classes as balanced as possibile. The resulting set of thresholds is $\{0, 8, 13, 19, \infty\}$, which defines suitable distance intervals corresponding to short ($[0, 8)$), medium ($[8, 13)$, $[13, 19)$) and long-ranged interactions among amino acids. A potential improvement be-

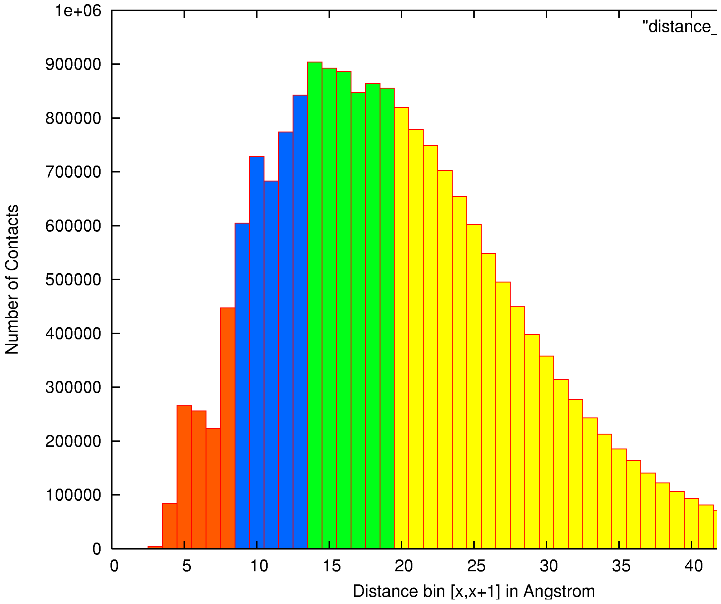

Figure 5: Distribution of contacts in [0,50] distance bins with trivial $|i - j| \leq 3$ residue contacts ignored. The classes were chosen in order to retain good distance constraints and balanced classes. Class 1 ([0,8)Å) corresponds, as a first approximation, to physical contacts.

yond this choice is to automatically determine an optimal set of thresholds based on some criteria, e.g. the reconstruction ability on a set of benchmarking proteins.

# 3 Modelling Structures with Contact Maps

We predict protein models by solving a global optimisation problem, where a function (pseudo-energy) is minimised by searching the configurational space of 3D structures. The pseudo-energy function we use to guide the search is designed in a way that allows us to solve an unconstrained minimisation problem by a simple simulated annealing protocol. More specifically, the pseudo-energy function measures the degree of match of a protein conformation to the constraints encoded in the contact map (binary or multi-class) predicted in the first stage. In the following, we describe the set of moves used to explore the configurational space and the different forms of potential functions used respectively for binary and multi-class contact maps.

## 3.1 Optimisation Algorithm

The algorithm we use for the reconstruction of the coordinates of protein C$_\alpha$ traces is organised in two sequential phases, *bootstrap* and *search*. The function of the first phase is to generate an initial physically realisable configuration. A random structure is created using a self-avoiding random walk and explicit modelling of predicted helices, by adding C$_\alpha$ positions one after the other until an initial draft of the whole backbone is obtained. More specifically, this part runs through a sequence of $N$ steps, with $N$ being the length of the input chain. At stage $i$, the position of the $i$-th C$_\alpha$ is computed as $r_i = r_{i-1} + d\frac{r}{|r|}$ where $d \in [3.733, 3.873]$ and $r$ is a random direction vector. Both $d$ and $r$ are uniformly sampled. If the $i$-th residue is predicted at the beginning of an helix, all the following residues in the same segment are modelled as an ideal helix with random orientation.

In the *search* step, the algorithm refines the initial bootstrapped structure by global optimisation of a pseudo-potential function using local moves and a simulated annealing protocol. Simulated annealing is a good choice in this case, since the constraints obtained from various predictions are in general not realisable and contradictory. Hence the need for using a "soft" method that tries to enforce as many constraints as possible never terminating with failure, and is robust with respect to local minima caused by contradictions. The search strategy is similar to that in [4], but with a number of modifications. At step $t$ of the search, a randomly chosen C$_\alpha$ atom at position $r_i^{(t)}$ is displaced to the new position $r_i^{(t+1)}$ by a crankshaft move (Figure 6), leaving all the other C$_\alpha$ atoms of the protein in their original position. Secondary structure elements are displaced as a whole, without modifying their geometry (Figure 7). The move in this case has one further degree of freedom in the helix rotation around its axis. This is assigned randomly, and uniformly distributed. A new set of coordinates $\mathcal{S}^{(t+1)}$ is accepted as the best next candidate with probability $p = min(1, e^{\Delta \mathrm{C}/T^{(t)}})$ defined by the annealing protocol, where $\Delta\mathrm{C} = \mathrm{C}(\mathcal{S}^{(t)}, \mathcal{M}) - \mathrm{C}(\mathcal{S}^{(t+1)}, \mathcal{M})$ and $T^{(t)}$ is the temperature at stage $t$ of the schedule.

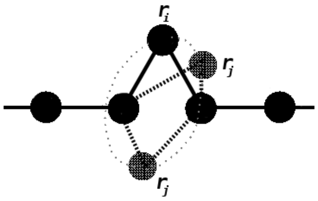

Figure 6: Crankshaft move: a randomly chosen C$_\alpha$ atom at position $r_i^{(t)}$ is displaced to the new position $r_i^{(t+1)}$ leaving all the others C$_\alpha$ atoms of the protein in their original position.

### 3.2 Pseudo-energy function

Let $\mathcal{S}_n = \{r_i\}_{i=1\dots n}$ be a sequence of $n$ 3D coordinates, with $r_i = (x_i, y_i, z_i)$ the coordinates of the $i$-th C$_\alpha$ atom of a given conformation related to a protein $p$. Let $\mathcal{D}_{\mathcal{S}_n} = \{d_{ij}\}_{i<j}, d_{ij} = \|r_i - r_j\|_2$, be the corresponding set of $n(n-1)/2$ mutual distances between C$_\alpha$ atoms. A first set of constraints comes from the (predicted) contact map and depends on the type of contact maps, i.e. binary (see section 3.2.1) or multi class maps (see section 3.2.2). The representation of protein models induces the constraints $\mathcal{B} = \{d_{ij} \in [3.733, 3.873], |i-j| = 1\}$, encoding bond lengths, and another set $\mathcal{C} = \{d_{ij} \geq D_{HC}, i \neq j\}$ for clashes. The set $\mathcal{M} = C \cup \mathcal{B} \cup \mathcal{C}$ defines the configurational space of physically realisable protein models.

#### 3.2.1 Binary contact map constraints

When using binary contact maps the set of contraints coming from the predicted maps can be represented as a matrix $C = \{\mathrm{c}_{ij}\} \in \{0,1\}^{n^2}$. Let $\mathcal{F}_0 = \{(i,j) \mid d_{ij} > d_T \wedge c_{ij} = 1\}$ denote the pairs of amino acid in contact according to $C$ (binary case) but not in $\mathcal{S}_n$ ("false negatives"). Similarly, define $\mathcal{F}_1 = \{(i,j) \mid d_{ij} \leq d_T \wedge c_{ij} = 0\}$ as the pairs of amino acids in contact in $\mathcal{S}_n$ but not according to $C$ ("false positives"). The objective function is then

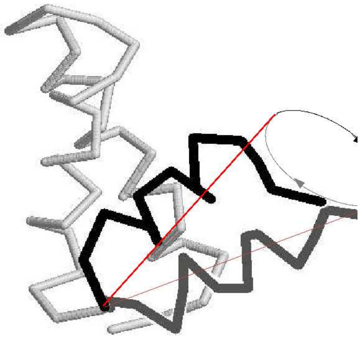

Figure 7: Secondary structure elements are displaced as a whole, without modifying their geometry.

defined as:

$$\begin{aligned} \mathrm{C}(\mathcal{S}_n, \mathcal{M}) &= \alpha_0\{1 + \sum_{(i,j)\in\mathcal{F}_0} (d_{ij}/D_T)^2 + \sum_{(i,j):d_{ij}\notin\mathcal{B}} (d_{ij} - D_B)^2\} \\ &+ \alpha_1|\mathcal{F}_1| + \alpha_2 \sum_{(i,j):d_{ij}\notin\mathcal{C}} e^{(D_{HC}-d_{ij})} \end{aligned} \quad (1)$$

### 3.2.2 4-class contact map constraints

In the case of 4-class contact maps, the constraint derived from the predicted map assumes a slightly different form. Since contacts between pairs of C$\alpha$ are here predicted in four classes, a contact is penalised not only if it is not present in the predicted map, but also depending on its distance to the boundaries of the correspoding class: $\mathcal{F}_k = \{(i, j) \mid D_k < d_{ij} < D_{k+1} \wedge c_{ij} \neq k\}$ with $D_k$ being the distance thresholds that define the classes.

Let $D_k^{'} = (D_k + D_{k+1})/2$, then the objective function is defined as:

$$\begin{aligned} \mathrm{C}(\mathcal{S}_n, \mathcal{M}) &= \alpha_0\{1 + \sum_k \sum_{(i,j)\in\mathcal{F}_k} (d_{ij}/D_k^{'})^2 \\ &+ \sum_{(i,j):d_{ij}\notin\mathcal{B}} (d_{ij} - D_B)^2\} + \alpha_1 \sum_{(i,j):d_{ij}\notin\mathcal{C}} e^{(D_{HC}-d_{ij})} \end{aligned} \tag{2}$$

### 3.3 Experiments and Results

The protein data set used in reconstruction simulations consists of a non redundant set of 258 protein structures (S258) showing no homology to the sequences employed to train the contact map predictors (see below). This set includes proteins of moderate size (51 to 200 amino acids) and diverse topology as classified by SCOP (Structural Classification of Proteins database) [27] (all-$\alpha$, all-$\beta$, $\alpha/\beta$, $\alpha + \beta$, surface, coiled-coil and small). No two proteins in this set share more than 25% sequence identity.

In all the experiments, we run the annealing protocol using a non linear (exponential decay) schedule with initial (resp. final) temperature proportional to the protein size (resp. 0). Pseudo energy parameters are set to $\alpha_0 = 0.2$ (false non-contacts), $\alpha_1 = 0.02$ (false contacts) and $\alpha_2 = 0.05$ (clashes) for binary maps and $\alpha_0 = 0.005$ and $\alpha_1 = 0.05$ (clashes) for multi-class maps, so that the conformational search is biased towards the generation of compact clash-free structures and with as many of the predicted contacts realised.

In the first set of simulations we compare the quality of reconstructions based on binary maps and multi-class maps for the case in which experimental constraints are known, that is the maps are native. We use binary maps at 12Å , since these are more informative than a number of alternative we tested (tests not shown).

In order to assess the quality of predictions, two measures are considered here: root mean square deviation (RMSD) and TM-score [28] between the predicted structure and the native one.

For each protein in the test set, we run 10 folding simulations and select the best one. The results for the best simulations are then averaged over all the 258 proteins in the set and are reported in Table 1.

| **Maps** | **RMSD** | **TM-score** |
|---|---|---|
| **Binary** | 4.01 | 0.65 |
| **4-Class** | 2.23 | 0.83 |

Table 1: Reconstruction algorithm results for the best models derived from binary and multi-class true contact maps.

# 4 Contact Map Prediction

Only a small number of algorithms have being developed for the prediction of distance maps [11, 29]. Far more common are methods for the prediction of binary contact maps [18–24], with distance cutoffs of 6Å , 8Å , 10Å , or 12Å usually chosen to define the threshold between a contact and a non-contact. At the Critical Assessment of Protein Structure Prediction, CASP [30], maps are evaluated with a distance threshold of 8Å between $C_\beta$ atoms ($C_\alpha$ in the case of Gly). There is a wide range of machine learning techniques for predicting contacts: hidden markov models [31], recursive neural networks [9], multi-layer perceptrons [18, 19, 24], support vector machines [22, 23], and self-organizing maps [21] are just a few.

Predictors of contact maps are virtually always ab initio, meaning that they do not rely directly on similarity to proteins of known structure. In fact, often, much care is taken to try to exclude any detectable similarity between training and test set instances when gauging predictive performances of structural feature predictors.

The method we present here is based on recursive neural networks, in particular 2-dimensional recursive neural networks (2D-RNNs). We predict both binary and multi-class maps. The system presented is an update of the system which took part in CASP7 [30]. The most significant update is the addition of homology information from the PDB [32]. In the following sections we give a detailed overview of the system and show that homology information greatly increases the performance of the predictor, even in the difficult [0,30)% sequence identity homology zone.

## 4.1 2D-RNNs

2D-RNNs were previously described in [20] and [33]. This is a family of adaptive models for mapping two-dimensional matrices of variable size into matrices of the same size.

If $o_{j,k}$ is the entry in the $j$-th row and $k$-th column of the output matrix (in our case, it will represent the estimated probability of residues $j$ and $k$ belonging to a particular class), and $i_{j,k}$ is the input in the same position, the input-output mapping is modeled as:

$$\begin{aligned}
o_{j,k} &= \mathcal{N}^{(O)}\left(i_{j,k}, h^{(1)}_{j,k}, h^{(2)}_{j,k}, h^{(3)}_{j,k}, h^{(4)}_{j,k}\right)\\
h^{(1)}_{j,k} &= \mathcal{N}^{(1)}\left(i_{j,k}, h^{(1)}_{j-1,k}, .., h^{(1)}_{j-s,k}, h^{(1)}_{j,k-1}, .., h^{(1)}_{j,k-s}\right)\\
h^{(2)}_{j,k} &= \mathcal{N}^{(2)}\left(i_{j,k}, h^{(2)}_{j+1,k}, .., h^{(2)}_{j+s,k}, h^{(2)}_{j,k-1}, .., h^{(2)}_{j,k-s}\right)\\
h^{(3)}_{j,k} &= \mathcal{N}^{(3)}\left(i_{j,k}, h^{(3)}_{j+1,k}, .., h^{(3)}_{j+s,k}, h^{(3)}_{j,k+1}, .., h^{(3)}_{j,k+s}\right)\\
h^{(4)}_{j,k} &= \mathcal{N}^{(4)}\left(i_{j,k}, h^{(4)}_{j-1,k}, .., h^{(4)}_{j-s,k}, h^{(4)}_{j,k+1}, .., h^{(4)}_{j,k+s}\right)\\
&\quad j,k = 1,\ldots,N\\
&\quad s = 1,\ldots,S
\end{aligned}$$

where $h_{j,k}^{(n)}$ for $n = 1, \ldots, 4$ are planes of hidden vectors transmitting contextual information from each corner of the matrix to the opposite corner. We parametrise the output update, and the four lateral update functions (respectively $\mathcal{N}^{(O)}$ and $\mathcal{N}^{(n)}$ for $n = 1, \ldots, 4$) using five two-layered feed-forward neural networks, as in [33]. Stationarity is assumed for all residue pairs $(j, k)$, that is the same parameters are used across all $j = 1, ..., N$ and $k = 1, ..., N$. Each of the 5 neural network contains its own individual parameters, that are not constrained to the ones of the other networks.

We use 2D-RNNs with *shortcut connections*. The best way to think of shortcuts is to think of a simple recurrent network in a 1-dimensional (1D) case. The standard definition of 1D recurrent neural networks prescribe an explicit dependency between the input being processed now (here), at time (position) $j$, and the item processed previously, $j - 1$, resulting in an implicit dependency between $j$ and all previous items. Most algorithms lack the power to extract information from the implicit dependencies (especially when using gradient learning) beyond the span of a few steps, because of the well known problem of vanishing gradient [34]. Therefore allowing shortcuts is an extension of this idea where in addition to simply having a a direct dependency on the previous item, $j - 1$, there is also a direct dependency on the previous $j - s$ for all $s > 1, ..., S$. Indeed, shortcut connections can be placed at any of the previous inputs $j - s$ for any $s \in 1, .., S$. The latter placement of shortcuts between $j$ and $S$ was used to produce near perfect secondary structure predictions in a bidirectional recurrent neural network when $(j, s)$ are native contacts [35]. Notice that increasing the number of shortcuts increases the parameters resulting in a model that may overfit on the data. Extending this idea to the 2D case in any direction in the matrix is straightforward (in fact any dimension can be processed). Shortcut directions and patterns are not strictly constrained (so long as cycles are not introduced in the directed graph representing the network) and may even be learned. With the addition of shortcuts the span of contextual information analysed by a recursive network can be extended, although this may come at the price of increased noise reaching the input, and increased potential for overfitting the examples.

The choice of input $i_{j,k}$ is an important factor for the algorithm. In the case of contact map prediction the simplest input is the amino acid symbols at $(j, k)$. Different input signals can be constructed to improve the algorithm. For example, contact density was used in [8] to improve contact map prediction accuracy significantly. In section 4.4 the design of the input will be discussed.

### 4.2 Training

Learning proceeds by gradient descent by minimising the relative cross entropy between target and output. Careful management of the gradient must take place, not letting it be too small or too large: the absolute value of each component of the gradient is kept within the [0.1,1] range, meaning that it is set to 0.1 if it is smaller than 0.1, and to 1 if it is greater than 1. The learning rate is set to 0.3 divided by the the total number of proteins in the dataset. The weights of the networks are initialised randomly.

Learning is slow due to the complexity of the problem. Each 2D-RNN contains 5 neural networks, replicated $N^2$ times for a protein of length $N$. During each training epoch forward and back-propagation has to occur in each of the $5 \times N^2$ networks, for all $P$ proteins in the training set. The neural network forward and back-propagation have a complexity proportional to $\mathcal{O}(\theta)$ where $\theta$ is the number of parameters in the network. Learning generally converges at about 300-350 epochs. Although the complexity of an epoch is polynomial at $\mathcal{O}(\theta N^2 P)$, the large size of the training set, and especially the quadratic term in the length of the proteins make learning quite time-consuming. Training of all systems (binary, multi-class; ab initio, template-based) took approximately three months on a cluster of 10 2.8GHz CPUs.

However, during prediction only one forward propagation needs to run for each instance, meaning that predictions for a set may be run in roughly 3 orders of magnitude less time than a training on the same set. For instance, maps for 1000 proteins of average length 120 amino acids can be predicted in approximately 13 hours on a single 2.8GHz CPU, and genomic-scale predictions are possible even on a small cluster of machines.

### 4.3 Architecture

In each of the 5 neural networks used to parameterise the functions, $\mathcal{N}^{(O)}$ and $\mathcal{N}^{(n)}$ for $n = 1, \ldots, 4$, we use a single hidden layer. Let $N_{hh}$ and $N_{ho}$ denote the number of units associated with the hidden layer and the output layer of the hidden contextual networks respectively. From the definition of the 2D-RNN we see that each hidden network has $I$ regular input units and $2 \times N_{ho} + S \times N_{ho}$ contextual inputs, where $S$ are the total number of shortcuts allowed. Thus, including the usual bias terms in each layer, the total number of parameters in one of the four hidden networks is: $(I + 2 \times N_{ho} + S \times N_{ho}) \times N_{hh} + N_{hh} + N_{hh} \times N_{ho} + N_{ho}$. The output network also contains $I$ regular inputs but it takes contextual inputs from the four hidden networks $4 \times N_{ho}$ resulting in: $(I + 4 \times N_{ho}) \times N_h + N_h + D \times Nh + D$ parameters, where $N_h$ are the number of units in the hidden layer of the output network and $D$ is the number of classes. The activation functions used are softmax and tanh. Only the output units of the output network have softmax functions in order to estimate Bayesian posterior probability of class membership. All other units have tanh transfer functions.

No overfitting avoiding techniques such as early stopping or weight decay were applied given the very large size of the datasets, and the fact that we ensemble many networks in the final predictor (see section 4.5.2).

Due to the large computational power needed to train one model we ensemble networks both from different trainings and from different stages of the same training. Networks are saved every 5 epochs, and for each training the last 3 networks are ensembled. Three networks with different architectural parameters ($N_{hh} = N_{ho} = N_h = 13, 14, 15$) are trained for each predictor. Results for network performances in this work are reported for these ensembles of $3 \times 3 = 9$ models. Ensembling leads to significant classification performance improvements over single models.

All results are in 5-fold cross validation, meaning that, in fact, 5 times 9 models are available for each system. For the reconstruction results (see next section) only the final networks for each training are ensembled, for a total of $1 \times 3 \times 5 = 15$ for each system.

The number of classes is $D = 2$ or $D = 4$ depending on the problem (binary vs. multi-class). For all networks the number of shortcuts is $S = 2$, with more sophisticated shortcut placements to be investigated in the future.

## 4.4 Input Design

Input $i_{j,k}$ associated with the $j$-th and $k$-th residue pair contains primary sequence information, evolutionary information, structural information, and direct contact information derived from the PDB templates:

$$i_{j,k} = (i_{j,k}^{(E)}, i_{j,k}^{(T)}) \tag{3}$$

where, assuming that $e$ units are devoted to evolutionary sequence information and structural information in the form of secondary structure [36, 37], solvent accessibility [36, 38] and contact density [8]:

$$i_{i,j}^{(E)} = (i_{j,k}^{(1)^{(E)}}, \ldots, i_{j,k}^{(e)^{(E)}}) \tag{4}$$

Template information is placed in the remaining $t$ units:

$$i_{j,k}^{(T)} = (i_{j,k}^{(1)^{(T)}}, \ldots, i_{j,k}^{(t)^{(T)}}) \tag{5}$$

Hence $i_{j,k}$ contains a total of $e + t$ components.

As in [8] $e = 418$, consisting of a sparse $20 \times 20$ matrix corresponding to the frequency of all pairs of amino acids observed in the two columns $j$ and $k$ of the multiple sequence alignment - this was chosen in order to capture information about correlated mutations. Structural information in the form of secondary structure (three classes), solvent accessibility (two classes), and contact density (four classes) for residue $j$ and $k$ are placed in the remaining 6,4 and 8 input units respectively.

For the template units we use $t = 3$ for binary maps and $t = 5$ for multi class maps, representing weighted contact class information from the templates and one template quality unit. For example, in the case of multi class maps the first four input units contain the weighted average contact class frequency in the PDB templates, while the last unit encodes the average quality of the template column. Assume that $d_{j,k}^{(p)}$ is a 4-component binary vector encoding the contact class of the $j$-th and $k$-th residue pair in the $p$-th template. Then, if $P$ is the total number of templates for a protein:

$$(i_{j,k}^{(1)^{(T)}}, \ldots, i_{j,k}^{(4)^{(T)}}) = \frac{\sum_{p=1}^{P} w_p d_{j,k}^{(p)}}{\sum_{p=1}^{P} w_p} \tag{6}$$

where $w_p$ is the weight attributed to the $p$-th template. If the sequence identity between template $p$ and the query is $id_p$ and the quality of a template (measured as X-ray resolution

+ R-factor/20 or 10 for NMR hits, as in [39]) is $q_s$, then the weight is defined as:

$$w_p = q_p id_p^3 \tag{7}$$

Taking the cube of the identity between template and query allows to drastically reduce the contribution of low-similarity templates when good templates are available. For instance a 90% identity template is weighed two orders of magnitude more than a 20% one. In preliminary tests (not shown) this measure performed better than a number of alternatives.

The final unit of $i_{j,k}$, the quality unit, encodes the weighted average coverage and similarity of a column of the template profile as follows:

$$i_{j,k}^{(5)^{(T)}} = \frac{\sum_{p=1}^{P} w_p c_p}{\sum_{p=1}^{P} w_p} \tag{8}$$

where $c_p$ is the coverage of the sequence by template $p$ (i.e. the fraction of non-gaps in the alignment). Encoding template information for the binary maps is similar.

Ab initio based predictions use only the first part of the input, $i_{j,k}^{(E)}$ from equation 4, including secondary structure, solvent accessibility and contact density, although these are predicted ab initio. The template based predictions use the complete $i_{j,k}$ as input.

## 4.5 Experiments

### 4.5.1 Problem definition

The main objective of the experiments is to compare ab initio systems (PDB templates are assumed unavailable) and template-based systems. When very reliable PDB information (e.g. sequence identity to the query greater than 30-35%) is available we expect template-based predictions to be substantially better, and in fact, to nearly exactly replicate the maps of the best template. More interesting questions are: whether template-based predictions improve on ab initio ones in the so called twilight zone of sequence similarity (less than 30%); whether, in this same region, template-based predictions are better than can be obtained by simply copying the map of the best template, or a combination of the maps of the templates.

The 4 systems that we test are 12 Å ab intio contact maps $(12_{AI})$, 12 Å contact maps with templates $(12_{TE})$, multi-class ab intio $(M_{AI})$ and multi- class with templates $(M_{TE})$.

### 4.5.2 Dataset

The dataset used in the present simulations is extracted from the December 2003 25% pdb_select list[1]. We use the DSSP program [40] (CMBI version) to assign relevant structural features (secondary structure and relative solvent accessibility). $C_\alpha$ coordinates, directly available from the PDB, are used to calculate contact density [8]. Sequences for

[1] http://homepages.fh-giessen.de/~hg12640/pdbselect

|  | class 0 | class 1 | class 2 | class 3 |
|---|---|---|---|---|
| 12Å | 4,062,483 | 15,755,172 |  |  |
| Multi class | 1,623,411 | 3,205,472 | 5,176,584 | 9,812,188 |

Table 2: Number of residues contained in 12Å binary classes and the four classes in the Multi class definition.

which DSSP does not produce an output due, for instance, to missing entries or format errors are removed. For computational reasons, and to focus on single domains, proteins which have more than 200 amino acids are also removed. After processing by DSSP and the removal of long proteins, the set contains 1602 proteins and 163,379 amino acids. All the tests reported in this paper are run in 5-fold cross validation. The 5 folds are of roughly equal sizes, composed of 318-327 proteins. The datasets are available upon request.

Evolutionary information in the form of Multiple sequence alignments have long being shown to improve prediction of protein structural features [20, 33, 37, 41–45]. Multiple sequence alignments for the 1602 proteins 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 [46] with parameters $b = 3000$, $e = 10^{-3}$ and $h = 10^{-10}$.

Table 2 shows the class distribution of both types of map in the dataset. What is immediately obvious from this table is that the class distribution is more balanced in the 4 class problem and therefore should be easier to learn.

### 4.5.3 Template generation

For each of the 1602 proteins we search for structural templates in the PDB. We base our search on PDBFINDERII [47] as available on August 22 2005. An obvious problem arising is that all proteins in the set are expected to be in PDB (barring name changes), hence every protein will have a perfect template. To avoid this, we exclude from PDBFINDERII every protein that appears in the set. We also exclude all entries shorter than 10 residues, leading to a final 66,350 chains. Because of the PDBFINDERII origin, only one chain is present in this set for NMR entries.

To generate the actual templates for a protein, we run two rounds of PSI-BLAST against the version of the redundancy-reduced NR database described above, with parameters $b = 3000$ (maximum number of hits), $e = 10^{-3}$ (expectation of a random hit) and $h = 10^{-10}$ (expectation of a random hit for sequences used to generate the PSSM). We then run a third round of PSI-BLAST against the PDB using the PSSM generated in the first two rounds. In this third round we deliberately use a high expectation parameter ($e = 10$) to include hits that are beyond the usual Comparative Modelling scope ($e < 0.01$ at CASP6 [25]). We further remove from each set of hits thus found all those with sequence similarity exceeding

95% over the whole query, to exclude PDB resubmissions of the same structure at different resolution, other chains in N-mers and close homologues.

The distribution of sequence similarity of the best template, and average template similarity is plotted in figure 8. Roughly 14% of the proteins have no hits at more than 10% sequence similarity. About 19% of all proteins have at least one very high quality (better than 90% similarity) entry in their template set. Although the distribution is not uniform, all similarity intervals are adequately represented: for about 41% of the proteins no hit is above 30% similarity; for nearly 24% of the proteins the best hit is in the 30-50% similarity interval. The average similarity for all PDB hits for each protein, not surprisingly, is generally low: for roughly 73% of all proteins the average identity is below 30%.

It should be noted that template generation is an independent module in the systems. We are currently investigating whether more subtle strategies for template recognition would still benefit contact map predictions, with or without retraining the systems on the new template distributions.

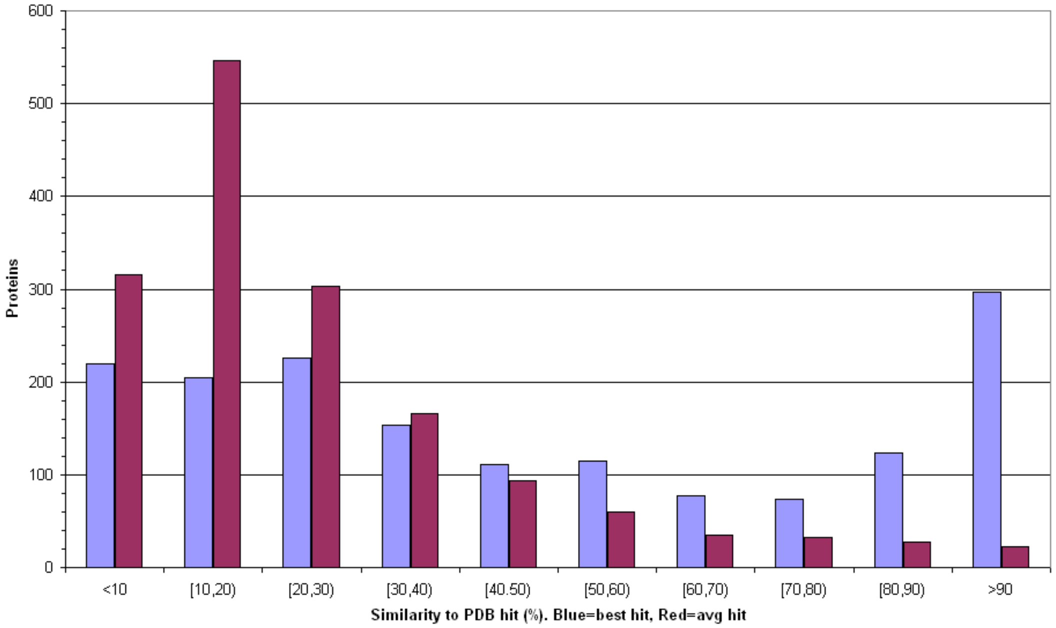

Figure 8: Distribution of best-hit (blue) and average (red) sequence similarity in the PSI-BLAST templates for the S2171 set. Hits above 95% sequence similarity excluded.

### 4.5.4 Training/testing protocol

The predictors of contact maps rely on predictions of secondary structure, solvent accessibility and contact density [8]. True structural information was used for training in both ab

| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | $\geq$ 90 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $12_{AI}$ | 85.9 | 87.5 | 86.8 | 85.6 | 87.2 | 86.5 | 86.2 | 86.1 | 86.4 | 87.3 | 86.8 |
| $12_{TE}$ | 85.3 | 87.8 | 91.3 | 93.6 | 95.7 | 96.0 | 95.8 | 96.4 | 97.0 | 97.3 | 93.2 |

Table 3: Percentage of classified predicted residue pairs for the ab initio ($12_{AI}$) and template based 12 Å predictor ($12_{TE}$) as a function of sequence identity to the best template. Template sequence identity 10 means all proteins that have a best hit template in the identity range [0, 10) %, All is the complete set.

| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | $\geq$ 90 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $12_{AI}$ | 85.8 | 87.6 | 88.1 | 89.9 | 92.0 | 90.8 | 93.1 | 90.5 | 94.0 | 94.0 | 87.9 |
| $12_{TE}$ | 85.3 | 87.1 | 88.4 | 91.4 | 92.8 | 92.8 | 94.0 | 94.3 | 94.8 | 94.4 | 87.7 |

Table 4: Identical to table 2 except only calculated for non template regions of the map.

initio and template based systems. For testing, we used predictions from our servers: Porter, PaleAle and BrownAle predicting secondary structure, solvent accessibility and contact density respectively. The ab initio models use ab initio secondary structure, solvent accessibility and contact density predictions. The template models use template-based secondary structure and solvent accessibility and ab initio contact density predictions (template-based contact density remains to be investigated).

All our experiments are carried out in 5-fold cross validation. The same dataset and multiple alignments are used to train the ab initio and template based secondary structure predictor Porter, solvent accessibility predictor PaleAle and the contact density predictor BrownAle. By design, these were trained using the same 5 fold split as the map predictors, therefore removing a trained fold while testing was a simple procedure and all 1D predictions are by models that were trained on a dataset split independent on the query.

The accuracy measure for all classes is calculated in order to compare the ab initio and template based models:

$$Accuracy = \sum_{c=0}^{C-1} \frac{correct_c}{total_c} \tag{9}$$

where $C$ is the total number of classes. All the accuracy values are calculated as a function of the best hit template found in the PDB to the query sequence. The best hit was determined by sequence identity between a template sequence and the query sequence.

### 4.6 Results and discussion

Table 3 reports the comparison between 12Å ab initio and template based predictions ($12_{AI}$ vs. $12_{TE}$) as a function of sequence identity to the best PDB hit. The only decrease in per-

| | 10 | 20 | 30 |
|---|---|---|---|
| $12_{TE}$ | 79.2 | 86.8 | 92.0 |
| Baseline | 84.0 | 89.2 | 92.1 |

Table 5: Percentage of classified predicted residue pairs for $12_{TE}$ when only considering the residues covered by the best template. Baseline is a predictor that copies the contact assignment from the best hit template.Template sequence identity 10 means all proteins that have a best hit template in the identity range [0, 10) %.

| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | ≥ 90 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $M_{AI}$ | 59.3 | 59.4 | 58.4 | 57.3 | 58.3 | 57.4 | 58.3 | 58.5 | 58.2 | 59.9 | 58.8 |
| $M_{TE}$ | 60.2 | 64.2 | 75.9 | 82.5 | 87.8 | 88.8 | 88.1 | 89.7 | 91.5 | 92.1 | 80.8 |

Table 6: Percentage of classified predicted residue pairs for the ab initio ($M_{AI}$) and template based Multi class predictor ($M_{TE}$) as a function of sequence identity to the best template. Template sequence identity 10 means all proteins that have a best hit template in the identity range [0, 10) %, All is the complete set.

| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | ≥ 90 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $M_{AI}$ | 59.0 | 58.3 | 61.8 | 64.8 | 71.6 | 69.4 | 75.0 | 71.4 | 75.7 | 75.5 | 61.1 |
| $M_{TE}$ | 60.3 | 60.7 | 65.7 | 71.2 | 76.4 | 75.5 | 80.3 | 82.1 | 80.2 | 79.4 | 63.8 |

Table 7: Identical to table 5 except only calculated for non template regions of the map.

| | 10 | 20 | 30 |
|---|---|---|---|
| $M_{TE}$ | 60.2 | 69.8 | 78.8 |
| Baseline | 54.8 | 67.1 | 78.6 |

Table 8: Percentage of classified predicted residue pairs for $M_{TE}$ when only considering residues covered by the best template. Baseline is a predictor that copies the class assignment from the best hit template. Template sequence identity 10 means all proteins that have a best hit template in the identity range [0, 10) %

formance is in the [0,10)% identity range, where the accuracy slightly decreases by 0.6%. However, the same results for multi class maps show that there is never a decrease in performance (Table 6). A role in this is played by the quality of predictions in regions not covered by the templates (reported in Tables 4 and 7). In these areas, for a sequence similarity of 20% and greater both $12_{TE}$ and $M_{TE}$ perform better than, respectively, $12_{AI}$ and $M_{AI}$. However, for lower similarity, $12_{AI}$ outperforms $12_{TE}$ on areas not covered by templates, while $M_{TE}$ still improves on $M_{AI}$. This may be either due to more balanced nature of the problem, easier contextual propagation in the multi-class case (the narrower class ranges impose stricter distance constraints among neighbours), or a combination of both. Ultimately, templates improve multi-class predictions in all regions of sequence similarity (including [0,10)%), both for regions covered and regions not covered by templates.

Tables 5 and 8 report the comparison between template based predictions and a baseline for 12Å and multi- class respectively. The baseline simply calculates the class for position $(i, j)$ from the coordinates in the best template. We also tested different baselines in which, instead of just the top template, the top 10 templates and all templates were used to get the class by a majority vote among the templates covering each template. We tested both an unweighed vote and a vote in which each template is weighed by its sequence similarity to the query, cubed. The latter weighing scheme is identical to the one used to present the templates to the neural networks (see equation 7). In all cases the baseline is worse than the best hit baseline, therefore the results are not reported here. We only report the predictions vs. baseline for the [0,30)% templates, since above 30% identity, as expected, the results are essentially the same. In this twilight region, where it is difficult to extract information from templates, $M_{TE}$ outperforms the baseline, however $12_{TE}$ does not.

The multi-class results are clearly encouraging, outperforming the baseline (Table 8), always improving on non-template regions (Table 7) and overall maps (Table 6). Figure 10 and 11 show an example of a map predicted for a low best hit sequence identity of 22.7%.

### 4.7 Modelling protein structures from predicted maps

In Figure 9, the average RMSD vs sequence length is shown for models for set S258 derived from true 4-class contact maps (stars), from $M_{TE}$ maps (squares) and from $M_{AI}$ maps (Xs), together with the baseline (crosses). The baseline represents a structure collapsed into its center of mass. Note that no templates are allowed that show a sequence identity greater than 90% to the query. Hence, the $M_{TE}$ results are based on a mixture of good, bad and no templates, akin to the distribution one would expect when presented with a protein sequence that is not in the PDB. The distribution of templates for S258 (not reported) resembles closely the one for the training/testing set, reported in Figure 8. It is also important to note that the results are an average of 10 reconstructions. If more reconstructions were run and, especially, if these were ranked, the results would likely improve considerably. The average reconstruction RMSD for $M_{TE}$ is 9.46Å and the average TM score 0.51. If the best of the 10 reconstructions is picked, these improve to 8.59Å and 0.55, respectively.

Reconstructions based on 4-class maps are significantly better than those from binary maps. Tested on both *ab initio* and homology-based 4-class maps, results show that

| **Maps** | **RMSD** | **TM-score** |
|---|---|---|
| $M_{AI}$ | 14.60 | 0.27 |
| $M_{TE}$ | 9.46 | 0.51 |

Table 9: Reconstruction algorithm results for models derived from multi-class predicted contact maps with ($M_{TE}$) and without ($M_{AI}$) allowing homology information. Note that, since no templates are allowed that show a sequence identity greater than 90% to the query, the $M_{TE}$ results are based on a mixture of good, bad and no templates (see Figure 8 for a sample distribution of template quality). The reported values are the average over the 10 runs of simulated annealing.

homology-based predictions are generally more accurate than *ab initio* ones even when homology is dubious. For sequence similarity above 30% the predictions' TM-score is on average slightly above 0.7 indicating high reliability, is approximately 0.45 in the 20-30% interval, and 0.27 in the region below 20%. If reconstruction performances are measured on the S258 set without allowing homology information at any stage (pure *ab initio* predictions) the average TM-score is 0.27, with 43 of the 258 structures above a TM-score of 0.4.

# 5 Conclusions

In this work we have described a machine learning pipeline for high-throughput prediction of protein structures, and have introduced a number of novel algorithmic ideas.

First, based on the observation that protein binary contact maps are somewhat lossy representations of the structure and yield only relatively low-resolution models, we have introduced multi-class maps, and shown that, via a simple simulated annealing protocol, these lead to much more accurate models, with an average RMSD to the native structure of just over 2Å and a TM score of 0.83.

Secondly, extending on ideas we have developed for predictors of secondary structure and solvent accessibility [36] we have presented systems for the prediction of binary and multi-class maps that use structural templates from the PDB to yield far more accurate predictions than their ab initio counterparts. We have also shown that multi-class maps lead to a more balanced prediction problem than binary ones. Although it is unclear whether because of this, or because of the nature of the constraints encoded into them, template-based systems for the prediction of multi-class maps we tested are capable of exploiting both sequence and structure information even in cases of dubious homology, significantly improving over their ab initio counterpart well into and below the twilight zone of sequence identity. This turns out to be only partly true, at least in our tests, for binary contact map predictors. Moreover, multi-class map predictions are far more accurate than the maps of the best templates for all the twilight and midnight zone of sequence identity, including the case in which only templates with less than 10% sequence identity to the query are available.

Conversely, for binary contact maps, the best template is on average more accurate than the prediction for all the [0%,30%) region of sequence identity.

Finally we have shown that template-based predictions of multi-class maps lead to fair to good predictions of protein structures, with an average TM score of 0.7 or higher to the native when good templates are available (sequence identity greater than 30%), and of 0.45 in the [20%, 30%) identity region. Ab initio predictions are still, on average, poor, at an average TM score of 0.27. Nevertheless, it is important to note how the component for homology detection in this study is basic (PSI-BLAST), and entirely modular, in that it may be substituted by any other method that finds templates without substantially altering the pipeline. Whether more subtle homology detection or fold recognition components could be substituted to PSI-BLAST, with or without retraining the underlying machine learning systems, is the focus of our current studies. The overall pipeline, including the template-based component, is available at the URL: http://distill.ucd.ie/distill/. Protein structure predictions are based on multi-class maps, and templates are automatically provided to the pipeline when available.

## 6 Acknowledgments

This work is supported by Science Foundation Ireland grant 05/RFP/CMS0029, grant RP/2005/219 from the Health Research Board of Ireland and a UCD President's Award 2004.

## References

[1] M. Adams, A. Joachimiak, G. T. Kim, R. Montelione, and J. Norvell. Meeting review: 2003 nih protein structure initiative workshop in protein production and crystallization for structural and functional genomics. *J. Struct. Funct. Genomics*, 5:1–2, 2004.

[2] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268:209–225, 1997.

[3] P. Larranaga, B. Calvo, R. Santana, Bielza C., J. Galdiano, I. Inza, and J. A. Lozano. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[4] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2:295–306, 1997.

[5] J. M. G. Izarzugaza, O. Grana, M. L. Tress, A. Valencia, and N. D. Clarke. Assessment of intramolecular contact predictions for casp7. *Proteins*, 69(S8):152–158, 2007.

[6] D.J. Osguthorpe. Ab initio protein folding. *Current Opinion in Structural Biology*, 10(2):146–152, 2000.

[7] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph theory algorithm for protein side-chain prediction. *Protein Science*, 12:2001–2014, 2003.

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

[9] G. Pollastri, P. Baldi, A. Vullo, and P. Frasconi. Prediction of protein topologies using giohmms and grnns. *Advances in Neural Information Processing Systems (NIPS) 15, MIT Press*, 2003.

[10] D.A. Debe, M.J. Carlson, J. Sadanobu, S.I. Chan, and W.A. Goddard. Protein fold determination from sparse distance restraints: the restrained generic protein direct monte carlo method. *J. Phys. Chem.*, 103:3001–3008, 1999.

[11] A. Aszodi, M.J. Gradwell, and W.R. Taylor. Global fold determination from a small number of distance restraints. *J. Mol. Biol.*, 251:308–326, 1995.

[12] E.S. Huang, R. Samudrala, and J.W. Ponder. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.*, 290:267–281, 1999.

[13] J. Skolnick, A. Kolinski, and A.R. Ortiz. Monsster: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, 265:217–241, 1997.

[14] P.M. Bowers, C.E. Strauss, and D. Baker. De novo protein structure determination using sparse nmr data. *J. Biomol. NMR*, 18:311–318, 2000.

[15] W. Li, Y. Zhang, D. Kihara, Y.J. Huang, D. Zheng, G.T. Montelione, A. Kolinski, and J. Skolnick. Touchstonex: Protein structure prediction with sparse nmr data. *Proteins: Structure, Function, and Genetics*, 53:290–306, 2003.

[16] D. Bau, Pollastri. G., and A. Vullo. *Analysis of Biological Data: A Soft Computing Approach*, chapter Distill: a machine learning approach to ab initio protein structure prediction. World Scientific, 2007.

[17] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio. Ft-comar: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, 24(10):1313–1315, 2008.

[18] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12(1):15–21, 1999.

[19] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835–439, 2001.

[20] 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.

[21] R.M. McCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20, Suppl. 1:224–231, 2004.

[22] Y. Zhao and G. Karypis. Prediction of contact maps using support vector machines. *3rd international conference on Bioinformatics and Bioengineering (BIBE)*, pages 26–33, 2003.

[23] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinfomatics*, 8:113, 2007.

[24] M. Punta and B. Rost. Profcon: novel prediction of long-range contacts. *Bioinformatics*, 21:2960–2968, 2005.

[25] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP) - round 6. *Proteins*, 7:3–7, 2005.

[26] V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudan, A. Fiser, F. Pazos, A. Valencia, A. Sali, and B. Rost. Eva: continuous automatic evaluation od protein structure prediction servers. *Bioinformatics*, 17:1242–1251, 2001.

[27] Andreeva A., D. Howorth, S.E. Brenner, Hubbard T.J.P., C. Chothia, and A.G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.*, 32:D226–D229, 2004.

[28] J. Skolnick Y. Zhang. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57:702–710, 2004.

[29] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak. Protein distance contraints predicted by neural networks and probability density functions. *Pro. Eng.*, 10:1241–1248, 1997.

[30] Casp home page, http://predictioncenter.org/.

[31] Y. Shao and C. Bystroff. Predicting interresidue contacts using templates and pathways. *Proteins*, 53:487–502, 2003.

[32] 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. *Nucleic Acids Research*, 28:235–242, 2000.

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

[34] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5:157–166, 1994.

[35] A Ceroni, P Frasconi, and G Pollastri. Learning protein secondary structure from sequential and relational data. *Neural Networks*, 18(8):1029–39, 2005.

[36] G. Pollastri, A.J.M. Martin, C. Mooney, and A. Vullo. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, 8(201):12, 2007.

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

[38] G. Pollastri, P. Fariselli, R. Casadio, and P. Baldi. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47:142–235, 2002.

[39] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–24, 1994.

[40] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22 (12):2577–637, 1983.

[41] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, 1994.

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

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

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

[45] 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(S1):S62–S70, 2002.

[46] 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.

[47] E. Krieger, R.W.W. Hooft, S. Nabuurs, and G. Vriend. Pdbfinderii - a database for protein structure analysis and prediction. *Submitted*, 2004.

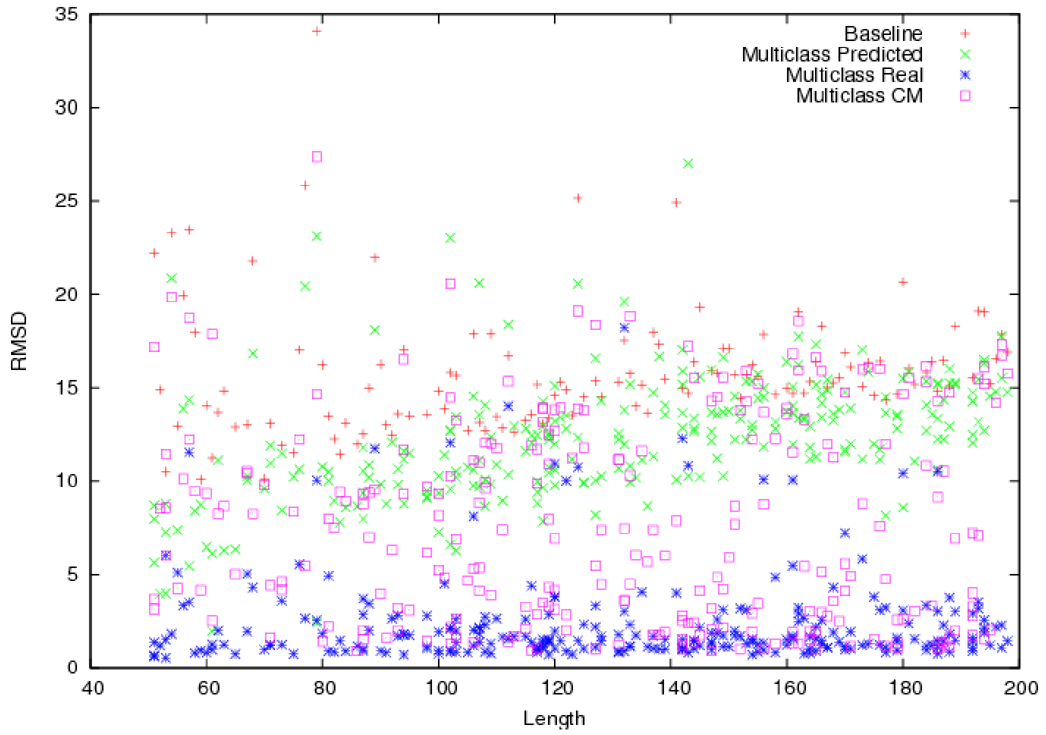

Figure 9: **4-class contact maps**: average RMSD vs sequence length is shown for models derived from true contact maps (blue stars), from predicted contact maps using information derived from homologues ($M_{TE}$) (purple squares) and from *ab initio* predicted contact maps (green Xs), together with the baseline (red crosses). Note that, since no templates are allowed that show a sequence identity greater than 90% to the query, the $M_{TE}$ results are based on a mixture of good, bad and no templates (see Figure 8 for a sample distribution of template quality).

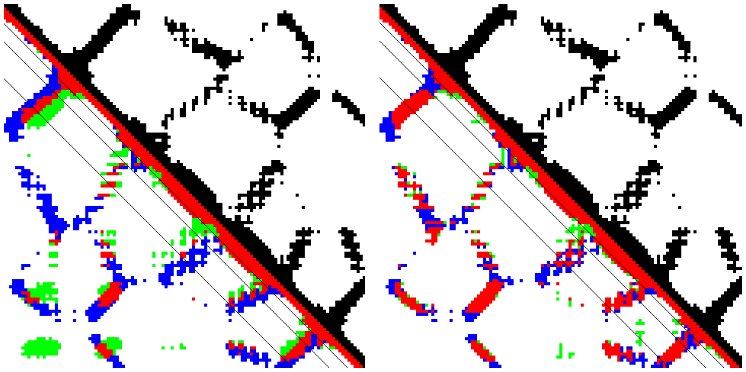

Figure 10: Protein 1B9LA 12 Å contact maps for ab initio (left) and template-based (right) predictions. The best template sequence identity is 22.7%. The top right of each map is the true map and the bottom left is predicted. In the predicted half white and red are true negative and positive respectively, blue and green are false negative and positive respectively. The three black lines correspond to $|i - j| \geq 6, 12, 24$.

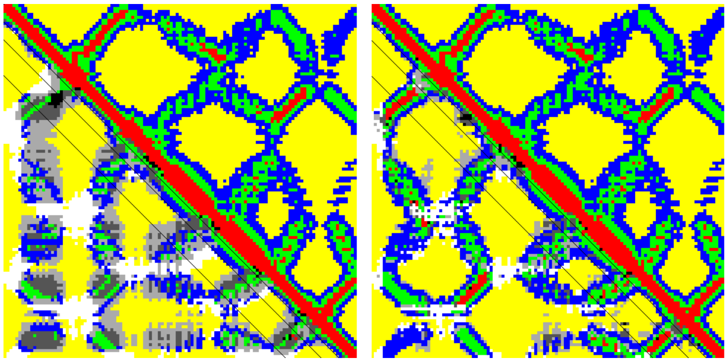

Figure 11: Protein 1B9LA Multi class contact maps for ab initio (left) and template-based (right) predictions. The best template sequence identity is 22.7%. The top right of each map is the true map and the bottom left is predicted. In the predicted half red, blue, green and yellow correspond to class 0, 1, 2 and 3 respectively. The greyscale in the predicted half corresponds to falsely predicted classes. The three black lines correspond to $|i - j| \geq 6, 12, 24$.