CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization

BY Y. WEI, J. THOMPSON AND C. A. FLOUDAS*

Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA

Most of the protein structure prediction methods use a multi-step process, which often includes secondary structure prediction, contact prediction, fragment generation, clustering, etc. For many years, secondary structure prediction has been the workhorse for numerous methods aimed at predicting protein structure and function. This paper presents a new mixed integer linear optimization (MILP)-based consensus method: a Consensus scheme based On a mixed integer liNear optimization method for seCOndary stRucture preDiction (CONCORD). Based on seven secondary structure prediction methods, SSpro, DSC, PROF, PROFphd, PSIPRED, Predator and GorIV, the MILP-based consensus method combines the strengths of different methods, maximizes the number of correctly predicted amino acids and achieves a better prediction accuracy. The method is shown to perform well compared with the seven individual methods when tested on the PDBselect25 training protein set using sixfold cross validation. It also performs well compared with another set of 10 online secondary structure prediction servers (including several recent ones) when tested on the CASP9 targets (http://predictioncenter.org/casp9/). The average Q3 prediction accuracy is 83.04 per cent for the sixfold cross validation of the PDBselect25 set and 82.3 per cent for the CASP9 targets. We have developed a MILP-based consensus method for protein secondary structure prediction. A web server, CONCORD, is available to the scientific community at http://helios.princeton.edu/CONCORD.

Keywords: secondary structure prediction; mixed integer optimization; consensus method

1. Background

During the past decade, significant progress has been made in protein structure prediction (Floudas et al. 2006; Kolodny et al. 2006; Floudas 2007; Zhang 2008). These include comparative modelling (Petrey & Honig 2005; Dunbrack 2006; Ginalska 2006; Cheng 2008), fold recognition and threading (Przybylski & Rost 2004; Wang et al. 2005; Wu & Zhang 2007; Xu et al. 2008), first principles prediction with database information (Zhang et al. 2003; Kolinski 2004; Rohl

*Author for correspondence (floudas@titan.princeton.edu).

et al. 2004) and first principles prediction without database information (Floudas 2000; Lee et al. 2000; Srinivasan & Rose 2002; Klepeis & Floudas 2003a; Klepeis et al. 2005; Subramani et al. 2009; McAllister & Floudas 2010; Rajgaria et al. 2010). At the same time, there are many limitations that are universal for all the methods. These include an accurate force field to describe the interactions between different atoms of molecules and between the molecule and its environment, a more effective sampling method for searching the conformational space and a better method of selecting the nearest-native structure. Other limitations exist for certain methods of protein structure prediction. For methods that require distance constraints, a good residue contact prediction is necessary because wrong predictions might guide the protein tertiary structure prediction to the wrong basin of the rugged energy landscape. For the protein structure and function prediction methods that use secondary structure information, secondary structure prediction plays a very important role (Rost 2001; Pollastri et al. 2002; Karypis 2006). The false prediction of secondary structure elements, especially beta strands, can result in wrong beta sheet topology; even for the case when the prediction of secondary structure elements is right but the locations of some elements are off, the overall topology for this protein can be twisted (Meiler & Baker 2003). Accurate secondary structure prediction also provides better understanding of protein structures when the structural homologies are unknown (Wray & Fisher 2007).

The earliest secondary structure prediction methods are based on amino acid statistics of a single protein sequence. Statistical properties, such as residue propensities and residue physico-chemical properties, are extracted from the given protein databases (Robson & Pain 1971; Chou & Fasman 1974; Lim 1974; Robson & Suzuki 1976; Garnier et al. 1978; Rost & Sander 1993a). Rules are deduced from the statistics and used for secondary structure prediction. One example is Chou–Fasman method (Chou & Fasman 1974), which is based on analysing the statistics of amino acids in alpha-helix, beta-strand and loops. The parameters derived from such statistical information are used to predict the secondary structure type. The work by Robson & Pain (1971) tries to predict helical residues using an information theory approach. Based on the single residue properties, some non-helical residues are predicted as helical. It was found that using neighbour residue information can reduce the false predictions by studying a limited data set of proteins. Some methods improve the prediction accuracy using additional statistical properties from a window of amino acids with the central amino acid at the prediction site. One example is the Garnier Osgothorpe Robson (GOR) method (Garnier et al. 1978). This method uses amino acid propensity information and the statistical information of 17 amino acid segment, and the likelihood of being one of the secondary structural stats is evaluated for the central amino acid. An accuracy of 63 per cent is reported when relative abundance of secondary structural elements is known. These methods predict protein secondary structure with a Q3 accuracy below 70 per cent (Rost 2003). This is because these early methods do not include the long-range residue interactions.

With the growing number of available protein sequences and the advances in the area of multiple sequence alignment, secondary structure prediction methods using evolutionary information emerged (Rost & Sander 1993b; Levin 1997; Cuff et al. 1998; Jones 1999; Hua & Sun 2001; Kloczkowski et al.
Secondary structure prediction via MILP

2002; Pollastri et al. 2002; Kim & Park 2003; Pollastri & McLysaght 2005; Mooney & Pollastri 2009). Usually the evolutionary information is used in the form of a position-specific scoring matrix by using PSI-BLAST. Profile network from HeiDelberg (PHD) is one such method that uses two-layer neural networks with evolutionary information (Rost & Sander 1993b). It is reported on average at least four percentage point improvement for all secondary structure elements and ten percentage point improvement for strand elements over any previous methods. Similarly, the SIMPA96 method reported a five percentage point improvement when using the evolutionary information over the same method when only single sequence information is used (Levin 1997). PSIPRED (Jones 1999) is another method that uses PSI-BLAST-obtained evolutionary information. It is based on a neural network method and found to achieve an average Q3 accuracy of 76.5–78.3%, depending on secondary structure element definition. This generation of methods using evolutionary information improved substantially the secondary structure prediction accuracy and they will continue to benefit from improved search strategies and larger databases (Rost 2001).

Recently, many attempts have been made for improving protein secondary structure prediction. Gassend et al. (2007) developed a method, using support vector machine (SVM), to find the biophysically motivated free energies for secondary structure prediction of α-helical proteins (Gassend et al. 2007), whereas Kountouris & Hirst (2009) used the SVM technique for secondary structure prediction and backbone dihedral angle prediction (Kountouris & Hirst 2009). Others attempt to combine different algorithms or introduce new ideas for secondary structure prediction. Won et al. (2007) combined hidden Markov model (HMM) with a genetic algorithm so that the topology of HMM can be designed automatically (Won et al. 2007). Yao et al. (2008) used the dynamic Bayesian networks (DBN), method which models the PSI-BLAST profile with multi-variate Gaussian functions (Yao et al. 2008), and showed that the DBN method generates better results than other pure HMM methods. It is also shown that, when combined with the neural network methods, it generates even more accurate results. Green et al. (2009) introduced parallel cascade identification (PCI), a tool from the field of nonlinear system identification, for secondary structure prediction (Green et al. 2009). Madera et al. (2010) developed a framework for secondary structure prediction that considers long-range interactions and it is described as a k-mer order model (Madera et al. 2010). This model can be used on top of any secondary structure methods. Three-dimensional structural information is also used, in addition to the sequence information, in some of the recent protein secondary structure prediction methods (Pollastri et al. 2002, 2007; Pollastri & McLysaght 2005; Montgomerie et al. 2006, 2008; Mooney & Pollastri 2009). Montgomerie et al. (2006) combines a structure-based sequence alignment with traditional sequence-based secondary structure methods (Montgomerie et al. 2006), whereas Pollastri et al. (2002) exploited protein three-dimensional structural information in terms of a simple structural frequency profile (Pollastri et al. 2007). These methods reported higher prediction accuracies than other methods that use only sequence-based information.

The majority of the secondary structure prediction methods use machine learning techniques, such as neural network, support vector machine, etc. Other methods do exist. One example is the first principle methods that predict the
primary structure by minimizing the free energy of the protein. Klepeis & Floudas (2002) introduced an approach based on the minimization of free energy of oligopeptide for the prediction of helical segments (Klepeis & Floudas 2002), and Klepeis & Floudas (2003a,b) introduced an approach based on combinatorial optimization for the prediction of beta sheet topology (Klepeis & Floudas 2003b).

Combining multiple methods can improve the prediction accuracy because non-systematic errors can be removed by correctly combining them. This idea has been widely used in different areas of computational biology, such as secondary structure prediction (Cuff et al. 1998; Cheng et al. 2007; Gupta et al. 2009), fold recognition and threading (Ginalska et al. 2003; Xu et al. 2004; Wu & Zhang 2007), protein disorder prediction (Kumar & Carugo 2008; Xue et al. 2010), prediction of gene coding sequences (Kang et al. 2007), clustering and functional interpretation of gene-expression data (Swift et al. 2004), etc. Kumar & Carugo (2008) developed a consensus method for predicting protein conformational disorder with an accuracy of 81.4 per cent based on 12 methods, using least-squares optimization. 3D-jury is a consensus web server that generates protein structures from many different existing servers (Ginalska et al. 2003); LOMETS is an online server for protein structure prediction using consensus method (Wu & Zhang 2007). One example of consensus methods for secondary structure prediction is the consensus data mining (CDM) method. It combines two complementary methods having different strengths: GOR and fragment database mining (FDM) methods. It is claimed that this combined method overcomes the cross validation prediction accuracy barrier of 80 per cent (Cheng et al. 2007). Multi-variate linear regression combination (MLRC) is a consensus method for secondary structure prediction combining GorIV (Garnier et al. 1996), SIMPA96 (Levin 1997) and SOPMA (Geourjon & Deleage 1995). Jpred (Cuff et al. 1998), a consensus secondary structure prediction server, processes results from different neural network methods (Cuff & Barton 2000).

This paper aims at developing a new MILP-based consensus method (CONCORD) for protein secondary structure prediction. This is the first consensus method for protein secondary structure prediction using MILP. This method uses seven selected secondary structure prediction methods, including DSC (King & Sternberg 1996), PROF (Ouali & King 2000), PROFphd (Rost 2000), PSIPRED (Jones 1999), Predator (Frishman & Argos 1996), GorIV (Garnier et al. 1996) and SSpro (Pollastri et al. 2002). The selections of the methods depend on the availabilities of stand-alone programs. CONCORD maximally combines the strengths of seven secondary structure prediction methods by optimizing the number of correctly predicted amino acids in the training set. A consensus prediction score based on the confidence scores of amino acid predictions of the seven methods is introduced to evaluate the likelihood of an amino acid being at one of the secondary structure states. The presented method performs well compared with all the seven individual methods when tested by a sixfold cross validation on the PDBselect25 protein set, and an average prediction accuracy (Q3 score) of about 83.04 per cent is obtained. When compared with another set of 10 secondary structure prediction servers on the recent CASP9 targets, the consensus method achieves a Q3 accuracy of 82.3 per cent. Sequence similarity is checked between the training protein set (PDBselect25) and the test proteins (CASP9) using PISCES.
(Wang & Dunbrack 2003) to ensure the independence. Of 107 CASP9 proteins, 102 showed less than 30 per cent sequence similarity to the PDBselect25 proteins, and 96 out of 107 CASP9 proteins had less than 25 per cent sequence similarity to the PDBselect25 proteins. The Q3 accuracy of CONCORD for CASP9 proteins with 25 per cent sequence similarity is 82.2 per cent, and the Q3 accuracy of CONCORD for CASP9 proteins with 30 per cent sequence similarity is 82.3 per cent.

2. Methods

(a) Nomenclature

This section lists the sets, parameters and variables used in this model.

(i) Indices and sets

$I$: set of amino acid positions of a protein, $i \in I$;
$P$: set of the training proteins, $[1 \ldots 3000]$, $p \in P$;
$M$: set of the methods used, $[1, 2, 3, 4, 5, 6, 7]$, $m \in M$. Seven methods are used in this consensus method: $m = 1$ indicates the DSC method; $m = 2$ indicates the PROF method; $m = 3$ indicates the PROFphd method; $m = 4$ indicates the SSpro method; $m = 5$ indicates the Predator method; $m = 6$ indicates the GorIV method; and $m = 7$ indicates the PSIPRED method; and

$\text{subsetPI}(P, I)$: subset indicating the number of amino acids for each protein $p$.

(ii) Parameters

$\text{confS}(p, i, m)$: the confidence score predicted by method $m$ for amino acid $i$ of protein $p$, $p \in P$, $i \in \text{subsetPI}(P, I)$, $m \in M$; and

$\text{predSS}(p, i, m)$: the prediction results of method $m$ for amino acid $i$ of protein $p$; It equals 1 for a true prediction, 0 for a false prediction, $p \in P$, $i \in \text{subsetPI}(P, I)$, $m \in M$.

(iii) Binary variables

$y(p, i)$: equals 1 if the sum of the scores (see §2d) of the correct secondary structure predictions is higher than the sum of the incorrect ones for amino acid $i$ of protein $p$ by at least $\epsilon(p, i)$, $p \in P$, $i \in \text{subsetPI}(P, I)$; and

$y2(p)$: equals 1 if the sum of the scores of the correct predictions of all amino acids of a protein $p$ is greater than the sum of the scores of the incorrect predictions of the same protein $p$ by at least $\epsilon2(p)$, $p \in P$, $i \in \text{subsetPI}(P, I)$.

(iv) Positive variables

$\lambda(m)$: the weight variables for different methods, $0 \leq \lambda(m) \leq 1$, $m \in M$;

$\epsilon(p, i)$: a soft margin variable for the binary variable $y(p, i)$, $p \in P$, $i \in \text{subsetPI}(P, I)$ (see §2a(iii)); and

$\epsilon2(p)$: a soft margin variable for the binary variable $y2(p)$, $p \in P$ (see §2a(iii)).
(b) The training objective function

The training objective function of the MILP model takes the following format:

$$\text{MAX} \left\{ \sum_{(p,i)} \frac{y(p,i)}{139} + \sum_p y2(p) - \sum_{(p,i)} e(p,i) - \sum_p e2(p) \right\} \forall (p, i) \in \text{subsetPI}(P, I).$$  \hspace{1cm} (2.1)

The first term in the objective function is to maximize the total number of amino acids whose correct prediction has a higher score than the incorrect ones by at least $e(p,i)$. The second term in the objective function is used to maximize the total number of proteins whose sum of the scores of the correct predictions is higher than that of the incorrect predictions by at least $e2(p)$. The third and fourth terms are included here to minimize the sum of soft margins. Note that in the first term, 139 is the average length of proteins in the training set, and it is used to balance the weights of the first and second terms.

This training objective function operates on the individual amino acids and the whole protein. The basis of this function is that some secondary structure prediction methods have better prediction performance in some local regions of a protein than other methods. By operating on the confidence scores for each amino acids of proteins, the consensus method aims at identifying the correct secondary structure predictions for amino acids from different methods. The second term focuses on the global performance of the consensus prediction.

(c) The training constraints

There are three basic constraints used in the consensus method for secondary structure prediction. The first constraint makes sure that, for an amino acid of a protein, when the difference between the sum of the scores of correct secondary structure predictions and the sum of the scores of incorrect predictions from different methods is lower than $e(p,i)$, the binary variable $y(p,i)$ is equal to zero; it takes the following form:

$$\sum_m \lambda(m) * \text{conf}S(p, i, m) * (1 - \text{predSS}(p, i, m)) - \sum_m \lambda(m) * \text{conf}S(p, i, m) * \text{predSS}(p, i, m) + e(p, i) < 1 - y(p, i) \forall (p, i) \in \text{subsetPI}(P, I), \ m \in M. \hspace{1cm} (2.2)$$

Similar to the first constraint, the second constraint forces $y(p)$ to be zero if the difference between the sum of the scores of correct secondary structure predictions of amino acids for a protein $p$ and the sum of the scores of incorrect secondary structure predictions of amino acids for the same protein $p$ is smaller than $e2(p)$. The functional form of this constraint is expressed as

$$\sum_{m,i} \lambda(m) * \text{conf}S(p, i, m) * (1 - \text{predSS}(p, i, m)) - \sum_{m,i} \lambda(m) * \text{conf}S(p, i, m) * \text{predSS}(p, i, m) + e2(p) < 1 - y(p) \forall (p, i) \in \text{subsetPI}(P, I), \ m \in M. \hspace{1cm} (2.3)$$

The final constraint used in the model normalizes the weights terms, \( \lambda(m) \), of the seven methods:

\[
\sum_m \lambda(m) = 1, \quad \lambda(m) \geq 0, \quad m \in M.
\]  \( \text{(2.4)} \)

\( (d) \) The prediction score

Based on the confidence scores of the secondary structure predictions from the seven individual methods for each amino acid of a protein, a consensus prediction score is written in order to score the secondary structure predictions of the consensus method. It is expressed as

\[
S(p, i) = \sum_m \lambda(m) \times \text{conf}S(p, i, m), \quad p \in P, \quad i \in \text{subsetPI}(P, I), \quad m \in M.
\]  \( \text{(2.5)} \)

This prediction score is calculated for true predictions and false predictions. For each amino acid of a protein, different methods may predict different secondary structure types, of which only one is the correct type (a true prediction; DSSP is used for determining the secondary structure type; Kabsch & Sander 1983). The prediction score evaluates the confidence levels of all the secondary structure types for each amino acid of a protein; By carefully selecting the weight variable, \( \lambda(m) \), of each individual method through the MILP optimization-based approach, the consensus method is developed to score higher for the correct predictions than for the incorrect predictions from different methods.

\( (e) \) Training process and prediction

The training process of the MILP model aims to maximize the training objective function using CPLEX (ILOG CPLEX 8.0 reference manual), and by which the weight parameters \( \lambda(m) \) are obtained. The training process for each fold during cross-validation takes about two weeks. Once \( \lambda(m) \) for each method is obtained and all the secondary structure predictions of the seven individual methods are finished, the protein secondary structure prediction is obtained by using the prediction score as defined in equation (2.5). The prediction scores of each amino acid of a protein for helical, strand and coil types are computed and compared, and the secondary structure type with the highest prediction score is taken as the prediction. The time for each protein secondary structure prediction, including the time for obtaining the predictions of the seven methods, is about 5–10 min.

\( (f) \) Prediction post-processing

The secondary structure prediction for a protein by combining the seven predictions using equation (2.5) could generate a secondary structure element that has only one strand amino acid or two helical amino acids in isolation, thus some post-processing is needed. When one coil amino acid is surrounded by helical amino acids (strand amino acids) and the coil amino acid is hydrophobic (LIMCYVFW), this amino acid is converted to helical (strand) amino acid. For the case where only one strand amino acid is surrounded by two helical (or coil) amino acids, this strand amino acid is changed into a helical (or coil) amino acid;
If this strand amino acid is surrounded by a helical amino acid and a coil amino acid, then it is converted to a coil amino acid. Similarly, the same rule applies to the case when one helical amino acid is isolated.

When two helical amino acids are isolated by coils, these two helical amino acids are converted into coil amino acids. If they are surrounded by strands, and the hydrophobic one(s) of these two helical amino acids are converted into strand amino acids, non-hydrophobic amino acids are converted to coil amino acids. If these two amino acids are surrounded by coil amino acids and strand amino acids, then they are converted to coil amino acids; except for the case when the amino acid close to strand amino acids is hydrophobic, it is converted to a strand amino acid.

The results shown in the paper are the CONCORD predictions after post-processing, except for the cases that are explicitly stated.

\[(g)\] **Protein data sets, cross validation and tests**

From PDBselect25 (October 2008 release), 3000 proteins are selected as the training set. Proteins with incomplete backbone information are excluded from the original set. The list of 3000 proteins is included in the electronic supplementary material, table S1.

The training protein set is divided into six parts, with each having 500 proteins. These six parts are used for the sixfold cross validation. Every training process uses a combined set of proteins from five parts and uses the remaining one part for testing purposes. For each training process, a set of optimal parameters from the MILP model are generated and these parameters are then used to predict the secondary structures for the corresponding testing set. Thus, an average prediction accuracy is obtained for this testing set. This process continues till all six accuracies have been obtained. Through the optimization process, better average accuracies are achieved over the other seven individual methods for all the testing sets.

In order to predict the secondary structure of proteins, a training process is conducted based on all the 3000 proteins, and an optimal parameter set \((\lambda(m)\) in row 2 of table 1) is achieved through the optimization process of the MILP model. These parameters are the weights for all the seven methods that should be used when building the consensus secondary structure prediction method.

A second testing case is based on the CASP9 protein targets (http://predictioncenter.org/casp9/). Table S2 of electronic supplementary material presents the list of proteins. We compared the performance of our secondary structure prediction method with another set of 10 secondary structure prediction methods, YASSPP (Karypis 2006), PORTER (Pollastri & McLysaght 2005), Jpred (Cuff et al. 1998), gorV (Kloczkowski et al. 2002), CDM (Cheng et al. 2007), symPred (Lin et al. 2010), FDM (Cheng et al. 2005), DISSpred (Kountouris & Hirst 2009), PCISS (Green et al. 2009) and PROTEUS (Montgomerie et al. 2006). By testing the performance of the secondary structure prediction methods on CASP9 targets, the problem of overestimating the performance can be avoided (Rost 2005). Thus, an evaluation on the CASP9 targets for the consensus method provides very valuable information. To ensure the independence of the test set, sequence similarity check for CASP9 proteins against PDBselect25 is performed using PISCES (Wang & Dunbrack 2003). Sequence similarity
Table 1. Parameters $\lambda(m)_{v1}$ (for the seven method CONCORD) obtained from the overall training process are shown in row 2. Parameters $\lambda(m)_{v2}$ (for the five method CONCORD, GorIV and Predator are excluded) obtained from the overall training process are shown in row 3. Rows 4–9 show the average Q3 scores of the individual methods and CONCORD for every test sets. Row 10 shows the average Q3 scores of the six test sets. Rows 11–16 are the average SOV scores of the individual methods and CONCORD for every test sets. Row 17 shows the average SOV scores over the six test sets. CONCORD $V_1$ (column 9) denotes the CONCORD method based on all the seven methods, whereas CONCORD $V_2$ (column 10) indicates the CONCORD method based on the five methods, excluding GorIV and Predator.

<table>
<thead>
<tr>
<th>methods</th>
<th>GorIV</th>
<th>DSC</th>
<th>predator</th>
<th>PROFphd</th>
<th>PROF</th>
<th>PSIPRED</th>
<th>SSpro</th>
<th>CONCORD $V_1$</th>
<th>CONCORD $V_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda(m)_{v1}$</td>
<td>0.00065</td>
<td>0.03074</td>
<td>0.001361</td>
<td>0.01</td>
<td>0.084316</td>
<td>0.749907</td>
<td>0.123026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda(m)_{v2}$</td>
<td>0.098206</td>
<td>0.041</td>
<td>0.05265</td>
<td>0.54737</td>
<td>0.1108</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 test 1</td>
<td>61.66</td>
<td>71.02</td>
<td>62.25</td>
<td>68.65</td>
<td>75.03</td>
<td>81.68</td>
<td>82.9178</td>
<td>83.34</td>
<td>83.25</td>
</tr>
<tr>
<td>Q3 test 2</td>
<td>61.84</td>
<td>70.76</td>
<td>62.41</td>
<td>67.82</td>
<td>74.10</td>
<td>81</td>
<td>82.2736</td>
<td>82.95</td>
<td>82.92</td>
</tr>
<tr>
<td>Q3 test 3</td>
<td>62.53</td>
<td>70.90</td>
<td>63.38</td>
<td>68.72</td>
<td>74.78</td>
<td>81.42</td>
<td>82.7626</td>
<td>83.05</td>
<td>83.06</td>
</tr>
<tr>
<td>Q3 test 4</td>
<td>61.94</td>
<td>70.70</td>
<td>62.73</td>
<td>68.34</td>
<td>74.11</td>
<td>80.99</td>
<td>82.5668</td>
<td>82.97</td>
<td>82.97</td>
</tr>
<tr>
<td>Q3 test 5</td>
<td>61.42</td>
<td>71.46</td>
<td>63</td>
<td>68.79</td>
<td>75.33</td>
<td>81.82</td>
<td>82.3342</td>
<td>82.94</td>
<td>82.94</td>
</tr>
<tr>
<td>Q3 test 6</td>
<td>61.29</td>
<td>70.30</td>
<td>62.35</td>
<td>68.47</td>
<td>74.46</td>
<td>81.39</td>
<td>82.58</td>
<td>82.98</td>
<td>82.97</td>
</tr>
<tr>
<td>ave Q3</td>
<td>61.78</td>
<td>70.86</td>
<td>62.69</td>
<td>68.47</td>
<td>74.69</td>
<td>81.39</td>
<td>82.57</td>
<td>83.04</td>
<td>83.02</td>
</tr>
<tr>
<td>SOV test 1</td>
<td>57.50</td>
<td>67.18</td>
<td>59.84</td>
<td>65.14</td>
<td>73.38</td>
<td>78.83</td>
<td>77.33</td>
<td>80.49</td>
<td>80.49</td>
</tr>
<tr>
<td>SOV test 2</td>
<td>56.76</td>
<td>65.67</td>
<td>59.39</td>
<td>64.20</td>
<td>71.20</td>
<td>77.74</td>
<td>76.43</td>
<td>79.44</td>
<td>79.40</td>
</tr>
<tr>
<td>SOV test 3</td>
<td>58.08</td>
<td>66.68</td>
<td>60.52</td>
<td>65</td>
<td>72.02</td>
<td>78.44</td>
<td>77.03</td>
<td>80.45</td>
<td>80.45</td>
</tr>
<tr>
<td>SOV test 4</td>
<td>57.56</td>
<td>67.14</td>
<td>60.45</td>
<td>65.22</td>
<td>72.18</td>
<td>78.50</td>
<td>77.68</td>
<td>80.98</td>
<td>80.95</td>
</tr>
<tr>
<td>SOV test 5</td>
<td>57.09</td>
<td>66.36</td>
<td>60.66</td>
<td>64.56</td>
<td>72.26</td>
<td>78.28</td>
<td>76.79</td>
<td>79.88</td>
<td>79.88</td>
</tr>
<tr>
<td>SOV test 6</td>
<td>57.44</td>
<td>66.22</td>
<td>59.96</td>
<td>65.61</td>
<td>72.43</td>
<td>78.79</td>
<td>77.60</td>
<td>80.48</td>
<td>80.48</td>
</tr>
<tr>
<td>ave SOV</td>
<td>57.41</td>
<td>66.54</td>
<td>60.14</td>
<td>64.96</td>
<td>72.25</td>
<td>78.43</td>
<td>77.14</td>
<td>80.29</td>
<td>80.28</td>
</tr>
</tbody>
</table>
thresholds of 25 and 30 per cent are used for deriving the independent sets of proteins. CONCORD is also tested on the independent protein sets of CASP9.

\[(h)\] True secondary structure assignments

DSSP (Kabsch & Sander 1983) is used to determine the protein secondary structure from the experimental structure. There are eight secondary structure classes by DSSP standard: H (α helix), G (3_{10} helix), I (π helix), E (β strand), B (isolated β bridge), T (turn), S (bend) and the rest. Since most of the secondary structure prediction use Q3 score for the prediction evaluation, the eight classes are reduced to three classes (H for helix, E for strand and C for coil) under most applications. This reduction can affect the prediction accuracy (Cuff & Barton 1999). In this paper, the reduction method used is the same as in CASP assessment: H, G and I to H; B and E to E; all other states to C (Lesk et al. 2001).

\[(i)\] Prediction accuracy evaluation

Q3 score is the three state (H, E and C states) overall prediction accuracy defined as the ratio between the number of correctly predicted amino acid sites and the number of total amino acids in the protein. It is the most widely used metric for evaluating the performance of secondary structure prediction methods. Its formula is expressed as

$$Q3 = 100 \times \frac{\sum_i \text{Pred}_i}{N},$$

where \(N\) is the total number of amino acids in the protein, and \(\text{Pred}_i\) is 1 only when the prediction of amino acid is correct.

There is another way to evaluate the performance of secondary structure prediction method. SOV strands for segment overlap measure. It measures the average overlap between the actual and predicted segments instead of using single amino acid prediction accuracy (Zemla et al. 1999). The formula expressing the definition of SOV is written as

$$\text{SOV} = 100 \times \frac{1}{N} \sum_{i \in \{H, E, C\}} \sum_{S_i} \left( \frac{\text{minov}(S_1, S_2) + \delta(S_1, S_2)}{\text{maxov}(S_1, S_2)} \right) \times \text{len}(S_1),$$

where \(N\) is the total sum of \(N_i\) over the three states, \(S_i\) is the set of all overlapping pairs of segments \((s1, s2)\) in conformation state \(i\), \(\text{len}(S_1)\) is the number of residues in segment \(S_1\), \(\text{minov}(S_1, S_2)\) is the length of the actual overlap and \(\text{maxov}(S_1, S_2)\) is the total extent of the segment. \(\delta(S_1, S_2)\) is the minimum of \(\text{maxov}(S_1, S_2) - \text{minov}(S_1, S_2)\), \(\text{minov}(S_1, S_2)\), \(\text{int}[0.5 \times \text{len}(S_1)]\) and \(\text{int}[0.5 \times \text{len}(S_2)]\).

3. Results

\[(a)\] Sixfold cross validation

Based on the 3000 proteins obtained from PDBselect25 (October 2008 release), a sixfold cross validation has been performed. The weight parameters obtained

from the optimization process of the MILP model are listed in table 1. Prediction accuracies and SOV scores for each individual method and the consensus method are also shown in table 1.

The weights for GorIV and Predator among the seven methods (numbers are listed in row 2 of table 1) are the smallest, while the PSIPRED’s weight is the largest in the consensus method. The fact that different methods show different weights is owing to the different prediction accuracies of each method. This explains why GorIV and Predator methods have the smallest weights. On the other hand, note that even though the SSpro method has the highest accuracy for secondary structure prediction, its weight is the second largest after PSIPRED. This is because many of these seven secondary structure prediction methods use PSI-BLAST profiles (e.g. PSIPRED, SSpro, etc.), and the prediction results of the different methods may correlate with each other in some fashion. This could result in multiple sets of the weights that will generate similar results. Indeed during the training process of the MILP model, another parameter set (0.01 for GorIV; 0.038565 for DSC; 0 for Predator; 0.072998 for PROFphd; 0.216558 for PROF; 0.424866 for PSIPRED and 0.237014 for SSpro) is found with slightly worse performance than current proposed weights.

As can be seen in table 1, CONCORD, using seven methods, generates the highest Q3 and SOV scores compared with the seven individual methods. For the cross validation test, CONCORD obtains a Q3 score of 83.04 per cent. Among the seven methods, SSpro is the best in terms of Q3 score, with a Q3 value of 82.57 per cent; whereas PSIPRED generates the highest SOV score among the seven methods (78.43%). The prediction accuracies (Q3) of all the methods have the ranking as follows: SSpro > PSIPRED > PROF > DSC > PROFphd > Predator > GorIV. The ranking of SOV scores is similar to Q3, with PSIPRED leading SSpro by 1 per cent. Because multiple sequence alignment information is not used for current PROFphd method, the performance of PROFphd is greatly affected. The performance of other methods can also be different from the reported accuracies owing to various reasons, such as the different test sets of proteins, the secondary structure assignment method difference, etc.

A sixfold cross validation is also performed for the CONCORD method trained using five methods, excluding GorIV and Predator. Because GorIV and Predator have the lowest weights among all the seven methods, they are excluded in the second step to allow us to compare the performances of two cases for the CONCORD method. The individual weights for the five methods used for CONCORD are listed in row 3 of table 1. PSIPRED has again the highest weight, while the weight of SSpro ranks second. As shown in the last column of table 1, the average Q3 and SOV scores of the cross validation for the CONCORD method using five methods are similar to the CONCORD method, using all of the seven methods.

Binary prediction accuracies for H, E and C residues are listed in the electronic supplementary material, table S3. The prediction accuracy for binary predictions is defined as $\frac{TP + TN}{TP + TN + FP + FN}$, where TP is true positive, TN is true negative, FP is false positive and FN is false negative. The prediction accuracy of CONCORD for Q(H/∼H), Q(E/∼E) and Q(C/∼C) are 91 per cent, 91.7 per cent and 84.3 per cent for sixfold cross validation, respectively.
(b) Testing on independent CASP9 targets

Critical assessment of protein structure prediction (CASP) is a biennial worldwide competition for protein structure prediction, and the competition is in a double blind fashion: The structures of the target proteins are unknown to the predictors and the organizers (http://predictioncenter.org/casp9/). Although there is no secondary structure prediction category in the competition, the secondary structure prediction is involved in many parts of the event. Many protein tertiary structure predictions—for example, BioSerf, Astro-Fold, Circle and ZAMDP, etc.—and some contact prediction methods use secondary structure prediction (http://predictioncenter.org/casp9/doc/Abstracts.pdf).

In this section, we have tested CONCORD on 107 proteins in CASP9 using the proposed parameter set in table 1 (weights \(\lambda(m)_{v1}\) in row 2). The prediction results of CONCORD, together with the other seven individual methods, are shown in table 2. CONCORD generates the highest prediction accuracy (82.3%). Among all the individual methods, PSIPRED performs best with a Q3 value of 81.7 per cent. The overall ranking of each method is very similar to the ranking of sixfold cross validation except that PSIPRED shows better prediction accuracy than SSpro.

The prediction accuracy (Q3) of CONCORD is reduced from the average accuracy of 83.04 per cent for sixfold cross validation to 82.3 per cent for the CASP9 targets. SSpro method has the largest decrease in accuracy, from 82.6 to 79.9 per cent. On the other hand, some methods perform better for CASP9 targets than the sixfold cross validation. For example, PROFphd’s prediction accuracy is improved from 68.5 per cent for cross validation to 69.4 per cent for CASP9 proteins. All other methods produced very similar prediction accuracies.

In order to compare with some other methods that are not used for developing CONCORD, another set of 10 online secondary structure prediction servers are chosen: YASSPP (Karypis 2006), PORTER (Pollastri & McLysaght 2005), Jpred (Cuff et al. 1998), gorV (Kloczkowski et al. 2002), DISSpred (Kountouris & Hirst 2009), PCISS (Green et al. 2009), PROTEUS (Montgomerie et al. 2006), CDM (Cheng et al. 2007), symPred (Lin et al. 2010) and FDM (Cheng et al. 2005). The performance of secondary structure prediction is realized by an automatic Perl script that submits the secondary structure predictions to each of the 10 servers and analyses the received email prediction results. There are consensus methods among the 10 servers. For example, the CDM method is a consensus method combining gorV and FDM (Cheng et al. 2007); Jpred is also a consensus method combining DSC (King & Sternberg 1996), PHD (Rost & Sander 1993b), Predator (Frishman & Argos 1996), NNSSP (Salamov & Solovyev 1995), ZPRED (Zvelebil et al. 1987) and MULPRED (G. J. Barton 1988, unpublished data).

The results of the comparison between CONCORD with other 10 online servers for secondary structure prediction is shown in table 2. As can been seen, PORTER has the highest Q3 score among the 10 servers. Its 81.6 per cent accuracy is very close to the PSIPRED method. YASSPP and Jpred also show prediction accuracies above 80 per cent. CONCORD with a Q3 score of 82.3 per cent performs well compared with all the servers.

Some proteins in the CASP9 test set may have homologous proteins in the training protein set (PDBselect25); thus, it is also important to check how CONCORD performs on the independent CASP9 protein sets compared
Table 2. Prediction accuracies (Q3) on CASP9 targets for CONCORD and 17 tested methods, including seven locally installed methods (PSIPRED, SSpro, PROFphd, PROF, Predator, DSC and GorIV) and 10 online servers (PORTER, Jpred, YASSPP, gorV, CDM, symPred, FDM, PCCSS, DISSpred and PROTEUS). Consensus methods are indicated in bold face. Also shown in the table are the Q3 and SOV results for CASP9 proteins that show less than 25% sequence similarity to the training proteins (PDBselect25) and 30% sequence similarity to the training proteins (PDBselect25), respectively.

<table>
<thead>
<tr>
<th>method</th>
<th>CASP9</th>
<th></th>
<th>CASP9: 25% sequence similarity</th>
<th></th>
<th>CASP9: 30% sequence similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q3</td>
<td>SOV</td>
<td>Q3</td>
<td>SOV</td>
<td>Q3</td>
</tr>
<tr>
<td>CONCORD</td>
<td>82.3</td>
<td>76.6</td>
<td>82.2</td>
<td>76.2</td>
<td>82.3</td>
</tr>
<tr>
<td>PSIPRED</td>
<td>81.7</td>
<td>76.6</td>
<td>81.6</td>
<td>76.2</td>
<td>81.7</td>
</tr>
<tr>
<td>SSpro</td>
<td>79.9</td>
<td>71.9</td>
<td>80</td>
<td>72</td>
<td>80.2</td>
</tr>
<tr>
<td>PROF</td>
<td>74.9</td>
<td>70.2</td>
<td>75</td>
<td>70</td>
<td>75.2</td>
</tr>
<tr>
<td>DSC</td>
<td>70.9</td>
<td>64.9</td>
<td>71.1</td>
<td>65.3</td>
<td>71.2</td>
</tr>
<tr>
<td>PROFphd</td>
<td>69.4</td>
<td>63.5</td>
<td>69.7</td>
<td>63.5</td>
<td>69.3</td>
</tr>
<tr>
<td>predator</td>
<td>62</td>
<td>58.7</td>
<td>62.2</td>
<td>58.8</td>
<td>61.9</td>
</tr>
<tr>
<td>gorIV</td>
<td>62</td>
<td>56.7</td>
<td>62.7</td>
<td>57.5</td>
<td>62</td>
</tr>
<tr>
<td>PORTER</td>
<td>81.6</td>
<td>74.4</td>
<td>81.6</td>
<td>74</td>
<td>81.7</td>
</tr>
<tr>
<td>YASSPP</td>
<td>80.8</td>
<td>75</td>
<td>81</td>
<td>75</td>
<td>81</td>
</tr>
<tr>
<td>Jpred</td>
<td>80.2</td>
<td>74.9</td>
<td>79.9</td>
<td>74.4</td>
<td>80.1</td>
</tr>
<tr>
<td>DISSpred</td>
<td>78.4</td>
<td>69.1</td>
<td>78.3</td>
<td>69.1</td>
<td>78.4</td>
</tr>
<tr>
<td>symPred</td>
<td>76.3</td>
<td>65.9</td>
<td>76.1</td>
<td>65.5</td>
<td>76.2</td>
</tr>
<tr>
<td>PCCSS</td>
<td>74.2</td>
<td>68.1</td>
<td>74.9</td>
<td>69.4</td>
<td>74.9</td>
</tr>
<tr>
<td>PROTEUS</td>
<td>73.1</td>
<td>67.8</td>
<td>73.5</td>
<td>67.9</td>
<td>73.1</td>
</tr>
<tr>
<td>gorV</td>
<td>67.4</td>
<td>63.2</td>
<td>67.4</td>
<td>63.4</td>
<td>67.4</td>
</tr>
<tr>
<td>CDM</td>
<td>65.3</td>
<td>66.3</td>
<td>65.7</td>
<td>61</td>
<td>65.7</td>
</tr>
<tr>
<td>FDM</td>
<td>62.1</td>
<td>54.3</td>
<td>62.8</td>
<td>54.7</td>
<td>62.7</td>
</tr>
</tbody>
</table>

with other methods. Two independent CASP9 test sets are generated against PDBselect25 by PISCES (Wang & Dunbrack 2003) using 25 per cent and 30 per cent sequence similarity as thresholds, respectively. Table 2 shows the results of various methods on the CASP9 25 per cent sequence similarity test set and 30 per cent sequence similarity test set. As shown in the table, the prediction accuracy Q3 of CONCORD remains the same (82.3%) for the CASP9 30 per cent sequence similarity test set and is 82.2 for the 25 per cent sequence similarity test set. The relative ranking of each method on these two sets of proteins remains the same. The results of the SOV scores shown in table 2 are similar to the Q3 results. CONCORD and PSIPRED have the highest SOV scores.

Binary prediction accuracies for Q(H/∼H), Q(E/∼E) and Q(C/∼C) are 90.5 per cent, 90.9 per cent and 90.9 per cent for the CASP9 test set, respectively (electronic supplementary material, table S3).

(c) Effect of post-processing

In order to study the effect of post-processing on the CONCORD predictions, we have compared the CONCORD predictions before the post-processing step.
Table 3. Effect of the post-processing step on the CONCORD performance. Q3a is the Q3 prediction accuracy after post-processing is applied, whereas Q3b is the Q3 prediction accuracy before post-processing is applied. Similarly, SOVa and SOVb are the SOV values after and before the post-processing step. CONCORD\textsubscript{V1} denotes the CONCORD method using all the seven methods, while CONCORD\textsubscript{V2} indicates the CONCORD method using five methods, excluding GorIV and Predator.

<table>
<thead>
<tr>
<th>method</th>
<th>Q3b</th>
<th>Q3a</th>
<th>SOVb</th>
<th>SOVa</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCORD\textsubscript{V1} (PDBselect25)</td>
<td>82.84</td>
<td>83.04</td>
<td>78.20</td>
<td>80.29</td>
</tr>
<tr>
<td>CONCORD\textsubscript{V2} (PDBselect25)</td>
<td>82.84</td>
<td>83.02</td>
<td>78.20</td>
<td>80.28</td>
</tr>
<tr>
<td>CONCORD\textsubscript{V1} (CASP9)</td>
<td>81.4</td>
<td>82.3</td>
<td>75.71</td>
<td>76.6</td>
</tr>
<tr>
<td>CONCORD\textsubscript{V2} (CASP9)</td>
<td>82.16</td>
<td>82</td>
<td>75.87</td>
<td>76.52</td>
</tr>
</tbody>
</table>

and the CONCORD predictions after the post-processing step. Both the sixfold cross validation and CASP9 results are presented. All the results are shown in table 3. Also shown in the table are the results for CONCORD based on five methods.

The results in table 3 indicate that post-processing slightly increases the prediction accuracy Q3 for CONCORD, from 82.84 to 83.04 per cent for sixfold cross validation and from 81.4 to 82.3 per cent for CASP9. However, the aim of post-processing is to eliminate the isolated predictions such as, one strand residue is surrounded by coil–helical residues, two helical residues are surrounded by coil–strand residues. This can be evaluated by the SOV score. The SOV score for sixfold cross validation is increased by 2.1 per cent, and the SOV score for CASP9 is increased by 0.9 per cent.

A similar trend is observed for CONCORD using five methods. The SOV score is increased more than Q3 scores after the post-processing step. However, for the CASP9 prediction, the post-processing step decreases the Q3 accuracy slightly, although the SOV score is increased. Much of the improvement seems to come from the post-processing step for eliminating unlikely segments.

4. Discussion

CONCORD uses information from the seven individual methods, such as the weight of each method and the prediction result (including the predicted amino acid type, H, E or C, and the confidence score for each amino acid type prediction); thus, the prediction accuracy of the consensus method is highly dependent on the individual method, especially those that have the larger weights in the consensus method, such as SSpro and PSIPRED.

CONCORD uses the confidence scores of the predicted amino acids of the seven methods and the weights of the seven methods to determine the secondary structure type for each amino acid. Because the SSpro method does not provide such confidence scores for its secondary structure predictions, its average prediction accuracy is used as the confidence score for the amino acids. The secondary structure prediction of CONCORD is based on the sum of the products
between the confidence score and the weight term over all methods, thus the consensus method can also provide the confidence of the prediction for each amino acid.

By performing the secondary structure predictions for the 18 methods on 107 CASP9 proteins, a consistent comparison for these methods is performed. For all the server predictions, the results reflect their most current capabilities; however, for the seven locally installed methods, this may not be true. For example, the PROFphd method, without using the multiple sequence alignment, generated lower accuracy than reported. Overall, there are five methods (PSIPRED, PORTER, YASSPP, Jpred and CONCORD) that have average accuracies above 80 per cent. Among all the other consensus methods (PROTEUS, CDM and Jpred), Jpred has the highest prediction accuracy after CONCORD. Among all the non-consensus methods, PSIPRED and PORTER generate the highest Q3 scores.

Although the results for CONCORD using seven methods and those for CONCORD using five methods are similar (table 3), CONCORD with seven methods shows slightly better performance consistently. The analysis of the effect of post-processing on the CONCORD performance shows that post-processing is an important step to generate a higher SOV score, while its impact on prediction accuracy is small.

We have also tested the performance of each method on two independent CASP9 test sets. One corresponds to the CASP9 proteins that show less than 25 per cent sequence similarity to PDBselect25, whereas the other one corresponds to the CASP9 proteins that show less than 30 per cent sequence similarity to PDBselect25. Both the Q3 and SOV results on these two test sets show similar performance on all CASP9 proteins. It can be seen from tables 1 and 2 that the SOV score of CONCORD is about 2 per cent better than the best of the individual methods for the sixfold cross validation. Although other methods, such as PSIPRED, PROF, etc., also generate lower SOV scores for the CASP9 targets than those for the sixfold cross validation, SSpro has the greatest decrease in SOV.

One possible reason that CONCORD achieved slightly better results than PSIPRED and SSpro is that the obtained weights from the training process (two weeks run) are suboptimal (longer training process causes memory issues).

It is claimed that 88 per cent accuracy is the theoretical limit for protein secondary structure prediction (Rost 2001). Further improvement of the secondary structure prediction accuracy may be due to new methods, or better consensus methods, all of which will benefit from the expanding PDB. The expanding PDB will possibly include more diverse set of protein sequences and structures, which allows better training.

5. Conclusions

A new consensus method, CONCORD, based on seven methods for protein secondary structure prediction has been proposed. This consensus method is based on a MILP model that produces the parameters for secondary structure prediction.

The test on PDBselect25 dataset by a sixfold cross validation and the test on the CASP9 target proteins (including the independent test sets of CASP9
proteins) show that CONCORD performs well compared with other individual methods according to the prediction accuracy. A comparison is also performed between CONCORD and another 10 secondary structure prediction servers (PORTER, Jpred, YASSPP, gorV, CDM, symPred, FDM, PCiSS, DISSpred and PROTEUS) by testing on the CASP9 target proteins. The results show that CONCORD performs well compared with all single servers (including several most recent ones). A web server, CONCORD, is available to the scientific community at http://helios.princeton.edu/CONCORD. The results file that the user receives by email also provides the confidence scores for each amino acid prediction.

C.A.F. gratefully acknowledges financial support from National Science Foundation, National Institutes of Health (R01 GM52032; R24 GM069736) and US Environmental Protection Agency, EPA (GAD R832721-010). Although the research described in the article has been funded in part by the US Environmental Protection Agency’s STAR programme through grant (R832721-010), it has not been subjected to any EPA review and does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References


Secondary structure prediction via MILP


Rajgaria, R., Wei, Y. & Floudas, C. A. 2010 Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction


