Amino Acids (2011) 40:963–973
DOI 10.1007/s00726-010-0721-1

ORIGINAL ARTICLE

iFC^2: an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content

Ke Chen · Wojciech Stach · Leila Homaeian · Lukasz Kurgan

Received: 9 March 2010 / Accepted: 6 August 2010 / Published online: 21 August 2010

Abstract Several descriptors of protein structure at the sequence and residue levels have been recently proposed. They are widely adopted in the analysis and prediction of structural and functional characteristics of proteins. Numerous in silico methods have been developed for sequence-based prediction of these descriptors. However, many of them do not have a public web-server and only a few integrate multiple descriptors to improve the predictions. We introduce iFC^2 (integrated prediction of fold, class, and content) server that is the first to integrate three modern predictors of sequence-level descriptors. They concern fold type (PFRES), structural class (SCEC), and secondary structure content (PSSC-core). The server exploits relations between the three descriptors to implement a cross-evaluation procedure that improves over the predictions of the individual methods. The iFC^2 annotates fold and class predictions as potentially correct/incorrect. When tested on datasets with low-similarity chains, for the fold prediction iFC^2 labels 82% of the PFRES predictions as correct and the accuracy of these predictions equals 72%. The accuracy of the remaining 28% of the PFRES predictions equals 38%. Similarly, our server assigns correct labels for over 79% of SCEC predictions, which are shown to be 98% accurate, while the remaining SCEC predictions are only 15% accurate. These results are shown to be competitive when contrasted against recent relevant web-servers. Predictions on CASP8 targets show that the content predicted by iFC^2 is competitive when compared with the content computed from the tertiary structures predicted by three best-performing methods in CASP8. The iFC^2 server is available at http://biomine.ece.ualberta.ca/1D/1D.html.

Keywords Protein structure classification · Secondary structure · Structural class · Fold type · SCOP

Abbreviations

CASP	Critical assessment of techniques for protein structure prediction
FASTA	FAST-all
iFC^2	Integrated prediction of fold class and content
iFC^2-FT	iFC^2 cross-evaluation for fold type
iFC^2-SSC	iFC^2 cross-evaluation for secondary structure content
iFC^2-SC	iFC^2 cross-evaluation for structural class
MAE	Mean absolute error
PSSM	Position-specific scoring matrix
PSSC-core	Prediction of secondary structure content through comprehensive sequence representation
SCEC	Prediction of structural class using evolutionary collocation
PDB	Protein data bank
PFRES	Protein fold recognition using evolutionary information and predicted secondary structure
SCOP	Structural classification of proteins
SVM	Support vector machine
3D	Tertiary

Electronic supplementary material The online version of this article (doi:10.1007/s00726-010-0721-1) contains supplementary material, which is available to authorized users.

K. Chen · W. Stach · L. Homaeian · L. Kurgan (✉)
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
e-mail: lkurgan@ece.ualbeta.ca

Introduction

The growing gap between the number of solved protein structures and the number of discovered protein sequences motivates research towards the development of computational methods for prediction of the tertiary structure. The performance of these methods is assessed biannually in the critical assessment of techniques for protein structure prediction (CASP) competition. Although the results of CASP competitions show that the prediction quality continues to rise, the predictions require substantial computing resources and in the case of new folds (sequences with no structural templates in the PDB) the predicted structure models are still relatively poor (Jauch et al. 2007). Recently, several descriptors of protein structure were suggested as an alternative way to study and predict the tertiary protein structure. These descriptors are either at the sequence or at the residue level. The sequence-level descriptors include protein family and superfamily type, fold type, structural class, and secondary structure content. The residue-level descriptors encompass secondary structure, solvent accessibility, residue depth, contact numbers and contact order, backbone dihedral and torsion angles, and B-factor. Numerous computational methods have been proposed for sequence-based prediction of the family and superfamily (Enright et al. 2002; Melvin et al. 2007; Gewehr et al. 2007), the fold type (Ding and Dubchak 2001; Hvidsten et al. 2003; Shen and Chou 2006; Jeong et al. 2006; Taguchi and Gromiha 2007; Melvin et al. 2007; Chen and Kurgan 2007; Shamim et al. 2007; Chen et al. 2008d; Damoulas and Girolami 2008; Shen and Chou 2009), the structural class (Chou 2000a, b, 2005a, b; Chou and Cai 2004; Chen et al. 2006, 2008a, b, 2009; Kedarisetti et al. 2006; Kurgan and Chen 2007; Lin and Li 2007; Kurgan et al. 2008; Xiao et al. 2008a, b; Zhang et al. 2008b; Li et al. 2008; Yang et al. 2009; Mizianty and Kurgan 2009a), and the secondary structure content (Chou 1999; Liu and Chou 1999; Cai et al. 2003a, b; Ruan et al. 2005; Lee et al. 2006; Homaeian et al. 2007). Similarly, the predictors of the residue-level descriptors include methods for prediction of the secondary structure (Jones 1999; McGuffin et al. 2000; Rost 2001, 2005, 2008; Rost et al. 2004; Pollastri and McLysaght 2005; Ofer and Yaoqi 2007; Montgomerie et al. 2008; Kurgan 2008; Ding et al. 2009), tight turns (Chou 1997, 2000a, b), the relative solvent accessibility (Ahmad and Gromiha 2002; Ahmad et al. 2003; Kim and Park 2004; Garg et al. 2005; Nguyen and Rajapakse 2006; Dor and Zhou 2007; Chen et al. 2008c; Mooney and Pollastri 2009), the residue depth (Yuan and Wang 2008; Zhang et al. 2008a), the contact order and number (Pollastri et al. 2001; Kinjo et al. 2005; Yuan 2005; Song and Burrage 2006; Shi et al. 2008), the backbone angles (Dor and Zhou 2007; Xue et al. 2008; Faraggi et al. 2009), and the B-factor (Radivojac et al. 2004; Yuan et al. 2005; Schlessinger and Rost 2005; Zhang et al. 2009). Another emerging descriptor of protein sequence is the pseudo amino acid composition, which comes in multiple forms and which was recently used to predict a number of different protein attributes (Chou 2001, 2005, 2009). The usefulness of these descriptors is supported by a recent study which shows that the tertiary structure can be recovered from several residue-level descriptors (Kinjo and Nishikawa 2005a, b). A study by Rangwala and Karypis (2006) reveals that sequence-level descriptors are useful in solving the remote homology detection problem. The knowledge of these descriptors was applied in a vast number of areas including reduction of the search space of possible conformations of the tertiary structures (Bahar et al. 1997), identification of domain boundaries in multi-domain protein structures (Redfern et al. 2007), analysis of interactions between CapZ protein and cell membranes (Smith et al. 2006), analysis of prion proteins (Concepcion et al. 2005), discrimination of outer membrane proteins (Gromiha 2005a, b; Gromiha and Suwa 2005), target selection for structural genomics (Mizianty and Kurgan 2009b; Kurgan and Mizianty 2009), and in prediction of tertiary structure (Chou 2004; Wu and Zhang 2008), residue–residue contacts (Björkholm et al. 2009), beta and gamma turns (Zhang et al. 2005; Zheng and Kurgan 2008; Hu and Li 2008), coding and non-coding RNAs (Liu et al. 2006), folding and unfolding rates (Gong et al. 2003; Ivankov and Finkelstein 2004; Gromiha 2005; Gromiha et al. 2006; Gromiha and Selvaraj 2008; Jiang et al. 2009; Shen et al. 2009; Chou and Shen 2009a; Gao et al. 2010), folding transition-state position (Huang and Cheng 2007), functional residues (Fischer et al. 2008), DNA-binding sites (Kuznetsov et al. 2006), RNA-binding residues (Wang et al. 2008), disordered proteins (Sethi et al. 2008), and enzyme proteins and their class (Dobson and Doig 2003, 2005).

However, many of the abovementioned prediction methods do not provide a web-server for public use and only a few of them integrate multiple descriptors. The integrated servers include PredictProtein (Rost et al. 2004), which predicts two residue-level descriptors, secondary structure and solvent accessibility, Real-SPINE (Dor and Zhou 2007; Faraggi et al. 2009) that also predicts two residue-level descriptors, solvent accessibility and backbone angles, a recent method by Pollastri and colleagues (Mooney and Pollastri 2009) that combines prediction of secondary structure, solvent accessibility, backbone motifs, and contact density, and a suite of predictors called SCRATCH (Cheng et al. 2005), which covers the predictions of secondary structure, solvent accessibility, disorder, strand residues, residue contact, B cell epitopes, solubility and antigenicity. We observe that some of the existing

sequence- and residue-level descriptors are closely related and thus they could be used to evaluate and/or re-predict the values of other descriptors. For instance, structural class depends on the secondary structure content (Kurgan et al. 2008) and different fold types belong to the same classes (Murzin et al. 1995). Motivated by the above observations we propose an integrated server which combines prediction of three sequence-level descriptors including fold type, structural class, and secondary structure content. Our server is the first to incorporate a cross-evaluation procedure, in which predictions of the three related descriptors are combined together to perform a cross check and a re-prediction. By applying the cross-evaluation, the predictions provided by our server are shown to be of higher quality than the predictions of the individual methods.

Materials and methods

Datasets

Five datasets were used to build and evaluate the proposed server. The data which were used to design and test the iFC^2 server were taken from Chen and Kurgan (2007). It includes 908 protein sequences that concern the 27 most populated folds in SCOP database. The sequence identity between any two sequences in this dataset is below 40%. Since the "small inhibitors toxins lectins" fold does not belong to the four structural classes that are predicted by SCEC, it was excluded from the dataset. As a result, 858 protein sequences are used and they cover 26 folds. The detailed description of the 26 folds and the four structural classes, including coding of the fold names, is given in Table 1 (Supplement). The 858 sequences are randomly divided into two equal-size subsets. One subset is used to design the iFC^2 server (*training dataset*), and the other is used to evaluate predictions generated by the server (*test dataset 1*). Among the 429 sequences in test dataset 1, 96 sequences share 25–40% sequence identity to one or more sequence of the training set. After removing the 96 sequences, the remaining 333 sequences constitute the *test dataset 2* in which no sequence shares above 25% similarity to any sequence in the training dataset. The importance of the reduction of the sequence similarity for the purpose of testing-related predictors was recently discussed in Chou and Shen (2006), (2007). The design is performed based on tenfold cross-validation tests on the training set to avoid overfitting and to assure that the results obtained on the training dataset would translate well into other, unseen data. We chose this type of the test which randomizes the selection of the ten folds, instead of the jackknife cross-validation which leads to a unique non-randomized result (Chou and Shen 2007, 2008) and which was recently used in related works (Li et al. 2009; Lin et al. 2009; Nanni and Lumini 2009; Vilar et al. 2009; Yang et al. 2009; Zeng et al. 2009), to reduce the computational time. The *test dataset 3* was used exclusively to test the prediction of the secondary structure content. The sequences were taken from the CASP8 competition; majority of these chains lack class and fold annotations and therefore they could not used to evaluate these predictions. Several structures in CASP8 are incomplete, i.e., they do not provide coordinates for some sequence segments. We removed targets that lack coordinates for over 20% of the residues, and as a result the test dataset 2 includes 97 proteins. Although majority of the CASP8 targets could be matched against templates in PDB at the time of the CASP8 competition, several targets were classified as template-free, i.e., they could not be predicted using a structural template. The eight template-free targets include T0397, T0416, T0443, T0465, T0482, T0496, T0510, and T0514. Since RAPTOR (one of the top-performing 3D-structure prediction methods that was compared against our server) did not provide a complete model for T0443 and as a result DSSP (Kabsch and Sander 1983) could not assign the secondary structure for this target, this target was excluded from our dataset. The *test dataset 4* consists of the remaining seven template-free targets.

Description of the PFRES, SCEC, and PSSC-core predictors

The iFC^2 server combines structural class predictions provided by the SCEC method (Chen et al. 2008b), fold type predictions generated by PFRES method (Chen and Kurgan 2007), and secondary structure content predicted with PSSC-core (Homaeian et al. 2007).

PFRES (Chen and Kurgan 2007) is a computational method for the sequence-based classification of the 27 most populated folds in the SCOP database (Murzin et al. 1995). It uses two sources of information for the prediction. The first source concerns evolutionary information. It is represented by a PSSM profile generated with PSI-BLAST (Altschul et al. 1997), which is utilized to compute PSI-BLAST profile-based composition vector. The detailed description of PSSM profile is included in Chen et al. (2008b). The other source is the secondary structure predicted with PSI-PRED (McGuffin et al. 2000). This structure is encoded into a set of features, which include secondary structure content, number of secondary structure segments, and arrangement of secondary structure segments; detailed definitions of these features are provided in Chen and Kurgan (2007). PFRES adopts a voting procedure which ensembles the predictions from three individual classifiers, including Support Vector Machine (SVM),

random forest and instance-based Kstar. PFRES achieves 66.4 and 68.4% accuracies for two test datasets that share below 35% sequence identity with the training set and was shown to outperform other relevant predictors at the time of its publication (Chen and Kurgan 2007).

SCEC (Chen et al. 2008b) predicts four types of the structural classes (all-α, all-β, α/β, and $\alpha + \beta$), as defined in SCOP (Murzin et al. 1995), based on a protein sequence. The all-α and all-β classes represent protein structures that consist of mainly α-helices and β-strands, respectively. The α/β and $\alpha + \beta$ classes contain both α-helices and β-strands which are mainly interspersed and segregated, respectively. In SCEC, protein sequence is represented by evolutionary collocation of amino acid (AA) pairs (Chen et al. 2008b). A total of 50 AA pairs are used and the frequencies of these AA pairs constitute the input into SVM classifier. SCEC achieves 61–96% accuracies when tested on several datasets characterized by different levels of sequence identity and it was demonstrated to provide competitive predictive quality at the time of its publication (Chen et al. 2008b).

PSSC-core (Homaeian et al. 2007) performs sequence-based prediction of secondary structure content. The features used to encode the input sequence, which are used in PSSC-core, include amino acid composition and composition moment vectors, frequency of tetra-peptides associated with helical and strand conformations, various property-based groups such as exchange groups, chemical groups of the side chains and hydrophobic group, auto-correlations based on hydrophobicity, side-chain masses, and hydropathy, and conformational patterns for β-sheets. PSSC-core uses a multiple linear regression model and achieves a mean absolute error (MAE) of 0.105 for helix content and 0.078 for strand content using a dataset of below 30% sequence identity, and it was shown to provide improved quality of predictions when compared with relevant existing methods at the time of the publication (Homaeian et al. 2007).

Architecture of the iFC^2 server

The architecture of the iFC^2 server is shown in Fig. 1. For a given protein sequence, PSSC-core uses only the sequence to generate its inputs, while the other two methods, SCEC and PFRES, also utilize sequence-derived information. The SCEC computes PSSM profile using PSI-BLAST while the inputs to PFRES are based on the protein sequence, the PSSM profile (the same as the one used by SCEC) and secondary structure predicted with PSI-PRED, which also utilizes PSSM profile as its input. The PSSC-core, SCEC, and PFRES methods generate their inputs and perform their predictions separately, although we reuse the PSI-PRED predictions and the PSI-BLAST outputs. Predictions from the three methods together with the secondary structure

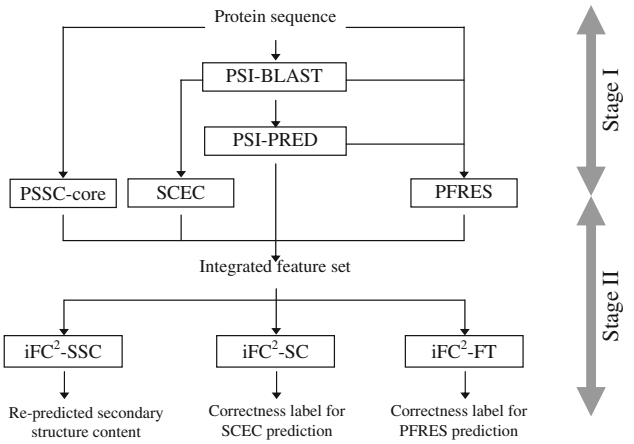

Fig. 1 Architecture of the iFC^2 method. In the first stage, the features based on the input protein sequence, PSSM profile generated with PSI-BLAST, and the secondary structure predicted with PSI-PRED are inputted into the individual prediction methods, i.e., PSSC-core, SCEC, and PFRES. In the second stage, the predictions made by individual predictors and the content values calculated from the secondary structure predicted with PSI-PRED are integrated into a new feature set. The new feature set is inputted into iFC^2-SSC, iFC^2-SC and iFC^2-FT predictors

content calculated from the output of PSI-PRED are fed into iFC^2's prediction models, which perform cross-evaluation and assessment of the consistency between the predictions made by three methods. For instance, given that PSSC-core would predict the helix, strand, and coil contents as 0.8, 0.0, and 0.2, respectively, and that SCEC would predict the structural class as all-β, the iFC^2 could recognize the conflict between the two predictions and fix the inconsistency. For the predictions of the structural class and the fold type, iFC^2 utilizes iFC^2-SC (cross-evaluation for structural class) and iFC^2-FT (cross-evaluation for fold type) prediction models that provide a binary output, i.e., "correct" or "incorrect" label is assigned to the predictions performed by SCEC and PFRES, respectively. These labels allow annotation of potentially incorrect predictions generated by these two methods. In the case of the secondary structure content prediction, iFC^2 uses the iFC^2-SSC (cross-evaluation for secondary structure content) model to re-predict content values by combining the knowledge of the predicted structural class, the predicted fold type, and the secondary structure content predicted by PSSC-core and PSI-PRED. The inputs to the iFC^2's prediction models (iFC^2-SSC, iFC^2-SC, and iFC^2-FT), which are denoted as the integrated feature set in Fig. 1, include (1) helix and strand contents predicted by PSSC-core, (2) helix and strand contents calculated from secondary structure predicted with PSI-PRED, (3) structural class predicted with SCEC, and (4) fold type predicted by PFRES. The four content values are continuous (real-values) while the predicted structural class and the fold type are discrete. The parameterization of the

classifiers in the iFC^2 server was performed based on tenfold cross-validation on the training dataset. Detailed results of the parameterization are given in Tables 2 and 3 (Supplement).

The iFC^2 server was implemented as a web-server that is available at http://biomine.ece.ualberta.ca/1D/1D.html. Detailed procedure of how to use the web-server is given in the Supplement. As pointed out recently by Chou and Shen (2009b), the user-friendly and publicly accessible web-servers represent the future direction for developing practically useful predictors.

iFC^2-SSC

iFC^2-SSC is a linear regression model with two linear functions, one for prediction of the helix content and the other for the strand content.

$$
\begin{aligned}
&predicted\_strand\_content = \\
&0.159 * PSSCcore\_strand\_content \\
&+0.624 * PSIPRED\_strand\_content \\
&+0.126 * (fold\_type == g) \\
&+0.106 * (fold\_type == h) \\
&+0.104 * (fold\_type == i) \\
&+0.170 * (fold\_type == j) \\
&+0.212 * (fold\_type == k) \\
&+0.140 * (fold\_type == l) \\
&+0.129 * (fold\_type == n) \\
&+0.212 * (fold\_type == o) \\
&+0.104 * (fold\_type == z) \\
&+\ldots\ldots + 0.005
\end{aligned}
$$

$$
\begin{aligned}
&predicted\_helix\_content = \\
&0.100 * PSSCcore\_helix\_content \\
&+0.560 * PSIPRED\_helix\_content \\
&-0.260 * PSIPRED\_strand\_content \\
&+0.252 * (fold\_type == a) \\
&+0.100 * (fold\_type == b) \\
&+0.179 * (fold\_type == c) \\
&+0.252 * (fold\_type == d) \\
&+0.179 * (fold\_type == e) \\
&+0.179 * (fold\_type == f) \\
&+0.102 * (fold\_type == i) \\
&+0.102 * (fold\_type == n) \\
&+0.160 * (fold\_type == p) \\
&+0.130 * (fold\_type == q) \\
&+0.160 * (fold\_type == r) \\
&+0.160 * (fold\_type == s) \\
&+0.160 * (fold\_type == t) \\
&+0.190 * (fold\_type == u) \\
&+0.160 * (fold\_type == v) \\
&+0.160 * (fold\_type == w) \\
&+0.222 * (fold\_type == x) \\
&+0.130 * (fold\_type == z) \\
&+\ldots\ldots + 0.079
\end{aligned}
$$

The less significant inputs for which the regression coefficient values are between -0.1 and 0.1 are omitted in the above formulas. We observe that the prediction model for the helix content incorporates all six fold types that correspond to the all-α class (folds a, b, c, d, e, and f) and majority of the fold types that correspond to the α/β and $\alpha + \beta$ classes (folds p, q, r, s, t, u, v, w, x, and z) with the corresponding coefficients greater than 0.1. Moreover, such high coefficient values are assigned to only two fold types that concern the all-β class (folds i and n). This agrees with the underlying biology, i.e., the all-α class contains larger amount of helices than α/β and $\alpha + \beta$ classes, and the latter two classes should contain more helical residues than the all-β class. In the case of the prediction model for the strand content, seven out of eight fold types that correspond to the all-β class are assigned coefficients greater than 0.1, and only one fold that corresponds to the remaining three structural classes has the coefficient greater than 0.1. This again agrees with the fact that the all-β class includes protein chains that have more strand residues than the other three structural classes. The secondary structure contents calculated from the secondary structure predicted by PSI-PRED are assigned the largest coefficients in both formulas and this suggests that these two inputs contribute the most to the prediction of the secondary structure contents.

iFC^2-SC

This predictor is implemented based on an SVM classifier. The use of the SVM classifier in PFRES, SCEC, and to implement the iFC^2 server is motivated by the prior successful applications of this method in related areas (Chou and Cai 2002; Cai et al. 2003; Melvin et al. 2007). The classification model takes the predictions of PSSC-core, SCEC, PFRES, and PSI-PRED as the input and produces a binary output, i.e., either "correct" or "incorrect", which is attached to the structural class predicted by SCEC. The aim of the iFC^2-SC model is to flag the "incorrect" predictions provided by SCEC (resulting in diminished user's confidence in these predictions) and to validate (increasing the user's confidence) the "correct" predictions. The SVM model was parameterized based on tenfold cross-validation performed on the training dataset. We considered selection of a kernel function [we evaluated polynomial and radial basis function (RBF) kernels], parameterization of the kernel function (the value of the width of the RBF kernel γ, and the exponent of the polynomial kernel e), and selection of the value of the soft margin constant C. The experimental results shown in Table 2 (Supplement) resulted in selection of the RBF kernel with $\gamma = 1.5$ and $C = 1$; this configuration provides the highest relative operating characteristic (ROC) value.

iFC^2-FT

This model is also implemented as an SVM-based classifier, which is designed to verify the predictions of the PFRES method. Similar to iFC^2-SC, it uses the predictions of PSSC-core, SCEC, PFRES, and PSI-PRED as the input and it outputs a binary value that annotates the fold type predicted by PFRES as either "correct" or "incorrect". As in the case of the iFC^2-SC, the SVM used in the iFC^2-FT was parameterized based on the tenfold cross-validation on the training dataset. The parameterization results that are

summarized in Table 3 (Supplement) led to selection of $C = 5$ and the RBF kernel with $\gamma = 2$.

Results and discussion

Results for the secondary structure content, fold and structural class predictions

The performance of the iFC^2 server was evaluated on the test datasets 1 and 2. The results generated by the server were compared against predictions generated by the individual input prediction methods, i.e., PSSC-core, PFRES and SCEC, as well as against other recent web-servers for the fold and structural class predictions that were published after the publication of the methods (PSSC-core, PFRES and SCEC) that are integrated into the server. We note that the comparative analysis includes only the methods that are provided as web-servers since the proposed work is also implemented this way, and that we could not find recent web-servers for the content prediction. Instead, we compare the content predictions against results generated by modern tertiary structure predictions, which is summarized in the following section. The comparison includes *multiple classifiers* (Chen et al. 2009) and *MODAS* (Mizianty and Kurgan 2009a) servers for the structural class prediction, which are available at http://chemdata.shu.edu.cn/protein_st/ and at http://biomine.ece.ualberta.ca/MODAS/, respectively, and the *SVM-Fold* (Melvin et al. 2007) and *PFP-FunDSeqE* (Shen and Chou 2009) servers that can be accessed at http://svm-fold.c2b2.columbia.edu and http://www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/, respectively.

Using the dataset 1, the MAE values of helix/strand content predicted by PSSC-core and iFC^2-SSC equal 0.125/0.085 and 0.06/0.049, respectively (see Table 1). The error reduction rates achieved by the proposed server are $0.065/0.125 = 52\%$ for the helix content prediction and $0.036/0.085 = 42\%$ for the strand content prediction. The PCC values are improved from 0.75 to 0.94 and from 0.72 to 0.89 for the helix and strand content predictions, respectively. Similar results are observed for test dataset 2 that has a reduced similarity to the training dataset (see Table 1). The corresponding MAE values of helix/strand content predicted by PSSC-core and iFC^2-SSC equal 0.119/0.086 and 0.061/0.048, respectively. The results suggest that the iFC^2-SSC obtains comparable results on datasets with the 25 and 40% sequence identity to the training sequences.

For the structural class prediction, the iFC^2-SC assigns "correct" label for 79.3% of predictions made by the SCEC for the chains in the dataset 1 (see Table 2). Among these "correct" predictions, the accuracy of the SCEC equals 98.2%, while the accuracy of the SCEC for the predictions deemed as "incorrect" by the iFC^2-SC equals

Table 1 Comparison of the quality of predictions generated with iFC^2-SSC and the PSSC-core

Algorithm	Test dataset 1 (40% identity)				Test dataset 2 (25% identity)			
	Helix content		Strand content		Helix content		Strand content	
	MAE	PCC	MAE	PCC	MAE	PCC	MAE	PCC
PSSC-core	0.125	0.75	0.085	0.72	0.119	0.76	0.086	0.72
iFC^2-SSC	0.060	0.94	0.049	0.89	0.061	0.94	0.048	0.89

The quality of the secondary structure content prediction is measured with mean absolute error (MAE) and Pearson correlation coefficient (PCC)

Table 2 Comparison of the quality of predictions generated with iFC^2-SC, SCEC and with two recent web-servers for prediction of structural classes, multiple classifiers and MODAS

Algorithm	Test dataset 1 (40% identity)						Test dataset 2 (25% identity)					
	Accuracy (%)					Coverage (%)	Accuracy (%)					Coverage (%)
	All-α	All-β	α/β	$\alpha + \beta$	Overall		All-α	All-β	α/β	$\alpha + \beta$	Overall	
SCEC	80.00	83.58	87.83	41.46	80.89	100	78.00	82.69	87.76	40.63	80.18	100
Multiple classifiers	81.54	85.82	85.71	80.49	84.62	100	84.00	84.62	84.35	84.38	84.38	100
MODAS	90.77	86.57	91.01	80.49	88.58	100	90.00	85.58	91.16	81.25	88.29	100
iFC^2-SC	100	98.32	98.80	88.89	98.20	79.3	100	97.62	98.45	92.86	98.12	79.9

We report the coverage (defined as the percentage of SCEC predictions that are predicted by iFC^2-SC as correct) and accuracy = TP/(TP + FP), where TP and FP are the number of true-positive and false-positive predictions, respectively, and where TP refers to the correct predictions of SCEC that are also predicted as correct by iFC^2-SC and FP refers to incorrect predictions of SCEC that are predicted as correct by iFC^2-SC. The methods are sorted by the increasing values of the overall accuracy

only 14.6%. When contrasted with the accuracy of the SCEC for the entire test dataset 1, which equals 80.9%, the post-screening performed with iFC^2-SC improves the accuracy by 17.3% as a trade-off for removing (labeling as "incorrect") 20.7% of the predictions. When compared with the recent web-servers for the structural class prediction, the multiple classifiers and MODAS, the iFC^2-SC achieves 10–14% higher accuracy and improves results for all four classes. The accuracies of the individual predictors on the two test datasets, i.e., test datasets 1 and 2, are relatively similar (see Table 2).

Similarly, in the case of the fold type prediction, iFC^2-FT labels 81.8% of the PFRES predictions for the sequences from the dataset 1 as "correct" and the accuracy of these predictions equals 71.8% (see Table 3). At the same time, the accuracy of the PFRES for the predictions identified by iFC^2-FT as "incorrect" is only 38.5%. When compared with the accuracy of 65.7% that was obtained by PFRES for the entire test dataset 1, the removal of 18.2% of the predictions which were labeled as "incorrect" by iFC^2-FT leads to the improvement of the PFRES accuracy by 6.1%. When contrasted against the two recent web-servers for the prediction of the fold types, the iFC^2-FT achieves accuracy that is comparable with the SVM-Fold and improves over the PFP-FunDSeqE by about 18% (see Table 3). We observe that the iFC^2-FT achieves relatively similar accuracies for the different folds, which is in contrast to the SVM-Fold. More specifically, for the test dataset 1 the SVM-Fold obtains 100% accuracy for ten folds and 0% accuracy for another ten folds, while the

Table 3 Comparison of the quality of predictions generated with iFC^2-FT, PFRES and with two recent web-servers for prediction of fold types, PFP-FunDSeqE and SVM-Fold

Folds	Test dataset 1 (40% identity)				Test dataset 2 (25% identity)			
	PFRES	iFC^2-FT	PFP-FunDSeqE	SVM-Fold	PFRES	iFC^2-FT	PFP-FunDSeqE	SVM-Fold
a.1	83.33	66.67	100.00	100.00	100.00	100.00	100.00	100.00
a.3	100.00	100.00	40.00	0.00	100.00	100.00	40.00	0.00
a.4	75.86	80.00	51.72	100.00	66.67	72.22	47.62	100.00
a.24	23.08	25.00	23.08	100.00	18.18	20.00	18.18	100.00
a.26	100.00	100.00	100.00	0.00	100.00	100.00	100.00	0.00
a.39	77.78	87.50	66.67	88.89	71.43	83.33	57.14	85.71
b.1	91.67	94.64	75.00	96.67	91.67	93.33	72.92	95.83
b.6	60.00	66.67	70.00	100.00	50.00	57.14	62.50	100.00
b.121	66.67	100.00	33.33	100.00	80.00	100.00	40.00	100.00
b.29	60.00	60.00	30.00	0.00	50.00	50.00	25.00	0.00
b.34	53.85	66.67	23.08	100.00	42.86	75.00	14.29	100.00
b.40	45.00	50.00	30.00	95.00	50.00	50.00	37.50	93.75
b.42	55.56	55.56	44.44	100.00	42.86	42.86	28.57	100.00
b.47	25.00	33.33	25.00	0.00	33.33	50.00	0.00	0.00
b.60	100.00	100.00	50.00	0.00	100.00	100.00	50.00	0.00
c.1	89.71	92.42	76.47	97.06	90.00	93.75	74.00	100.00
c.3	58.33	58.33	66.67	0.00	54.55	54.55	63.64	0.00
c.23	40.00	50.00	35.00	100.00	35.71	50.00	28.57	100.00
c.2	46.15	52.17	73.08	0.00	42.86	50.00	66.67	0.00
c.37	46.15	54.55	53.85	0.00	45.00	52.94	50.00	0.00
c.47	68.75	62.50	68.75	100.00	66.67	57.14	66.67	100.00
c.55	14.29	0.00	0.00	85.71	14.29	0.00	0.00	85.71
c.69	53.85	63.64	0.00	0.00	45.45	55.56	0.00	0.00
c.93	100.00	100.00	100.00	0.00	100.00	100.00	100.00	0.00
d.15	30.00	28.57	10.00	100.00	28.57	25.00	0.00	100.00
d.58	61.29	65.52	29.03	90.32	60.00	65.22	28.00	88.00
Overall	65.73	71.79	53.38	73.89	63.36	69.63	49.85	72.07
Average per fold	62.55	65.91	49.05	59.76	60.87	65.31	45.23	60.04
Coverage	100	81.8	100	100	100	81.1	100	100

We report the coverage and accuracy (the definitions are given in Table 2). The overall accuracy refers to the accuracy for the entire dataset and the average per fold accuracy corresponds to the average of the accuracies for the individual folds

iFC^2-FT achieves 100% accuracy for five folds and 0% accuracy for only one fold. As a result, the average accuracy per fold for the iFC^2-FT equals 65.9%, which is 6% higher than the average accuracy per fold for the SVM-Fold. Table 3 shows that the accuracies of the four individual prediction methods are about 2–4% lower on the test dataset 2 when compared with the test dataset 1, but the overall conclusions concerning differences between the methods are consistent for both dataset.

Comparison with leading 3D-structure predictors on prediction of secondary structure content

The three top-performing 3D-structure prediction methods in the CASP8 competition, Zhang-Server (Zhang 2007), RAPTOR (Xu et al. 2003) and Pro-sp3-TASSER (Zhou et al. 2007), are compared against the iFC^2-SSC server in the context of the secondary structure content prediction (see Table 4). The 3D-structures predicted by the three methods were processed using DSSP program and the secondary structure contents were calculated from the assigned secondary structure. For the iFC^2-SSC predictions, the coil content is computed as 1 − (strand_content + helix_content), and is rounded to zero if negative. For the test dataset 3 (all CASP8 sequences), the iFC^2-SSC obtains the lowest MAE for the strand content and second lowest MAE for the coil content. Overall the iFC^2-SSC is surpassed only by the Zhang-Server and provides competitive predictions when compared with the Pro-sp3-TASSER and RAPTOR. For the test dataset 4 (new folds in CASP8) our server obtains the lowest MAE for the strand and coil contents. The MAEs of the iFC^2-SSC predictions are similar for both datasets, while the 3D-structure predictors produce lower quality models for the template-free sequences. The larger improvements obtained for the new folds could be explained by the fact that the iFC^2 server is independent of the availability of structural templates. These results suggest that the strand and coil contents predicted by iFC^2-SSC could be useful for the refinement of the 3D-structure predictions of modern 3D-structure predictors for the template-free modeling.

Table 4 Comparison of the quality of the secondary structure content predictions generated with the iFC^2 server and the leading 3D structure prediction methods

Methods	MAE on test dataset 3			MAE on test dataset 4		
	Helix	Strand	Coil	Helix	Strand	Coil
PSSC-core	0.105	0.098	0.107	0.108	0.124	0.115
iFC^2-SSC	0.088	0.061	0.075	0.073	0.062	0.065
Zhang-Server	0.043	0.066	0.074	0.042	0.166	0.136
Pro-sp3-TASSER	0.053	0.086	0.125	0.080	0.162	0.194
RAPTOR	0.104	0.100	0.108	0.064	0.147	0.149

Conclusions

The iFC^2 server integrates three related modern predictors of the sequence-level descriptors that include SCEC for the structural class, PFRES for the fold type, and PSSC-core for the secondary structure content. The server is fully automated and requires only the FASTA formatted sequences as the input. It includes classification and prediction models that improve content predictions, when compared with the PSSC-core, and provides useful annotations for the predicted class and fold. The structural class and fold type predictions generated by the server are shown to be competitive when compared with relevant recent web-servers on datasets with low sequence identity. The content values predicted by the iFC^2 server are also shown to be competitive or better with the content extracted from the predictions of the state-of-the-art tertiary protein structure predictors. Our server provides higher quality predictions for the strand content although these structures are more difficult to predict than helices and coils due to the underlying non-local, with respect to the sequence, interactions involved in their formation. The strand and coil contents predicted by iFC^2 could be useful for the refinement of the tertiary structure predictions for the template-free modeling.

Acknowledgments KC and WS research was supported by the Alberta Ingenuity and iCORE Scholarships. LK acknowledges support from NSERC Canada.

References

Ahmad S, Gromiha MM (2002) NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18:819–824

Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins 50:629–635

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 17:3389–3402

Bahar I, Atilgan AR, Jernigan RL, Erman B (1997) Understanding the recognition of protein structural classes by amino acid composition. Proteins 29:172–185

Björkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, Hvidsten TR (2009) Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contacts. Bioinformatics 25:1264–1270

Cai YD, Liu XJ, Chou KC (2003a) Prediction of protein secondary structure content by artificial neural network. J Comput Chem 24:727–731

Cai YD, Zhou GP, Chou KC (2003b) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84:3257–3263

Chen K, Kurgan L (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23:2843–2850

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448

Chen C, Chen LX, Zou XY, Cai PX (2008a) Predicting protein structural class based on multi-features fusion. J Theor Biol 253:388–392

Chen K, Kurgan L, Ruan J (2008b) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem 29:1596–1604

Chen K, Kurgan M, Kurgan L (2008c) Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values. J Biomed Sci Eng 1:1–9

Chen Y, Chen Q, Chen F, Zhao Y (2008d) Protein fold recognition based on error correcting output codes and SVM. Protein Pept Lett 15:443–447

Chen L, Lu L, Feng K, Li W, Song J, Zheng L, Yuan Y, Zeng Z, Feng K, Lu W, Cai Y (2009) Multiple classifier integration for the prediction of protein structural classes. J Comput Chem 30:2248–2254

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76

Chou KC (1997) Prediction and classification of alpha-turn types. Biopolymers 42:837–853

Chou KC (1999) Using pair-coupled amino-acid composition to predict protein secondary structure content. J Protein Chem 18:473–480

Chou KC (2000a) Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1:171–208

Chou KC (2000b) Prediction of tight turns and their types in proteins. Anal Biochem 286:1–16

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255

Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC (2005a) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6:423–436

Chou KC (2005b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6:262–274

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009

Chou KC, Shen HB (2006) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347:150–157

Chou KC, Shen HB (2007) Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Shen HB (2009a) Recent advances in developing web-servers for predicting protein attributes. Nat Sci 2:63–92

Chou KC, Shen HB (2009b) FoldRate: a web-server for predicting protein folding rates from primary sequence. Open Bioinform J 3:31–50

Concepcion GP, David MP, Padlan EA (2005) Why don't humans get scrapie from eating sheep? A possible explanation based on secondary structure predictions. Med Hypotheses 64:919–924

Damoulas T, Girolami MA (2008) Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics 24:1264–1270

Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17:349–358

Ding YS, Zhang TL, Gu Q, Zhao PY, Chou KC (2009) Using maximum entropy model to predict protein secondary structure with single sequence. Protein Pept Lett 16:552–560

Dobson PD, Doig AJ (2003) Distinguishing enzyme structures from non-enzymes without alignments. J Mol Biol 330:771–783

Dobson PD, Doig AJ (2005) Predicting enzyme class from protein structure without alignments. J Mol Biol 345:187–199

Dor O, Zhou Y (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 68:76–81

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584

Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74:847–856

Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613–620

Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. Proteins 78(9):2114–2130

Garg A, Kaur H, Raghava GP (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 61:318–324

Gewehr JE, Hintermair V, Zimmer R (2007) AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings. Bioinformatics 23:1203–1210

Gong H, Isom DG, Srinivasan R, Rose GD (2003) Local secondary structure content predicts folding rates for simple, two-state proteins. J Mol Biol 327:1149–1154

Gromiha M (2005a) Motifs in outer membrane protein sequences: applications for discrimination. Biophys Chem 117:65–71

Gromiha M (2005b) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 45:494–501

Gromiha M, Selvaraj S (2008) Bioinformatics approaches for understanding and predicting protein folding rates. Curr Bioinform 3:1–9

Gromiha M, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 21:961–968

Gromiha M, Selvaraj S, Thangakani AM (2006) A statistical method for predicting protein unfolding rates from amino acid sequence. J Chem Inf Model 46:1503–1508

Homaeian L, Kurgan L, Ruan J, Cios KJ, Chen K (2007) Prediction of protein secondary structure content for the twilight zone sequences. Proteins 69:486–498

Hu X, Li Q (2008) Using support vector machine to predict beta- and gamma-turns in proteins. J Comput Chem 29:1867–1875

Huang JT, Cheng JP (2007) Prediction of folding transition-state position (T) of small, two-state proteins from local secondary structure content. Proteins 68:218–222

Hvidsten TR, Kryshtafovych A, Komorowski J, Fidelis K (2003) Novel approach to fold recognition using sequence-derived

properties from sets of structurally similar local fragments of proteins. Bioinformatics 19(Suppl 2):ii81–ii91
Ivankov DN, Finkelstein AV (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 101:8942–8944
Jauch R, Yeo HC, Kolatkar PR, Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. Proteins 69(Suppl 8): 57–67
Jeong J, Berman P, Przytycka T (2006) Fold classification based on secondary structure—how much is gained by including loop topology? BMC Struct Biol 6:3
Jiang Y, Iglinski P, Kurgan L (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 30:772–783
Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988
Kim H, Park H (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 54:557–562
Kinjo AR, Nishikawa K (2005a) Recoverable one-dimensional encoding of protein three-dimensional structures. Bioinformatics 21:2167–2170
Kinjo AR, Nishikawa K (2005b) Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. Biophysics 1:67–74
Kinjo AR, Horimoto K, Nishikawa K (2005) Predicting absolute contact numbers of native protein structure from amino acid sequence. Proteins 58:158–165
Kurgan L (2008) On the relation between the predicted secondary structure and the protein size. Protein J 27:234–239
Kurgan L, Chen K (2007) Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun 357:453–460
Kurgan L, Mizianty M (2009) Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. Nat Sci 1(2):93–106
Kurgan L, Cios K, Chen K (2008a) SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinform 9:226
Kurgan L, Zhang T, Zhang H, Shen S, Ruan J (2008b) Secondary structure based assignment of the protein structural classes. Amino Acids 35:551–556
Kuznetsov IB, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 64:19–27
Lee S, Lee BC, Kim D (2006) Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62:1107–1114
Li ZC, Zhou XB, Lin YR, Zou XY (2008) Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35:581–590
Li S, Li H, Li M, Shyr Y, Xie L, Li Y (2009) Improved prediction of lysine acetylation by support vector machines. Protein Pept Lett 16:977–983
Lin H, Li QZ (2007) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466
Lin ZH, Wang HL, Zhu B, Wang YQ, Lin Y, Wu YZ (2009) Estimation of affinity of HLA-A*0201 restricted CTL epitope based on the SCORE function. Protein Pept Lett 16:561–569
Liu W, Chou KC (1999) Prediction of protein secondary structure content. Protein Eng 12:1041–1050
Liu J, Gough J, Rost B (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2:529–536
McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405
Melvin I, Ie E, Kuang R, Weston J, Stafford WN, Leslie C (2007) SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. BMC Bioinform 8(Suppl 4):S2
Mizianty M, Kurgan L (2009a) Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinform 10:414
Mizianty M, Kurgan L (2009b) Meta prediction of protein crystallization propensity. Biochem Biophys Res Commun 390(1): 10–15
Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart D (2008) PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res 36:W202–W209
Mooney C, Pollastri G (2009) Beyond the Twilight Zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins 77: 181–190
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
Nanni L, Lumini A (2009) A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease. Protein Pept Lett 16:163–167
Nguyen MN, Rajapakse JC (2006) Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins 63:542–550
Ofer D, Yaoqi Z (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 66:838–845
Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21: 1719–1720
Pollastri G, Baldi P, Fariselli P, Casadio R (2001) Improved prediction of the number of residue contacts in proteins by recurrent neural networks. Bioinformatics 1:S234–S242
Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK (2004) Protein flexibility and intrinsic disorder. Protein Sci 13:71–80
Rangwala H, Karypis G (2006) Building multiclass classifiers for remote homology detection and fold recognition. BMC Bioinform 7:455
Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput Biol 3:e232
Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134:204–218
Rost B (2005) Prediction of protein structure in 1D—secondary structure, membrane regions, and solvent accessibility. In: Bourne PE, Weissig H (eds) Struct Bioinform 44:559–587
Rost B (2008) Prediction of protein structure in 1D—secondary structure, membrane regions, and solvent accessibility. In: Bourne PE, Weissig H (eds) Structural Bioinformatics. Wiley, New York

Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32:W321–W326

Ruan J, Wang K, Yang J, Kurgan L, Cios KJ (2005) Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. Artif Intel Med 35:19–35

Schlessinger A, Rost B (2005) Protein flexibility and rigidity predicted from sequence. Proteins 61:115–126

Sethi D, Garg A, Raghava GP (2008) DPROT: prediction of disordered proteins using evolutionary information. Amino Acids 35:599–605

Shamim MT, Anwaruddin M, Nagarajaram HA (2007) Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23:3320–3327

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22:1717–1722

Shen HB, Chou KC (2009) Predicting protein fold pattern with functional domain and sequential evolution information. J Theor Biol 256:441–446

Shen HB, Song JN, Chou KC (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng 2:136–143

Shi Y, Zhou J, Arndt D, Wishart DS, Lin G (2008) Protein contact order prediction from primary sequences. BMC Bioinform 9:255

Smith J, Diez G, Klemm AH, Schewkunow V, Goldmann WH (2006) CapZ–lipid membrane interactions: a computer analysis. Theor Biol Med Model 3:33–37

Song JN, Burrage K (2006) Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinform 7:425

Taguchi Y, Gromiha M (2007) Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinform 8:404

Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E (2009) A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J Theor Biol 261:449–458

Wang Y, Xue Z, Shen G, Xu J (2008) PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids 35:295–302

Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 72:547–556

Xiao X, Lin WZ, Chou KC (2008a) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29:2018–2024

Xiao X, Wang P, Chou KC (2008b) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254:691–696

Xu J, Li M, Kim D, Xu Y (2003) RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 1:95–117

Xue B, Dor O, Faraggi E, Zhou Y (2008) Real-value prediction of backbone torsion angles. Proteins 72:427–433

Yang JY, Peng ZL, Yu ZG, Zhang RJ, Anh V, Wang D (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J Theor Biol 257:618–626

Yuan Z (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinform 6:248

Yuan Z, Wang ZX (2008) Quantifying the relationship of protein burying depth and sequence. Proteins 70:509–516

Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B-factor profiles. Proteins 58:905–912

Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259:366–372

Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69(8):108–117

Zhang Q, Yoon S, Welsh WJ (2005) Improved method for predicting β-turn using support vector machine. Bioinformatics 21: 2370–2374

Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L (2008a) Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinform 9:388

Zhang TL, Ding YS, Chou KC (2008b) Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L (2009) On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 76:617–636

Zheng C, Kurgan L (2008) Prediction of ß-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinform 9:430

Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, Skolnick J (2007) Analysis of TASSER-based CASP7 protein structure prediction results. Proteins 69(8):90–97