Journal of Theoretical Biology 402 (2016) 117–128

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids

Gaurav Raicar [a,*], Harsh Saini [a], Abdollah Dehzangi [b], Sunil Lal [c], Alok Sharma [a,d,e]

[a] The University of the South Pacific, Fiji
[b] University of Iowa, USA
[c] Massey University, New Zealand
[d] IIIS, Griffith University, Australia
[e] RIKEN, Japan

HIGHLIGHTS

- A Forward Consecutive Search (FCS) scheme is proposed.
- Physicochemical attributes are strategically selected.
- Physicochemical-based features supplement existing feature extraction techniques.
- Improvements in prediction accuracies after utilizing physicochemical information.

ARTICLE INFO

Article history:
Received 21 January 2016
Received in revised form
20 April 2016
Accepted 2 May 2016
Available online 7 May 2016

Keywords:
Protein fold recognition
Structural class prediction
Physicochemical properties
Syntactical-based features
Evolutionary-based features
Forward consecutive search scheme

ABSTRACT

Predicting the three-dimensional (3-D) structure of a protein is an important task in the field of bioinformatics and biological sciences. However, directly predicting the 3-D structure from the primary structure is hard to achieve. Therefore, predicting the fold or structural class of a protein sequence is generally used as an intermediate step in determining the protein's 3-D structure. For protein fold recognition (PFR) and structural class prediction (SCP), two steps are required – feature extraction step and classification step. Feature extraction techniques generally utilize syntactical-based information, evolutionary-based information and physicochemical-based information to extract features. In this study, we explore the importance of utilizing the physicochemical properties of amino acids for improving PFR and SCP accuracies. For this, we propose a Forward Consecutive Search (FCS) scheme which aims to strategically select physicochemical attributes that will supplement the existing feature extraction techniques for PFR and SCP. An exhaustive search is conducted on all the existing 544 physicochemical attributes using the proposed FCS scheme and a subset of physicochemical attributes is identified. Features extracted from these selected attributes are then combined with existing syntactical-based and evolutionary-based features, to show an improvement in the recognition and prediction performance on benchmark datasets.

1. Introduction

In the field of bioinformatics and biological sciences, predicting the three-dimensional (3-D) structure of a protein plays a crucial role. The functions of protein, being closely linked to its structure enable us to further understand the cellular functions, protein-protein interactions and aids the development of new drug designs and therapies (Chmielnicki and Stapor, 2012). The multitude of protein sequences generated due to large-scale sequencing projects are significantly higher than the known 3-D protein structure. Computational techniques have to be employed to determine the structure of a protein quickly and efficiently.

Directly predicting the protein 3-D structure from its sequence is hard to achieve. However, classifying protein sequences to their fold or structural class is a transitional stage in determining the 3-D structure of a protein. In order to determine the fold or structural class of a protein sequence, two steps are required: 1)

* Corresponding author.
E-mail addresses: raicar_g@usp.ac.fj (G. Raicar), saini_h@usp.ac.fj (H. Saini), i.dehzangi@gmail.com (A. Dehzangi), s.lal@massey.ac.nz (S. Lal), sharma_al@usp.ac.fj (A. Sharma).

http://dx.doi.org/10.1016/j.jtbi.2016.05.002
0022-5193/

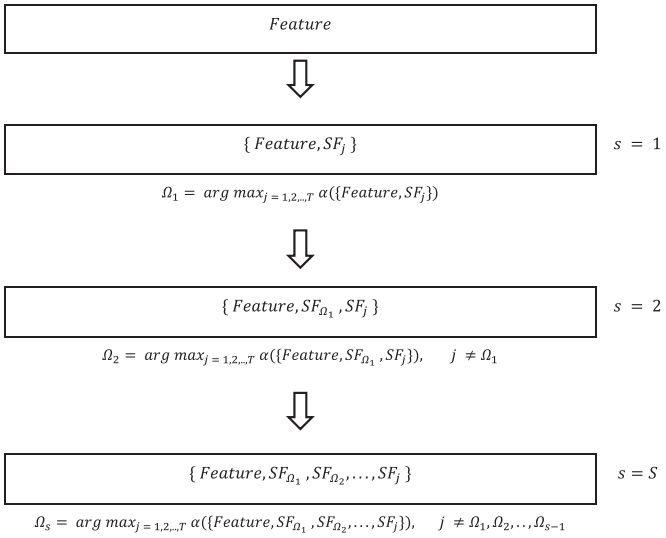

Fig. 1. Forward Consecutive Search (FCS) Scheme.

feature extraction step and 2) classification step. In feature extraction step, informative features are extracted from primary protein sequences. These features are further used in the classification step for protein fold recognition (PFR) and structural class prediction (SCP). If the extracted features are well discriminative, it can help improving the recognition and prediction rate. This makes feature extraction a crucial step in the overall procedure (Dehzangi et al., 2013a, 2013b, 2013c, 2013d, 2014a, 2014c; Deschavanne and Tuffery, 2009; Dong et al., 2009; Kavousi et al., 2011; Lyons et al., 2014, 2015, 2016; Paliwal et al., 2014b; Sharma et al., 2013a, 2014; Saini et al., 2014, 2015).

A lot of research has been done in the domain of protein SCP (Chou and Zhang, 1994; Chou, 1995; Bahar et al., 1997; Zhou, 1998; Chou and Maggiora, 1998; Zhou and Assa-Munt, 2001; Heffernan et al., 2015a, 2015b). One of the important progresses made in this domain was a study conducted by Chou and Cai (2004). They proposed a scheme whereby the feature vector of a protein sample was represented by its functional domain composition to formulate the predictor. The validation was made on a very stringent benchmark dataset which covers the following 7 classes: (i) all-alpha, (ii) all-beta, (iii) alpha/beta, (iv) alpha+beta, (v) multi-domain, (vi) small protein, and (vii) peptide. The cutoff threshold was 20%, meaning that none of proteins included in the benchmark dataset has greater than 20% pairwise sequence identity to any other in a same subset. For such an extremely stringent benchmark dataset, the overall jackknife success rate by the "Functional Domain Composition" method was over 90%. The pseudo amino acid composition approach has also been widely used by many investigators (Chen et al., 2006b, 2012; Sahu and Panda, 2010; Zhang et al., 2014; Qin et al., 2012) for predicting protein structural classes.

In the literature, many feature extraction techniques have been developed and used for PFR and SCP. Features are generally extracted by utilizing syntactical-based, evolutionary-based and physicochemical-based information. Features which are dependent on physicochemical attributes can reveal global properties of proteins (Bulashevska and Eils, 2006; Chinnasamy et al., 2005). These features are able to maintain high discriminatory information even

Table 1
DD Dataset ***n***-fold cross validation.

	Feature	Baseline Accuracy (***n***=10) (%)	Improved Accuracy (***n***=10) (%)	Rank
PFR	**PF1**	50.6	62.3	537, 339, 199, 317, 466
	PSSM+PF1	66.4	69	314, 453, 351, 469, 1
	O	51	65.6	12, 535, 314, 70, 1
	PSSM+O	64.9	70.6	537, 179, 399, 440, 1
	Bigram	74.1	74.7	463, 394, 151, 205, 471
	Separated dimers (*K*=7)	76	77.1	463, 536, 16, 1, 203
SCP	**PF1**	71.8	79.1	179, 216, 84, 466, 340
	PSSM+PF1	81.8	83.7	239, 461, 442, 1, 340
	O	67.8	80.8	12, 537, 179, 346, 1
	PSSM+O	77.1	82.9	537, 345, 70, 472, 1
	Bigram	83.3	84.4	463, 114, 308, 1, 2
	Separated dimers (*K*=7)	86.4	87.5	84, 536, 114, 394, 350

n-fold cross-validation was carried out 100 times for statistical stability.
Improved Accuracy refers to the *n*-fold cross-validation accuracy of the combination of features {*Feature*, SF}.

Table 2
TG Dataset ***n***-fold cross validation.

	Feature	Baseline Accuracy (***n***=10) (%)	Improved Accuracy (***n***=10) (%)	Rank
PFR	**PF1**	38.8	50.4	532, 341, 199, 461, 340
	PSSM+PF1	52.7	59	180, 343, 465, 463, 440
	O	36.3	51.3	535, 199, 349, 490, 491
	PSSM+O	46.7	57.3	512, 348, 461, 1, 2
	Bigram	68.1	70.5	494, 222, 205, 147, 81
	Separated dimers (*K*=3)	73.5	74.5	151, 347, 460, 471, 1
SCP	**PF1**	69.9	80.3	209, 314, 346, 151, 443
	PSSM+PF1	77.2	84.7	209, 355, 442, 346, 205
	O	63.6	81.3	537, 209, 351, 442, 199
	PSSM+O	73.4	84.3	199, 348, 343, 442, 463
	Bigram	81.5	86.8	494, 351, 469, 217, 244
	Separated dimers (*K*=3)	87.7	89.3	211, 63, 483, 3, 488

n-fold cross-validation was carried out 100 times for statistical stability.
Improved Accuracy refers to the *n*-fold cross-validation accuracy of the combination of features {*Feature*, SF}.