Mol Divers (2008) 12:171–179
DOI 10.1007/s11030-008-9093-9

FULL LENGTH PAPER

Prediction of protein structural classes using hybrid properties

Wenjin Li · Kao Lin · Kaiyan Feng · Yudong Cai

Received: 5 August 2008 / Accepted: 25 September 2008 / Published online: 25 October 2008

Abstract In this paper, amino acid compositions are combined with some protein sequence properties (physiochemical properties) to predict protein structural classes. We are able to predict protein structural classes using a mathematical model that combines the nearest neighbor algorithm (NNA), mRMR (minimum redundancy, maximum relevance), and feature forward searching strategy. Jackknife cross-validation is used to evaluate the prediction accuracy. As a result, the prediction success rate improves to 68.8%, which is better than the 62.2% obtained when using only amino acid compositions. Therefore, we conclude that the physiochemical properties are factors that contribute to the protein folding phenomena and the most contributing features are found to be the amino acid composition. We expect that prediction accuracy will improve further as more sequence information comes to light. A web server for predicting the protein structural classes is available at http://app3.biosino.org:8080/liwenjin/index.jsp.

Keywords Protein structural class · Nearest neighbor algorithm · mRMR (Minimum Redundancy, Maximum Relevance) · Physiochemical properties · Amino acid compositions

Electronic supplementary material The online version of this article (doi:10.1007/s11030-008-9093-9) contains supplementary material, which is available to authorized users.

W. Li · K. Lin · Y. Cai (✉)
CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
e-mail: cyd@picb.ac.cn

K. Feng
Division of Imaging Science & Biomedical Engineering, The University of Manchester, Room G424, Stopford Building, Manchester, M13 9PT, UK

Introduction

Three-dimensional (3-D) structures of proteins are closely related to their primary structure—their amino acid sequence. For many years scientists have tried to develop a prediction model to correlate amino acid sequences to the protein structures. Although early prediction studies only used amino acid compositions without amino acid sequence data [1–7], one study was capable of predicting a 3-D protein structure with 84% accuracy [8]. With the realization that amino acid sequences contribute factors to increase the accuracy of protein folding predictions, many researchers added more information related to the sequence, such as the pseudo amino acid compositions which involve not only amino acid compositions but also the sequence-order and length information [9], hydropobicity, polarity and distribution of certain amino acids [10–13] to enhance the prediction capability. Recently, Ding [14] employed eight physiochemical features to construct pseudo amino acid compositions and managed to gain 92.6% prediction accuracy using dual-layer fuzzy support vector machine (FSVM) network. Ding's work only covered 204 proteins and four structural classes: all-α all-β, α/β and $\alpha+\beta$.

This paper does not aim for an increase in the prediction rate. Instead, we are investigating whether combining the physiochemical properties and amino acid compositions are better than using amino acid compositions alone in predicting the protein structural classes. We will also test the hypothesis that the physiochemical properties, derived from the amino acid sequence arrangement, contribute more to the prediction of protein structural classes than using the amino acid compositions alone. We will demonstrate that, by combining the mRMR (Minimum Redundancy, Maximum Relevance) [15] and forward feature selections, we are able to first optimize the prediction model and second increase the efficiency of

building the prediction model. Instead of the traditional four structural classes we extend them to seven structural classes, which include all-α, all-β, α/β, $\alpha+\beta$, multi-domain proteins, membrane and cell surface proteins, small proteins. We also use a fairly large data set with 12520 proteins to minimize the defect of getting a high correct prediction rate by chance. Nearest neighbor algorithm (NNA) [16] is used to predict which structural class a query protein should be placed.

Materials and methods

Dataset

According to the SCOP ("Structural Classification of Proteins") [17–19], proteins belong to seven structural classes: all alpha proteins (all-α), all beta proteins (all-β), alpha and beta proteins (α/β), alpha and beta proteins ($\alpha+\beta$), multi-domain proteins (γ), membrane and cell surface proteins (δ), and small proteins (ζ). The dataset we are using, including 2299 all-α, 3334 all-β, 3086 α/β, 2870 $\alpha+\beta$, 224 γ, 227 δ, and 984 ζ, is released by ASTRAL (release 1.71, 2007, http://astral.berkeley.edu/) [20–22]. After removing the proteins whose sequence containing unnatural amino acids, we get the refined dataset with 2270 all-α, 3199 all-β, 2842 α/β, 2812 $\alpha+\beta$, 213 γ, 222 δ, and 962 ζ—totally 12520 proteins (refer to Table 1 and supplemental material 1).

Combined protein sequence descriptors

Each protein is represented as a 111-dimensional vector which consists of 20 amino acid compositions and 91 physiochemical features. The physiochemical features include the properties of: hydrophobicity, normalized Van Der Waals volume, polarity, polarizability and solvent accessibility.

(1) Amino acid compositions: the percentage of each of the normal 20 amino acids occurring in the whole sequence.

(2) Hydrophobicity, normalized Van Der Waals volume, polarity and polarizability: global description of the amino acid sequence can be used to obtain 21 features for each of these properties [10,23].

The method for obtaining global properties such as hydrophobicity are as follows: First, each amino acid is classified into three categories – polar, neutral and hydrophobic amino acid [23]. For a given protein sequence, the polar amino acids, R,K,E,D,Q,N, are substituted by character P, the neutral amino acids, G,A,S,T,P,H,Y, are substituted by character N, and the hydrophobic amino acids, C,V,L,I,M,F,W, are substituted by character H. Thus each protein property sequence is a sequence of P, N, and H, instead of amino acids. Then, composition (C) is taken as the percentage of P, N or H; transition (T) is defined as the changing frequency between two different properties (such as the transition from P to N, or P to H, or H to N); distribution (D) is defined as how much of the protein sequence is needed to contain 25%, 50%, 75% and 100% of the Ps, Ns and Hs, respectively. An example below shows how these percentages and how all 21 features are obtained.

Suppose a property sequence under examine contains 9 Ps, 16 Ns and 11 Hs (totally 36), as showed in Fig. 1. The compositions (C) for P, N and H are $(9/36)*100\% = 25\%$, $(16/36)*100\% = 44.4\%$ and $(11/36)*100\% = 30.6\%$, respectively. The numbers of transitions between P and N, between P and H and between N and H are 9, 7, and 10, respectively. Therefore, the three transitions (T) between P, N and H are $(9/36)*100\% = 25\%$, $(7/36)*100\% = 19.4\%$ and $(10/36)*100\% = 27.8\%$, respectively. In this case, we search P in the property sequence from N terminal to C terminal. The first P appears in the 4th of the property sequence, i.e. the length of the first segment is 4. Therefore, the first distribution (D) value for P is $(4/36)*100\% = 11.1\%$. When the 25% of Ps are included, the length of the segment is 5. The second distribution (D) value for P is thus $(5/36)*100\% = 13.9\%$, and so forth. The third, forth and fifth distribution (D) values are $(13/36)*100\% = 36.1\%$, $(22/36)*100\% = 61.1\%$ and $(34/36)*100\% = 94.4\%$, respectively. Similarly, the five distribution values for N are 2.8%, 27.8%, 47.2%, 75%, 97.2%, and the five distribution values for H are 5.6%, 8.3%, 41.7%, 80.6%, 100%. Therefore, the property sequence derived from hydrophobicity property can produce 21 features: 3 for composition C(0.25, 0.444, 0.306), 3 for transition T(0.25, 0.194, 0.278) and 15 for distribution D(0.111, 0.139, 0.361, 0.611, 0.944, 0.028, 0.278, 0.472, 0.75, 0.972, 0.056, 0.083, 0.417, 0.806, 1).

Using the same way described above, the 21 features can be obtained from each of the other three physiochemical properties: normalized Van Der Waals volume, polarity and polarizability, respectively.

(3) Solvent accessibility: residues of a protein can be divided into two group (buried and exposed). A residue is considered as exposed if the percentage of the exposed surface is larger than 20%. The solvent accessibility of a protein

Table 1 Description of the dataset

Class	Original dataset	Dataset after refinement
All-α	2299	2270
All-β	3334	3199
α/β	3086	2842
$\alpha+\beta$	2870	2812
γ	224	213
δ	227	222
ζ	984	962
Overall	13006	12520

The original dataset is from ASTRAL SCOP release 1.71 and the refined dataset used in this study

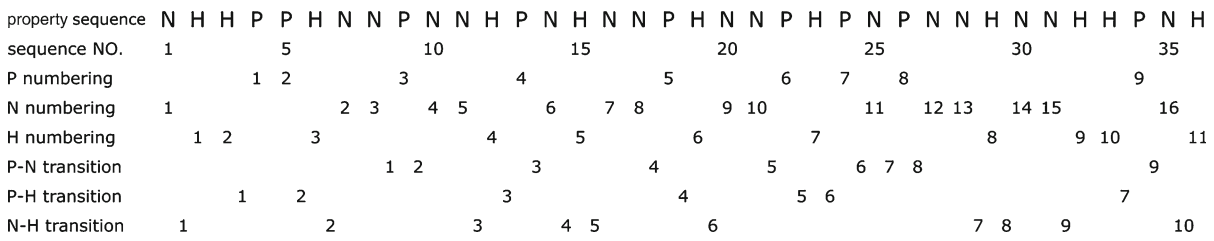

Fig. 1 Analyzing property sequence derived from hydrophobicity property

is obtained by PredAcc [24] in this study. We denote the exposed residues and exposed residues with gamma risks as E, and the hidden residues and hidden residues with gamma risks as H, from which we can obtain a property sequence with only E and H. Only composition for H, transition between H and E, and five distributions for H are considered as useful features (totally 7) to avoid feature redundancy [10].

In conclusion, the total number of features is $20+21*4+7=111$ (see Table 2). Each protein sequence is transferred into 111-dimensional vector (see supplemental materials 2), which needs to be normalized by the following equations:

$$\mu_j = \sum_{i=1}^{n} v_{ij}/n \tag{1}$$

$$S_j = \sqrt{\sum_{i=1}^{n} (v_{ij} - \mu_j)^2/(n-1)} \tag{2}$$

$$V_{ij} = \frac{v_{ij} - \mu_j}{S_j} \tag{3}$$

where v_{ij} is the value of the jth feature in the ith protein, n is the total number of proteins, S_j is the standard deviation of the jth feature, and V_{ij} is the normalization value of jth feature in the ith protein.

Minimum redundancy-maximum relevance (mRMR) [15,25]

Feature selection can very effectively reduce the feature dimensions, improve a learning machine's generalization, facilitating the data mining task. We choose mRMR feature selection algorithm because it is able to balance the minimum redundancy and the maximum relevance in a simple and elegant mathematical way. The maximum relevance part looks for features that contribute most to the classification, and the minimum redundancy part tries to exclude the features whose prediction capability has been included by the already selected features. The algorithm is described briefly as below:

Given a dataset Ω including all features, we are seeking a subset S of the features to satisfy both the minimum redundancy and the maximum relevance conditions.

Firstly, the mutual information I of two variables x and y is defined as:

$$I(x, y) = \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{4}$$

where $P(x_i, y_j)$ is the joint probabilistic distribution of x_i and y_j; $P(x_i)$ and $P(y_j)$ are the marginal probabilities of x_i and y_j, respectively.

The minimum redundancy is calculated as:

$$\min_{S\subseteq\Omega} W_I, W_I = \frac{1}{|S|^2} \sum_{i,j\in S} I(x_i, x_j) \tag{5}$$

where $|S|$ is the number of features in$S$, $W_I$is the minimum redundancy value.

For the targeted classes $h = \{h_1, h_2, \ldots, h_k\}$ (i.e. the class variable), the relevance between feature i and the targeted class variable h can be quantified by the mutual information $I(h, x_i)$ between h and the feature variablex_i. The maximum relevance can be obtained by the following formula:

$$\max_{S\subseteq\Omega} V_I, V_I = \frac{1}{|S|} \sum_{i\in S} I(h, x_i) \tag{6}$$

where V_I is the maximum relevance value.

In order to optimize both Eqs. 5 and 6, mRMR is accomplished by the following steps:

(1) Select a single most relevant feature according to Eq. 6, i.e. select feature i such that $I(h, x_i)$ is higher than other features.

$$\max_{i\in\Omega} I(x_i, h) \tag{7}$$

Table 2 Feature distribution

Sequence properties	The number of features in certain property			Total
	C	T	D	
Amino acid composition		20		20
Hydrophobicity	3	3	15	21
Normalized Van Der Waal volume	3	3	15	21
Polarity	3	3	15	21
Polarizability	3	3	15	21
Solvent accessibility	1	1	5	7

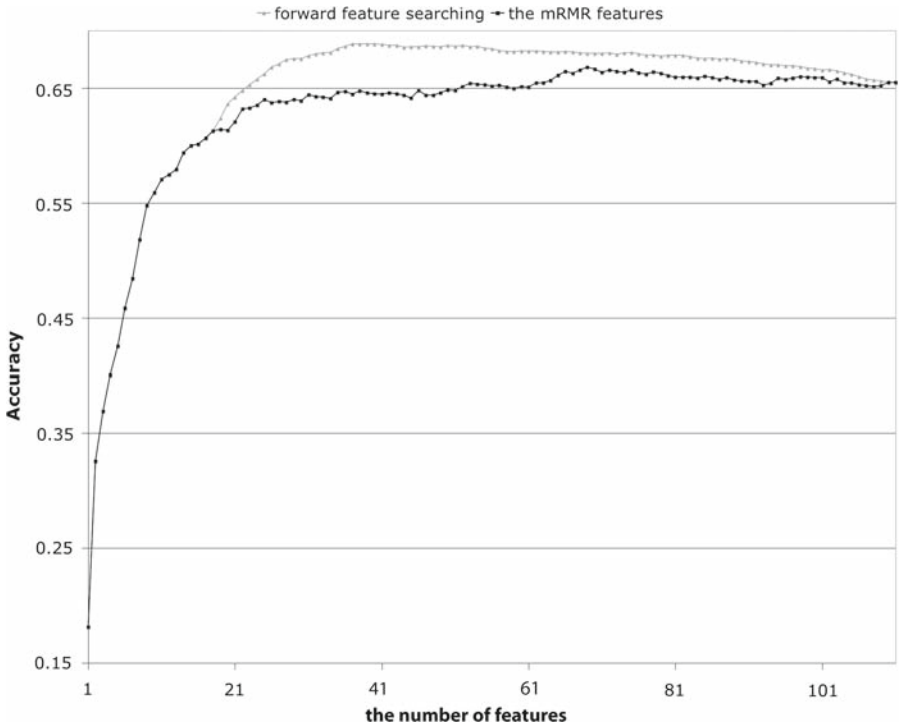

Fig. 2 The accuracy curves produced by the mRMR features (Black Square) and the forward feature searching method from the 19th feature (Grey Triangle)

where Ω is the whole feature set.

(2) The rest features are selected by adding one additional feature i each time to S to satisfy either of the two conditions in (8) and (9):

$$\max_{i\in\Omega_S}[I(h,x_i)-\frac{1}{|S|}\sum_{j\in S}I(x_i,x_j)] \tag{8}$$

$$\max_{i\in\Omega_S}[I(h,x_i)/\frac{1}{|S|}\sum_{j\in S}I(x_i,x_j)] \tag{9}$$

where $\Omega_s = \Omega - S$, representing the set of features that are yet to be selected. The Eq. 8 is called MID (mutual information difference criterion) selection criterion, while the Eq. 9 is called MIQ (mutual information quotient criterion) selection criterion. In this research, we choose Eq. 8 as the selection criterion.

The mRMR feature selection is fulfilled without the involvement of a prediction model, which can be performed very quickly. However, the optimization of feature selection through mRMR does not guarantee the selected features are also best for a particular prediction model. We describe a common feature selection strategy which involves a prediction model as below.

Forward feature searching strategy

Given a whole feature set Ω, and an initially selected feature subset $S(S \subseteq \Omega)$, the rest features can be selected by adding an additional feature i, such that $S\cup\{i\}$ satisfies the following condition:

$$\max_{i\in\Omega_S} A(S\cup\{i\}) \tag{10}$$

where $A(S\cup\{i\})$ means the prediction accuracy obtained by the prediction model evaluated by an evaluation method such as the jackknife test.

Results and discussion

As described in Sect. "Combined protein sequence descriptors", each protein is represented by a 111-dimentional space. Thus the 12520 proteins can be expressed by a $12520 * 111$ matrix (shown in supplemental material 2). Using mRMR method (see Sect. "Minimum redundancy-maximum relevance"), we obtain two feature lists (see supplemental material 3): the first list, named as maxRel features, showing the features in maximum relevance order, and the second list, named as mRMR features, showing the features selected by Eqs. 7 and 8 in a selection order. In order to find out how many foremost features in the mRMR feature list should be included for the prediction model, we add one feature at a time from the list in order and obtain the prediction accuracy for the selected features using jackknife test. In this study, we choose to run all the features in the list and gain the optimized feature subset that achieves the highest prediction accuracy. A scheme is designed to adjust between finding global prediction accuracy and local prediction accuracy if one encounters a large feature set. As a result, the highest accuracy is found to be 66.8% and it takes place when the

Table 3 The order and name of mRMR features with the prediction accuracies evaluated by jackknife cross-validation test

1	AA_composition_C	0.18131	57	Hydrophobicity_transition_PH	0.652476
2	Solvent_composition_H	0.325799	58	Polarity_distribution_N-1.0	0.650799
3	AA_composition_L	0.369329	59	Polarizability_distribution_N-1.0	0.649361
4	Polarity_composition_H	0.400639	60	Hydrophobicity_distribution_P-0.75	0.651118
5	AA_composition_V	0.425479	61	Polarity_distribution_H-0.0	0.650639
6	AA_composition_T	0.458626	62	Hydrophobicity_transition_PN	0.654712
7	VanDerWaal_composition_N	0.483946	63	Polarity_composition_N	0.654633
8	AA_composition_A	0.518211	64	AA_composition_Y	0.656629
9	AA_composition_G	0.547764	65	Polarity_distribution_P-0.25	0.661022
10	AA_composition_S	0.558866	66	AA_composition_R	0.664297
11	AA_composition_M	0.570607	67	Hydrophobicity_distribution_H-1.0	0.66262
12	VanDerWaal_distribution_N-0.25	0.574601	68	AA_composition_K	0.665575
13	Solvent_distribution_H-0.0	0.579073	69	Hydrophobicity_distribution_H-0.75	0.668211
14	AA_composition_W	0.59369	70	Polarity_distribution_H-1.0	0.666613
15	VanDerWaal_composition_P	0.6	71	Hydrophobicity_distribution_N-1.0	0.663578
16	Solvent_distribution_H-0.75	0.601118	72	Polarity_transition_PN	0.665575
17	VanDerWaal_distribution_H-0.0	0.606629	73	Polarizability_distribution_N-0.5	0.664457
18	AA_composition_E	0.612859	74	AA_composition_F	0.663419
19	VanDerWaal_distribution_N-1.0	0.613898	75	Polarity_distribution_N-0.75	0.665575
20	Solvent_transition_HE	0.613099	76	Hydrophobicity_distribution_H-0.25	0.663259
21	Solvent_distribution_H-0.25	0.620687	77	VanDerWaal_transition_PN	0.661901
22	AA_composition_P	0.631629	78	Polarity_distribution_H-0.5	0.664137
23	VanDerWaal_distribution_N-0.0	0.632428	79	AA_composition_Q	0.66278
24	VanDerWaal_distribution_N-0.5	0.635064	80	Polarity_transition_NH	0.660863
25	Hydrophobicity_composition_H	0.640096	81	VanDerWaal_distribution_P-0.0	0.659505
26	Hydrophobicity_distribution_N-0.0	0.637141	82	Polarizability_distribution_P-0.25	0.659505
27	Polarity_distribution_P-0.0	0.638578	83	Polarizability_composition_P	0.659425
28	Hydrophobicity_composition_P	0.63746	84	Hydrophobicity_distribution_P-1.0	0.658786
29	AA_composition_I	0.640096	85	Polarity_distribution_H-0.25	0.660304
30	Solvent_distribution_H-0.5	0.638658	86	Polarity_distribution_P-0.5	0.658866
31	VanDerWaal_transition_NH	0.644089	87	Polarizability_composition_N	0.657508
32	VanDerWaal_distribution_N-0.75	0.642572	88	Polarity_distribution_N-0.5	0.658946
33	Polarizability_distribution_P-1.0	0.642013	89	Polarity_transition_PH	0.656869
34	VanDerWaal_distribution_H-1.0	0.640974	90	Polarity_distribution_H-0.75	0.65599
35	AA_composition_D	0.646086	91	VanDerWaal_distribution_H-0.25	0.655911
36	Polarizability_distribution_P-0.0	0.646965	92	Polarizability_distribution_N-0.75	0.655751
37	Polarizability_distribution_N-0.0	0.644728	93	Polarity_distribution_N-0.25	0.652476
38	Hydrophobicity_composition_N	0.647684	94	Hydrophobicity_distribution_N-0.75	0.653994
39	Polarizability_distribution_H-0.0	0.645767	95	AA_composition_N	0.658626
40	Solvent_distribution_H-1.0	0.644808	96	VanDerWaal_distribution_H-0.75	0.657268
41	VanDerWaal_transition_PH	0.644728	97	Polarizability_composition_H	0.658466
42	Hydrophobicity_distribution_P-0.0	0.645927	98	Polarizability_transition_PH	0.659744
43	Polarizability_distribution_N-0.25	0.644968	99	VanDerWaal_distribution_P-0.25	0.659665
44	Hydrophobicity_transition_NH	0.64369	100	Hydrophobicity_distribution_P-0.5	0.658946
45	Hydrophobicity_distribution_H-0.0	0.641134	101	Polarizability_transition_PN	0.659185
46	Polarity_distribution_P-0.75	0.647923	102	VanDerWaal_distribution_P-1.0	0.655351
47	Polarity_distribution_N-0.0	0.64377	103	Hydrophobicity_distribution_N-0.5	0.657827
48	Polarizability_distribution_P-0.75	0.64361	104	VanDerWaal_distribution_P-0.75	0.654712
49	Hydrophobicity_distribution_P-0.25	0.645927	105	VanDerWaal_distribution_H-0.5	0.654393
50	Polarity_composition_P	0.648163	106	VanDerWaal_distribution_P-0.5	0.653035
51	Polarity_distribution_P-1.0	0.647843	107	Hydrophobicity_distribution_N-0.25	0.651997
52	AA_composition_H	0.651278	108	Polarizability_distribution_H-0.75	0.651358
53	Polarizability_distribution_H-0.25	0.654073	109	VanDerWaal_composition_H	0.651837
54	Hydrophobicity_distribution_H-0.5	0.653195	110	Polarizability_transition_NH	0.654872
55	Polarizability_distribution_H-1.0	0.652955	111	Polarizability_distribution_H-0.5	0.654952
56	Polarizability_distribution_P-0.5	0.651757			

AA_composition_C means the amino acid composition of amino acid C, solvent_composition_H is the H compositon obtained from the solvent accessibility of proteins, and so forth. The highest accuracy is 66.8%, which takes place when the 69th feature is added in

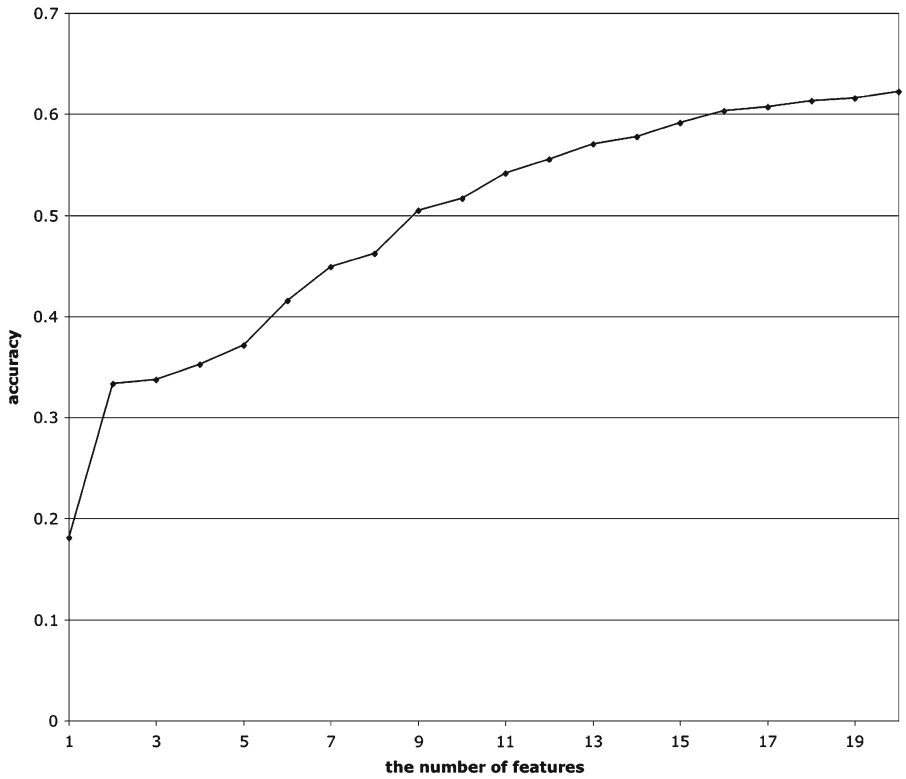

Fig. 3 The accuracy curve using only amino acid compositions: one feature is added each time in the order listed in the mRMR features

69th feature is added (see Fig. 2, Table 3). We also carry out a prediction using pure amino acid compositions in the same way as the 111-dimensional features. The highest accuracy is 62.2% (see Fig. 3) when all 20 amino acid compositions are included. We conclude that physiochemical properties do provide some extra prediction capability to the original amino acid compositions.

As we mentioned above, the mRMR method does not involve the mathematical prediction model. An issue is then brought forth how to combine mRMR method with forward feature searching method. The mRMR method is good at fast computation while the forward feature searching method is more accurate and often provides better selection results. The mRMR method can provide an initial optimized feature selection quickly, and the time for processing the forward feature selection is polynomial to the size of a feature set. In this study, we choose to select the foremost features using mRMR method and the rest features using forward feature selection method. Our choice may not be optimized as there are more ways to combine them to balance between speed and accuracy. Our method best compares the forward feature selection and the mRMR in parallel curves as described below. We choose the first 18 features of mRMR features (see Table 3) as the initial feature subset, which has reduced some computation burden left for the forward feature selection method. The first 18 mRMR features are chosen mainly because they continuously give an ascending accuracy curve until it achieves an accuracy of 61.3%. Then, we search the rest of the features using forward feature searching strategy, and gain an accuracy of 68.8% when 37 features are selected (see Fig. 2; Table 4). If more than 37 features are selected, the prediction model suffers more from the overfitting problem and the prediction accuracy deteriorates as the number of features increases. By comparing the two curves after the 18 features are selected in the x axis in Fig. 2, the forward feature selection method does provide better feature selection results as the compensation of more computation. Could one divide the features from mRMR method into several fragments and use the forward feature selection method in each of those fragments? This aspect will be investigated in the future.

Table 5 lists the 22 foremost features taken from the maxRel features (i.e. the maximum relevant part in the mRMR feature selection). The foremost 18 mRMR features can all be found in these 22 maxRel features. This tells that the maximum relevance is weighted strongly towards building the mRMR features. One could introduce a weighting parameter in Eqs. 8 or 9 to adjust the balance between maximum relevance and the minimum redundancy. In this study, the maxRel features were weighted strongly. We will leave the study of the balance between maximum relevance and the minimum redundancy in the future.

The arrangement of the amino acids in the protein impart more information than the amino acid composition alone. Therefore, we hypothesize that the physiochemical properties derived from the sequence arrangement should contribute more to the prediction of protein structural classes. Along with testing the hypothesis, we have also done the feature

Table 4 The order and names of the features selected by the forward feature searching method with the prediction accuracies evaluated by the jackknife cross-validation

1	AA_composition_C	0.18131	57	Hydrophobicity_composition_H	0.682748
2	Solvent_composition_H	0.325799	58	VanDerWaal_distribution_N-0.5	0.681789
3	AA_composition_L	0.369329	59	Hydrophobicity_distribution_H-0.25	0.681949
4	Polarity_composition_H	0.400639	60	Hydrophobicity_distribution_N-0.5	0.682189
5	AA_composition_V	0.425479	61	Polarity_distribution_H-0.25	0.682029
6	AA_composition_T	0.458626	62	Polarity_distribution_P-0.25	0.682109
7	VanDerWaal_composition_N	0.483946	63	Polarizability_distribution_P-0.25	0.681949
8	AA_composition_A	0.518211	64	Polarity_distribution_P-0.5	0.68131
9	AA_composition_G	0.547764	65	Polarity_distribution_P-1.0	0.68123
10	AA_composition_S	0.558866	66	AA_composition_I	0.681629
11	AA_composition_M	0.570607	67	Hydrophobicity_distribution_N-0.75	0.68123
12	VanDerWaal_distribution_N-0.25	0.574601	68	Polarity_distribution_H-1.0	0.680431
13	Solvent_distribution_H-0.0	0.579073	69	VanDerWaal_transition_NH	0.680272
14	AA_composition_W	0.59369	70	Polarizability_distribution_N-0.5	0.680032
15	VanDerWaal_composition_P	0.6	71	Hydrophobicity_distribution_H-1.0	0.680351
16	Solvent_distribution_H-0.75	0.601118	72	Hydrophobicity_transition_NH	0.680671
17	VanDerWaal_distribution_H-0.0	0.606629	73	Hydrophobicity_distribution_H-0.75	0.679473
18	AA_composition_E	0.612859	74	Hydrophobicity_distribution_N-1.0	0.680511
19	AA_composition_R	0.623802	75	VanDerWaal_distribution_H-0.5	0.68099
20	AA_composition_F	0.636422	76	VanDerWaal_transition_PH	0.679553
21	AA_composition_D	0.641933	77	Polarizability_composition_H	0.678674
22	AA_composition_Y	0.647843	78	Polarizability_transition_NH	0.678754
23	AA_composition_P	0.653035	79	Hydrophobicity_transition_PN	0.677556
24	AA_composition_Q	0.657987	80	Polarizability_distribution_P-1.0	0.678594
25	Hydrophobicity_distribution_P-0.25	0.66238	81	Hydrophobicity_distribution_H-0.5	0.678434
26	Hydrophobicity_transition_PH	0.668291	82	Polarizability_transition_PN	0.678355
27	Hydrophobicity_distribution_P-1.0	0.671086	83	Polarity_distribution_H-0.5	0.677077
28	AA_composition_N	0.67492	84	VanDerWaal_distribution_H-0.75	0.676198
29	Solvent_distribution_H-1.0	0.675399	85	Solvent_distribution_H-0.5	0.675639
30	Hydrophobicity_composition_P	0.675399	86	VanDerWaal_distribution_P-0.25	0.676198
31	Polarity_transition_PN	0.677955	87	Polarizability_distribution_N-1.0	0.67516
32	Polarizability_transition_PH	0.679792	88	Polarizability_distribution_P-0.0	0.675399
33	Solvent_transition_HE	0.680751	89	VanDerWaal_distribution_P-0.5	0.67524
34	Polarity_distribution_H-0.75	0.68107	90	Solvent_distribution_H-0.25	0.673642
35	Polarizability_distribution_P-0.5	0.684185	91	Polarity_distribution_N-1.0	0.673083
36	AA_composition_H	0.686262	92	VanDerWaal_composition_H	0.672204
37	Polarizability_distribution_H-0.0	0.688419	93	Polarizability_distribution_H-0.5	0.671166
38	VanDerWaal_distribution_N-1.0	0.688419	94	Polarity_distribution_N-0.25	0.670367
39	Polarity_composition_N	0.688339	95	Hydrophobicity_distribution_H-0.0	0.670288
40	Hydrophobicity_distribution_P-0.75	0.688259	96	Polarizability_distribution_N-0.0	0.669409
41	Polarity_distribution_P-0.75	0.688179	97	Polarizability_distribution_N-0.25	0.669249
42	Hydrophobicity_distribution_P-0.0	0.686981	98	VanDerWaal_distribution_H-1.0	0.66845
43	Hydrophobicity_distribution_P-0.5	0.686981	99	VanDerWaal_distribution_N-0.75	0.667332
44	Polarity_distribution_H-0.0	0.685623	100	Polarizability_composition_N	0.666773
45	Polarizability_distribution_H-1.0	0.685863	101	Polarizability_distribution_N-0.75	0.665575
46	Polarity_transition_PH	0.685783	102	VanDerWaal_distribution_P-0.0	0.665895
47	Polarity_composition_P	0.686661	103	VanDerWaal_distribution_P-0.75	0.664697
48	Polarity_distribution_P-0.0	0.686422	104	Polarizability_distribution_P-0.75	0.663259
49	Hydrophobicity_distribution_N-0.0	0.685942	105	Polarity_distribution_N-0.0	0.661821
50	AA_composition_K	0.686981	106	Polarizability_distribution_H-0.75	0.659984
51	Polarity_distribution_N-0.75	0.686182	107	Polarity_transition_NH	0.657907
52	Polarity_distribution_N-0.5	0.687061	108	Hydrophobicity_distribution_N-0.25	0.657029
53	VanDerWaal_distribution_N-0.0	0.686102	109	VanDerWaal_distribution_P-1.0	0.65599
54	Hydrophobicity_composition_N	0.686342	110	VanDerWaal_distribution_H-0.25	0.655032
55	Polarizability_composition_P	0.684585	111	Polarizability_distribution_H-0.25	0.654952
56	VanDerWaal_transition_PN	0.684105			

The first 18 features come from mRMR features. The highest accuracy is 68.8%, which takes place when the 37th feature is added in. Please refer to the footnote in Table 3 for the meaning of the feature names

analysis. The 18 most relative features mainly consists of amino acid composition, solvent accessibility and normalized Van Der Waals volume, with the number being 9, 4 and 3, respectively (see Table 5). There is no doubt that amino acid composition is vital for maintaining the function of proteins [26,27]. From Table 5, the most contributing features are amino acid compositions: in the first 5 features, 3 of them are amino acid compositions; in the first 10 features, 7 of them are amino acid compositions; in the first 15 features, 9 of them are amino acid compositions. Among the selected 37 features, 18 of them are amino acid compositions. In all cases, the amino acid compositions occupy about half or more than half of the features. Therefore, we conclude that the amino acid compositions contribute more to the prediction than the physiochemical features, and the hypothesis is false. As for solvent accessibility, it has been known that residues located in the surface of a protein usually function as active sites interacting with other molecules and ligands. For example, DNA-binding proteins prefer to bind with residues with higher solvent accessible area [28]; in the mutants of Hsc70 proteins, the reduction in solvent accessibility results a higher hydrophobic free energy level [29]. Thus solvent accessibility was widely used for drug design and fold recognition [30,31]. The solvent-accessible surface contributes to the stability of a ligand receptor complex [32], protein crystal stability [33] and was used to model side chain conformations [34]. And we also know that the solvent-accessible surface is derived from the normalized Van Der Waals volume [35]. Undoubtedly they are strongly related to each other, and together they help to predict the protein structural classes.

Table 5 Distribution of the first 18 mRMR features in the first 22 MaxRel features

Order	Name	Score
1	*AA_composition_C*	*0.267*
2	*Solvent_composition_H*	*0.192*
3	*VanDerWaal_composition_N*	*0.145*
4	*AA_composition_L*	*0.138*
5	*VanDerWaal_composition_P*	*0.125*
6	*Polarity_composition_H*	*0.117*
7	Hydrophobicity_composition_N	0.102
8	*AA_composition_T*	*0.102*
9	*AA_composition_V*	*0.101*
10	*Hydrophobicity_composition_P*	*0.101*
11	*AA_composition_E*	*0.088*
12	Polarity_composition_N	0.086
13	*AA_composition_G*	*0.085*
14	VanDerWaal_transition_NH	0.084
15	*AA_composition_A*	*0.081*
16	*AA_composition_S*	*0.079*
17	*VanDerWaal_distribution_N-0.25*	*0.074*
18	*Solvent_distribution_H-0.75*	*0.073*
19	*Solvent_distribution_H-0.0*	*0.072*
20	*AA_composition_W*	*0.071*
21	Polarity_distribution_P-0.0	0.070
22	*Solvent_transition_HE*	*0.069*

The italicized ones are the mRMR features also appearing in the final selected 37 features. The first 18 mRMR features mainly consist of amino acid compositions, solvent accessibility and normalized Van Der Waal volume with the number being 9, 4 and 3, respectively

Conclusions

By combining the NNA, mRMR and feature forward searching strategy, we are able to investigate the role of physiochemical properties in predicting the protein structural classes. The results are summarized as followings. Physiochemical features do provide extra prediction capacity to the original amino acid compositions. However, the hypothesis that physiochemical properties provide more prediction capacity is proved to be false. Therefore, more researches should be carried out to search more profound information derived from amino acid sequences. Solvent accessibility and normalized Van Der Waals volume contribute more to the prediction of the protein structural classes than other physiochemical features in the prediction. The combination of mRMR and forward feature searching method provides a way to effectively and efficiently develop the predicting models on the large dataset with hundreds of or more features. A combination method shall be chosen as to achieve good balance between computation speed and the successfulness in the feature selection. Finally, the predictor developed in this study is available on: http://app3.biosino.org:8080/liwenjin/index.jsp

Acknowledgements The authors thank Ziliang Qian, a PhD student in SIBS of CAS, for his help to develop the web-server.

References

1. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349. doi:10.3109/10409239509083488
2. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45–148
3. Klein P, Delisi C (1986) Prediction of protein structural class from the amino acid sequence. Biopolymers 25:1659–1672. doi:10.1002/bip.360250909
4. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162
5. Zhang CT, Chou KC (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Sci 1:401–408
6. Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11:523–538. doi:10.1093/protein/11.7.523
7. Chou KC, Zhang CT (1992) A correlation-coefficient method to predicting protein-structural classes from amino acid compositions. Eur J Biochem 207:429–433. doi:10.1111/j.1432-1033.1992.tb17067.x

8. Cai YD, Feng KY, Lu WC, Chou KC (2006) Using LogitBoost classifier to predict protein structural classes. J Theor Biol 238:172–176. doi:10.1016/j.jtbi.2005.05.034
9. Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27:478–482. doi:10.1002/jcc.20354
10. Dubchak I, Muchnik I, Holbrook SR, Kim SH (1995) Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 92:8700–8704. doi:10.1073/pnas.92.19.8700
11. Wang ZX, Yuan Z (2000) How good is prediction of protein structural class by the component-coupled method? Proteins 38:165–175. doi:10.1002/(SICI)1097-0134(20000201)38:2<165::AID-PROT5>3.0.CO;2-V
12. Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009. doi:10.1016/j.bbrc.2004.07.059
13. Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7:1–6. doi:10.1186/1471-2105-7-20
14. Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815. doi:10.2174/092986607781483778
15. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. doi:10.1109/TPAMI.2005.159
16. Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Commun 305:407–411. doi:10.1016/S0006-291X(03)00775-7
17. 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
18. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 30:264–267. doi:10.1093/nar/30.1.264
19. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32:D226–D229. doi:10.1093/nar/gkh039
20. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2002) ASTRAL compendium enhancements. Nucleic Acids Res 30:260–263. doi:10.1093/nar/30.1.260
21. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, et al (2004) The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192. doi:10.1093/nar/gkh034
22. Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 28:254–256. doi:10.1093/nar/28.1.254
23. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. Proteins 35:401–407. doi:10.1002/(SICI)1097-0134(19990601)35:4<401::AID-PROT3>3.0.CO;2-K
24. Mucchielli-Giorgi MH, Hazout S, Tuffery P (1999) PredAcc: prediction of solvent accessibility. Bioinformatics 15:176–177. doi:10.1093/bioinformatics/15.2.176
25. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3:185–205. doi:10.1142/S0219720005001004
26. Weng Z, Rickles RJ, Feng S, Richard S, Shaw AS, Schreiber SL et al (1995) Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions. Mol Cell Biol 15:5627–5634
27. Hansen JC, Lu X, Ross ED, Woody RW (2006) Intrinsic protein disorder, amino acid composition, and histone terminal domains. J Biol Chem 281:1853–1856. doi:10.1074/jbc.R500022200
28. Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486. doi:10.1093/bioinformatics/btg432
29. Kumarevel TS, Gromiha MM, Ponnuswamy MN (1998) Solvent accessibility analysis on the mutants of Hsc70 ATPase fragment. Biophys Chem 71:99–111. doi:10.1016/S0301-4622(97)00137-3
30. Gromiha MM, Ahmad S (2005) Role of solvent accessibility in structure based drug design. Curr Comput-Aided Drug Des 1:223–235. doi:10.2174/1573409054367664
31. Liu S, Zhang C, Liang S, Zhou Y (2007) Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 68:636–645. doi:10.1002/prot.21459
32. Froeyen M, DeWinter H, Herdewijn P (2006) Conformational analysis, solvent-accessible surface and geometric extent of inhibitors and substrates. Collect Czech Chem Commun 71:842–858. doi:10.1135/cccc20060842
33. Islam SA, Weaver DL (1990) Molecular interactions in protein crystals: solvent accessible surface and stability. Proteins 8:1–5. doi:10.1002/prot.340080103
34. Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V (2004) Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. J Comput Chem 25:712–724. doi:10.1002/jcc.10420
35. Connolly ML (1996) Molecular surfaces: A review. Solvent Accessible Surfaces http://www.netsci.org/Science/Compchem/feature14e.html