ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition III (IIII) III–III

PATTERN RECOGNITION

THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy

Lukasz A. Kurgan*, Leila Homaeian

Department of Electrical and Computer Engineering, University of Alberta, Canada

Received 5 July 2005; received in revised form 8 February 2006; accepted 14 February 2006

Abstract

This paper addresses computational prediction of protein structural classes. Although in recent years progress in this field was made, the main drawback of the published prediction methods is a limited scope of comparison procedures, which in same cases were also improperly performed. Two examples include using protein datasets of varying homology, which has significant impact on the prediction accuracy, and comparing methods in pairs using different datasets. Based on extensive experimental work, the main aim of this paper is to revisit and reevaluate state of the art in this field. To this end, this paper performs a first-of-its-kind comprehensive and multi-goal study, which includes investigation of eight prediction algorithms, three protein sequence representations, three datasets with different homologies and finally three test procedures. Quality of several previously unused prediction algorithms, newly proposed sequence representation, and a new-to-the-field testing procedure is evaluated. Several important conclusions and findings are made. First, the logistic regression classifier, which was not previously used, is shown to perform better than other prediction algorithms, and high quality of previously used support vector machines is confirmed. The results also show that the proposed new sequence representation improves accuracy of the high quality prediction algorithms, while it does not improve results of the lower quality classifiers. The study shows that commonly used jackknife test is computationally expensive, and therefore computationally less demanding 10-fold cross-validation procedure is proposed. The results show that there is no statistically significant difference between these two procedures. The experiments show that sequence homology has very significant impact on the prediction accuracy, i.e. using highly homologous datasets results in higher accuracies. Thus, results of several past studies that use homologous datasets should not be perceived as reliable. The best achieved prediction accuracy for low homology datasets is about 57% and confirms results reported by Wang and Yuan [How good is the prediction of protein structural class by the component-coupled method?. Proteins 2000;38:165–175]. For a highly homologous dataset instance based classification is shown to be better than the previously reported results. It achieved 97% prediction accuracy demonstrating that homology is a major factor that can result in the overestimated prediction accuracy.

Keywords: Protein structural class; SCOP; Machine learning; Homology; Prediction; Secondary protein structure

1. Introduction

Proteins consist of an amino acid (AA) sequence, which is organized into three major types of secondary structures: helices (α structure), strands (β structure), and coils. The first definition of protein structural classes is accredited to Levitt and Chothia [1]. Based on their pioneering work four structural classes of globular proteins are usually distinguished: (1) all-α class, which includes proteins with only small amount of strands, (2) all-β class with proteins with only small amount of helices, (3) α/β class with proteins that include both helices and strands and where strands are mostly parallel, and (4) $\alpha + \beta$ class, which includes proteins with both helices and strands and where strands are

* Corresponding author. Tel.: +1 780 492 5488; fax: +1 780 492 1811.
E-mail addresses: lkurgan@ece.ualberta.ca, leila@ece.ualberta.ca (L.A. Kurgan).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.02.014

mostly antiparallel. The knowledge of structural classes of proteins is useful for the broader problem of protein structure prediction. For instance, the accuracy of secondary structure prediction [2] and reduction of the search space of possible conformations of the tertiary structure [3,4] can be significantly improved by incorporating the knowledge of structural classes. Another factor that motivates this research is availability of the Structural Classification of Proteins (SCOP) database [5]. This popular database contains proteins that are manually annotated and classified into the structural classes, which are used to perform numerous protein structure related studies. Final motivation comes from substantial gap between number of protein for which structure is known and thus structural class can be assigned manually (only about 30 000 proteins stored in the Protein Data Bank) and the total number of currently known protein sequences (NCBI database contains well over 2 million proteins). Thus, development of a reliable method for prediction of structural classes for new and undetermined protein sequences is of pivotal importance.

The structural class assignment is currently performed mostly manually on the basis of known secondary structure and the annotated sequences are stored in SCOP. Over the last two decades numerous computational prediction methods were proposed, starting from early works in 1980s [6,7], through advancements made in 1990s [8–12], and finally to the most recent methods [13–17]. Early methods were very simple and were tested on very limited protein sets, which resulted in their poor performance. On the other hand, recent outbreak of methods results in mixture of results ranging from relatively poor (about 50% accuracy) to almost perfect (about 95% accuracy). These papers propose methods that are often tested on small datasets characterized by different characteristics, such as sequence homology, which is shown to have a significant impact on the prediction accuracy. They usually do not perform reliable comparison with other methods on common data, and sometimes use improper procedures that boost the accuracy. Lack of truly comprehensive study, which would summarize the field, address the problem of testing standardization and point new research directions is evident. An exception is a study done by Wang and Yuan [13], which as a major result points out accuracy limit of 60% when Bayesian classification and composition vector based protein sequence representation is used. Major weaknesses of this paper are inconsistent comparison with competitive methods, which was based on different and small datasets for different methods, focus on a single prediction method, and controversies that followed its release [18,19]. To this end, this paper describes a comprehensive and well defined multi-objective study that: (1) tests eight prediction algorithms, which include methods that were not used in the past, (2) investigates three test procedures, i.e. resubstitution, 10-fold cross-validation and jackknife, (3) studies impact of sequence homology on prediction accuracy, and (4) proposes new representation of protein sequences, which is compared with two previously used representations, i.e. composition vector and autocorrelation functions.

2. Background and related work

2.1. Structural classes definitions

Structural class definitions were initially developed in 1980s and redefined multiple times since then, see Table 1. The main differences were in the thresholds used to define amount of strands for all-α proteins, and amount of helices for all-β proteins. In 1986 Nakashima and colleagues defined five structural classes [6]. This was followed in 1995 by Chou who proposed classification into again five classes, but using different thresholds [20]. The change was due to Nakashima's classification, which set the thresholds for all-α proteins and all-β proteins that were not large enough to reflect the real features of the two structural classes. Chou also defined content of the secondary structures using the Dictionary of Secondary Structure of Proteins (DSSP) [21]. Another definition, which merges the $\alpha+\beta$ and the α/β classes into so-called mixed class and thus considers only four classes, was proposed by Eisenhaber and colleagues in 1996 [22]. All above classifications consider irregular, which are also called ξ, proteins that are small in numbers and therefore are omitted from classification.

The threshold based classifications were deemed obsolete in the late 1990s and were replaced by the manually performed SCOP classification. SCOP database includes description of the structural and evolutionary relationships of proteins from the Protein Data Bank (PDB) [24]. The SCOP classifies proteins on multiple levels including structural classes, but also as belonging to different families, superfamilies and containing different domains. Domain is defined as a structurally conserved part of a protein sequence, and together with the entire sequences is currently a target of structure prediction. The SCOP's classification does not incorporate hardcoded rules for structural classes. Intuitively, it makes decisions based on structural elements that are located in individual domains that constitute the protein. Researchers claim that the SCOP classification is more "natural" and provides more reliable information to study protein structural classes when compared to classification based on the percentage amounts of the secondary structures [5,13,25]. The SCOP classification currently includes 11 classes [26]: (1) all-α proteins; (2) all-β proteins; (3) α/β proteins; (4) $\alpha+\beta$ proteins; (5) multi-domain proteins; (6) membrane and cell surface proteins; (7) small proteins; (8) coiled coils proteins; (9) low resolutions proteins; (10) peptides; and (11) designed proteins. Usually, only the first four categories are considered for computational prediction purposes as they include significant majority of the protein sequences.

A number of other structural classification problems, which are out of the scope of this paper, are also defined and

Table 1
Structural class definitions

Reference	Structural class	Helix (α) amount	Strand (β) amount	Additional constrains and comments
[6]	α proteins	$>15\%$	$<10\%$	
	β proteins	$<15\%$	$<10\%$	
	$\alpha+\beta$ proteins	$>15\%$	$>10\%$	Contains dominantly antiparallel β-sheets
	α/β proteins	$>15\%$	$>10\%$	Contains dominantly parallel β-sheets otherwise
	Irregular			Otherwsie
[20]	α proteins	$\geqslant 40\%$	$\leqslant 5\%$	
	β proteins	$\leqslant 5\%$	$\geqslant 40\%$	
	$\alpha+\beta$ proteins	$\geqslant 15\%$	$\geqslant 15\%$	More than 60% antiparallel β-sheets
	α/β proteins	$\geqslant 15\%$	$\geqslant 15\%$	More than 60% parallel β-sheets
	ξ proteins	$\leqslant 10\%$	$\leqslant 10\%$	
[22,23]	α proteins	$>15\%$	$<10\%$	
	β proteins	$<15\%$	$>10\%$	
	Mixed proteins	$>15\%$	$>10\%$	
	Irregular			Otherwise
SCOP [5]	α proteins	N/A	N/A	Manual classification
	β proteins			
	$\alpha+\beta$ proteins			
	α/β proteins			
	+7 other classes			

investigated by the researchers. For instance, prediction of the protein folds (SCOP families) is one of the areas of intensive research that applies classification algorithms, such as support vector machines, neural networks, regression, etc. [27–30].

2.2. Related work

Prediction of the protein structural classes is usually performed as a two step procedure. First, sequences of different length are represented by a fixed length feature vector and next the feature values are fed into a classification algorithm. Early prediction methods used simple composition vector based sequence representation and threshold based class definitions, and applied discriminant analysis with different distance measures. The composition vector is a 20-dimensional vector, which represents the occurrence frequencies of the 20 AAs. Example distance measures include Euclidean distance [6], Hamming distance [7], and Mahalanobis distance [10]. Next generation of the prediction methods used more complex classification algorithms, and the same composition vector based representation. Examples include algorithms based on the maximum component coefficient principle [8], least correlation angle algorithm [31], fuzzy clustering [32], artificial neural network [9,11,12], vector decomposition [22], component coupled geometric classification algorithm [25], Bayesian classification [13], and most recently support vector machines [15]. The most noticeable progress among these algorithms was done by including the coupling effect among different AA components [3,25]. Recent works improve structural class prediction by using alternative sequence representation. Examples include auto-correlation functions based on non-bonded AA energy [33], polypeptide composition [16,34], and functional domain composition [17].

Detailed comparison of recent prediction methods, which includes information about classification algorithms, representations, class definitions, and accuracy of the prediction, is given in Table 2. The table also provides details about datasets used for testing, including the corresponding sequence homology, size and inclusion of domains.

Analysis of the above table reveals that the prediction algorithms were tested on often very small datasets characterized by unknown and most likely high sequence homology, which is shown to have significant impact on the prediction accuracy. They usually did not perform reliable comparison with other algorithms on common datasets, and in some cases used incorrect procedures that boosted the accuracy. Detailed discussion of these issues follows.

2.3. Important factors related to structural class prediction

2.3.1. Evaluation

Quality of prediction (classification) of sequences into structural classes is measured using two tests: resubstitution and jackknife. The resubstitution tests the prediction on the training data, while jackknife is a leave-one-out test procedure, which also can be seen as n-fold cross-validation where n is the number of data points. Although it is commonly recognized that resubstitution test leads to unrealistically high accuracies, traditionally this results is still being reported. In contrast, the jackknife test is perceived as very rigorous and reliable to evaluate classification accuracy and generalization abilities of the tested algorithms [3,25,33–35].

Table 2
Comparison of state-of-the-art structural class prediction methods

Classification algorithm	Representation	Classes	Dataset				Classification accuracy		
			Size	Homology	Domains	Reference	Resub	Jackknife	Reference
Vector decomposition	AA compos. vector	3 classes [22,23]	260	Unknown	No	[22,23]	60.8	57.7	[22,23]
			471	Unknown	No		58.2	57.3	
Geometric classification	AA compos. vector	4 classes SCOP	359	Unknown, but homologous	Yes	[25]	94.3	**84.1**	[25]
Component coupled geometric classification	AA compos. vector	4 classes SCOP	359	Unknown, but homologous	Yes		94.4	**84.7**	[33]
	Auto-correlation functions	4 classes SCOP	359	Unknown, but homologous	Yes		96.7	**90.5**	
Bayes classification	AA compos. vector	4 classes [6]	131	Unknown	No	[6]	99.2	42.7	[13]
		4 classes [20]	120	Unknown	No	[20]	100	53.3	
		3 classes [22,23]	260	Unknown	No	[22,23]	86.5	62.7	
			471	Unknown	No		79.6	66.7	
		4 classes SCOP	1189	40%	Yes	[13]	63.8	*53.8*	
			675	30%	Yes		66.7	*48.0*	
Discriminant analysis	AA and polypeptide compos. vector	4 classes SCOP	1054	40%	Yes	[34]	91.7	75.2	[34]
	AA compos. vector	4 classes SCOP	1054	40%	Yes		66.2	*55.8*	
Information discrepancy based classification	polypeptides	4 classes SCOP	359	Unknown, but homologous	Yes	[25]	—	**95.8**	[16]
		4 classes SCOP	1401	30%	Yes	[16]	–	75.0	
Support vector machines	AA compos. vector	4 classes SCOP	359	Unknown, but homologous	Yes	[25]	93.0	**95.2**	[15]
	AA compos. vector	7 classes SCOP	1601	Unknown, but homologous	Yes		87.0	84.1	
Intimate sorting classification	AA compos. vector, functional domain composition	7 classes SCOP	2230	20%	Yes	[17]	—	98.8	[17]

Although we do not argue that the jackknife test is reliable, at the same time it is computationally very expensive. Recent chapter written by Rost and Sander, experts in proteins structure prediction, discusses the issue by stating that "...a misunderstanding is often spread in the literature: the more separations (the larger n) the better. However, the exact number of n is not important provided the test set is representative and comprehensive and the cross-validation results are not misused to again change the parameters" [36]. To this end, we argue that n-fold cross validation can be substituted by 10-fold cross-validation, which is commonly used to test classification algorithms and is computationally much less demanding. At the same time, we should make sure that algorithm parameters are not adjusted to take advantage of a too low number of folds, i.e. for a 2-fold cross-validation the overfitting is much easier to achieve than for the 10-fold cross-validation. Using 10 instead of n folds allows to ease experimental comparison of different prediction algorithms, and thus to possibly establish better standards when it comes to comprehensiveness of future experimental studies.

2.3.2. Sequence homology and prediction accuracy

Sequence homology is one of the main factors that significantly impact prediction accuracy. Homology is defined as the percentage of AAs in the protein sequence that are identical after aligning the sequence with other sequences from a given dataset (gaps between consecutive AAs may be introduced during alignment, if necessary). Although homology is known to impact the prediction accuracy, no standards are imposed when it comes to performing tests. Just as an example, one of the most often used test datasets, i.e. datasets of 359 sequences, is highly homologous and thus the corresponding results show over 80% accuracy (shown in bold in Table 2), while low accuracies of often about 50% are shown for sets of low homology sequences (shown in italics in Table 2). Additionally, according to the classification in SCOP, all protein domains with more than 30% homology belong to the same protein family and should be classified as the same structural class. Prediction of the structural class of a new protein sequence or domain, which is homologous 30% or more to a protein of known structure, can be performed using sequence alignment. Therefore, some researchers state that the prediction method should aim only at proteins with lower than 30% homology [13].

Despite a few cases where high prediction accuracy for low homology sets is achieved, in general Wang and Yuan have shown that prediction of the four SCOP classes using Bayesian classification and composition vector based sequence representation is limited to about 60% [13].

Although these results were considered controversial by some researchers [18,19], they show that low accuracy should be expected, especially that the authors claim that their algorithm is one of the most powerful. There are three results shown in Table 2 using underscore that require more detailed attention:

- In the [33] a large low homology dataset was used and accuracy of 75.2% was achieved when composition of custom computed polypeptides was used, but the authors did not follow common standards during the tests. The polypeptides were computed using the entire dataset and then applied in the jackknife manner, which means that information about the tested protein was used to perform its prediction.
- Another prediction method achieved almost perfect, 98.8%, accuracy on a large non-homologous dataset [17]. Again, an improper procedure was used to boost the results. The sequence representation included functional domain decomposition, which consists of about 7800 features. Majority of them denote close similarity of one or more out of 7785 functional domains with the tested sequence. High accuracy of results is a result of high homology between the functional domains and the tested sequences, which again means that the information about the test sequences (in terms of the structural domains similar to the tested sequences) was used when evaluating the results on the test data.
- A large non-homogenous dataset was used and accuracy of 75% was achieved in [16]. Based on the published paper, we believe that this result is due to application of a novel information discrepancy based classification and a polypeptide based sequence representation. At the same, the authors used a highly dimensional representation that includes 8000 features, while the aim of this paper is to use representations that incorporate relatively small number of features. Additionally, our recent analysis revealed that the high accuracy was achieved by improper implementation. Similarly as in the previous two cases, information about the test sequences, including the to-be-predicted class, was used during the test. Our reimplementation of this method that closely follows the paper and avoids the implementation pitfalls shows about 63% accuracy for the same dataset.

In short, we conclude that high accuracy is due to high homology or improper procedures. In this paper we aim to verify prediction accuracy when multiple different classification algorithms are used on two different large and low homology datasets. The results are verified against the findings of Wang and Yuan [13] and results published in [16].

2.3.3. Classification algorithms

The analysis of Table 2 reveals that although several different classification algorithms were used, many other algorithms were never tried. This particularly applies to machine learning algorithms, and includes decision trees, rule based, and regression based algorithms. Also, no previous study applied several algorithms simultaneously on the same data to directly compare their quality. The published studies performed comparison on different datasets for different pairs of algorithms and therefore no reliable and comprehensive comparison between different methods can be performed. To this end, this paper uses three datasets to comprehensively compare quality of eight classification algorithms. Detailed description of the selected classification algorithms is provided later in the paper.

2.3.4. Sequence representation

One of the first results that hinted a possible sequence representation was that the structural class is related to the AA composition of the corresponding sequence [6]. Since then composition vector was used in numerous protein structure studies, including structural class, content, and structure predictions [13,15–17,22,23,25,33,34,37–43]. Therefore most of the existing structural class prediction methods are based on the composition vector. However, researchers also pointed out that the AA composition does not sufficiently utilize the sequence information [8,33,44] and some controversies over predictive accuracies of the methods based on composition vector arose in the past years [14,18,19,22,24,45]. The best prediction accuracies for low homology datasets when composition vector is used are about 55%, see results for [13,16,34] in Table 2.

The only other successfully applied representations are based on polypeptides [16] and auto-correlation functions computed for individual AA [33]. This paper studies an alternative approach, which does not use polypeptides due to their large number, but introduces a new comprehensive representation based on characteristics of individual AAs. It includes composition vector and other features that are related to position of AAs in the sequence, their hydrophobicity, chemical composition and weight. This comprehensive representation is compared with the composition vector and the auto-correlation functions.

3. Methods and goals

Based on a comprehensive experimental study this paper aims to address factors related to prediction algorithms, homology, sequence representation and test procedures. First, a detailed description of the considered experimental scenario is given, and next specific goals are defined.

3.1. Experimental scenario

This paper performs comprehensive experimental comparison of different prediction algorithms using datasets of low-homology, different sequence representations and test

procedures. We start with a detailed description of the considered datasets.

3.1.1. Datasets

The datasets include two previously used sequences sets, i.e. the 359 sequences datasets [25] and 1189 sequences dataset [13]. Although the first dataset (denoted as *359*) is relatively small, it was the most extensively used in the past studies. It includes 359 highly homologous domains and sequences. The CD-HIT program, which clusters protein databases at given sequence homology threshold, was used to estimate homology of this dataset [46,47]. This program is used by UniProt, PDB, EBI, and TIGR to filter highly homologous sequences. The results show that among the 359 sequences only 214 are below 100% homology threshold, clearly indicating that over 100 sequences in the dataset are virtually identical. Running CD-HIT with 95%, 80%, 60% and 40% homology threshold reveals that only 143, 133, 132, and 127 sequences below the respective homology threshold can be found. Due to the high homology this set is not used to show true classification accuracy, but rather to investigate the impact of homology on estimation of the classification accuracy.

The second set contains sequences with low 40% homology (denoted as *1189*), and is selected due to its prior application in the most comprehensive study of the structural class prediction methods. The original 1189 dataset was processed and filtered using the latest, 1.67, version of SCOP, and PDB release as of February 2005. As a result a dataset with 1092 domains and sequences was created. The main problems with the original dataset include conflicting domain ranges for 298 domains, which were replaced with 265 domains. This was due to some PDB sequences that do not have multiple domains in SCOP anymore, and that now have fewer domains in SCOP than at the time when the original dataset was created. The replacements were performed by selecting the same proteins with the closest domain range. Also, 55 domains had illegal symbols in their primary structures, 24 domains were indexed out of the corresponding protein sequences, and 18 PDB sequences were obsolete and either replaced with the new sequences or removed from the set. Finally, new version of the SCOP reclassifies 18 domains from the original dataset into SCOP classes 4–7, which are not considered for prediction. The final set includes 223 all-α, 294 all-β, 334 α/β, and 241 $\alpha+\beta$ domains and sequences.

Additionally, a new larger dataset with low homology sequences is created and used to perform experiments. The dataset is selected based on the 25% PDBSELECT list [48], which is about 15 times smaller than the PDB and includes only high quality non-homologous proteins, i.e. proteins scanned with high resolution and with low on average 25% homology (the homology ranges between 22% and 45%). Using PDB release as of February 2005, 2340 sequences and domains were extracted based on 25% PDBSELECT list. Among them 443 are all-α, 443 are all-β, 346 α/β, and 441 $\alpha+\beta$, while for the remaining sequences the SCOP classes are missing or belong to the other seven SCOP classes. The final 25PDB dataset (denoted as *25PDB*) contains 1673 proteins and domains.

These two low homology datasets are used to provide two independent sources of experimental results to evaluate classification accuracy. In order to enable other researchers to use these datasets, lists of all sequences and domains that constitute the 25PDB and 1189 datasets are given in Appendix A.

3.1.2. Representation

The paper also proposes a novel representation that includes variety of features related to AA composition, position, hydrophobicity, weight, and chemical composition including:

- composition vector (denoted as A) due to its extensive prior use;
- first order composition moment vector (denoted as B), which was successfully used for the protein secondary content prediction [43,49];
- autocorrelation functions based on the Oobatabe–Ooi AA energy index (denoted as C) [50], which were used for structural class prediction [33];
- autocorrelation based on the cumulative Eisenberg's hydrophobicity index (denoted as D) [51], which was successfully used for the protein secondary content prediction [40];
- chemical group composition (denoted as E), which was used for the protein structure prediction [42] and for the protein secondary content prediction [49];
- and the sequence molecular weight (denoted as F), which was used for the protein content prediction [49,52].

The above features were selected based on their prior successful application in protein structure prediction. The composition vector and composition moment vector are defined based on the count and position of AAs in the sequence [43]

$$x_i^{(k)} = \frac{\sum_{j=1}^{c_i} n_{ij}^k}{\prod_{d=0}^{k}(N-d)},$$

where $i = 1, 2, \ldots, 20$ is the AA index, k is the order of the composition moment vector (for $k = 0$ it reduces to composition vector), N is the length of the protein sequence, n_{ij} is the jth position of the ith AA, and c_i is the count (composition) of the ith AA in a sequence. Hydrophobicity is used to represent a protein sequence by using a hydrophobic scale, where each AA is replaced by its hydrophobic index value h_i, see Table 3. Alternatively AAs can be replaced by their corresponding energy index values o_i, see Table 3.

Autocorrelation function r_n is defined as [40]

$$r_n = \frac{\sum_{j=1}^{N-n} index_{i,j}\, index_{i,j+n}}{N-n},$$

Table 3
Hydrophobicity index, energy based index and molecular weight of AAs

AA	A/M	C/N	D/P	E/Q	F/R	G/S	H/T	I/V	K/W	L/Y
Eisenberg's hydro-phobicity index h_i	0.62	0.29	−0.90	−0.74	−1.19	0.48	−0.40	1.38	−1.50	1.06
	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26
Oobatabe–Ooi energy index o_i	−9.475	−12.210	−12.144	−13.815	−20.504	−7.592	−17.550	−15.608	−12.366	−15.728
	−15.704	−12.480	−11.893	−13.689	−16.225	−10.518	−12.369	−13.867	−26.166	−20.232
Molecular weight m_i	71	103	115	129	147.1	57	137.1	113.1	128.1	113.1
	131	114	97	128.1	156.1	87	101	99.1	186.1	163.1

Table 4
Chemical groups associated with AAs

AA	Associated chemical groups
A	CH, CO, NH, CH_3
C	CH, CO, NH, CH_2, SH
D	CH, CO, NH, CH_2, CO, COO^-
E	CH, CO, NH, CH_2, CH_2, CO, COO^-
F	CH, CO, NH, CH_2, CAROM, CHAROM, CHAROM, CHAROM, CHAROM, CHAROM
G	CH_2, CO, NH
H	CH, CO, NH, CH_2, CAROM, CHAROM, N, CHAROM, NH
I	CH, CO, NH, CH_2, CH, CH_3, CH_3
K	CH, CO, NH, CH_2, CH_2, CH_2, CH_2, NH_3^+
L	CH, CO, NH, CH_2, CH, CH_3, CH_3
M	CH, CO, NH, CH_2, CH_2, S, CH_3
N	CH, CO, NH, CH_2, CO, C, NH2
P	CHRING, CO, NHRING, CH_2RING, CH_2RING, CH_2RING
Q	CH, CO, NH, CH_2, CH_2, CO, C, NH2
R	CH, CO, NH, CH_2, CH_2, CH_2, NH, C, NH2, NH_2^+
S	CH, CO, NH, CH_2, OH
T	CH, CO, NH, CH, CH_3, OH
V	CH, CO, NH, CH, CH_3, CH_3
W	CH, CO, NH, CH_2, CAROM, CAROM, CAROM, NH, CHAROM, CHAROM, CHAROM, CHAROM, CHAROM
Y	CH, CO, NH, CH_2, CAROM, CHAROM, CHAROM, CHAROM, CHAROM, CAROM, OH

where $index_{i,j}$ is the index value for the ith AA at the jth position in the sequence, and n is the number of autocorrelation functions. Based on the prior results, $n = 6$ and $index_{i,j} = h_i$ were used for the Eisenberg's hydrophobicity index [40], and $n = 30$ and $index_{i,j} = o_i$ were used for the Oobatabe–Ooi AA energy index [33]. Another AA property that was found useful for structure prediction is the chemical composition of their side chains. There are 19 chemical groups, which constitute the side chains, and some chemical groups are associated with multiple different side chains, see Table 4.

The count (composition) of each of the chemical groups is computed for a protein sequence, i.e. all AAs that have a given chemical group are counted, and the resulting vector constitutes a set of 19 features for the prediction. Finally, the sequence molecular weight refers to the sum of the atomic weight of AAs that constitute the sequence, see Table 3. This feature is defined as

$$M_{avg} = \frac{\sum_{j=1}^{N} m_{ij}}{N},$$

where m_i is the molecular weight of the ith AA at the jth position in the sequence and N is the sequence length.

A feature selection study was conducted to select a subset of the most relevant, with respect to structural class prediction, features. The 25PDB and 1189 datasets were used together with three feature selection methods:

1. Feature Subset Consistency (*FSC*) feature selection method, which selects subset of features using a probabilistic filter-based approach that uses Las Vegas algorithm to search through different feature subsets [53].
2. Wrapper Subset Selection (*WSS*) feature selection method, which is a classification based wrapper that uses Naı¨ve Bayes algorithm [54].
3. Feature Correlation (*FC*) feature selection method, which selects subset of features based on their correlation with the class while maintaining low intercorrelation between the selected features [55].

The experiments apply 10-fold cross-validation test procedure. Table 5 summarizes the results, where the selected features are those that were picked up by the feature

Table 5
Feature selection results

Dataset	Selection method	# Features selected from each of the feature sets						Total # selected features
		A	B	C	D	E	F	
25PDB	FC	10	5	0	1	4	1	21
	FSC	14	5	0	1	5	1	26
	WSS	3	2	0	0	2	1	8
1189	FC	13	4	0	1	6	1	25
	FSC	10	7	0	1	4	1	23
	WSS	5	0	0	0	3	1	9
Average # selected features		9.2	3.8	0	0.7	4	1	N/A
Average % selected features		46%	19%	0%	11%	21%	100%	N/A
Average # folds feature selected		*26.1*	*14.4*	*0.3*	*10*	*11.1*	*59*	*N/A*
Features never selected		*0*	*0*	*22*	*0*	*7*	*0*	*N/A*
Total /# features		20	20	30	6	19	1	96

A—composition vector, B—first order comp moment vector, C—auto-correlation based on Oobatake–Ooi energy index, D—autocorrelation based on Eisenberg's hydrophobic index, E—chemical group composition, F—molecular weight.

selection algorithms for at least 5 out of 10 cross-validation folds (results shown in italics correspond to individual folds). The "Total # selected features" gives, for each dataset and each feature selection methods, the number of features that are selected in at least five folds.

The "Average # folds feature selected" row shows the average number of individual folds for which features belonging to a given feature set were selected. The maximum value is 60 since each selection executed 10 folds, and three selection methods were applied on two datasets. The "# Features never selected" row gives the number of features from a given feature sets that were not selected in any fold of any of the three selection methods and the two datasets. Two patterns are observed based on these two rows:

- Each feature set contains some features that have been selected although for some feature sets there are some features that were never selected. Chemical composition (and energy index autocorrelations feature sets have some features that are never selected, but also have features that were selected).
- For most of the feature sets all features were selected.

These results give motivation to select entire feature sets, rather than individual features. Additionally, using the entire feature sets allows to preserve complete description of the sequence with respect to a given sequence property, such as AA composition, hydrophobicity, chemical composition, etc.

The results between different feature selection algorithms are consistent, and show that composition vector and molecular weight are strongly related to structural classes. They also show that first order composition vector, autocorrelation based on hydrophobicity and chemical groups are related to structural classes, while no features based on the energy index autocorrelations were selected. The chemical composition feature set contains seven features that were never selected, but at the same time its average number of folds is similar to the number for first order composition vector and autocorrelation based on hydrophobicity feature sets, and thus it was also selected. Finally, energy index autocorrelation set is relatively less valuable since its average number of folds is at least an order of magnitude lower when compared with other sets. At the same time, eight features from this set were selected by in at least one selection fold, which shows that it still can be applied for this prediction task.

Therefore the proposed representation includes 66 features: 20-dimensional composition vector, 20-dimensional composition moment vector, 19-dimensional chemical group composition vector, 6-dimensional hydrophobic autocorrelations, and the sequence molecular weight. This representation is compared with two published representations: (1) 20-dimensional composition vector and (2) 30-dimensional energy autocorrelations.

3.1.3. Classification algorithms

Eight different classification algorithms are used to perform structural class prediction. The algorithms are selected to include all major families of classification algorithms and select high quality, with respect to relatively low complexity and good predictive accuracy, methods. They include some of the previously used algorithms, such as Bayesian classification (Naı¨ve Bayes is chosen), nearest neighbor (instance based learning algorithm is chosen), and support vector machines. We also introduce new to this field algorithms such as decision trees (C4.5 and random forest are chosen), rule based algorithms (RIPPER is chosen), neural networks (RBF network is chosen), and logistic regression; see Table 6. The WEKA 3.4 environment [56] was used to perform experiments with the above classifiers. Each of the classifiers was optimized with respect to their parameter for each of the three datasets and the three representations. The optimization was performed for each corresponding dataset-representation pair using $\frac{2}{3}$ of the data for training and $\frac{1}{3}$ of

Table 6
Summary of the applied classification algorithms

Classifier	Reference	Short description
Naı¨ve Bayes (*NB*)	[57]	Simple, scalable, due to assumption that all features are independent, and popular probabilistic classifier
Radial basis function neural network (*RBF*)	[58]	Scalable neural network classifier that applies k-means clustering to generate Gaussian radial basis functions and uses logistic regression to perform learning
Instance based classifier (*IB1*)	[59]	Simple, lazy learner that uses the idea of nearest neighbor classification
C4.5 (*C4.5*)	[60]	The most popular decision tree classifier
Random forest (*RF*)	[61]	One of the newest decision tree based classifiers, which constructs ensemble of decision trees based on dissimilarity between data points in the training data
Repeated incremental pruning to produce error reduction (*RIP*)	[62]	A scalable and one of the most accurate rule-based classifiers, which generates propositional rules using divide-and-conquer approach and a greedy set-covering based search
Support vector machine (*SVM*)	[63]	An accurate classifier that applies sequential minimal optimization algorithm to generate a support vector machine with polynomial or RBF kernels
Logistic regression (*LR*)	[64]	A multinomial logistic regression classifier that applies a ridge estimator

the data for testing. The parameters that gave maximal accuracy were selected. The considered parameters, their values and optimal setup for each classification algorithm are given in Appendix B.

3.2. Goals

The paper addresses the four following goals:

Goal 1. Comparison of classification accuracy of different classification algorithms. The eight classifiers are tested and compared using the low homology 25PDB and 1189datasets.

Goal 2. Impact of sequence homology on the classification accuracy. The 8 classifiers are compared with respect to results on the high homology 359 dataset and on the 2 low homology datasets to investigate the differences in accuracy.

Goal 3. Impact of sequences representation on the classification accuracy. The prediction accuracies when the three representations are used for the three considered datasets are compared.

Goal 4. Comparison between 10-fold cross-validation, jackknife and resubstitution test procedure with respect to estimation of prediction accuracy. Experiments for all prediction algorithms and for all datasets are performed and compared using the three test procedures.

4. Experiments and results

Summaries of the structural class prediction results for the eight classification algorithms, three sequence representations, and three test types are shown in Table 8 for 25PDB dataset, in Table 9 for 1189 dataset, and in Table 10 for 359 dataset. The results show average accuracy lift, and weighted by the class sizes sensitivity and specificity. The lift is defined as a difference between the achieved accuracy and the base-line accuracy, i.e. frequency of the largest class in the dataset, which corresponds to a classifier that always chooses the most frequent class. The corresponding base-line accuracies of the three datasets are 26.5% for the 25PDB dataset, 30.6% for the 1189 dataset, and 28.3% for the 359 dataset. The accuracy, sensitivity and specificity are defined based on a confusion matrix, see Table 7.

Table 7
Confusion matrix for the protein structural class prediction

Actual structural class	Predicted structural class			
	All-α	All-β	α/β	$\alpha+\beta$
All-α	*a*	*ab*	*ac*	*ad*
All-β	*ba*	*b*	*bc*	*bd*
α/β	*ca*	*cb*	*c*	*cd*
$\alpha+\beta$	*da*	*db*	*dc*	*d*

a, *b* and *c* are # of correct predictions for the respective four structural classes; *ab* is the number of incorrect predictions where all-α protein is predicted as all-β protein, *ba* is the number of incorrect predictions where all-β protein is predicted as all-α protein, etc.

The *accuracy* is defined as ratio between the number of correct predictions and n, which is the total number of predictions (proteins):

$$accuracy = \frac{a+b+c+d}{n}.$$

The *sensitivity* is the ratio between the correct and all predictions for a given structural class (all-α, all-β, α/β, and $\alpha+\beta$):

$$sensitivity_{all\text{-}\alpha} = \frac{a}{a+ab+ac+ad},$$
$$sensitivity_{all\text{-}\beta} = \frac{b}{b+ba+bc+bd},$$
$$sensitivity_{\alpha/\beta} = \frac{c}{c+ca+cb+cd},$$
$$sensitivity_{\alpha+\beta} = \frac{d}{d+da+db+dc}.$$

The *specificity* is the ratio between the correct and all predictions for proteins that should be excluded for a given

structural class:

$$specificity_{all\text{-}\alpha} = \frac{b+c+d+bc+bd+cb+cd+db+dc}{b+c+d+bc+bd+cb+cd+db+dc+ba+ca+da},$$

$$specificity_{all\text{-}\beta} = \frac{a+c+d+ac+ad+ca+cd+da+dc}{a+c+d+ac+ad+ca+cd+da+dc+ab+cb+db},$$

$$specificity_{\alpha/\beta} = \frac{a+b+d+ab+ad+ba+bd+da+db}{a+b+d+ab+ad+ba+bd+da+db+ac+bc+dc},$$

$$specificity_{\alpha+\beta} = \frac{a+b+c+ab+ac+ba+bc+ca+cb}{a+b+c+ab+ac+ba+bc+ca+cb+ad+bd+cd}.$$

The reported specificity and sensitivity are computed using a weighted, by the respective class sizes, average:

$$\begin{aligned} sensitivity = {} & \frac{a+ab+ac+ad}{n} sensitivity_{all\text{-}\alpha} \\ & + \frac{b+ba+bc+bd}{n} sensitivity_{all\text{-}\beta} \\ & + \frac{c+ca+cb+bd}{n} sensitivity_{\alpha/\beta} \\ & + \frac{d+da+db+dc}{n} sensitivity_{\alpha+\beta}, \end{aligned}$$

$$\begin{aligned} specificity = {} & \frac{a+ab+ac+ad}{n} specificity_{all\text{-}\alpha} \\ & + \frac{b+ba+bc+bd}{n} specificity_{all\text{-}\beta} \\ & + \frac{c+ca+cb+bd}{n} specificity_{\alpha/\beta} \\ & + \frac{d+da+db+dc}{n} specificity_{\alpha+\beta}. \end{aligned}$$

Best average accuracy lift and lifts within the 1% difference to the best with respect to either 10-fold cross-validation (10CV) or jackknife tests are shown in bold. Best results for each dataset are shown using underscore (Tables 8–10). Next, the results are analyzed with respect to the defined goals.

4.1. Goal 1—comparison of classification algorithms

The classification accuracy lift based on jackknife test for different classifiers, the proposed representation and the low-homology datasets are compared. The classifiers are ranked based on the average results for the two datasets. The results for the high-homology, 359 dataset are omitted due to their high and overestimated accuracy. The ranked algorithms and their accuracy are shown in Fig. 1.

The average classification accuracy ranges between 38.5% for the RIP and 55.5% for the LR, which shows that manual assignment of SCOP classes is very difficult to predict, especially when homology is low. In general, the results agree with Wang and Yuan who concluded about 60% prediction accuracy limit for their method [13]. The SVM and LR classifiers are best for both datasets and their average lift values differ by 1.5%. The differences between the second best SVM and the remaining classifiers are 4.5% for the third best RF, 5.4% for NB, 6.7 for RBF, 10.1 for C4.5, 12.4 for IB1, and 15.4 for the worst performing RIP. These differences are substantial as they constitute between 20% and 60% of the maximal lift. The results for the 1189 datasets are compared with those reported in [13], see Table 11.

The results are very close, and show that a new to the field logistic regression classifier is the same accurate as the leading Bayesian classifier. The results also show high quality of the support vector machines, but unlike the previously published results [15], they are shown on low homology data and are properly compared with other methods.

The confusion matrices for best results of the jackknife test for the 25PDB and 1189 dataset are shown in Table 12. Both best results were achieved for the proposed 66 features representation and by the logistic regression algorithm.

The matrices show that significant portion of data belonging to a given class is predicted as this class, i.e. for all-α, all-β, and α/β about 60% and for $\alpha+\beta$ about 40% on average. The least accurately predicted class is the $\alpha+\beta$ class, while the best results are achieved for the all-α class.

Analysis of sensitivity and specificity achieved by different algorithms shows a consistent pattern. The sensitivity is always significantly smaller compared to specificity, and the latter measure goes up to 86% for the best results for the low homology datasets. High average specificity means that false positives are relatively low and thus low accuracy is a result of relatively low sensitivity. Thus, the classification algorithms generate selective models that potentially can be further improved by pruning or constrain relaxation.

4.2. Goal 2—sequence homology

The classification accuracy lifts based on jackknife test using the proposed representation and the eight classifiers are compared between the low homology and high homology datasets, see Fig. 2. The classifiers are ranked based on the results for the low homology datasets.

The results shows that significantly higher accuracy lifts are achieved for the highly homologous, 359 dataset. The

Table 8
Summary of structural class prediction results for 25PDB dataset

Representation	Test	Accuracy lift (baseline accuracy = 26.5%)									Weighted sensitivity/specificity							
		NB	RBF	IB1	C4.5	RF	RIP	SVM	LR	Avg	NB	RBF	IB1	C4.5	RF	RIP	SVM	LR
Composition vector (20)	Resubstit	24.4	27.2	73.5	50.8	73.5	24.1	27.6	27.1	41.0	51/84	54/85	100	77/93	100	51/83	54/84	53/84
	10CV	21.5	22.8	11.3	11.8	21.1	12.6	**25.5**	**24.5**	18.9	48/83	49/83	38/80	38/79	48/82	39/79	52/83	51/83
	jackknife	22.5	23.5	11.4	15	23.8	4.3	**25.1**	**24.8**	18.8	49/84	50/84	38/80	42/80	50/83	31/76	52/83	51/83
Auto-correlation (30)	Resubstit	1.8	1.8	73.5	16.4	73.3	4.9	11.1	10.9	24.2	28/75	29/74	100	43/80	100	32/75	38/78	37/78
	10CV	0.9	0.9	3.2	0	2.6	0.4	**8.5**	**7.6**	3.0	27/75	27/74	30/77	27/74	29/76	27/74	35/77	34/77
	jackknife	1	1.8	4	−8.7	2.3	−21.2	**7.6**	**7.3**	−0.7	27/75	29/75	31/77	18/72	29/76	5/66	34/76	34/76
66 features	Resubstit	26.3	34.2	73.5	44.3	73.5	30.3	33.1	35.7	43.9	53/84	51/84	100	71/90	100	57/85	60/86	63/87
	10CV	21.4	20.8	12.7	16.8	24.5	16.2	28.6	**30.2**	21.4	48/83	47/83	39/80	43/81	51/83	43/80	55/85	57/85
	jackknife	22	21.1	12.5	17.2	20.6	5.3	29.3	**30.6**	19.8	49/83	48/83	39/80	44/81	47/82	32/76	56/85	57/86

Table 9
Summary of structural class prediction results for 1189 dataset

Representation	Test	Accuracy lift (baseline accuracy = 30.6%)									Weighted sensitivity/specificity							
		NB	RBF	IB1	C4.5	RF	RIP	SVM	LR	Avg	NB	RBF	IB1	C4.5	RF	RIP	SVM	LR
Composition vector (20)	Resubstit	23.6	28.3	69.4	68.6	69.4	24.3	25.1	22.9	41.5	54/82	59/84	100	99/100	100	55/81	56/83	53/83
	10CV	**20.8**	**21**	12.1	9	18.9	11.8	**21**	**21**	17.0	51/82	52/82	43/80	40/80	50/82	43/77	52/81	52/82
	jackknife	**21**	**21**	13.4	10.1	19.5	13.7	**21.7**	20.5	17.6	52/82	52/82	44/80	41/79	50/82	44/78	52/82	51/82
Auto-correlation (30)	Resubstit	0	7.3	69.4	60.4	69.4	4	14.7	12	29.7	31/69	38/77	100	91/97	100	35/71	45/78	43/78
	10CV	0	2	−1.4	−3.5	3.2	0.4	**6.4**	5.2	1.5	31/69	33/75	29/75	27/75	34/76	31/71	37/75	36/76
	jackknife	0	1.1	−1.1	−0.7	2.6	0.6	**7.1**	5.5	1.9	31/69	32/75	30/74	30/76	33/76	31/71	38/75	36/76
66 features	Resubstit	20.9	23	69.4	38.1	69.4	23.5	25.3	31.4	37.6	52/82	54/83	100	69/89	100	54/82	56/82	62/86
	10CV	18.4	17.5	12.8	13.7	19.6	13.9	21	**23.2**	17.5	49/81	48/81	44/80	44/81	50/82	45/79	52/81	54/84
	jackknife	18.1	16.3	13.5	13.4	21.3	14.7	21.5	**23.3**	17.8	49/81	47/80	44/80	44/81	52/83	45/79	52/81	54/84

Table 10
Summary of structural class prediction results for 359 dataset

Representation	Test	Accuracy lift (baseline accuracy = 28.3%)									Weighted sensitivity/specificity							
		NB	RBF	IB1	C4.5	RF	RIP	SVM	LR	Avg	NB	RBF	IB1	C4.5	RF	RIP	SVM	LR
Composition vector (20)	Resubstit	38.6	68.4	71.7	70.8	71.7	68.1	71.4	39.5	62.5	67/89	97/99	100	99/100	100	96/99	100	68/89
	10CV	34.7	60.3	**66.3**	62.1	63.9	45.2	65.1	28	53.2	63/87	88/96	94/98	90/97	92/97	74/91	94/98	56/85
	jackknife	34.9	60.6	**67.2**	65.1	**66.9**	52.7	**67.5**	30.7	55.7	63/87	89/96	96/98	94/98	95/98	81/93	96/99	59/86
Auto-correlation (30)	Resubstit	8.1	44.6	71.7	71.7	71.7	53.3	71.1	39.5	54.0	36/80	73/91	100	100	100	82/93	100	68/89
	10CV	4.5	21.4	**59.3**	52.1	54.8	31.3	**60.3**	27.4	38.9	33/78	50/83	88/96	81/93	83/94	60/86	89/96	56/85
	jackknife	5.1	21.7	**62.4**	53.6	56.3	36.5	**62.4**	24.7	40.3	33/79	50/83	91/97	82/94	84/95	65/88	91/97	53/84
66 features	Resubstit	57.2	71.4	71.7	71.4	71.7	69.6	71.7	71.7	69.6	86/95	100	100	100	100	98/99	100	100
	10CV	40.7	59.7	**67.8**	63.3	65.4	53.6	**66.9**	59.7	59.6	69/89	88/96	96/99	91/97	94/98	82/94	95/99	88/96
	jackknife	42.5	59	**68.7**	64.5	67.5	56.1	**68.7**	60.9	61.0	71/90	87/96	97/99	93/97	96/98	84/95	97/99	89/96

lift values are better for all classifiers, therefore showing a consistent pattern. The average accuracy lift for 25PDB is 19.8%, for 1189 dataset is similar and equal to 17.8%, while for the 359 dataset is 59.0% for the jackknife test. The best 97% accuracy for the 359 datasets was achieved by the IB1 and SVM classifiers. At the same time, IB1 achieves relatively poor accuracy for the low homology datasets. A paired t-test between the results achieved by the eight algorithms on the 25PDB and 359 datasets and using jackknife test and the 66 features representation gave t-score of 10.0 and between the 1189 dataset and 359 dataset gave t-score of 13.0. Both scores indicate that the difference in accuracy lift is statistically significant. At the same time, the paired t-test between the results for the 25PDB and 1189 datasets resulted in t-score of 1.0, which shows that the difference is statistically not significant.

The best results for the 359 dataset are compared with best results of other researchers, see Table 13.

The table shows that the best past results achieved with information discrepancy based classification algorithm and 8000 features are beaten by a simple IB1 classifier that uses 66 features. Among the six results above 95%, four are achieved using algorithms and representations introduced in this paper. We caution the reader that these results should not be taken as reliable indicator for the general task of structural class prediction. They are shown to demonstrate that high sequences homology results in higher accuracy.

4.3. Goal 3—sequence representation

The average, over the two low homology datasets, classification accuracy lifts for the jackknife test, the three different representations, and the eight classifiers are compared, see Fig. 3.

The results show that autocorrelation based representation performs worse than the remaining two representations. The differences range between 10% and 25% for different classifiers. This result does not agree with the comparison published in Ref. [33], where autocorrelations performed better than the composition vector. We conclude that the previous results were likely to be unreliable due to testing on the highly homologous dataset. At the same time, the autocorrelation based representation achieved 91% accuracy on the highly homologues 359 dataset, which shows that this representation is a viable alternative when high homology data are used. Similar outcome was reported in Ref. [33], see Tables 2 and 10. On the other hand, using autocorrelations with datasets containing low homology sequences resulted in significantly worse results. A paired t-test between the accuracy lift results achieved with different representations and for the eight classifiers and over the three datasets and jackknife test was performed, see Table 14.

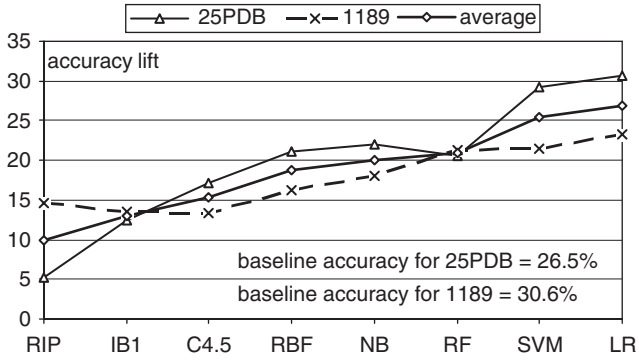

Fig. 1. Accuracy of the eight classifiers for the 25PDB dataset, 1189 datasets and average of the two; classifiers are ranked from the worst on the left to the best on the right.

Table 12
Confusion matrices for best results: (a) for the 25PDB dataset; (b) for the 1189 dataset

Predicted class	True class			
	All-α	All-β	α/β	$\alpha+\beta$
All-α	306	27	40	70
All-β	33	273	43	94
α/β	31	52	208	55
$\alpha+\beta$	85	121	66	169
All-α	127	14	53	29
All-β	14	185	43	52
α/β	39	37	216	42
$\alpha+\beta$	31	70	79	61

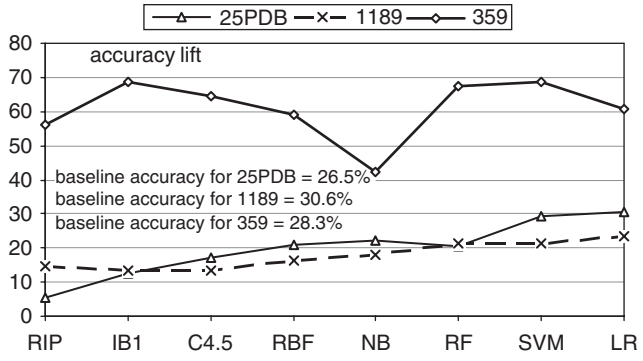

Fig. 2. Accuracy of the eight classifiers for the 25PDB, 1189, and 359 datasets.

The results show that for the three datasets the autocorrelation based representation is statistically significantly worse than the two other representations. At the same time, the t-test shows that there is no statistically significant difference between the accuracy of classification when using the 66 features and the composition vector based representations. The results confirm high quality of the composition vector with respect to structural class prediction. At the same time, for high quality classifiers, such as SVM and LR, the newly proposed representation provides significant benefits, which are balanced by similar or slightly worse results for the lower quality classifiers. In case of SVM

Table 11
Comparison of best results for low homology 1189 datasets

Classification algorithm	Representation	Classification accuracy		
		Resubstitution	Jackknife	Reference
Support vector machine	AA composition vector	57.8	52.3	This paper (second best result)
Bayes classification	AA composition vector	63.8	53.8	[13]
Logistic regression	66 features	62.0	53.9	This paper (best result)

Table 13
Comparison of best results for high homology 359 datasets

Classification algorithm	Representation	Classification accuracy		
		Resubstit	Jackknife	Reference
Component coupled geometric classification	AA composition vector	94.3	84.1	[25]
Component coupled geometric classification	AA composition vector	94.4	84.7	[33]
	Auto-correlation functions	96.7	90.5	
Support vector machines	AA composition vector	93.0	95.2	[15]
Support vector machine	AA composition vector	100.0	95.8	This paper (second best result)
Random forest	66 features	100.0	95.8	This paper (second best result)
Information discrepancy based classification	Polypeptides	–	95.8	[16]
Support vector machine	66 features	100.0	97.0	This paper (best result)
Instance based classifier	66 features	100.0	97.0	This paper (best result)

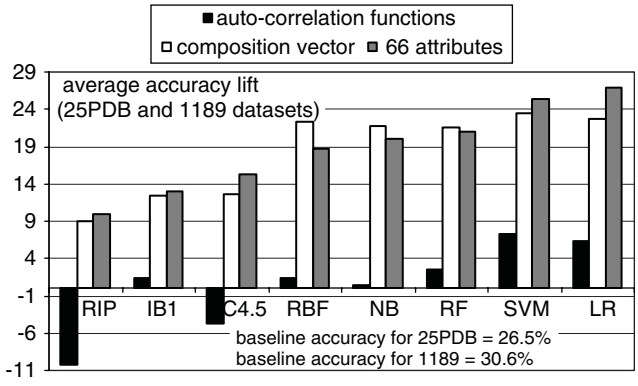

Fig. 3. Accuracy of the eight classifiers for the 25PDB and 1189, and the three representations.

the average, over the two low homology datasets, increase of accuracy lift due to using the new representation instead of composition vector is 2.0% and for LR is 4.3%. Considering the overall range of reported accuracy lifts and that the improvements concern the most accurate classifiers, the result is considered significant and shows that the proposed representation is better when compared with the most commonly used composition vector.

4.4. Goal 4—test procedures

The two most commonly used test procedures with respect to protein structural class prediction are resubstitution and jackknife. The resubstitution test results in overestimation of the prediction accuracy. For instance, 43.9%, 37.6%, and 69.6% average accuracy lift for the proposed representation was achieved using the resubstitution test for the 25PDB, 1189 and 359 datasets, respectively. To compare, jackknife test resulted in average 19.8, 17.8, and 61.0% corresponding accuracy lifts. Thus the resubstitution test should not, and, in many cases, is not used to estimate accuracy. At the same time, execution of the jackknife test requires substantial computational time, while a much less demanding and commonly performed 10-fold cross-validation test could be used instead. For instance, tests for the LR classifiers and the 25PDB datasets using 10-fold cross-validation requires about 50 min, while about 8400 min was required for the jackknife test. To evaluate if there is any statistically significant difference between the two test procedures, a paired t-test between the accuracy lift results for the 10-fold cross-validation and the jackknife tests and the eight classification algorithms, and over the three datasets and the three representations was performed, see Table 15.

The t-test results show that for most of the cases there is no statistically significant difference between the accuracy lift reported based on the 10-fold cross-validation and the jackknife tests. In several instances, i.e. for the 1189 dataset and the composition vector representation, and for the 359 dataset and both the composition vector and the proposed 66 features representations, the accuracies computed using jackknife test are statistically significantly better than those computed using 10-fold cross-validation. This means that using jackknife test sometimes may result in overestimation of the accuracy. The study shows that not only 10-fold cross-validation seems to be at least as reliable as the currently performed jackknife test, but at the same time is it substantially less computationally demanding. Therefore, feature studies should perform 10-fold cross-validation tests.

5. Summary and conclusions

Prediction of protein structural classes is a very important and challenging problem. Over the last three decades many attempts, with varying degrees of success and novelty, were made to propose such prediction methods. Based on the past works we conclude that in general a progress was being made. At the same time, there was no comprehensive study that would point out some of the existing problems, which include testing standardization, introduction of new classification algorithms, and alternative sequence representations. The main drawback of the past papers was their

Table 14
Paired t-test results between results achieved using different protein representations

Dataset	25PDB			1189			359		
	t-test result	t-score	Confidence level	t-test result	t-score	Confidence level	t-test result	t-score	Confidence level
66 compared with CV	=	0.9	N/A	=	0.2	N/A	=	1.4	N/A
66 compared with AC	++	10.3	> 99.9%	++	22.8	> 99.9%	++	4.1	> 99.7%
CV compared with AC	++	9.8	> 99.9%	++	13.1	> 99.9%	++	3.5	> 99.5%

++ denotes that the first representation is statistically significantly better than the second representation,—indicates that the first representation is statistically significantly worse, and = indicates that there is no significant difference; the following abbreviations are used for the corresponding representations: composition vector (CV), auto-correlations (AC), and 66 features (66).

Table 15
Paired t-test results between 10-fold cross-validation and jackknife test results

Dataset representation		25PDB		1189			359			
		CV	AC	66	CV	AC	66	CV	AC	66
10-fold cross-validation compared with jackknife	t-test result	=	=	=	—	=	=	—	=	–
	t-score	0.1	1.3	1.1	2.5	0.9	0.8	3.0	1.8	3.8
	Confidence level	N/A	N/A	N/A	> 97%	N/A	N/A	> 99%	N/A	> 99.5%

++ denotes that the 10-fold cross-validation gives statistically significantly higher values than the jackknife test,—indicates that 10-fold cross-validation gives statistically significantly lower values, and = indicates that there is no significant difference; the following abbreviations are used for the corresponding representations: composition vector (CV), auto-correlations (AC), and 66 features (66).

limited scope. Usually a new method was proposed and compared with competing methods on individual and different, in terms of the size and the sequence homology, datasets and using different sequence representations. This results in unreliable conclusions and difficulty in evaluation of the true state of the art for this prediction problem.

To this end, this paper performs comprehensive, multi-goal study that addresses comparison of eight classification algorithms on three common datasets, using three sequence representations and three test types. The tested algorithms include those used in the past and several new ones. Similarly a new sequence representation and test procedure are proposed and compared with those used in the past. Based on extensive experimental study several important discoveries and conclusions are made:

- First, sequence homology is found to significantly affect prediction accuracy. Algorithms should not be compared using datasets of unknown and different homology, as the results for highly homologous datasets are shown to be statistically significantly higher than those for the datasets with low homology. The tests should be performed using low homology and standard (benchmark) datasets. This paper provides two datasets that can be used for future comparative studies.
- Second, a new to the field prediction algorithm based on logistic regression is found to generate results that are competitive or better when compared with the past results. Also, high quality of the previously used support vector machine classifiers is confirmed.
- Third, results confirm the 60% accuracy limit first discussed by Wang and Yuan [13]. Higher accuracy of some other competing methods was achieved by using highly homologous datasets and/or by application of improper procedures. We show that for eight considered prediction algorithms, state-of-the-art sequences representation and low, about 30%, homologous dataset, the best results are in the range of 57% accuracy.
- Fourth, we show that the newly proposed sequence representation is beneficial for high quality prediction algorithms, i.e. for logistic regression and support vector machines, while it does not help to improve accuracy of other algorithms.
- Finally, the resubstitution tests are shown to significantly overestimate the prediction accuracy, and the commonly performed jackknife test procedure leads to unnecessarily high computational demand. The experimental results revealed that a significantly simpler, in terms of the computational load, 10-fold cross-validation test is shown to be statistically not significantly different than the jackknife test. Therefore, we recommend that this test type should be used in the future studies.

Acknowledgement

This research was supported in part by the Canadian Natural Science and Engineering Research Council (NSERC).

Appendix A

List of sequences and domains from the 25PDB dataset; values after semicolon denote domain ranges and the fifth character denotes specific chain (if missing then the sequence has only one chain) (see Table A.1).

Table A.1

25PDB dataset

All-α

1A1W, 1A56, 1A6M, 1AB3, 1ABV, 1ADUB:180-265, 1AIPH:3-53, 1AJ3, 1AK0, 1ALU, 1AOY, 1ASH, 1AVOA, 1B0NA:74-108, 1B0NA:1-68, 1B0NB, 1B22A, 1B28A, 1B4UA, 1B8ZA, 1BAL, 1BAX, 1BBHA, 1BBN, 1BC9, 1BEA, 1BG8A, 1BGF, 1BH8B, 1BH9A, 1BK6A, 1BKRA, 1BL0A:63-124, 1BL0A:9-62, 1BO9A, 1BP3A, 1BR0A, 1BSHA:87-138, 1BT6A, 1BU2A:22-148, 1BUYA, 1BW6A, 1C1KA, 1C20A, 1C53, 1C75A, 1C9IA:331-357, 1CF7A, 1CF7B, 1CIF, 1CMBA, 1CNT4, 1COKA, 1COO, 1COPD, 1CTJ, 1CY5A, 1CZ2A, 1D2VA, 1D2ZB, 1D5VA, 1D8BA, 1D8JA, 1D8LA:65-140, 1DGNA, 1DIZA:100-282, 1DK8A, 1DNYA, 1DP3A, 1DP5B, 1DP7P, 1DPUA, 1DQEA, 1DU6A, 1DVKB, 1DVOA, 1E29A, 1E52A, 1E6BA:88-220, 1E6IA, 1E7LA:104-157, 1EB7A:165-323, 1EB7A:1-164, 1ECA, 1ECIA, 1EF4A, 1ELKA, 1ELRA, 1ENWA, 1EO0A, 1EOQA, 1ERD, 1ETED, 1EUMA, 1EXJA:3-120, 1EYHA, 1F4IA, 1F5QB:147-252, 1F5QB:6-146, 1F6VA, 1F7CA, 1FADA, 1FAFA, 1FEXA, 1FF1A, 1FFKS, 1FIPA, 1FLIA, 1FP2A:8-108, 1FPOC:1-76, 1FPOC:77-171, 1FQKA:61-181, 1FR2A, 1FS9A, 1FYJA, 1FZPB, 1G03A, 1G1EB, 1G6IA, 1G7OA:76-215, 1G8EA, 1G8QA, 1GA3A, 1GAKA, 1GC6A:88-198, 1GJTA, 1GKMA, 1GNC, 1GOTG, 1GSCA:85-217, 1GSQ:76-202, 1GU2B, 1GUMA:81-220, 1GUXB, 1GVDA, 1GXMB, 1GYZA, 1GZSB, 1H0TB, 1H1JS, 1H31B, 1H3LB, 1H4JB, 1H4LD, 1H6OA, 1H8EI, 1H97A, 1H99A:54-168, 1H9EA, 1HBKA, 1HCIA:272-396, 1HCRA, 1HD6A, 1HE8A:525-725, 1HFES, 1HH8A, 1HKQA, 1HLOA, 1HM7A, 1HMWA:26-335, 1HNS, 1HQ1A, 1HQBA, 1HRYA, 1HS5A, 1HS7A, 1HX8B:22-162, 1HX8B:167-299, 1HXGA:221-548, 1HXGA:15-220, 1HZ4A, 1I1SA, 1I27A, 1I2TA, 1I4ZA, 1IAPA, 1IB1A, 1ICHA, 1IE9A, 1IFYA, 1IG6A, 1IIEA, 1IIOA, 1IJYA, 1IK7B, 1IRDB, 1IRG, 1IRJD, 1IRL, 1IRQA, 1IRZA, 1IT2A, 1ITF, 1ITHA, 1IUFA:76-141, 1IUYA, 1IW8D, 1IX9A:1-90, 1J0PA, 1J0TA, 1J2JB, 1J75A, 1J7QA, 1J9IA, 1JEIA, 1JFBA, 1JFIA, 1JFIB, 1JGCA, 1JGSA, 1JHGA, 1JIGA, 1JJRA, 1JJSA, 1JKUA, 1JKW:11-161, 1JKW:162-287, 1JL7A, 1JLI, 1JNIA, 1JOYA, 1JQJD:213-333, 1JR5A, 1JR8A, 1JUMA:2-72, 1JUMA:73-187, 1JVR, 1JW2A, 1JYBA:2-147, 1K04A, 1K0MA:92-240, 1K1VA, 1K3XA:125-213, 1K5OA, 1K61D, 1K6KA, 1K8KE, 1K94A, 1K99A, 1KA8A, 1KANA:126-253, 1KBHA, 1KEYC, 1KF6B:106-243, 1KFTA, 1KGZB:12-80, 1KHOA:1-249, 1KJS, 1KO9A:136-323, 1KOYA, 1KQMB, 1KS8A, 1KWFA, 1KX7A, 1L3PA, 1L9LA, 1LB3A, 1LBU:1-83, 1LD8A, 1LDDA, 1LEA, 1LIAA, 1LJ9A, 1LMB3, 1LQ1A, 1LRIA, 1LS1A:1-88, 1LWBA, 1LYCA, 1M12A, 1M15A:2-95, 1M1EB, 1M1QA, 1M5NS, 1M70A:1-92, 1M70A:93- 190, 1M8YA, 1M9XC, 1MC2A, 1MDYB, 1MHZG, 1MKDA, 1MN8D, 1MP1A, 1MR8A, 1MWBA, 1MZBA, 1N1FA, 1N32R, 1N3KA, 1N62D:82-160, 1N69B, 1N89A, 1N8VA, 1N9DA, 1NC5A, 1ND9A, 1NEQ, 1NG7A, 1NGNA, 1NH2B, 1NHM, 1NI8A, 1NK2P, 1NKD, 1NKL, 1NKUA, 1NLXA, 1NOM:91-148, 1NP7A:205-483, 1NQ4A, 1NS1A, 1NWNA, 1NY9A, 1NYAA, 1NZPA, 1O4XA:5-79, 1O4XA:110-163, 1O82A, 1O9RA, 1OAFA, 1OAIA, 1OCZE, 1OHZB, 1OMRA, 1ON7B, 1OOHA, 1OQPA, 1OR6A, 1OR7F, 1ORGA, 1OS6A, 1OSLA, 1OTKA, 1OTRA, 1OTWA, 1OYIA, 1OYKA, 1P22B:64-136, 1P3BA, 1P3BC, 1P3BF, 1P5SA, 1P6RA, 1P8CD, 1P94A, 1PC2A, 1PD3A, 1PFVA:389-550, 1PGYA, 1PN5A:59-151, 1PNBA, 1PNBB, 1PP7U, 1PRA, 1PSRA, 1PSYA, 1PUOA:93-164,1PUOA:5-73, 1PVHB, 1PZQA, 1PZRA, 1Q02A, 1Q08A, 1Q2ZA, 1Q8CA, 1QATA:206-298, 1QKSA:9-135, 1QNTA:92-176, 1QPMA, 1QQIA, 1QV1A, 1QWNA:412-522, 1QZ4A, 1R2AA, 1R4AE, 1R4GA, 1R5ID, 1R5RA, 1RES, 1RFBA, 1RKCA:129-258, 1RKCA:1-128, 1RQTA, 1RRTA:9-230, 1RSOA, 1RSOB, 1RSS, 1RYKA, 1S0PA, 1S7AA, 1SIG, 1SKNP, 1SLY:1-450, 1T5JA, 1TAFA, 1TBAA, 1TFB:111-207, 1UB9A, 1UCPA, 1UCRB, 1UCVA, 1UFIB, 1UK5A, 1UQVA, 1USTA, 1UTG, 1UW4B, 1UZCA, 1V38A, 1V3FA, 1V54H, 1V74B, 1V92A, 1VF6A, 1VF6C, 1VII, 1VLS, 1WJFA, 1WTUA, 1XBL, 1XO1A:186-290, 1YCQA, 1YTFD:5-54, 2A0B, 2BBY, 2CPGB, 2EIAA:17-147, 2ERL, 2EZI, 2EZL, 2ILK, 2LEFA, 2LFB, 2LISA, 2PVBA, 2SAS, 2TMVP, 3CSMA, 3HDHC:204-295, 3HTSB, 3YGSP, 4CTSA

All-β

1A1X, 1A8VA:48-118, 1A9V, 1AG4, 1AIW, 1AJW, 1AM2, 1AOL, 1AONO, 1AVGI, 1AX3, 1AYOA, 1B34B, 1B35A, 1B55A, 1B9XA, 1BAK, 1BBPA, 1BCI, 1BDO, 1BDYA, 1BHU, 1BJ8, 1BPV, 1BQHH, 1BR9, 1BSHA:1-86, 1BWMA:3-116, 1BYMA, 1C01A, 1C28A, 1C4RB, 1C5EA, 1C5FK, 1C5LL, 1C8CA, 1C9IA:3-330, 1C9OA, 1C9UB, 1CAWB, 1CDB, 1CI0A, 1CI5A:1-95, 1CID:106-177, 1CID:1-105, 1CPM, 1CQ3A, 1CQYA, 1CR5A:26-107, 1CTO, 1CUR, 1D1NA, 1D3BA, 1D7PM, 1D8LA:1-64, 1DCS, 1DDMA, 1DG6A, 1DJ7B, 1DQGA, 1DQIA, 1DQTA, 1DS1A, 1DXMA, 1DXWA, 1DZ1A, 1DZKA, 1E0LA, 1E44B, 1E5CA, 1E5UI:1-89, 1E9GA, 1E9YA:106-238, 1EAJB, 1EARA:1-74, 1EAZA, 1ED7A, 1EGXA, 1EHKB:41-168, 1EJFA, 1EO2A, 1EQRA:1-106, 1ERNB:10-116, 1ETHA:337-448, 1EUWA, 1EWIA, 1EXH, 1EXSA, 1EYSH:59-259, 1EZGA, 1F3UB, 1F53A, 1F6OA, 1F86A, 1F8EA, 1FEUD, 1FFKN, 1FG9E:13-109, 1FG9E:110-221, 1FHOA, 1FHRA, 1FI2A, 1FJRA, 1FL0A, 1FLMA, 1FLTY, 1FMMS, 1FOD1, 1FUJA, 1FVIA:190-293, 1FYC, 1G291:241-301, 1G291:302-372, 1G2BA, 1G3GA, 1G43A, 1G5VA, 1G6EA, 1G6ZA, 1G84A, 1G88A, 1G9OA, 1GC6A:199-297, 1GCQC, 1GGLA, 1GJXA, 1GL4B, 1GMIA, 1GNHA, 1GP0A, 1GPPA, 1GQHD, 1GQWB, 1GSGP:339-547, 1GUIA, 1GV9A, 1GVMF, 1GVP, 1GWMA, 1GXCA, 1GXEA, 1GYWB, 1H2CA, 1H2NA, 1H2WA:1-430, 1H3ZA, 1H4AX:1-85, 1H6FB, 1H6XA, 1HAVA, 1HCE, 1HCFX, 1HDKA, 1HE8A:353-524, 1HK6A, 1HKFA, 1HLCA, 1HM8A:252-459, 1HMWA:336-599, 1HMWA:600-699, 1HT6A:348-404, 1HTRP, 1HU8A, 1HWHB:131-237, 1HWHB:32-130, 1HXRB, 1HZEA, 1I07A, 1I16, 1I1JA, 1I40A, 1I4VA, 1I8AA, 1I9BA, 1IAOA:83-178, 1IARB:1-96, 1IARB:97-197, 1IB5A, 1IB8A:91-164, 1IBYA, 1IC1A:1-82, 1IC1A:83-190, 1IFC, 1IFRA, 1IGQA, 1IHWA, 1IISC:5-86, 1IISC:87-171, 1IKOP, 1ILFA, 1IM3D, 1IRSA, 1IS3A, 1IWNA, 1J0SA, 1J3RA, 1J7VR:101-206, 1J7VR:2-100, 1JER, 1JHJA, 1JJJA, 1JK4A, 1JM1A, 1JO8A, 1JOPA, 1JOVA, 1JQ7A, 1JSYA:176-399, 1JSYA:6-175, 1JT8A, 1JYTA, 1K0HA, 1K2FA, 1K3BA, 1K3XA:1-124, 1K45A, 1K4ZA, 1K5CA, 1K5JA, 1K5NA:182-276, 1K5NB, 1K8HA, 1K8KC, 1K9CA, 1KAWA, 1KD6A, 1KDMA, 1KHOA:250-370, 1KIKA, 1KJ2B, 1KNMA, 1KO6C, 1KQ1A, 1KQRA, 1KSR, 1KT6A, 1KUM, 1KV7A:31- 170, 1KV7A:171-335, 1KWAA, 1KXGA, 1KXLA, 1L1CA, 1L1NB, 1L1OB, 1L2HA, 1LB6A, 1LF7A, 1LIXB:160-261, 1LKTA, 1LM8V, 1LMIA, 1LPLA, 1LUGA, 1LUQB, 1M1FB, 1M30A, 1M4OA, 1M5ZA, 1M7EA, 1MAI, 1MDAH, 1ME6A, 1MFGA, 1MFMA, 1MGQA, 1MI8A, 1MJUL:108-214, 1MJUL:1-107, 1MNNA, 1MUZA, 1MVFD, 1MVXA, 1MY7B, 1MZKA, 1N0FC, 1N32L, 1N3JA, 1N6UA:110-212, 1N6UA:1-109, 1N8BA, 1N8KA:340-374,1N8KA:1-163, 1NCT, 1NE3A, 1NEPA, 1NGLA, 1NH0A, 1NH2C, 1NIVA, 1NKOA, 1NKR:6-101, 1NKR:102-200, 1NLS, 1NNXA, 1NOFA:31-43, 1NPUA, 1NQJA, 1NWBA, 1NXMA, 1NYCA, 1NZ9A, 1O1UA, 1O3SA:8-137, 1O4TA, 1O4YA, 1O5LA:1-129, 1O5PA, 1O6SB, 1O7IB, 1OD3A, 1ODMA, 1OEKA, 1OFZA, 1OGOX:202-574, 1OGOX:3-201, 1OH1A, 1OH4A, 1OIOA, 1OK0A, 1OP4A, 1OQKA, 1OU8A, 1OUXA, 1OY2A, 1P0SE, 1P1MA:331-404, 1P1MA:1-49, 1P35C, 1P3EA, 1P4PA, 1P9UA, 1PEX, 1PFBA, 1PFSA, 1PGS:141-314, 1PGS:4-140, 1PH7A:205- 328, 1PH7A:36-204, 1PHT, 1PINA:6-39, 1PJWA, 1PK6A, 1PKHB, 1PLC, 1PMS, 1PQ7A, 1PRTD, 1PRTF, 1PSE, 1PYBA, 1Q67B, 1QAUA, 1QDNA:1-85, 1QFOA, 1QKSA:136-567, 1QLEB:108-252, 1QOUB, 1QQP4, 1QREA, 1QW9A:5-17, 1QW9A:385-501, 1QWDA, 1QWNA:523-1044, 1QWYA, 1QXMA:149-286, 1QXMA:4-148, 1QY1A, 1R0UA, 1R21A, 1R2MA, 1R6JA, 1R6KA, 1R75A, 1RHI1, 1RI9A, 1RIP, 1RK8C, 1RKRA, 1RL1A, 1ROCA, 1RQWA, 1S2BA, 1S2EA, 1SE1A:1-125, 1SFP, 1SG3A:195-343, 1SG3A:1-187, 1SM4A:67-207, 1SR3A, 1SSXA, 1TFHB:107-210, 1TFHB:5-106, 1TIID, 1TIU, 1TL2A, 1TME1, 1TTG, 1TUL, 1UB4B, 1UCSA, 1UD8A:391- 480, 1UEPA, 1UFFA, 1UFXA, 1UG1A, 1UJVA, 1UJXA, 1ULP, 1UMIA, 1USCA, 1UT4B, 1UW7A, 1UZ0A, 1V27A, 1VIE, 1WBC, 1WHI, 1WKT, 1XNTA, 1YTFD:55-119, 2ARCB, 2BPA2, 2DYNA, 2HNTE, 2HRVA, 2ILA, 2NLRA, 2SNS, 2STV, 2TNFA, 3CHBD, 3DPA:125-218, 3DPA:1-124, 3EZMA, 3MSPA, 3NCMA, 3SEB:1-121, 3SIL, 3VUB, 4AAHA, 4HMGA, 4ULL

Table A.1 *Continued*

25PDB dataset

α/β

1ABA, 1AO3A, 1AY7B, 1AYL:228-540, 1AYL:1-227, 1B26A:179-412, 1B26A:4-178, 1B3OA:232-499, 1B3OA:10-109, 1B4UB, 1B8GB, 1B93A, 1BCRA, 1BCRB, 1BQCA, 1BRT, 1BVH, 1BX4A, 1BYI, 1BYKA, 1C25, 1CEN, 1CFZA, 1CP2A, 1CQGA, 1CUI, 1CXQA, 1D2HA, 1D3VA, 1D4OA, 1D5TA:389-431, 1DBWB, 1DCIA, 1DE5B, 1DIRA, 1DL3A, 1DO0A, 1DOSA, 1DQZA, 1E0JA, 1E5KA, 1E6BA:8-87, 1ECXA, 1EDG, 1EEXB, 1EFM:12-190, 1EFPA:2-184, 1EIWA, 1EIZA, 1EM8B, 1EO1A, 1EOMA, 1EQA, 1ES9A, 1ETHA:1-336, 1EXCA, 1F2TB, 1F51E, 1F61A, 1F9VA, 1FEZA, 1FFKC, 1FFKG, 1FFKL, 1FFKV, 1FO5A, 1FOVA, 1FP2A:109-352, 1FQKA:28-60, 1FSGA, 1FVKA, 1FVPA, 1FYEA, 1FZTA, 1G291:1-240, 1G5QA, 1G64A, 1G66A, 1G7EA, 1G7OA:1-75, 1G8AA, 1GA6A, 1GCI, 1GIN, 1GKLA, 1GLLO:2- 253, 1GLLO:254-499, 1GLV:1-122, 1GN1G, 1GPH1:235-465, 1GQOV, 1GRCA, 1GSCA:1-84, 1GSGP:8-338, 1GSQ:1-75, 1GUMA:4-80, 1GVFA, 1GWZ, 1H2WA:431-710, 1H6JA, 1H6VC:171-292, 1H6VC:14-170, 1H6VC:293-366, 1H75A, 1HD2A, 1HDOA, 1HG3A, 1HJQA, 1HLGA, 1HM8A:2-251, 1HQKA, 1HT6A:1-347, 1HTWA, 1HUXA, 1HXHA, 1I0DB, 1I24A, 1I2ZA, 1I4NA, 1I4WA, 1I69B, 1I7LA:113-214, 1IAQB, 1IBSB:167-315, 1IBSB:6-166, 1IIBA, 1IIWA, 1IN1A, 1IOIA, 1ITQA, 1IU9A, 1IXH, 1IZYA, 1J2RC, 1J5SA, 1JDNA, 1JF8A, 1JI3A, 1JIKA, 1JL1A, 1JLSB, 1JMKO, 1JMVA, 1JN0A:313-333, 1JON, 1JQ3C, 1JQJD:1-209, 1JR4A, 1JSXA, 1JTVA, 1JUBA, 1JXIA, 1K0MA:6-91, 1K7CA, 1K92A:1-188, 1KGDA, 1KGZB:81-344, 1KI9B, 1KICA, 1KJQB:2-112, 1KMVA, 1KNGA, 1KQPA, 1KR2F, 1KTE, 1L7AA, 1L8OA, 1LC7A, 1LIXB:262-439, 1LIXB:57-159, 1LK9A, 1LKXD, 1LL4A:36-292, 1LLFA, 1LQTB:2-108, 1LQTB:109-324, 1LQTB:325-456, 1LS1A:89-295, 1LU4A, 1M0IA, 1M1BB, 1M1NA, 1M1NB, 1M2DA, 1M2EA, 1M3GA, 1M4LA, 1M65A, 1M6BB:311-479, 1M6BB:6-165, 1M7GD, 1MAVA, 1MF7A, 1MJ5A, 1MLDA:1-144,1MOQ, 1MQ0A, 1MUWA, 1MWJA, 1MXIA, 1N1DA, 1N25A, 1N2OB, 1N32B, 1N3LA, 1N4WA:9-318, 1N55A, 1N7HB, 1N7IB, 1N8KA:164-339, 1N9KA, 1NBWB, 1NF9A, 1NH7A:1-210, 1NMPA, 1NN5A, 1NNFA, 1NNUC, 1NOFA:44-320, 1NOYA, 1NP6B, 1NP7A:1-204, 1NRJB, 1NW8A, 1NZJA, 1O08A, 1O58A, 1O7JA, 1O7QA, 1O8XA, 1OAA, 1OBOA, 1OC7A, 1OD6A, 1ODGA, 1ODZA, 1OFTA, 1OHEA:42-198, 1OHHG, 1OJRA, 1ON4A, 1OOYA:1-242, 1OOYA:261-481, 1ORHA, 1OT5A:123-460, 1OVYA, 1P1MA:50-330, 1P33C, 1P4CA, 1P5FA, 1P5ZB, 1P6OA, 1P73C, 1P74B:1-101, 1P74B:102-272, 1PB7A, 1PDO, 1PFVA:176-388, 1PFVA:4-140, 1PMOC, 1POIB, 1PWYE, 1PYOB, 1PZTA, 1Q1QA, 1Q7LA, 1Q7LD, 1Q92A, 1QC9A, 1QDLB, 1QFEA, 1QGEE, 1QGVA, 1QHHA, 1QHHB, 1QHHC, 1QJ4A, 1QKIB:11-199, 1QKIB:435-449, 1QLWB, 1QMLA, 1QNRA, 1QNTA:6-91, 1QO5K, 1QOPB, 1QTNB, 1QTWA, 1QW9A:18-384, 1QWNA:31-411, 1QZMA, 1R18A, 1R26A, 1R2QA, 1R5PB, 1R5XA, 1R5YA, 1R6DA, 1R6HA, 1RFLA, 1RFVA, 1RHQA, 1RKUA, 1RPA, 1RRF, 1RTQA, 1RYOA, 1S4PB, 1SFSA, 1SHUX, 1ST9A, 1SX5A, 1T2DA:1-150, 1THX, 1UD8A:1-390, 1UEHA, 1UG6A, 1UOCA, 1URSA, 1US0A, 1USLA, 1UWCA, 1UZBA, 1V2XA, 1V7RA, 1V8AA, 1VGUB, 1VHWF, 1VIMA, 1XO1A:19-185, 1YACA, 1YUB, 2AT2A:145-295, 2AT2A:1-144, 2PJRB, 2PTH, 2TPSA, 2TSYA, 3CLA, 3FUA, 3HDHC:12-203, 3PVIA, 4EUGA, 6PFKA, 7A3HA, 7MHTA, 8ABP

$\alpha+\beta$

169LA, 1A2N, 1A2PA, 1A67, 1A9ND, 1AA3, 1AF5, 1AIHB, 1AIPH:54-196, 1AKO, 1APS, 1APZA, 1AQ4A, 1AQZB, 1AVPA, 1AYYB, 1B04B, 1B10A, 1B33N, 1B3AA, 1B5EA, 1B65A, 1B69A, 1B6FA, 1B87A, 1B9LA, 1BNLA, 1BOB, 1BXYA, 1BY2, 1BYSA, 1BYWA, 1C05A, 1C7KA, 1CC8A, 1CJKB, 1CKJB, 1CKV, 1CQMA, 1CV8, 1CXYA, 1CZPA, 1D5TA:292-388, 1D8IA, 1D9UA, 1DCHA, 1DCJA, 1DEF, 1DI2B, 1DIZA:1-99, 1DOKA, 1DT4A, 1E0GA, 1E1HA, 1E1HD, 1E44A, 1E5UI:90-187, 1E7KA, 1E7LA:1-103, 1E87A, 1E9YA:1-105, 1EARA:75-142, 1EAYC, 1EB6A, 1ECSA, 1EF5A, 1EGGB, 1EGWA, 1EKTA, 1EL6A, 1EMWA, 1EQKA, 1EQRA:107-287, 1EQRA:288-420, 1EQRA:421-590, 1EUVA, 1EUVB, 1EV0A, 1EW4A, 1EXJA:121-277, 1F08A, 1F0ZA, 1F2RI, 1F32A, 1F40A, 1F51A, 1F60B, 1F7LA, 1F96A, 1F9YA, 1FFK1:1-79, 1FFK1:80-172, 1FFKD, 1FFKF, 1FFKP, 1FFKU, 1FJCA, 1FM0D, 1FPYA:1-100, 1FPYA:101-468, 1FU6A, 1FVIA:2-189, 1FW9A, 1FX4A, 1G61A, 1G71A, 1GC1G, 1GC6A:1-87, 1GD0A, 1GH8A, 1GHHA, 1GK9A, 1GK9B, 1GO1A, 1GPH1:1-234, 1GPQB, 1GTPA, 1GTQA, 1GW5S, 1GXUA, 1GXYA, 1GY7B, 1GYFA, 1GYXA, 1H0YA, 1H3QA, 1H5PA, 1H6HA, 1H6KY, 1H6VC:367-495, 1H8CA, 1HBNB:2-188, 1HE8A:144-321, 1HL6D, 1HMJA, 1HQ6A, 1HQI, 1HQZ1, 1HV2A, 1HYWA, 1HZ6B, 1HZTA, 1I0VA, 1I12A, 1I17A, 1I35A, 1I7EA, 1I9YA, 1IAD, 1IAJB, 1IAOA:1-82, 1IB8A:1-90, 1IBXA, 1ID0A, 1IDPA, 1IHRA, 1IJKC, 1IKM, 1IMUA, 1IOUA, 1IPBA, 1IPGA, 1IQSA, 1IQZA, 1IRYA, 1IS7K, 1ITPA, 1IU3C, 1IUJB, 1IV3A, 1IVZA, 1IX9A:91-205, 1J0GA, 1J27A, 1J3GA, 1J4WA:1-74, 1J4WA:104-174, 1J57A, 1J6RA, 1J8CA, 1JATA, 1JATB, 1JBIA, 1JC5B, 1JD21, 1JD2K, 1JD2L, 1JD2M, 1JFMA, 1JH6A, 1JHSA, 1JIDA, 1JIHA:390-509, 1JK3A, 1JKNA, 1JN0A:149-312, 1JNZB, 1JO0A, 1JOSA, 1JRKA, 1JRMA, 1JRUA, 1JW3A, 1JYOA, 1K0KA, 1K1GA, 1K3EA, 1K4IA, 1K5NA:1-181, 1K83K, 1K8BA, 1K8KF, 1K92A:189-444, 1KAFD, 1KANA:1-125, 1KCGC, 1KCQA, 1KF6B:1-105, 1KG0C, 1KJKA, 1KJQB:113-318,1KN0A, 1KN6A, 1KO9A:12-135, 1KOTA, 1KP6A, 1KPQA, 1KPTA, 1KQFB:2-245, 1KUFA, 1KVDB, 1KVEA, 1KZNA, 1L0OA, 1L1PA, 1L3GA, 1L3KA:103-181, 1L3KA:8-91, 1L4ZB, 1L5PA, 1L9AA, 1L9YA, 1LBU:84-213, 1LKKA, 1LL4A:293-354, 1LL8A, 1LNIA, 1LO7A, 1LQ9A, 1LTZA, 1LY7A, 1M0VA, 1M15A:96-357, 1M4JA, 1MBXD, 1MBYA, 1ME4A, 1MG4A, 1MG7A:14-187, 1MG7A:188-380, 1MHDA, 1MHMB, 1MK0A, 1MK4A, 1MKBA, 1ML8A, 1MLDA:145-313, 1MOGA, 1MOLA, 1MSZA, 1MW4A, 1MWPA, 1MWWB, 1N13C, 1N32C:107-207, 1N32C:2-106, 1N32I, 1N32J, 1N4WA:319-450, 1N62C:1-177, 1N62C:178-286, 1N62D:2-81, 1N6ZA, 1NEIA, 1NH7A:211-284, 1NKIA, 1NO5A, 1NR3A, 1NRJA, 1NSKL, 1NVJD, 1NWWB, 1NWZA, 1NXIA, 1NZ8A, 1O0PA, 1O26A, 1O2FB, 1O50A:77-145, 1O7BT, 1O7NB, 1O8RA, 1OCYA, 1ODHA, 1OF5A, 1OF5B, 1OFHG, 1OH0A, 1OJ5A, 1OJGA, 1OO5A, 1OPD, 1OPZA, 1OQJB, 1OQQA, 1OQVA, 1OQWA, 1OTFA, 1OTGA, 1OWTA, 1P0RA, 1P0ZA, 1P1TA, 1P22B:2-59, 1P32B, 1P4LD, 1P4OA, 1P65A, 1P9KA, 1PA4A, 1PAVA, 1PBA, 1PBUA, 1PC6B, 1PCFA, 1PIL, 1PINA:45-163, 1PQSA, 1PRTA, 1PRTB:4-89, 1PUGC, 1PVMB:65-142, 1PYTA, 1PZ4A, 1Q53A, 1Q5YB, 1Q8LA, 1Q8RA, 1QB3B, 1QDDA, 1QDNA:86-201, 1QFCA, 1QG7A, 1QHKA, 1QKFA, 1QKIB:200-434, 1QKIB:450-511, 1QKLA, 1QL0A, 1QMTA, 1QOLA, 1QR5A, 1QS1A:265-461, 1QS1A:60-264, 1QSOA, 1QSTA, 1QTOA, 1QXYA, 1QYMA, 1QYNA, 1R29A, 1R52B, 1R8HC, 1REGY, 1RFA, 1RJTA, 1RO2A, 1RRTA:231-360, 1RWZA:1-122, 1RWZA:123-244, 1RY9A, 1RYJA, 1S0YD, 1S0YE, 1S5FA, 1S5UB, 1S79A, 1S7JA, 1SB6A, 1SCJB, 1SF0A, 1SGOA, 1SJWA, 1SLY:451-618, 1SP4A, 1ST4A:38-145, 1ST4A:146-337, 1T0GA, 1T0YA, 1T1DA, 1T2DA:151-315, 1TBAB:61-155, 1TIG, 1TIIC, 1UB1A, 1UFYA, 1UNNC, 1UQ5A, 1USMA, 1UUTA, 1UUZB, 1UW4A, 1V2YA, 1V74A, 1VAZA, 1VCC, 1VHIB, 1VI8B, 1VIH, 1XXCA, 2ATCB:1-100, 2BOPA, 2FDN, 2FMR, 2IGD, 2JDXA, 2NEF, 2NMTA:34-218, 2PLEA, 2PROB:86-158, 2PROB:4-85, 2SAK, 2SXL, 2TBD, 2TLDI, 2U1A, 2VIL, 3GCC, 3LZT, 3SEB:122-238, 3ZNBA

List of sequences and domains from the 1189 dataset; values after semicolon denote domain ranges and the fifth character denotes specific chain (if missing then the sequence has only one chain) (see Table A.2).

Table A.2

1189 dataset

All-α

1AAB, 1AB3, 1ABV, 1ACA, 1ACP, 1ADR, 1AEP, 1AF8, 1AFRA, 1AGRE, 1AJ3, 1AK4C, 1ALLA, 1AN2A, 1AORA:211-605, 1AOY, 1ARU, 1ASH, 1BBHA, 1BBL, 1BCFA, 1BEO, 1BFMA, 1BGC, 1BIA:1-63, 1BIP, 1BUCA:233-383, 1BVP1:1-120, 1BVP1:255-349, 1C5A, 1CC5, 1CEM, 1CMBA, 1CNT1, 1COO, 1COPD, 1CPCA, 1CPCB, 1CPQ, 1CPT, 1CRKA:1-98, 1CSGA, 1CSH, 1CSMA, 1CUK:65-142, 1CYI, 1DNPA:201-469, 1DVH, 1ECA, 1ECIA, 1ECMA, 1ENH, 1ERC, 1ERD, 1ERP, 1ERY, 1ETPA:93-190, 1ETPA:1-92, 1FAPB, 1FCDC:81-174, 1FCDC:1-80, 1FIPA, 1FJLA, 1FLP, 1FOW, 1FPS, 1GAB, 1GKS, 1GLM, 1GLN:306-468, 1GLQA:79-209, 1HBG, 1HCRA, 1HDJ, 1HMCA, 1HME, 1HNR, 1HRZA, 1HSTA, 1HUEA, 1HULA, 1HUW, 1HVD, 1HYP, 1IHFB, 1ILK, 1IMQ, 1ITHA, 1JLI, 1JVR, 1LBD, 1LBU:1-83, 1LCCA, 1LEA, 1LFB, 1LH1, 1LIS, 1LKI, 1LLIA, 1LPE, 1LRE, 1LRV, 1MBD, 1MDYA, 1MHLA, 1MHLC, 1MMOB, 1MMOD, 1MMOG, 1MNGA:1-92, 1MNTA, 1MYKA, 1NER, 1NGR, 1NKL, 1OCCE, 1OCCH, 1OLGA, 1OPC, 1OSA, 1OXA, 1PBWA, 1PDNC, 1PHB, 1PNBA, 1PNBB, 1POA, 1POC, 1PPRM:1-156, 1PPRM:157-312, 1PRCC, 1PUEE, 1R69, 1RCD, 1REC, 1RES, 1RFBA, 1RGP, 1RIBA, 1ROM, 1RPO, 1RRO, 1SCMB, 1SETA:1-110, 1SIG, 1SLY:1-450, 1SRA, 1TAFA, 1TAFB, 1TCOB, 1TF4A:1-460, 1TNS, 1TPT:1-70, 1UTG, 1VII, 1VNC, 1VTMP, 1XGSA:195-271, 1XSM, 1YRNA, 1YRNB, 1YTFB, 256BA, 2ABK, 2BCT, 2BMHA, 2CCYA, 2CYP, 2END, 2GSTA:85-217, 2HMQA, 2HMX, 2HTS, 2INT, 2LHB, 2LIGA, 2MTAC, 2MYSB, 2PDE, 2PGD:177-473, 2SAS, 2SCPA, 2SPCA, 2WRPR, 351C, 3INKC, 3SDHA, 4ICB, 1ADT:176-265, 1AOFA:36-133, 1BMFA:380-510, 1BMFD:358-475, 1CGPA:138-205, 1CLC:135-575, 1CUK:156-203, 1DJXA:200-298, 1DPRA:65-136, 1DPRA:3-64, 1GNWA:86-211, 1GRJ:2-79, 1GRL:410-523, 1GRL:6-136, 1HC2:136-398, 1HC2:5-135, 1JKW:11-161, 1JKW:162-287, 1LLA:2-109, 1LLA:110-379, 1OCTC:5- 75, 1PNRA:3-58, 1RLR:10-221, 1RYT:2-147, 1SFE:93-176, 1TADA:57-177, 1TFR:183-305, 1YTFD:5-54, 1ZYMA:22-144, 2SBLB:150-839, 2TCT:2-67, 5EAS:221-548, 5EAS:24-220, 2LEFA, 1GH1A

All-β

1ABRB:1-140, 1ABRB:141-267, 1AGJA, 1AH9, 1AHSA, 1AIZA, 1ALY, 1AMY:347-403, 1ANU, 1AOL, 1AONO, 1AOZA:130-338, 1AOZA:1-129, 1AOZA:339-552, 1ARB, 1BBPA, 1BBT1, 1BBT3, 1BDO, 1BEBA, 1BOVA, 1BTKA, 1BTN, 1BTY, 1BVP1:121-254, 1BW3, 1CDCB, 1CDG:496-581, 1CDG:582-686, 1CID:106-177, 1CID:1-105, 1CKAA, 1CPN, 1CSKA, 1CTM:168-230, 1CTM:1-167, 1CTM:231-250, 1CTO, 1CUK:1-64, 1CUR, 1CWPA, 1CYX, 1DUPA, 1DUTA, 1DYNA, 1EAGA, 1EAL, 1EFT:213-312, 1EFT:313-405, 1EPBA, 1EPNE, 1ETA1, 1EUR, 1EXG, 1FGP, 1FIVA, 1FMB, 1FNA, 1FUIA:356-591, 1FYC, 1GEN, 1GHK, 1GLAF, 1GOF:151-537, 1GOF:1-150, 1GOF:538-639, 1GPC, 1GPR, 1GZI, 1HAVA, 1HBP, 1HCD, 1HGEA, 1HMS, 1HOE, 1HSQ, 1HTP, 1HXN, 1I1B, 1IDAA, 1IDK, 1IFC, 1IHWA, 1ILR1, 1IRSA, 1IYU, 1JDC:358-418, 1JER, 1KAPP:247-470, 1KCW:193-338, 1KCW:1-192, 1KNB, 1KSR, 1LAC, 1LCL, 1LTSD, 1LXA, 1MAI, 1MJC, 1MPP, 1MSAA, 1MSPA, 1MUP, 1NBCA, 1NCIA, 1NEU, 1NFA, 1NOA, 1NPOA, 1NSCA, 1OBPA, 1OCCB:91-227, 1OSPO, 1PCL, 1PDR, 1PEX, 1PFSA, 1PHT, 1PKYA:70-167, 1PLC, 1PLS,1PMI, 1PMS, 1PPI:404-496, 1PRR:91-173, 1PRR:1-90, 1PRTD, 1PRTF, 1PSE, 1PVC1, 1PVC2, 1PVC3, 1PYP, 1RIP, 1RSY, 1SACA, 1SCS, 1SE4:1-121, 1SEMA, 1SGC, 1SHCA, 1SHG, 1SLAA, 1SLUA, 1SMPI, 1SRIA, 1SRO, 1SSO, 1STMA, 1STY, 1SVA1, 1TDTA, 1TEN, 1TF4A:461-605, 1THJA, 1THW, 1TIE, 1TIID, 1TIU, 1TLK, 1TME1, 1TNFA, 1TNM, 1TNRA, 1TSP, 1TUL, 1TUPA, 1ULO, 1VCAA:91-199, 1VCAA:1-90, 1VFBA, 1VIE, 1VMOA, 1WAPA, 1WBA, 1WHI, 1WHO, 1WIU, 1WKT, 1XNB, 1XSOA, 1YAIA, 1YHB, 1YTFC, 1ZNCA, 1ZXQ:1-86, 1ZXQ:87-192, 2ALP, 2ARCA, 2AVIA, 2BBKH, 2BBVA, 2BPA1, 2BPA2, 2CAS, 2CBP, 2CPL, 2ENG, 2FGF, 2ILA, 2KAUB, 2MEV1, 2MEV2, 2NCM, 2PCDA, 2PCDM, 2PEC, 2PIA:1-103, 2PRD, 2RSPA, 2SIL, 2SNV, 2STV, 2TBVA, 2TRCB, 3CD4:98-178, 3CD4:1-97, 3NN9, 3ULLA, 4AAHA, 4GCR:86-174, 4GCR:1-85, 1AOFA:134-567, 1ASYA:68-204, 1BGLA:626-730, 1BGLA:3-219, 1BGLA:220-333, 1BGLA:731-1023, 1BHGA:226-328, 1BHGA:22- 225, 1BIA:271-317, 1BMFA:24-94, 1BMFD:9-81, 1BNCA:331-446, 1CD1A:186-279, 1CDG:407-495, 1CGPA:9-137, 1CIY:256-461, 1CKMA:239-327, 1CLC:35-134, 1CTN:24-132, 1DAR:283-400, 1DDT:381-535, 1DKGA:139-197, 1DLC:290-499, 1EBPA:10-116, 1ESFA:1-120, 1FDR:2-100, 1FNB:19-154, 1GGTA:516-627, 1GGTA:628- 729, 1GGTA:8-190, 1GTRA:339-547, 1HC2:399-653, 1KCW:347-553, 1KCW:554-705, 1KCW:706-884, 1KCW:892-1040, 1KEVA:1-139, 1KEVA:314-351, 1KIT:217-346, 1KIT:25-216, 1KIT:347-543, 1LLA:380-628, 1LYLA:14-153, 1MMD:34-79, 1PGS:141-314, 1PGS:4-140, 1QBA:28-200, 1QORA:292-327, 1QORA:2-112, 1RGS:113-244, 1SFTA:2-11, 1SFTA:245-383, 1SVB:303-395, 1YTFD:55-119, 2AAA:382-476, 2BB2:86-175, 2CND:11-124, 2HFT:107-211, 2HFT:1-106, 2KAUC:2-129, 2KAUC:423-475, 2OHXA:1-163, 2OHXA:340-374, 2PHLA:11-210, 2PHLA:220-381, 2SBLB:7-149, 3DPA:125-218, 3DPA:1-124, 3HHRB:32-130, 4KBPA:9-120, 4BCL, 2TSSA:1-93

α/β

1ABA, 1AD3A, 1ADD, 1ADEA, 1AG8A, 1AMP, 1AMY:1-346, 1ART, 1ASU, 1BAM, 1BLE, 1BMFG, 1BNCA:1-114, 1BROA, 1BRSD, 1BYB, 1CB2A, 1CBG, 1CEC, 1CFR, 1CHD, 1CSEE, 1CTT:151-294, 1CTT:1-150, 1CUS, 1CYDA, 1DAPA:1-118, 1DAPA:269-320, 1DCTA, 1DEAA, 1DHPA, 1DHR, 1DNPA:1-200, 1DORA, 1DOSA, 1DPGA:1-181, 1DPGA:413-426, 1DPPA, 1DRAA, 1DTS, 1DUBA, 1DXY:1-100, 1E2B, 1EAF, 1EBHA:142-436, 1ECEA, 1ECPA, 1EDE, 1EDG, 1EDT, 1EFT:1-212, 1EGO, 1ENY, 1ERIA, 1ESC, 1FCDA:256-327, 1FCDA:115-255, 1FCDA:1-114, 1FDS, 1FMCA, 1FUA, 1FUIA:1-355, 1GARA, 1GCA, 1GGGA, 1GHR, 1GLN:1-305, 1GLQA:1-78, 1GND:1-291, 1GND:389-430, 1GPB, 1GPH1:235-465, 1GYM, 1HDCA, 1HGXA, 1HJRA, 1HMPA, 1HMY, 1HRDA:1-194, 1HRDA:195-449, 1HURA, 1HVQ, 1ICEA, 1ICEB, 1IDM, 1IDO, 1IGS, 1ITG, 1JDC:1-357, 1KIFA:1-194, 1KIFA:288-339, 1KTE, 1LAM:1-159, 1LAM:160-484, 1LCT, 1LDM:1-160, 1LEHA:1-134, 1LEHA:135- 364, 1LFAA, 1LST, 1LUCA, 1LUCB, 1LVL:266-335, 1LVL:151-265, 1LVL:1-150, 1MEK, 1MIOA, 1MIOB, 1MPB, 1NAL1, 1NAR, 1NBAA, 1NFP, 1NHP:1-119, 1NHP:120- 242, 1NHP:243-321, 1NIPA, 1NOYA, 1NSJ, 1NSYA, 1NTR, 1NULA, 1NZYA, 1OBR, 1OFGA:1-160, 1OFGA:323-381, 1OPR, 1ORB:150-293, 1ORB:1-149, 1ORDA:108-569, 1ORDA:1-107, 1ORTA:151-335, 1ORTA:1-150, 1OYA, 1PAUA, 1PAUB, 1PBE:276-391, 1PBE:1-173, 1PBN, 1PBP, 1PDO, 1PEA, 1PFKA, 1PHP, 1PHR, 1PII:1-254, 1PII:255-452, 1PKYA:168-344, 1PKYA:1-69, 1POT,1PPI:1-403, 1PTA, 1PUD, 1PVUA, 1QRDA, 1RAAA:151-310, 1RAAA:1-150, 1RCF, 1RLAA, 1RPA, 1RVAA, 1RVVA, 1SBP, 1SCUA:1-121, 1SCUA:122-288, 1SRRA, 1TAHB, 1TCA, 1TDE:119-244, 1TDE:1-118, 1TDE:245-316, 1THTA, 1TIB, 1TLFA, 1TML, 1TPFA, 1TPLA, 1TPT:71-335, 1UDG, 1V39, 1VHRA, 1VID, 1VTK, 1WHTA, 1WHTB, 1XEL, 1XVAA, 1XYZA, 1YASA, 1YBVA, 1YPTA, 2ACR, 2ADMA, 2ANHA, 2AT2A:145-295, 2AT2A:1-144, 2BGU, 2CHR:127-370, 2CMD:1-145, 2CTB, 2DKB, 2DLN:1-96, 2DRI, 2EBN, 2FX2, 2GLT:1-122, 2GSTA:1-84, 2HNP, 2LBP, 2MASA, 2NACA:336-374, 2NACA:148-335, 2NACA:1-147, 2OLBA, 2PGD:1-176, 2PIA:104-223, 2RN2, 2RSLA, 2TMDA:341-489, 2TMDA:1-340, 2TMDA:490-645, 2TMDA:646-729, 2TPRA:286-357, 2TPRA:169-285, 2TPRA:1-168, 2TRXA, 2XIS, 3CHY, 3CLA, 3DFR, 3PGM, 3PMGA:191-303, 3PMGA:1-190, 3PMGA:304-420, 3TGL, 5NUL, 5P21, 7ICD, 8ABP,

Table A.2 *Continued*

1189 dataset
8DFR, 1ADJA:326-421, 1AK5:2-101, 1AK5:222-483, 1ATIA:395-505, 1AYL:228-540, 1AYL:1-227, 1BGLA:334-625, 1BMFA:95-379, 1BMFD:82-357, 1CDG:1-406, 1CHMA:2-156, 1COY:4-318, 1CTN:133-443, 1DAR:1-282, 1DIH:241-273, 1DIH:2-130, 1DIK:510-874, 1DIK:377-505, 1DSBA, 1DXY:101-299, 1FDR:101-248, 1FNB:155-314, 1GAL:3-324, 1GAL:521- 583, 1GD1O:313-333, 1GDHA:2-100, 1GDHA:101-291, 1GESA:3-146, 1GESA:263-335, 1GESA:147-262, 1GLAG:254-499, 1GLAG:4-253, 1GNWA:2-85, 1GPMA:3-207, 1GPMA:208-404, 1GRL:191-366, 1GSEA:2-80, 1GTMA:3-180, 1GTMA:181-419, 1GTRA:8-338, 1HLPA:21-162, 1HPLA:1-336, 1HPM:4-188, 1HPM:189-381, 1HYHA:21- 166, 1KEVA:140-313, 1KFD:324-518, 1LDB:15-162, 1LDG:18-163, 1LLDA:7-149, 1MLA:3-127, 1MLA:198-307, 1MMD:2-33, 1MMD:80-759, 1PDA:3-219, 1PKYA:351- 470, 1PNRA:59-340, 1POXA:183-365, 1POXA:9-182, 1PSDA:108-295, 1PSDA:296-326, 1PSDA:7-107, 1PVDA:2-181, 1PVDA:182-360, 1PXTA:28-293, 1QAPA:130-296, 1QBA:338-780, 1QORA:113-291, 1REQA:2-560, 1REQB:20-475, 1RLR:222-748, 1RNL:5-142, 1SCUB:239-388, 1SFE:12-92, 1SFTA:12-244, 1TADA:27-56, 1TFR:12-180, 1TRKA:3-337, 1TRKA:338-534, 1TRKA:535-680, 1YVEI:83-307, 1ZYMA:145-249, 1ZYMA:3-21, 2AAA:1-381, 2CND:125-270, 2KAUC:130-422, 2KAUC:476-567, 2OHXA:164-339, 2REB:3-268, 2TS1, 3RUBL:148-467, 5RUBA:138-457, 1BKSA, 1BKSB
$\alpha + \beta$
119L, 193L, 1AB8A, 1ABRA, 1ACF, 1AF5, 1AFI, 1AG2, 1AH6, 1AHQ, 1AIHA, 1AK7, 1AKO, 1AORA:1-210, 1APA, 1APS, 1APYA, 1APYB, 1AST, 1ATLA, 1BP1:1-217, 1BP1:218-456, 1BRNL, 1BV1, 1CBY, 1CEWI, 1CHKA, 1COAI, 1CRKA:99-380, 1CTF, 1DAPA:119-268, 1DCOA, 1DDT:1-187, 1DEF, 1DHMA, 1DMAA, 1DONA, 1DPGA:182-412, 1DPGA:427-485, 1EBHA:1-141, 1EFNB, 1EPS, 1ESL:1-118, 1FCA, 1FCDA:328-401, 1FD2, 1FJMA, 1FKD, 1FRD, 1FROA, 1FWP, 1FXRA, 1GBS, 1GCB, 1GMPA, 1GND:292-388, 1GPH1:1-234, 1GTPA, 1GTQA, 1GUAB, 1HFC, 1HQI, 1HUMA, 1IBA, 1IGD, 1KIFA:195-287, 1KPTA, 1KUH, 1KVDA, 1KVDB, 1LBA, 1LBU:84-213, 1LDM:161-329, 1LIT, 1LML, 1LTSA, 1LTSC, 1LVL:336-458, 1MAT, 1MKAA, 1MLI, 1MNGA:93-203, 1MOLA, 1MRJ, 1MSK, 1MUT, 1NAPA, 1NHP:322-447, 1NOX, 1NPK, 1OFGA:161-322, 1ORDA:570-730, 1OTFA, 1OTGA, 1OUNA, 1PBA, 1PBE:174-275, 1PIL, 1PLQ:127-258, 1PLQ:1-126, 1PMAA, 1PMAB, 1PNKA, 1PNKB, 1POH, 1PRTA, 1PTF, 1PUT, 1PYAA, 1QBEA, 1RAAB:1-100, 1REGX, 1RIS, 1SCEA, 1SE4:122-239, 1SEIA, 1SETA:111-421, 1SHAA, 1SLY:451-618, 1SMNA, 1SPBP, 1SRSA, 1STD, 1STFI,1STU, 1SVR, 1SXL, 1TBD, 1TFE, 1TIF, 1TIG, 1UAE, 1UBI, 1UDII, 1URNA, 1VCC, 1VHH, 1VHIA, 1VIG, 1VJW, 1XGSA:272-295, 1XGSA:1-194, 1XXAA, 1ZNBA, 2ACT, 2BAA, 2BOPA, 2CHR:1-126, 2CHSA, 2CMD:146-312, 2DLN:97-306, 2DNJA, 2KAUA, 2MS2A, 2PHY, 2PIA:224-321, 2PLDA, 2PNB, 2POLA:245-366, 2POLA:123-244, 2POLA:1-122, 2PTL, 2SICI, 2TPRA:358-482, 2U1A, 2VIK, 3FIB, 3PMGA:421-561, 3RUBS, 7RSA, 9RNT, 1ADJA:2-325, 1ATIA:1- 394, 1BIA:64-270, 1BNCA:115-330, 1CD1A:7-185, 1CKMA:11-238, 1COY:319-450, 1CTN:444-516, 1DAR:476-599, 1DAR:600-689, 1DIH:131-240, 1DIK:2-376, 1DIV:1-55, 1DIV:56-149, 1DLHA:3-81, 1ESFA:121-233, 1EZM, 1GD1O:149-312, 1GESA:336-450, 1GGTA:191-515, 1GPMA:405-525, 1GRJ:80-158, 1GRL:367-409, 1GRL:137-190, 1HAN:133-289, 1HAN:2-132, 1HTTA:4-325, 1HXPA:2-177, 1HXPA:178-348, 1KAPP:1-246, 1LGR:101-468, 1LGR:1-100, 1LLDA:150-319, 1LYLA:161-502, 1MBB:3-200, 1MBB:201-342, 1MLA:128-197, 1MXA:1-102, 1MXA:108-231, 1MXA:232-383, 1PDA:220-307, 1PKP:78-148, 1PKP:4-77, 1PMD:76-263, 1PREA:2-84, 1PRTB:4-89, 1QAPA:8-129, 1QBA:201-337, 1SCUB:1-238, 1TPT:336-440, 1UP1:7-92, 1UP1:99-182, 1VAOA:6-273, 1VAOA:274-560, 1YTBA:61-155, 2GLT:123-316, 2MNR:3-132, 2REB:269-328, 3RUBL:22-147, 4KBPA:121-432, 5RUBA:2-137, 2AAK, 1BVTA, 1AOP:81-145, 1AOP:346-425, 1AOP:149-345, 1O7BT, 1IQZA, 1CYO, 2TSSA:94-194

Appendix B

Results of the parameter optimization for the considered eight classification algorithms (see Table B.1).

Table B.1

Classification algorithm	Parameters	Considered values	Optimal setup for a given dataset and feature representation[a]	
Naïve Bayes (*NB*)	1. Supervised or unsupervised discretization	Superv, unsuperv	359+CV	1. Unsuperv
			359+AC	1. Unsuperv
			359+66	1. Superv
			1189+CV	1. Unsuperv
			1189+AC	1. Superv
			1189+66	1. Unsuperv
			25PDB+CV	1. Unsuperv
			25PDB+AC	1. Unsuperv
			25PDB+66	1. Superv
Radial basis function neural network (*RBF*)	1. Number of clusters per class	2, 3, 4, . . . , 10	359+CV	1. 8; 2. e^{-8}
	2. Ridge parameter for regression	$e^{-10}, e^{-9}, e^{-8}, e^{-7}, \ldots, e^{1} e^{2}, e^{3}$	359+AC	1. 8; 2. e^{-8}
			359+66	1. 8; 2. e^{-8}
			1189+CV	1. 3; 2. 100
	Regression performed always until convergence		1189+AC	1. 8; 2. e^{-8}
			1189+66	1. 3; 2. 100
			25PDB+CV	1. 2; 2. e^{-8}
			25PDB+AC	1. 2; 2. e^{-8}
			25PDB+66	1. 2; 2. e^{-8}

Table B.1. *Continued*

Classification algorithm	Parameters	Considered values	Optimal setup for a given dataset and feature representation[a]	
Instance based classifier (*IB1*)	None	None	N/A	
C4.5 (*C4.5*)	1. Pruning confidence	0.1, 0.2, 0.25, 0.4, 0.6	359+CV	1. 0.25; 2. 1
	2. Minimum number of instances per leaf	1, 2, 5, 10, 15	359+AC	1. 0.25; 2. 1
			359+66	1. 0.25; 2. 1
			1189+CV	1. 0.4; 2. 1
	Tree pruning always performed		1189+AC	1. 0.25; 2. 2
			1189+66	1. 0.25; 2. 10
			25PDB+CV	1. 0.2; 2. 5
			25PDB+AC	1. 0.2; 2. 10
			25PDB+66	1. 0.25, 2. 10
Random forest (*RF*)	1. Number of trees in the forest	1, 2, 5, 10, 20, 30, 40, 50	359+CV	1. 10
			359+AC	1. 30
			359+66	1. 20
			1189+CV	1. 40
			1189+AC	1. 20
			1189+66	1. 30
			25PDB+CV	1. 40
			25PDB+AC	1. 10
			25PDB+66	1. 30
Repeated incremental pruning to produce error reduction (*RIP*)	1. Number of optimization runs within a split	1, 2, 5, 10	359+CV	1. 2; 2. 2
			359+AC	1. 2; 2. 2
	2. Minimal weights of instances	1, 2, 5, 10, 20		
	Rule pruning always performed		359+66	1. 5; 2. 2
			1189+CV	1. 2; 2. 2
			1189+AC	1. 2; 2. 2
			1189+66	1. 5; 2. 2
			25PDB+CV	1. 5; 2. 20
			25PDB+AC	1. 2; 2. 2
			25PDB+66	1. 2; 2. 2
Support vector machine (*SVM*)	1. Complexity constant	1, 5, 10, 20, 50	359+CV	1. 10; 2. RBF; 3b. 1
	2. Kernel type	Polynomial, RBF	359+AC	1. 10; 2. RBF; 3b. 10
	3a. Exponent for the polynomial kernel	1, 2, 3	359+66	1. 10; 2. RBF; 3b. 1
	3b. Gamma for the RBF kernel	0.01, 0.1, 1, 10	1189+CV	1. 100; 2. RBF; 3b. 0.01
			1189+AC	1. 100; 2. RBF; 3b. 1
			1189+66	1. 1; 2. RBF; 3b. 0.1
			25PDB+CV	1. 20; 2. RBF; 3b. 0.1
			25PDB+AC	1. 20; 2. POLY; 3a. 1
			25PDB+66	1. 10; 2. POLY; 3a. 1
Logistic regression (*LR*)	1. Ridge parameter for the log-likelihood	$e^{-10}, e^{-9}, e^{-8}, e^{-7}, \ldots e^{1}, e^{2}, e^{3}$	359+CV	1. e^{-5}
	Regression performed always until convergence		359+AC	1. e^{-8}
			359+66	1. e^{-8}
			1189+CV	1. e^{-8}
			1189+AC	1. e^{-8}
			1189+66	1. e^{-8}
			25PDB+CV	1. e^{-8}
			25PDB+AC	1. 1
			25PDB+66	1. e^{-2}

[a]Datasets include 25PDB, 1189, and 359; feature representations include composition vector (CV), autocorrelation (AC), and 66 features (66); 1189+CV stands for 1189 dataset and composition vector representation.

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1996) 552–557.
[2] M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, Protein Eng. 11 (1998) 249–251.
[3] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[4] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Understanding the recognition of protein structural classes by amino acid composition, Proteins 29 (1997) 172–185.
[5] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of protein database for the investigation of sequence and structures, J. Mol. Biol. 247 (1995) 536–540.
[6] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 153–162.
[7] P. Klein, C. Delisi, Prediction of protein structural class from the amino-acid sequence, Biopolymers 25 (1986) 1659–1672.
[8] C.T. Zhang, K.C. Chou, An optimization approach to predicting protein structural class from amino-acid composition, Protein Sci. 1 (1992) 401–408.
[9] B.A. Metfessel, P.N. Saurugger, D.P. Connelly, S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, Protein Sci. 2 (1993) 1171–1182.
[10] K.C. Chou, C.T. Zhang, Predicting protein-folding types by distance functions that make allowances for amino-acid interactions, J. Biol. Chem. 269 (1994) 22014–22020.
[11] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein-folding class using global description of amino-acid sequence, Proc. Nat. Acad. Sci. 92 (1995) 8700–8704.
[12] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, S.H. Kim, Recognition of a protein fold in the context of the SCOP classification, Proteins 35 (1999) 401–407.
[13] Z.-X. Wang, Z. Yuan, How good is the prediction of protein structural class by the component-coupled method?, Proteins 38 (2000) 165–175.
[14] Y. Cai, Is it a paradox or misinterpretation?, Proteins 43 (2001) 336–338.
[15] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for prediction of protein domain structural class, J. Theor. Biol. 221 (2003) 115–120.
[16] L. Jin, W. Fang, H. Tang, Prediction of protein structural classes by a new measure of information discrepancy, Comput. Biol. Chem. 27 (2003) 373–380.
[17] K.C. Chou, Y.D. Cai, Prediction protein structural class by functional domain composition, Biochem. Biophys. Res. Commun. 321 (2004) 1007–1009.
[18] G. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins 44 (2001) 57–59.
[19] Z.-X. Wang, The prediction accuracy for protein structural class by the component-coupled methods is around 60%, Proteins 43 (2001) 339–340.
[20] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins 21 (1995) 319–344.
[21] W. Kabsch, C. Sander, Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[22] F. Eisenhaber, C. Frömmel, P. Argos, Prediction of secondary structural content of proteins from their amino acid composition alone, II the paradox with secondary structural class, Proteins 25 (1996) 169–179.
[23] F. Eisenhaber, et al., Prediction of secondary structural contents of proteins from their amino acid composition alone, I new analytic vector decomposition methods, Proteins 25 (2) (1996) 157–168.
[24] H.M. Berman, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[25] K.C. Chou, G.M. Maggiora, Domain structural class prediction, Protein Eng. 11 (1998) 523–538.
[26] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data, Nucleic Acid Res. 32 (2004) D226–D229.
[27] J. Grassmann, M. Reczko, S. Suhai, L. Edler, Protein fold class prediction—new methods of statistical classification, Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB), 1999, pp. 106–112.
[28] C.H. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics 17 (2001) 349–358.
[29] C. Leslie, E. Eskin, W. Stafford Noble, The spectrum kernel: a string kernel for SVM protein classification, Proceedings of the Pacific Symposium on Biocomputing, 2002, pp. 566–575.
[30] F. Markowetz, L. Edler, M. Vingron, Support vector machines for protein fold class prediction, Biometrical J. 45 (2003) 377–389.
[31] K.C. Chou, C.T. Zhang, A new approach to predicting protein folding types, J. Protein Chem. 12 (1993) 169–178.
[32] C.T. Zhang, K.C. Chou, G.M. Maggiora, Predicting protein structural classes from amino acid composition: application of fuzzy clustering, Protein Eng. 8 (1995) 425–435.
[33] W.S. Bu, Z.P. Feng, Z. Zhang, C.T. Zhang, Prediction of protein structural classes based on amino acid index, Eur. J. Biochem. 266 (1999) 1043–1049.
[34] R. Luo, Z. Feng, J. Liu, Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269 (2002) 4219–4225.
[35] K.C. Chou, W.M. Liu, G.M. Maggiora, C.T. Zhang, Prediction and classification of domain structural classes, Proteins 31 (1998) 97–103.
[36] B. Rost, C. Sander, Third generation prediction of secondary structure, in: D.M. Webster (Ed.), Protein Structure Prediction: Methods and Protocols, 2000, pp. 71–95.
[37] A.V. Filkenstein, O. Ptitsyn, Statistical analysis of the correlation among amino acid residues in helical, β-structural and non-regular regions of globular proteins, J. Mol. Biol. 62 (1971) 613–624.
[38] P.Y. Chou, U.D. Fasmad, Prediction of protein conformation, Biochemistry 13 (1974) 211–215.
[39] C.T. Zhang, et al., Prediction of helix/strand content of globular proteins based on their primary sequences, Protein Eng. 11:11 (1998) 971–979.
[40] Z.D. Zhang, Z.R. Sun, C.T. Zhang, A new approach to predict the helix/strand content of globular proteins, J. Theor. Biol. 208 (2001) 65–78.
[41] Z. Lin, X-M. Pan, Accurate prediction of protein secondary structural content, J. Protein Chem. 20 (3) (2001) 217–220.
[42] M.K. Ganapathiraju, et al., Characterization of protein secondary structure, IEEE Signal Process. Mag. (2004) 78–87.
[43] J. Ruan, K. Wang, J. Yang, L. Kurgan, K. Cios, Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences, Artif. Intell. Med. 35 (1–2) (2005) 19–35 (special issue Computational Intelligence Techniques in Bioinformatics).
[44] I.V. Grigoriev, S.H. Kim, Detection of protein fold similarity based on correlation of amino acid properties, Proc. Nat. Acad. Sci. 96 (1999) 14318–14323.
[45] G. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.
[46] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein database, Bioinformatics 17 (2001) 282–283.
[47] W. Li, L. Jaroszewsk, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (1) (2002) 77–82.

[48] U. Hobohm, C. Sander, Enlarged representative set of protein structures, Protein Sci. 3 (1994) 522.
[49] L.A. Kurgan, L. Homaeian, Prediction of secondary protein structure content from primary sequence alone—a feature selection based approach, Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2005), Leipzig, Germany, LNAI 4587, 2005, pp. 334–345.
[50] M. Oobatake, T. Ooi, An analysis of non-bonded energy of proteins, J. Theor. Biol. 67 (1997) 567–584.
[51] J. Cornette, et al., Hydrophobicity scales and computational techniques for detecting amphipathic structures in protein, J. Mol. Biol. 195 (1987) 659–685.
[52] S.M. Muskal, S.H. Kim, Predicting protein secondary structure content: a tandem neural network approach, J. Mol. Biol. 225 (1992) 713–727.
[53] H. Liu, R. Setiono, A probabilistic approach to feature selection—a filter solution, Proceedings of the 13th International Conference on Machine Learning, Italy, 1996, pp. 319–327.
[54] R. Kohavi, G. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[55] M.A. Hall, Correlation-based feature subset selection for machine learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.
[56] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.
[57] G.H. John, P. Langley P, Estimating continuous distributions in Bayesian classifiers, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, 1995, pp. 338–345.
[58] A. Saha, C.L. Wu, D.S. Tang, Approximation, approximation dimension reduction and nonconvex optimization using linear superpositions of gaussians, IEEE Trans. Comput. 42 (1993) 1222–1233.
[59] D. Aha, D. Kibler, Instance-based learning algorithms, Mach. Learn. 6 (1991) 37–66.
[60] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.
[61] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[62] W. Cohen, Fast effective rule induction, Proceeding of the 12th International Conference on Machine Learning, Lake Tahoe, CA, 1995, pp. 115–123.
[63] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (3) (2001) 637–649.
[64] S. le Cessie, J.C. van Houwelingen, Ridge estimators in logistic regression, Appl. Stat. 41 (1) (1992) 191–201.