行政院國家科學委員會專題研究計畫　成果報告

（子計畫三）：結構生物資訊核心

計畫類別： 整合型計畫
計畫編號： NSC91-3112-B-009-001-
執行期間： 91 年 05 月 01 日至 92 年 04 月 30 日
執行單位： 國立交通大學生物科技學系

計畫主持人： 黃鎮剛

報告類型： 完整報告
報告附件： 國外研究心得報告
出席國際會議研究心得報告及發表論文
處理方式： 本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中　華　民　國 92 年 8 月 17 日

91-3112-B-009-001-

Structural Bioinformatics Core

Fine-grained Protein Fold Assignment by Support Vector Machines using generalized n−peptide Coding Schemes and jury voting from multiple parameter sets

Jenn-Kang Hwang
Department of Biological Science and Technology
National Chiao Tung University, Hsin Chu 300, Taiwan

*Corresponding authors
TEL: +886-3-572-9287
FAX: +886-3-572-9288
E-mail: jkhwang@cc.nctu.edu.tw

Abstract

In the coarse-grained fold assignment of major protein classes, such as all-~, all- , ~+ , ~/ proteins, one can easily achieve high prediction accuracy from primary amino acid sequences. However, the fine-grained assignment of folds, such as those defined in the Structural Classification of Protein (SCOP) database, presents a challenge due to the larger amount of folds available. Recent study yielded reasonable prediction accuracy 56.0 % on an independent set of 27 most populated folds. In this report, we apply the support vector machine method (SVM), using a combination of protein descriptors based on the properties derived from the composition of *n* -peptide and jury voting, to the fine-grained fold prediction, and are able to achieve an overall prediction accuracy 69.6% on an same independent set, significantly higher than the previous results. On ten-fold cross validation, we obtained prediction accuracy 65.3%. Our results show that primary sequences contain SVM coupled with suitable global sequence coding schemes can significantly improve the fine-grained fold prediction and our approach should prove useful in structure prediction and modeling.

Keywords: support vector machines; fine-grained fold prediction; global sequence coding scheme; *n*-peptide

Introduction

Due to progress in experimental genomics, tremendous amounts of sequence data come out, and the increase in the number of putative protein sequences greatly exceeds that of three-dimensional structures of proteins. Hence, to extract three-dimensional structures from sequences becomes even more important nowadays. Roughly speaking, there are in general two kinds of approaches to structure prediction [1]. One is the *ab initio* method that predicts structures directly from the sequences based on the general physico-chemical principles[2-7]. The other one is the empirical method that relies on the empirical knowledge of proteins structures or sequences to assign the query sequences to the proper folds by either homology modeling, threading techniques or taxonometric approach[8-13]. Homology modeling identifies the possible template structures of the query sequences by aligning them with the sequences of known three-dimensional structures, based on the criteria that sequence identity higher than 25% usually have similar structures. Threading techniques find the possible folds by the sequence-structure alignment, without relying on the sequence homology between the query and target sequences. The taxonometric method, based on the assumption that the number of folds is limited, tries to predict protein structures in terms of the assignment of query sequences to the particular classification of protein folds. Proteins are said to have a common folding structure if their major secondary structures have similar arrangement and topological connections. The latter approach becomes increasingly important due to the fast growth of protein structures. Previous studies[11, 14-16] showed that, in the coarse-grain structural classification such as all- ~, all- , ~+ , ~/ and irregular folds [17] one can easily achieved 70% or better prediction accuracy from the amino acid composition. However, in order to obtain predict a high-resolution three-dimensional structure, one needs to be able to assign fine-grained folds for the query structures. The fine-grained assignment of folds, such as defined in the Structural Classification of Protein (SCOP) database, presents a challenge for structure prediction due to the larger number of folds. Recently, Ding and Dubchak (2001) applied the support vector machines (SVM) to the problem of fold assignment. They used six coding schemes[11, 12] to extract structural or physico-chemical properties from the primary sequences. They compressed 20 amino acids into three groups for the following attributes: the percent composition of amino acids, predicted secondary structure, normalized van der Waals volumes, hydrophobicity, polarity and polarizability. They then calculated three descriptors, i.e., "composition", "transition" and "distribution" for each attribute of these three groups of amino acids. Their approach yielded around 56.0% prediction accuracy for an independent set. Despite its seemingly lower prediction

accuracy than before, the prediction was made in the context of 27 fine-grained SCOP folds, about an order higher than the number of protein classes used in the earlier work. They achieved this by multi-class fold prediction system based on the jury votes from several parameter sets of structural or physico-chemical properties of the sequences described by three groups of amino acids. In this work, we use SVM coupled with more comprehensive protein descriptors based on n-peptide coding schemes and jury voting procedures, we can obtain a prediction accuracy significsntly higher than the previous study.

Methods

The SVM is a powerful classification method[18] that becomes popular in computational biology[13, 19, 20 21] and other areas. The original idea of SVM is to use a linear separating hyperplane to separate training data in two classes: Given training vectors $x_i,\ i=1,...,l$ and a vector y defined as: $y_i=1$ if x_i is in one class, and $y_i=-1$ if x_i is in the other class. The support vector technique tries to find the separating hyperplane $w^T x_i+b=0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane. This requirement is equivalent to the minimization of $\frac{1}{2}w^T w$ with respect to w and b under the constraint that $y_i\left(w^T x_i+b\right)\geq 1$. However, in practice, these data to be classified may not be linearly separable. To overcome this difficulty, SVM non-linearly transforms the original input space into a higher dimensional feature space by $\varphi(x)=(\varphi_1(x),\varphi_2(x),...)$ and tries to minimize $\frac{1}{2}w^T w+C\sum_{i=1}^{l}\xi_i$ with respect to w, b and ξ, under the constraint that $y_i\left[w^T\varphi(x_i)+b\right]\geq 1-\xi_i$ where $\xi_i\geq 0$. This procedure has the advantage of allowing training errors. It should be noted that only some of $x_i's$ are used to construct w and b and these data called support vectors.

Data sets and input coding schemes

We used the same data set as that of Ding and Dubchak (2001), which consist of 386 proteins of the most populated 27 SCOP folds in which the protein pairs have sequence identity below 35% for the aligned subsequences longer than 80 residues. These 27 proteins folds cover most major structural classes such as ~, , ~/ and ~+ [22], and have at least seven or more proteins in their classes. To successfully apply the machine learning techniques to the biological problems, one need to extract relevant input vectors from the biological data, i.e., in this case, the primary sequences.

In this work, our global sequence coding schemes cover the distribution of *n*-peptides for protein attributes. When *n* is one, it encodes the composition of amino acids, which has been useful discriminating the coarse-grained fold classes[14-16, 23]. When *n* is two, the input vector encodes the dipeptide composition, which has been successfully applied to predict *in vivo* stability of proteins[24]. We can extend *n* to three or more, but, in practice, it becomes impractical even in the case of $n=3$ (the size of the input vector becomes 8000). This can be overcome if we reduce the size of the input vectors by regrouping the amino acids into smaller number of classes according to their physico-chemical properties. In this work, we denote the coding schemes by X if all 20 amino acids are used, X' when the amino acids are classified as 4 groups: charged, polar, aromatic and nonpolar, and X'', if predicted secondary structures are used. We assign the symbol X the values of D, T, Q and P, denoting the distributions of dipeptides, 3-peptides and 4-peptides, respectively. Similar ideas making use of n-gram models have been successfully applied to protein family identification[25]. Since these parameters are built independently, one can apply machine learning techniques based on a single set of input vector or a combination of several sets. All the SVM calculations are performed using LIBSVM[26], a general library for support vector classification and regression. We use PREDATOR [27] to predict the secondary structure of the protein sequences.

Training and testing procedures

For SVM classifiers to perform a multi-class prediction, we followed two commonly used approaches[13]. The first approach is the "one-against-all" method where k SVM classifiers are constructed and the ith SVM is trained with proteins in the ith fold as positive, and all other proteins as negative. Each protein in the test set is tested against other proteins, and if tested positive, it will get a vote for the class. However, if tested negative, this protein will not get any vote for the class. The "one-against-all" method will give rise to the possibility of giving some proteins too few or even no votes for any fold. However, we can complement this method by the "one-against-one" method. Given F classes of proteins, we can construct $F(F-1)/2$ SVM classifiers and train with proteins from two different folds. Thus, in the current work, we constructed for 27 folds a total of $27(27-1)/2=351$ classifiers. In the "one-against-one" method, each protein in the test set will always get a vote for either one of the two folds. In the end, we used the jury voting to determine the final assignment of folds to each protein in the test set. Figure 1 shows the architecture of our SVM classifier. We use the standard Q_i percentage accuracy [13, 28, 29] for assessing the accuracy of protein fold identification $Q_i = c_i / n_i \times 100,$ where n_i is the number of test data in the i^{th} class and c_i the number correctly predicted. The overall Q is

given by $Q = \sum_i^F w_i Q_i$, where $w = n_i / N$.

We used two evaluation methods for the performance of the prediction system. First, we test the system against the independent set, which comprises 385 proteins of 27 folds from PDB-40D set [30] that have sequence identity below 40% within the testing set and below 35% compared with those of the training set. Secondly, we evaluate the classifiers by cross validation, which measure the prediction accuracy of them systematically by first excluding a few proteins during the training process and then testing the classifiers against these excluded proteins. In the ten-fold cross validation evaluation, each testing set comprises around 10% of the proteins. In addition to our parameter sets, we also used the following parameter sets of Ding & Dubchak [11, 12] - the attributes of amino acids (C), predicted secondary structure (S) and hydrophobicity (H).

Results

We compare the prediction accuracy of $n-$peptide coding schemes for the independent test set. Figure 2 gives the general trend of one-against-all prediction accuracies of isolate parameters sets: X, X' and X''. The parameter set M, The composition of 20 amino acids M, a useful coarse-grained fold discriminator, also gives the highest average prediction accuracy 59% for the 27 folds. The parameter D, the composition of dipeptides, gives much lower prediction accuracy. For the X' set, the composition of 4 classes of amino acids, the prediction accuracy displays the same monotonous decay when the length of the peptide fragments grows longer. It is interesting to note that M' gives much lower prediction accuracy than M, indicating the composition of 20 amino acids contains more useful information in discriminating protein folds than the compressed classes of amino acids. For the X'' set, the composition of predicted secondary structure, the prediction accuracy peaks at D' and then slowly flattens out. To obtain the best over-all prediction accuracy, we need a combination of parameters in both one-against-one and one-against-all classifiers. After some preliminary computations, we settled on the following parameter sets: M, D, T', Q', P' and T'' (using one-against-all classifiers), and C+S+H+D (using one-against-one classifier), from which the highest combined votes will determine the predicted folds. Here C, S and H are the percent composition of amino acids, predicted secondary structure and hydrophobicity, respectively. Table lists our results for the independent set. In the one-against-one method, all the parameter sets (M, D, T', Q', P' and T'') give average prediction accuracy greater than 40%. In the one-against-one method, the parameter set, M, the composition of 20 amino acids, gives the best prediction accuracy 59% in the context of one parameter set. Our

results are consistent with previous findings [14-16, 23] that M, the composition of amino acid, is a very good discriminator in the classification of the coarse-grained folds. However, we also find that M, as an isolate parameter set, is also very helpful in identifying the 27 fine-grained classes of fold. The parameter sets T', Q' and P' encode the distribution of tripeptide, 4-peptide and 5-peptide sequences defined by amino acids that are classified into four groups. T' performs best, Q' and P' give lower prediction accuracy. Amongst various combinations of parameter sets for the one-against-one method, we found that the C+S+H+D set gave the best prediction accuracy 63.1%, which is higher than the one-against-all method using M set by around 4%. The jury column in Table gives the final prediction accuracy 69.6% for each fold by the votes from the parameter sets, a 6.5% improvement on the one-against-one method, showing the effectiveness of the jury voting procedures[13]. In the break-down analysis, our approach gives excellent prediction accuracy (>80%) for the folds: ~$_1$ (globin-like ~-proteins), ~$_2$ (cytochrome *c* folds), ~$_5$ (4-helical cytokines), $_1$ (the immunoglobulin-like -sandwich fold), $_7$ (the trefoil fold), (~/)$_1$ (the TIM-barrel) and (~+)$_3$ (small proteins like inhibitors, toxins and lectins). On the other end of the prediction spectrum, our method gives poor results (accuracy < 50%) for folds like $_2$ (cupredoxins), $_6$ (OB-fold), (~/)$_3$ (flavodoxin-like), (~/)$_9$ (periplasmic binding protein-like) and (~+)$_1$ (-grasp or ubiquitin-like). These poor results reflect the consistent failures of recognizing the correct folds by almost all the parameter sets. Figure 3 compares the prediction accuracy for each fold (in white) of our approach with that of Ding and Dubchak 2001 (in black). Our final prediction accuracy 69.6% is a significant improvement on their result 56.0% by 12.5%. Our method gives better prediction for 24 folds, most noticeably the following folds: ~$_3$, $_3$, $_4$, $_7$, $_8$ and (~+)$_1$, where improvements are more than 50%. Both approaches give poor results for $_2$ and (~/)$_9$. Figure 4 shows the 10-fold cross validation of the PDB-40D set, which was done by randomly picking 10% of the protein as the test set during the training process and then tested the classifiers against the test sets. The cross validation gives quite consistent results as that of the independent set. The final overall average prediction accuracy for the cross validation is 65.3%, which also significantly improve the previous result 45.4%.

Discussion

The previous works showed that in the coarse-grained fold assignment of major protein classes, such as all-~, all- , ~+ , ~/ proteins, one could easily achieve high prediction accuracy (70%~80%) from amino acid composition. Ding and Dubchak (2001) showed that, in the fine-grained fold prediction, SVM combined with

jury voting from multiple parameter sets yielded prediction accuracy significantly higher than that of any single parameter set – they obtained 56% prediction accuracy on an independent test set and 45.4% on cross validation. We showed in this study that the amino acid composition M alone yield 59% prediction accuracy, which, though better than the current result, is still not yet practical in realistic applications. Using protein descriptors based on the properties derived from the composition of n -peptide and jury voting from a combination of parameter sets, we are able to achieve a prediction accuracy 69.6% on an independent set, an order of magnitude higher than the current results, and 65.3% on 10-fold cross-validation. The prediction accuracy is approaching to that for the coarse-grained fold classes. Our results show that SVM, novel global sequence coding schemes and proper combinations of input parameter sets should become an increasingly practical tool in structure modeling.

References

1. Baker D, Sali A. Protein structure prediction and structural genomics Science 2001;294:93-96
2. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome Proc Natl Acad Sci U.S.A. 2002;99:5993-5998
3. Xia Y, Levitt M, Huang ES, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J. Mol. Biol. 2000;300:171-185
4. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions J. Mol. Biol. 1999;290:267-281
5. Zhang CT, Hou J, Kim SH. Fold prediction of helical proteins using torsion angle dynamics and predicted restraints Proc Natl Acad Sci U.S.A. 2002;99:3581-3585
6. Srinivasan R, Rose GD. Ab initio prediction of protein structure using LINUS Proteins 2002;47:489-495
7. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J. Mol. Biol. 2001;306:1191-1199
8. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347-352
9. Russell A, Torda AE. Protein sequence threading: averaging over structures Proteins 2002;47:496-505
10. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement.

Proteins 2001;2001:133-149
11. Dubchak I, Muchnik I, Holbrook SR, Kim S-H. Prediction of protein folding class using global description of amino acid sequence Proc Natl Acad Sci U.S.A. 1995;92:8700-8704
12. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim S-H. Recognition of a protein fold in the context of the structural classification of proteins (SCOP) Proteins 1999;35:401-407
13. Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks Bioinformatics 2001;17:349-358
14. Chou KC, Liu WM, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes Proteins 1998;31:97-103
15. Chou KC, Zhang CT. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 1995;30:275-349
16. Dubchak I, Holbrook SR, Kim S-H. Prediction of protein folding class from amino acid composition. Proteins 1993;16:79-91
17. Levitt M, Chothia C. Structural patterns in globular proteins Nature 1976;261:552-558
18. Vapnik V The Nature of Statistical Learning Theory. New York: Springer; 1995.
19. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Ares M, Jr, Hasussler D. Knowledge-based analysis of microarray gene expression data by using Support Vector Machine Proc Natl Acad Sci U.S.A. 2000;97:
20. Jaakkola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies ISMB 1999;149-158
21. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach J. Mol. Biol. 2001;308:397-407
22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995;247:536-540
23. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition J. Biochem. 1986;99:152-162
24. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its peptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence Protein Engineering 1990;4:155-161
25. Wu CH, Zhao S, Chen HL, J. LC, McLarty J. Motif identification neural design for rapid and sensitive protein family search Comput Appl Biosci 1996;12:109-118
26.
27. Frishman D, Argos P. Knowledge-based secondary structure assignment Proteins:

Struct. Funct. Genet. 1995;23:566-579
28. Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412-424
29. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy J. Mol. Biol. 1993;232:584-599
30. Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of protein database Nucleic Acids Res. 2000;28:257-259

Table I Prediction accuracy Q_i (in %)for protein fold for the independent test set

	One-against-all					One-against-one	Jury
Folds[1]	M	D	T'	Q'	T"	C+S+H+D	Final
~1	83.3	83.3	66.7	100.0	66.7	83.3	83.3
~2	88.8	22.2	55.5	22.2	44.4	100.0	100.0
~3	55.0	30.0	55.0	40.0	40.0	40.0	70.0
~4	62.5	37.5	37.5	37.5	62.5	62.5	75.0
~5	100.0	66.7	55.5	44.4	66.7	100.0	100.0
~6	55.6	44.4	33.3	33.3	11.1	44.4	55.6
1	63.6	43.2	50.0	47.7	75.0	84.1	90.9
2	50.0	16.7	16.7	25.0	16.7	16.7	16.7
3	61.5	46.2	61.5	61.5	53.8	61.5	76.9
4	33.3	33.3	66.7	66.7	50.0	50.0	66.7
5	75.0	25.0	37.5	37.5	37.5	50.0	50.0
6	31.6	26.3	31.6	21.1	47.4	31.6	47.7
7	75.0	50.0	50.0	50.0	75.0	75.0	100.0
8	50.0	50.0	50.0	50.0	25.0	25.0	50.0
9	71.4	28.6	71.4	42.9	28.6	57.1	57.1
(~/)1	83.3	66.7	60.4	62.5	45.8	87.5	93.8
(~/)2	50.0	33.3	25.0	33.3	33.3	50.0	66.7
(~/)3	30.8	7.7	15.4	30.8	15.4	53.8	38.5
(~/)4	40.7	37.0	33.3	37.0	25.9	55.5	55.6
(~/)5	50.0	33.3	41.7	33.3	33.3	50.0	50.0
(~/)6	37.5	37.5	50.0	37.5	50.0	37.5	50.0
(~/)7	42.9	42.9	42.9	42.9	42.9	57.1	57.1
(~/)8	71.4	71.4	57.1	71.4	28.6	71.4	71.4
(~/)9	25.0	25.0	50.0	50.0	25.0	25.0	25.0
(~+)1	37.5	25.0	25.0	25.0	37.5	37.5	37.5
(~+)2	22.2	22.2	25.9	18.5	25.9	48.1	51.9
(~+)3	100.0	88.9	85.2	81.5	74.1	96.3	100.0

Avg	59.0	43.1	47.0	44.9	44.9	63.1	69.6

[1]The fold notations are: $\sim_{1\sim6}$ are all-$\sim$ proteins including globin-like, cytochrome C, DNA-binding 3-helical bundle, 4-helical up-and-down-bundle and 4-helical cytokines, EF-hand, respectively. $_{1\sim9}$ are all- beta proteins including immunoglobulin-like −sandwich, cupredoxins, viral coat and capsid proteins, ConA-like lectins/glucanases, SH3-like barrel, OB-fold, −trefoil, trypsin-like serine proteases and lipocalins. $(\sim/\ )_{1\sim9}$ are $\sim/$ proteins : Tim-barrel, FAD/NAD-binding motif, flavodoxin-like, NAD(P)-binding Rossmann-fold, P-loop containing nucleotide, thioredoxin-like, Ribonuclease H-like motif, hydrolases and periplasmic binding protein-like. $(\sim+\ )_{1\sim3}$ are $\sim+$ proteins including −Grasp, ferredoxin-like and small inhibitors, toxins or lectins.

Figure captions

Figure 1 The architecture of our SVM classifiers to predict the folds. The symbols *X, Y, Z,* … designate the parameter sets used in the one-against-all" classifiers, and the symbols *x, y, z,* … the parameter sets used in the one-against-all" classifiers. Each classifier casts one jury vote and the fold that gets the most votes is the predicted fold for the query sequence.

Figure 2 Comparison of the one-gainst-all prediction accuracies of X, X' and X'' parameter sets. The symbol M, D, T, Q and F represent n-peptide fragments with $n = 1 \sim 5$, respectively.

Figure 3 Comparison of the prediction accuracy $Q_i(\%)$ of this work (in white) with that of Ding and Dubchak 2001(in black) for the 27 folds in the independent test.

Figure 4 Comparison of the prediction accuracy $Q_i(\%)$ of this work (in white) with that of Ding and Dubchak 2001(in black) for the 27 folds in the 10-fold cross validation.

Figure 1

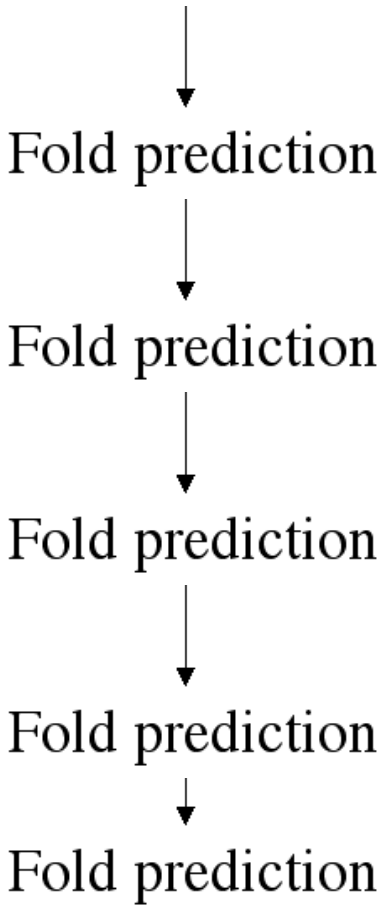

Fold prediction
Fold prediction
Fold prediction
Fold prediction
Fold prediction

Figure 2

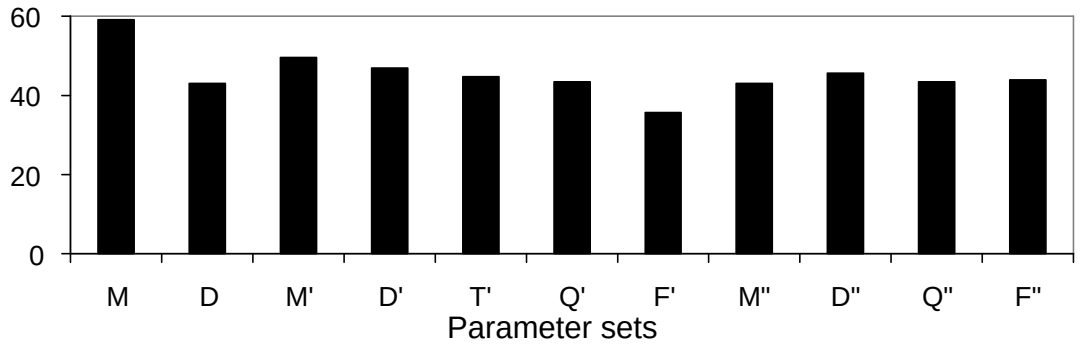


Figure 3

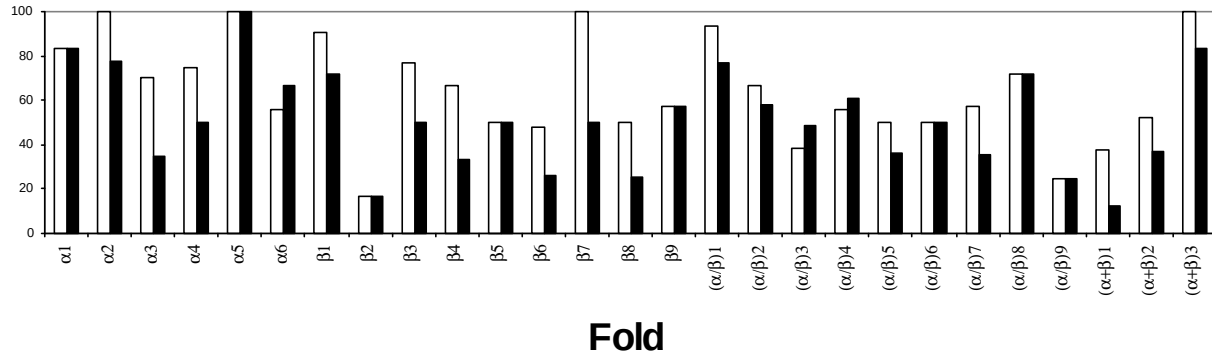
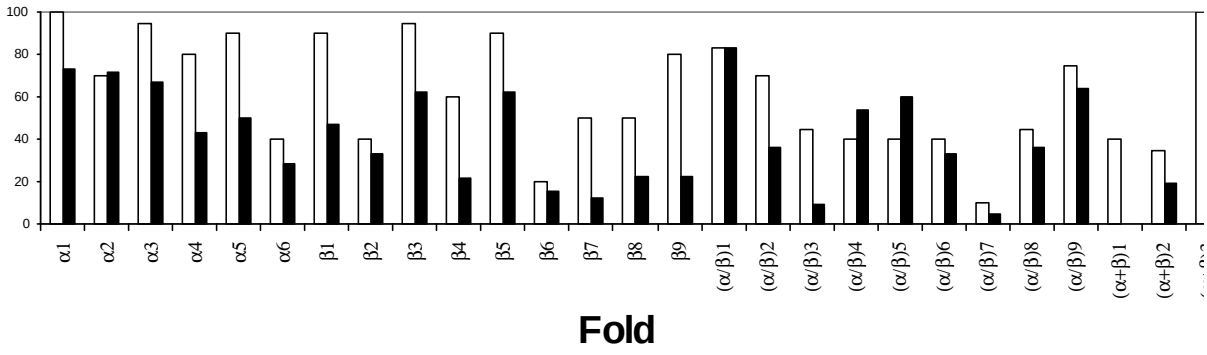

100
80
60
40
20
0
α1
α2
α3
α4
α5
α6
β1
β2
β3
β4
β5
β6
β7
β8
β9
(α/β)1
(α/β)2
(α/β)3
(α/β)4
(α/β)5
(α/β)6
(α/β)7
(α/β)8
(α/β)9
(α+β)1
(α+β)2
(α+β)3
Fold

Figure 4


100
80
60
40
20
0
α1
α2
α3
α4
α5
α6
β1
β2
β3
β4
β5
β6
β7
β8
β9
(α/β)1
(α/β)2
(α/β)3
(α/β)4
(α/β)5
(α/β)6
(α/β)7
(α/β)8
(α/β)9
(α+β)1
(α+β)2
Fold