DIMACS Technical Report 95-6
April 1995

Relation between Protein Structure, Sequence Homology and Composition of Amino Acids

by

Eddy N. Mayoraz[1]
RUTCOR–Rutgers University's Center for Operations Research,
P.O. Box 5062, New Brunswick, NJ 08903-5062

Inna Dubchak[2]
Department of Chemistry, University of California at Berkeley,
Berkeley, CA 94720

Ilya Muchnik[3]
RUTCOR–Rutgers University's Center for Operations Research,
P.O. Box 5062, New Brunswick, NJ 08903-5062

[1]First author gratefully acknowledges the support of RUTCOR and DIMACS.
[2]The second author is working at the Structural Biology Division of Lawrence Berkeley Laboratory (Division Director Prof. Sung-Hou Kim), supported by US Department of Energy.
[3]The third author gratefully acknowledges the support of DIMACS (grant NSF-CCR-91-1999) and RUTCOR (grant F49620-93-1-0041).

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

A method of quantitative comparison of two classifications rules applied to protein folding problem is presented. Classification of proteins based on sequence homology and based on amino acid composition were compared and analyzed according to this approach. The coefficient of correlation between these classification methods and the procedure of estimation of robustness of the coefficient are discussed.

1 Introduction

One of the most powerful methods of protein structure prediction is the model building by homology (Hilbert *et al.* 1993). Chothia and Lesk (Chothia & Lesk 1986) suggested that if two sequences can be aligned with 50% or greater residue identity they have a similar fold. This threshold of 50% is usually used as a "safe definition of sequence homology" (Pascarella & Argos 1992) and in conventional opinion grants a reasonable confidence that a protein sequence has chain conformation of the template excluding less conserved regions. Note that in biology, *homology* implies an evolutionary relationship which is not measurable, but in this paper, according to Pascarella & Argos 1992, we use this term to denote some measure of sequence similarity.

But it was shown that structure information can be transferred to homologous proteins only when sequence similarity is high and aligned fragments are long (Sander & Schneider 1991). It is known that homologous proteins can have completely different 3D structures. For example, ras p21 protein and elongation factor are identical in the topology of the chain fold and similar in overall structure, yet the two proteins are dissimilar in sequence with less than 20% identical residues (Sander & Schneider 1991). Opposite example (Kabsch & Sander 1984): octapeptides from subtilisin (2SBT) and immunoglobulin (3FAB) are dissimilar in structure yet 75% identical in sequences. The main obstacle to development of strict criteria for calculation of homology threshold is the limited size of empirical database.

Our study is based on the classification scheme 3D_ALI of Pascarella and Argos (Pascarella & Argos 1992) that merges protein structures and sequence information. It classifies the majority of the known X-ray three-dimensional (3D) structures (254 proteins and protein domains) into 83 folding classes, 38 of them having two or more representatives, and the other 45 containing only a single protein example. This grouping is based on a structural superposition among protein structures with a similar main-chain fold, either performed by the authors or collected from the literature.

Sequences from protein primary structure data bank (SWISSPROT) are associated with all 254 3D structures providing that they have not less than 50% sequence homology and at least 50% of 3D structure residues were alignable (Sander & Schneider 1991). Each of them is labeled by the number of one of 83 classes that includes a protein with maximum homology. This labeling can be considered as a classification procedure based on sequence homology.

Another simple measure of protein similarity often used is the distance function in the 20-dimensional vector space of amino acid composition. Several groups (Nakashima *et al.* 1986; Chou 1989; Dubchak *et al.* 1993) have shown that the amino acid composition of a protein provides significant correlation with its structural class, and a number of studies were devoted to a quantitative description of this relationship. In all these studies, a protein is characterized as a vector of 20-dimensional space where each of its component is defined by the composition of one of 20 amino acids. Recognition schemes based on percent composition are fairly effective for simple classifications where proteins are described in terms of the following structural classes: all α, all β, $\alpha + \beta$, α/β, and irregular. It is obvious that the difficulty of recognition grows rapidly with the number of classes. Even in the distinction between $\alpha + \beta$ and α/β classes, serious problems exist because parameter vectors of these types of structure are located too close in the parameter space. That is why amino acid composition is considered a crude representation of a protein sequence and not a tool for protein structure prediction in a context where more than four structural classes are to be distinguished.

In this study, we investigate the correlation between two classifications of protein sequences with unknown structure, which are obtained by the same simplest nearest neighbor association procedure (Duda & Hart 1973). These two classifications differ only in the measure of similarity of proteins: one is based on homology, while the other uses a distance in the space of amino acid composition. We define and calculate correlation coefficient between them and propose a technique to estimate the robustness of this coefficient of correlation. High correlation between these two classifications would show a hidden power of amino acid composition for folding class prediction and at the same time would cast some doubt on the application of homology to this prediction without a careful examination.

The remaining part of this paper is divided into 3 sections. Section 2 describes the methods and data used in this study. The results of our analysis are reported in section 3. The last section contains a brief discussion.

2 Material and methods

This section describes a method for the evaluation of correlations between two measures of protein sequence similarity as well as a technique to measure their robustness, by experiments based on two sets of data.

2.1 Material

As it was mentioned in the introduction, our first dataset, from now on referred to as *training set*, consists of 254 protein sequences with known 3D-structure. For each of the protein sequences in the training set we have:

- complete sequence of amino acids;
- a label in $\{1, \ldots, 83\}$ that denotes the class representing its 3D structure.

The second dataset, used as *testing set*, contains 2338 protein sequences. This set, disjoint from the training set, is the subset of all the sequences of the SWISSPROT database, whose homology with at least one of the 254 sequences of the training set is greater than or equal to 50%. For each of these 2338 sequences, the following items are available:

- complete sequence of amino acids;
- a pointer to the sequence in the training set which has the largest homology with this sequence;

- the value of the homology with this most similar sequence in the training set.

Homology, denoted d_h in this work, is a real number between 0 and 1 showing a proportion of homologous residues in a sequence. Besides homology, we will study simpler measures of similarity between sequences, based only on the rate of each amino acid in the sequence, and independent of the relative position of these amino acids. The *composition* of a protein sequence is defined as a 20-dimensional vector of coefficients in $[0, 1]$ indicating the rate (number of instances of the acid divided by the total length of the sequence) of each amino acid in the sequence. Very natural distance measures on the *composition space* $[0, 1]^{20}$ are provided by the norms L_p and will be denoted d_p:

$$d_p(\boldsymbol{x}, \boldsymbol{y}) \;=\; (\sum_{i=1}^{20} |x_i - y_i|^p)^{\frac{1}{p}}$$

In this study, we will retain only d_1 and $d_{\frac{1}{2}}$, since the former is easy to interpret, and the latter provides the highest correlation with d_h, as we observed empirically.

2.2 Scheme of analysis

In a first stage, 20-dimensional vectors of composition are constructed for each sequence of training and testing sets. For any distance measure d, the *closest neighbor* (according to d) in training set is identified for all sequences in the testing set. This operation induces a new classification of the testing set, where each sequence is associated with the class of its closest neighbor. Note that this procedure, executed here for d_1 and $d_{\frac{1}{2}}$, is identical to that performed by Pascarella and Argos (Pascarella & Argos 1992) for d_h.

Next step is the comparison of two different classifications based on d_h, d_1 and $d_{\frac{1}{2}}$. These comparisons are carried out by the computation of the *coefficient of agreement* between two classifications, *i.e.* the number of times when two classifications agree. More detailed comparisons containing a table indicating the number of times two classifications agree, versus the intensity of each of distance measures, was also performed.

The coefficients of agreement between different classifications obtained this way are random variables depending on the testing set. In order to make any valid conclusion, it is essential to estimate the robustness of these coefficients. To do so, we suggest to reduce the amount of information available in training and testing sets from whole space of amino acid composition to more and more simple spaces of this space. To be valid, each of these simplifications of the composition space must maintain the *consistency* of the training set, *i.e.* any two sequences of the training set belonging to two different classes must be different in the simpler space.

At every step of this simplification process, new distances are defined in the simpler space, the process of classifying by a new distance measure is executed, and the coefficient of agreement between new classification and the one induced by d_h is calculated. The ratio between each of these new coefficients of agreement and the original one (agreement between d_h and d_1 or $d_{\frac{1}{2}}$) is interpreted as a measure of the robustness of the original coefficient of agreement. Indeed, if a strong simplification of the composition data deteriorates only partially the agreement between the classification induced by homology and the one based on composition, we can conclude that our evaluation of the original agreement (with the complete information on composition) is robust.

It should be mentioned that the simpler spaces used in our approach are binary, and the distance used is the Hamming distance. Thus, it happens very often that the nearest neighbor is not unique, and even that several of them belong to different classes. If this is a case, the classification follows a simple *voting rule* associating the class with the largest number of nearest neighbors, and in case of ties, the new sequence is put into a dummy class and it will always be considered as a mistake of the classification.

The method used to reduce the information in composition space, maintaining a consistency of training set, is described in next section.

2.3 Consistent reduction of information

The binarization technique described in this section has been designed by E. Boros, V. Gurvich, P.L. Hammer, T. Ibaraki and A. Kogan in August 1993, for the extraction of minimal consistent information of a dataset, and it is part of a newly developed methodology called *Logical Analysis* in the field of *Machine Learning* (Crama, Hammer & Ibaraki 1988). The purpose of Logical Analysis is to treat data from any machine learning problem, using logical tools, and among other uses, it aims at the extraction of simple logical patterns able to explain the repartition of the samples of a training set between their various classes. Since the method essentially deals with logical variables, the first stage of the whole analysis is a binarization of given data.

In many practical problems, a large set of attributes is available for each dataset, some of them play a crucial role in the studied classification, while others are completely irrelevant. Therefore, the binarization described here tends to extract the minimal amount of information, while keeping consistency of the training set.

For easy interpretation of the encoding, each binary attribute corresponds to one particular threshold and this attribute is true for a particular data when the corresponding original attribute has a value equal or greater than this threshold. The binarization problem consists in finding a small number of such thresholds, so that the resulting encoding maintains the consistence of the training set.

More formally, let say that there are n original attributes consisting all of continuous values, and the training set $S \in I\!R^n$ is partitioned into c classes $S = S^1 \cup S^2 \cup \ldots \cup S^c$. An encoding e of this space $I\!R^n$ will be consistent with S if and only if it fulfills the following property:

$$\forall k, l \in \{1, \ldots, c\}, k \neq l \ \ \forall \boldsymbol{a} \in S^k \ \ \forall \boldsymbol{b} \in S^l \ \ e(\boldsymbol{a}) \neq e(\boldsymbol{b}). \tag{1}$$

If the binary encoding $e : I\!R^n \rightarrow \{0,1\}^m$ is obtained by m threshold values placed on some of the n continuous variables, then $e(\boldsymbol{a}) \neq e(\boldsymbol{b})$ if and only if there is one threshold $t \in I\!R$ along one variable $i \in \{1, \ldots, n\}$ such that

$$\text{either} \ \ a_i \geq t \text{ and } b_i < t \quad \text{or} \ \ a_i < t \text{ and } b_i \geq t. \tag{2}$$

Therefore, if $\{v_1, \ldots, v_{k_i}\}$ is the set of values that a continuous attribute $i \in \{1, \ldots, n\}$ takes over the training set S, the only interesting thresholds along attribute i are in the ranges v_j and v_{j+1} (let say $= v_{j+1}$ w.l.o.g.), such that there are two data from two different classes of the training set, one for which attribute i takes the value v_j, and the other for which it takes the value v_{j+1}. This simple rule, applied independently for each of the n original attributes, provides a finite list $T \subset I\!\!R \times \{1, \ldots, n\}$ of thresholds, candidates for the binarization.

The problem of extracting from this usually large list T, a small subset such that the resulting encoding fulfill condition (1), can be stated as an integer linear program in which a variable $x_{ti} \in \{0, 1\}$ is associated to each threshold $(t, i) \in T$:

$$
\begin{array}{lll}
\min & \sum_{(t,i)\in T} x_{ti} & \\
\text{s.t.} & \sum_{(t,i)\in T} s_{ti}^{ab} x_{ti} \geq 1 & \forall \boldsymbol{a} \in S^k, \boldsymbol{b} \in S^l,\ k \neq l \\
 & x_{ti} \in \{0,1\} & \forall (t,i) \in T
\end{array}
\tag{3}
$$

where $s_{ti}^{ab} = 1$ if $t, i, \boldsymbol{a}$ and $\boldsymbol{b}$ satisfy relation (2), and $s_{ti}^{ab} = 0$ otherwise. A subset of candidate thresholds corresponds to each binary vector $\boldsymbol{x} \in \{0, 1\}^{|T|}$, and clearly, such a subset fulfill relation (1) if and only if all the constraints of the linear program (3) are satisfied.

A simple interpretation of this information reduction procedure by binarization is the following. Some hyperplanes (thresholds) are placed in the original continuous space (here $I\!\!R^{20}$). Each of these hyperplanes is normal to a vector of the basis of the continuous space. These hyperplanes are chosen in such a way that each of the hyperboxes they delimit contains only sequences belonging to the same class. The smallest number of hyperplanes fulfilling this property is desired.

The integer linear program (3), known as the *set covering problem*, has been widely studied. It is well known to be hard to solve exactly (NP-Complete). However, a large number of heuristics are available in the literature, and provide solutions close to minimal for most of the real life problems (see for example Beasley 1990 for a good algorithm and for a survey). Moreover, in our application it is not essential to get the minimal subset of thresholds fulfilling (1), and a reasonably small subset is sufficient. It turned out that the simplest greedy heuristic described in figure 1 is giving some satisfactory results.

```
init:       x = (0,...,0), A is the matrix of constraints in (3)
main loop: while (A contains at least one row) loop
                j: index of the column of A with the more 1s
                x_j = 1
                suppress from A every row i with a_ij = 1
           end loop
postoptimization (optional):
           find the smallest vector ≤ x still solution of (3)
```

Figure 1: Simplest greedy algorithm for the set covering problem.

In our application $|S| = 254$ and the number of classes is $c = 83$. By applying the simple rule described above, the number of candidate thresholds is $|T| = 1608$ and the set covering problem has originally $31'058$ constraints. However, the size of the solution produced by the greedy algorithm was 12. This means in particular that the matrix of constraints has a high density making the set covering problem easy, and that the size of 12 is probably very close to the optimal.

Since a reduction from the continuous space $\mathbb{R}^{20}$ to the discrete space $\{0,1\}^{12}$ is substantial, we are going to explore intermediate reductions, obtained by increasing the right-hand-side of the constraints in (3), from 1 to 2, 3, up to 10. Note that in a solution of (3) with that right-hand-side set to k, every pairs of sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ from different classes can be distinguished by at least k different thresholds of the solution. The increase observed on the size of the solution was almost linear: around five additional thresholds were necessary for each next value of k.

3 Results

It is known that in SWISSPROT database, *i.e.* among $\approx 35'000$ proteins, nearly 90% have a homology less than 50% with any of proteins with known structure, and no conventional technique is known to predict their 3D-structure from their sequence.

In this paper we analyze only $\approx 7\%$ (2338 exactly) of all known protein sequences having relatively high homology to classified 3D-structures.

Table 1(a) (resp. (b)) shows the distribution of these 2338 sequences according to their distances to their nearest neighbor with respect to homology d_h along the rows and with respect to d_1 along the columns in (a) (resp. $d_{\frac{1}{2}}$ in (b)). The whole field of correlations between these two distributions is displayed in each of these two 10 by 10 tables, and high correlations can be observed at least in cases of high similarity in protein sequences.
On the right-hand-side of each of these tables, a 2 by 2 table reports this same field of correlations where the bipartitions are based on threshold values (0.85 for d_h, 22 for d_1, and 17×100 for $d_{\frac{1}{2}}$) chosen to highlight the fact that most of the data lies on the diagonal of these 2 by 2 tables.

The main result, presented in table 2, shows the percentage of identical classifications between the procedure based on homology (using d_h) and the one based on composition (using d_1 or $d_{\frac{1}{2}}$). These percentages are detailed for each of 69 non empty classes among 83 (a classification based on homology associates none of the 2338 with some 14 classes). The total percentages are then reported, and finally, the percentages of identical classification are computed for three groups of classes: in each class of the first group, at least 31 sequences are associated according to the homology classification; each class of the second group contains between 6 and 30 sequences; and classes of the third group contain not more than 5 sequences, always according to the classification based on homology. Throughout the whole table, each of the values are given once counting any sequences, once counting only the sequences with homology to their nearest neighbor between 0.5 and 0.7, and once counting only the sequences

(a) $100 d_h \in$	$d_1 \geq 55$	≥ 50	≥ 45	≥ 40	≥ 35	≥ 30	≥ 25	≥ 20	≥ 15	<15	all	>22	≤ 22
50–54	29	18	33	34	36	45	34	21	11	10	271		
55–59	19	12	25	44	54	54	50	20	13	3	294		
60–64	21	14	21	22	42	35	70	12	8	3	248		
65–69	16	15	12	26	35	63	60	27	19	8	281	1462	236
70–74	12	16	22	31	36	27	30	18	10	6	208		
75–79	10	12	18	15	23	23	28	26	9	2	166		
80–84	7	28	19	34	22	15	41	46	9	9	230		
85–89	4	5	14	22	10	15	24	57	33	54	238		
90–94	1	9	13	20	7	14	24	62	45	37	232	213	427
95–99	1	0	2	2	0	0	5	22	33	105	170		
all	120	129	179	250	265	291	366	311	190	237	2338		

(b) $100 d_h \in$	$\frac{d_{1/2}}{100} \geq 45$	≥ 40	≥ 35	≥ 30	≥ 25	≥ 20	≥ 15	≥ 10	≥ 5	<5	all	>17	≤ 17
50–54	98	37	54	29	17	9	12	5	6	4	271		
55–59	70	40	70	45	32	17	14	3	1	2	294		
60–64	53	23	43	47	39	24	11	5	2	1	248		
65–69	46	26	41	57	58	34	13	6	0	0	281	1542	156
70–74	47	11	25	41	33	26	14	8	1	2	208		
75–79	24	8	10	17	32	38	23	10	1	3	166		
80–84	28	6	9	16	23	40	49	37	19	3	230		
85–89	15	1	2	7	10	20	59	58	52	14	238		
90–94	23	3	1	4	7	9	23	53	85	24	232	150	490
95–99	3	0	1	0	2	1	5	5	38	115	170		
all	407	155	256	263	253	218	223	190	205	168	2338		

Table 1: Distribution of the distances to the closest neighbors according to d_1 and $d_{\frac{1}{2}}$ versus the distances to the closest neighbors according to d_h.

with homology to their nearest neighbor greater than 0.7.

The results are almost similar between the classification based on d_1 and the one based on $d_{\frac{1}{2}}$. The first one coincides 86% of the time with the classification based on homology, while the second one coincides for 87% of the sequences. We carried out similar experiments for many other distances d_p. For the Euclidian norm d_2, the global percentage of identical classification is 83%, and it drops down to 71% for the distance d_∞. Among 6 different values of $p < 1$, $p = 0.51$ gave the highest correlation (87.37%) with homology. The fact that distances d_p, $p < 1$ give a better result has an easy explanation. With a distance d_1, the gaps between any of the coefficients of two vectors are considered with the same strength, while in d_p, $p < 1$, more importance is accorded to the small gaps than to the big ones. The higher similarity with homology obtained for $p < 1$ reflects the fact that in the homology procedure, a higher attention is accorded to the amino acids whose cardinalities are about the same in the two sequences.

The coefficient of similarity of classification rules observed in table 2 is a statistical measure depending on the dataset, and thus, its stability depends on the number of samples. We found interesting to observe these coefficients along a partition of the classes according to their cardinalities. As expected, it turned out that the coefficient of similarity is much smaller for the small classes (77% for the classes with at most 5 elements, for $d_{\frac{1}{2}}$), than for the bigger classes (at least 87%). Moreover, for the 20 largest classes (more than 30 elements), some singularities should be observed: the coefficient of similarity is incredibly low for one class (13% for class 35); intermediate for three classes (from 69% to 79% for classes 13, 29 and 38); above 90% for the 16 other big classes; and even 100% for few of

class	# of proteins			% $d_h = d_1$			% $d_h = d_{\frac{1}{2}}$		
label	total	$d_h \leq 0.7$	> 0.7	total	$d_h \leq 0.7$	> 0.7	total	$d_h \leq 0.7$	> 0.7
1	1	0	1	100		100	100		100
2	30	20	10	90	85	100	90	85	100
3	14	7	7	93	86	100	93	86	100
4	4	1	3	100	100	100	100	100	100
5	14	9	5	100	100	100	100	100	100
6	70	42	28	90	83	100	94	90	100
7	84	24	60	87	63	97	94	83	98
8	98	74	24	94	92	100	93	91	100
9	2	1	1	100	100	100	100	100	100
10	7	3	4	86	67	100	86	67	100
11	2	1	1	50	0	100	50	0	100
12	3	2	1	67	50	100	67	50	100
13	69	34	35	62	29	94	70	44	94
14	41	34	7	85	82	100	98	97	100
15	534	182	352	90	70	100	92	75	100
16	280	221	59	95	93	100	96	95	100
17	12	6	6	83	67	100	75	67	83
18	141	15	126	90	27	98	95	60	99
19	39	14	25	87	71	96	97	93	100
20	76	50	26	92	94	88	91	92	88
21	5	5	0	40	40		40	40	
22	71	60	11	99	98	100	90	88	100
23	1	1	0	0	0		0	0	
24	36	19	17	89	79	100	100	100	100
25	10	7	3	90	86	100	90	86	100
27	2	0	2	100		100	100		100
28	7	2	5	100	100	100	100	100	100
29	67	52	15	72	65	93	69	60	100
30	60	46	14	97	96	100	93	91	100
31	41	29	12	100	100	100	100	100	100
32	8	0	8	100		100	100		100
33	2	1	1	100	100	100	100	100	100
34	9	1	8	89	0	100	89	0	100
35	122	14	108	10	7	10	13	7	14
36	4	3	1	100	100	100	100	100	100
37	18	3	15	94	67	100	94	67	100
38	47	26	21	72	50	100	79	62	100
40	1	1	0	100	100		100	100	
43	6	3	3	100	100	100	100	100	100
44	1	0	1	100		100	100		100
45	2	2	0	100	100		100	100	
46	2	1	1	50	0	100	50	0	100
48	2	2	0	50	50		0	0	
49	12	9	3	100	100	100	100	100	100
50	3	3	0	100	100		67	67	
52	18	3	15	83	67	87	89	67	93
54	9	7	2	89	86	100	89	86	100
56	1	1	0	0	0		0	0	
57	2	1	1	50	0	100	100	100	100
58	41	9	32	100	100	100	100	100	100
59	65	27	38	100	100	100	100	100	100
60	2	2	0	0	0		50	50	
62	3	0	3	100		100	100		100
63	4	0	4	100		100	100		100
64	11	10	1	91	90	100	73	70	100
65	2	2	0	0	0		0	0	
67	3	3	0	100	100		100	100	
68	3	3	0	33	33		33	33	
70	5	3	2	100	100	100	100	100	100
71	2	1	1	50	0	100	100	100	100
72	40	6	34	100	100	100	100	100	100
73	9	8	1	78	75	100	78	75	100
74	11	9	2	82	78	100	91	89	100
76	30	21	9	97	95	100	97	95	100
77	1	1	0	100	100		100	100	
78	4	1	3	100	100	100	100	100	100
80	7	0	7	100		100	100		100
81	2	1	1	50	0	100	50	0	100
82	3	2	1	67	50	100	100	100	100
all classes	2338	1151	1187	86	80	91	87	83	91
$card \leq 5$	74	45	29	74	58	100	77	62	100
$5 < card \leq 30$	242	128	114	92	86	98	91	85	98
$30 < card$	2022	978	1044	85	81	90	87	84	90

Table 2: Percentage of similar classification between the rule resulting from d_h and the ones obtained by d_1 and $d_{\frac{1}{2}}$.

them of cardinality 40 and more (classes 31, 58, 59 and 72).

Obviously, this coefficient of similarity depends strongly on the homology, as it is illustrated in table 2. It should be mentioned that we also carried out some experiments to evaluate this coefficient of similarity against the length of the protein sequences and our results, not reported in this short paper, agreed with Sander and Schneider's work (Sander & Schneider 1991) that higher similarities between these two classification procedures are found for longer sequences.

To conclude this section, let us consider the results obtained after reduction of the information contained in the composition space, according to our procedure described in section 2.3. Without giving details on which thresholds occur in the solution of the integer linear program (3) for each right-hand-side k, it is worth mentioning that the amino acids F, I, T and W were used most often (if the number of thresholds placed along each of these amino acids is summed up for 8 different solutions of (3), we got a total higher than 12 for each of these four amino acids), while E and Y were almost never used (only once each among 10 different solutions). This information should be considered with a lot of care, since each of these vectors solution of (3) are obtained by a greedy procedure, and it is concivable that a different procedure would make some different uses to various amino acids. However, a more robust evaluation of the frequency of each amino acids could yield some very usefull information for the design of new distance measures in the composition space. For example, one could weight each amino acid by its frequency.

As it was already mentioned, the coefficient of similarity between the classification based on homology, and the one based on composition is 87% (resp. 86%) with $d_{\frac{1}{2}}$ (resp. d_1) as distance measure on $\mathbb{R}^{20}$, the composition space. With the classification based on Hamming distance and *voting rule* described in section 2.2, this coefficient of similarity becomes 73%, 61% and 45% for solutions of (3) with right-hand-side $k = 10$, 3 and 1 respectively (the respective size of binary spaces are 59, 22 and 12). In other words, when we extract 59 binary variables from 20 continuous ones of the composition space, 86% of the similarity between the classification based on homology and the one based on composition is preserved; 70% is preserved when we express the composition space with nearly only one binary variable for each amino acid; and the maximal compression of information, which reduces $\mathbb{R}^{20}$ to $\{0,1\}^{12}$, still maintains 52% of this similarity. In conclusion, we claim that our estimation of this coefficient of similarity is robust.

4 Discussion

Molecular biology is noted for vast amount of empirical data with "hidden" correct classification. A good example of this is Structure Analysis of Proteins and Peptides. There is a huge amount of data on their amino acid sequences that need to be classified versus small set of sequences with known three-dimensional (3D) structures. There are several classifications of proteins with known 3D structure, and the problem is to find a good extension for sequences with unknown structure. A realistic 3D classification divides sequences into a large number of classes and it provides an additional difficulties for building an extension.

Another example. It is well-known that folding patterns of T-cell receptors are very similar to that of antibodies. In this case a classification includes only two classes, where one of these classes is complementary to another. This complementary class usually has a very complicated description for its recognition and the problem is to find a characteristic property that can recognize if a given sequence is a T-cell receptor or not.

As third example, one of the interesting problem in molecular biology is analysis of interactions between different proteins and peptides and one of the application problems is to classify sequence database on the ability of proteins to interact with some given proteins.

We expect that many more classification of this kind will appear in near future and it is necessary to develop an approach that allows to compare various methods of classification and to show their relationship. This approach should compare methods, enhance their improvement and create a classification of higher level.

In this paper we present a simple version of this approach. We focus on the problem of a comparison of two methods of protein folding classification with realistic number of folding classes. We define a coefficient of similarity among two given methods, and describe a procedure to estimate a robustness of the coefficient for two sets of data. We applied these coefficient and procedure to the study of the relation between different measures of similarity for protein sequences. We show that folding class prediction by sequence homology and amino acid composition are very close.

On the one hand, this result emphasized restrictions for folding class prediction by sequence homology. On the other hand, it shows a new opportunity for using amino acid composition in models of folding class prediction, since we found that this prediction can be improved using different weights for different amino acids.

Acknowledgments

Beside their support mentioned in the first page, the authors would like to thank one anonymous referee of ISMB'95 for his appropriate criticism and valuable comments.

References

Beasley,J.E. 1990. A Lagrangian heuristic for set-covering problems. *Naval Research Logistics*, **37**: 151–164.

Crama,Y., Hammer,P.L. and Ibaraki,T. 1988. Cause-Effect Relationships and Partially Defined Boolean Functions. *Annals of Operations Research*, **16**: 299–326.

Chothia,C., Lesk,A.M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, **5**: 823–826.

Chou,P.Y. 1989. Prediction of protein structural classes from amino acids. In Prediction of Protein Structure and Principles of Protein Conformation. Plenum Press, New York, 549–586.

Dubchak,I., Holbrook,S.R. and Kim,S.-H. 1993. Prediction of Protein Folding Class From Amino Acid Composition. *Proteins*, **16**: 79–91.

Dubchak,I., Holbrook,S.R. and Kim S.-H. 1993. Prediction of Protein Three-Dimensional Folding Classes from Amino Acid Composition. *Proteins*, **16**: 79–91.

Duda,R.O. and Hart,P.E. 1973. *Pattern Classification and Scence Analysis.* John Wiley & Sons, New York.

Hilbert,M., Bohm,G. and Jaenicke,R. 1993. Structural relationship of homologous proteins as a fundamental principle in homology modeling. *Proteins*, **17**: 138–151.

Kabsch,W. and Sander,C. 1984. Identical pentapeptides can have completely different conformations. *Proceedings of National Academy Sciences*, **81**: 1075–1078.

Nakashima,H., Nishikawa,K. and Ooi,T. 1986. The folding type of a protein is relevant to the amino acid composition. *Journal of Biochemistry*, **99**: 152–162.

Pascarella,S. and Argos,P. 1992. A Data Bank Merging Related Protein Structures and Sequences *Protein Engineering*, **5**: 121–137.

Sander,C. and Schneider,R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**: 56–68.