*Statistics in Molecular Biology*
IMS Lecture Notes – Monograph Series (1999) Volume 33

# PROTEIN FOLD CLASS PREDICTION IS A NEW FIELD FOR STATISTICAL CLASSIFICARION AND REGRESSION

BY LUTZ EDLER AND JANET GRASSMANN

*Biostatistics Unit, Research Program on Genome Research and Bioinformatics, German Cancer Research Center, Heidelberg, Germany and BRAIN², Germany*

Protein structure classification and prediction is introduced and elaborated for the application of standard and new statistical classification, discrimination and regression methods. With the sequence to structure to function paradigm in the background, methods of secondary and tertiary structure prediction will be reviewed and super-secondary classes and of fold classes will defined. We apply two branches of statistical classification - methods based on posterior probabilities and methods based on class conditional probabilities - and we will explore the role of artificial neural networks for the protein structure prediction. The procedures will be applied to a set of 268 previously described protein sequences for their statistical performance in the prediction of the four super-secondary classes and also in the prediction of 42 fold structure classes.

**1. Introduction.** Knowledge of the three-dimensional (3D) structure of a protein is essential for describing and understanding its function and for its use in molecular modeling [Fasman (1989)]. The impact of the structural knowledge for medical interventions and the understanding of diseases and their evolution has been clearly demonstrated [Branden and Tooze (1991), Gierasch and King (1990)]. Knowledge of the 3D structure of hemoglobine [e.g. Perutz (1978) Dickerson and Geis (1983)] enabled researchers to increase its oxygen capacity. This was the first and crucial step of a development which resulted in a synthetic hemoglobin substitute with consequences for blood transfusion [Mickler and Longnecker (1992)]. On the other hand, sickle cell anemia is caused by a single mutation in the amino acid sequence of hemoglobin, a change from *Glu* to *Val* on the surface of the globin fold [see Branden and Tooze (1991) p. 39] which causes movements of the $\alpha$-helices relative to each other and makes the cell membrane more permeable to potassium ions. The disease is lethal for homozygotes, but increases the resistance to malaria in heterozygotes by killing the parasite through the drop of the potassium ion concentration. Therefore, the determination of structure is useful in different aspects: altering an existing protein's function (protein engineering), creating a protein *de novo* (protein de-

*AMS* 1991 *subject classifications.* Primary 92A10; secondary 62H30.
*Key words and phrases.* Protein structure, fold class, classification and prediction, discriminants, regression, neural networks, cross-validation.

sign), or understanding the evolution of diseases. Understanding and predicting how sequence information translates into 3D structure and folding of the then biologically active protein (functional properties) has become one of the most challenging problems in current molecular biology [Sternberg (1996)]. A solution of the protein folding problem would have great implications on interpreting sequence data as those created by the Human Genome Project. It would improve gene function analysis with implications on understanding hereditary genetics and diseases and it would provide clues for drug design and biological engines with considerable commercial consequences (see e.g. Cambridge Healthtech Institute: *http://www.healthtech.com/*).

The three-dimensional structure of a protein is determined physically by the 3D coordinates of all atoms of the protein. It is mostly obtained by x-ray crystallography and for smaller proteins also by NMR (nuclear magnetic resonance), see Branden and Tooze (1991). This determination has been achieved at present only for a small percentage of known proteins. On the other hand, the extraordinary improvement of the efficiency of modern sequencing techniques creates a large gap between the number of sequenced proteins and the number of structurally 'explained' proteins. Figure 1 shows the sharp increase of the number of entries of proteins sequences and protein domains in the SWISSPROT data base compared to the much tardier increase of proteins of known 3D structure in the PDB data base [Bernstein et al. (1977), Bairoch and Apweiler (1997), Benson et al. (1997)]. Since the Human Genome Project will generate an enormous amount of protein sequences more over the next few years [Rowen et al. (1997)] this gap will increase rapidly. The need to bridge the gap has called for biochemical and biophysical methods for the determination of the 3D structure which circumvent x-ray and NMR and use the basic sequence and physical properties of the building blocks of proteins. The protein fold problem poses itself then as the question [Richards (1991)]: How to predict the 3D structure of a protein from its amino acid sequence? This question had been around in protein research since the seminal proposition of Anfinsen (1961) to predict the conformation of a protein on the basis of its linear amino acid sequence. From there originates the hypothesis that the sequence of amino acids of a protein is necessary and sufficient for the determination of the 3D structure and, consequently, for its function. Almost 40 years later this problem is still in the center of theoretical and practical biotechnological protein research. Although unsolved in its original sense, the question continues to elicit important research results of structural biology and molecular modeling.

Anfinsen's hypothesis suggests that the amino acid sequence together with physical and chemical principles should suffice to determine the forces responsible for the folding and determination of the ultimate 3D structure. One approach

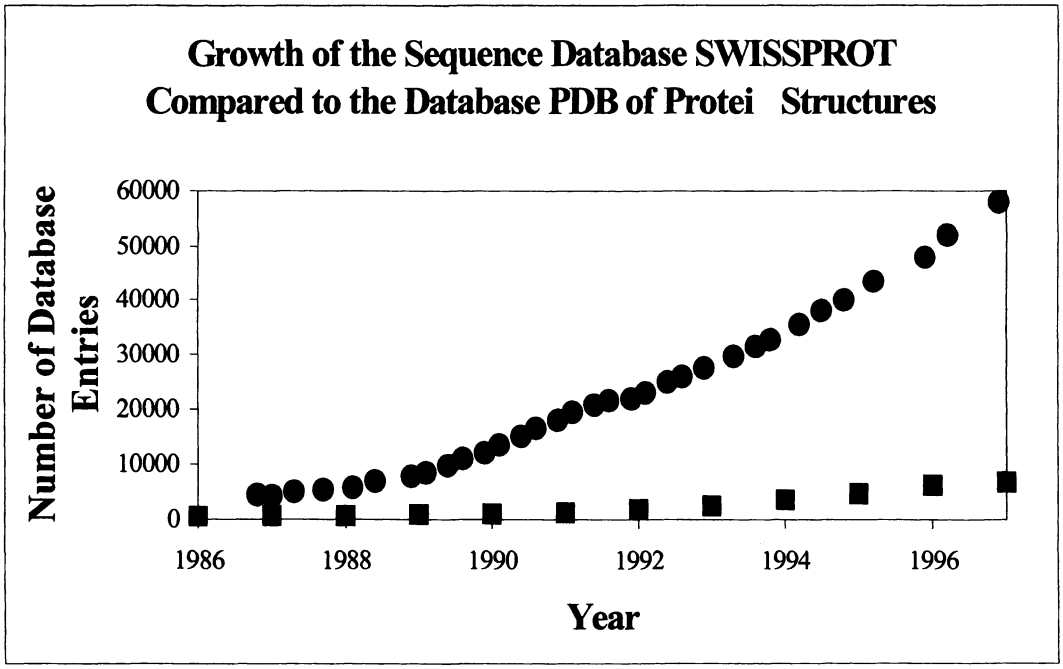


FIG. 1. *Increase of the number of entries of sequences of proteins and protein domain in SWISSPROT data base and of the sequences of known 3D structure in the PDB data bank.*

has been the *ab initio* calculations which combine biophysical, biochemical, and quantum mechanical calculations with molecular modeling in order to determine the native protein and its 3D structure from the denaturated and unfolded protein [Dinur and Hagler (1991), Defay and Cohen (1995)] by simulating dynamically nature's folding pathways [Creighton (1984, 1990)]. An increasing number of proteins could be constructed this way (see e.g. recent issues of the journal *Protein Engineering*) but the method is still far from being applicable to large proteins or to be used generally.

A second approach formulates the protein folding problem in terms of mathematical physics. This point of view has been reviewed excellently by Neumaier (1997) who provides also a survey of most of the relevant literature on chemical structure and local geometry of a protein and molecular mechanics. Focussing on mathematical models and molecular dynamics, quantitative 3D prediction applies (global) optimization of the potential-energy functions, directly making use of the physical forces between the atoms in the poly-peptide sequence of the protein. This method is computationally intensive and can succeed only if its able to find the global energy minimum which determines the ultimate 3D structure. The limitations of this method are in the computational complexity combined with the intrinsic problem to avoid local energy minima. A

third approach evolved recently from machine learning and artificial intelligence and claims to predict structural classes of proteins from the basic amino acid sequence and features derived from that [Holley and Karplus (1989), Lambert and Sheraga (1989), Friedrichs et al. (1991), Taylor (1992)]. Although success was rather limited [Schulz (1988)] this approach is appealing because of its new information theoretical access to the problem. It has attracted computer scientists, biomolecular modelers, bioengineers and biophysicists to use especially artificial neural networks (ANN) for the classification and prediction. The goals were lower in this case: not the quantitative prediction of the ultimate 3D structure but the prediction of the qualitative appertaining to a limited number of folding categories was searched. In principle, all three approaches could not reach their goal because of the problem's complexity which translates almost immediately into computing complexity as e.g. when biophysical methodology is combined with modern computer algorithms or when all possible configurations are screened for those with low energy [Neumaier (1997)]. Therefore, the search for better prediction methods is ongoing [see Defay and Cohen (1996) and the endeavors of the Ansilomar Conference from 1994].

So far, this type of protein prediction was without much interaction with statistics, although standard statistical methods such as regression, discriminant analysis, and cluster analysis have been applied to a variety of prediction problems with considerable success. To our knowledge, traditional statistical techniques have not been applied systematically to the protein folding problem. The protein structure prediction problem is usually complicated. This fact may have deterred researchers from using standard statistical methods to predict protein structure. However, if statistical methods could be applied to this problem it would be very interesting to compare their performance with that of machine learning and artificial neural networks. More attention could then be given to the assumptions of the procedures, the sampling of the data and the realism of the error probabilities and the prediction accuracy. By this work we want to add statistical classification and statistical methodology of pattern recognition to the toolbox for predicting protein structure directly from the sequence information, and we want to initiate the exchange and transfer of methods between the disciplines of protein research and applied statistics.

In principle, biophysical methodology should be able to define the unique structure of a protein from the atomic structure of the amino acids in the sequence. At present, it fails with the complexity of the calculations. Therefore, it may be a reasonable strategy at this stage to develop methods that exploit both types of information, the biophysical information from molecular mechanics and the statistical information from classification and regression on sequences. We can not be sure that this combination will lead soon to a breakthrough for this

knotty problem but it will require innovative collaboration between molecular biologists and statisticians. Interaction of the two disciplines could push the problem one step further to its solution. The following methods and ideas represent a statistical contribution to such a collaboration. In other words, the aim of this paper is to introduce statisticians to the subject, to provide some background of the problem and to improve the interaction between statisticians and computer scientists for approaching the statistical problem of classification and prediction of protein fold class.

Below, we will provide a short introduction to protein structure prediction and to some of the results achieved. We will introduce in the next section the biological problem of fold class prediction and review shortly previous methods of fold classification and prediction. This will comprise the data structure of the amino acid sequence and the definition of the statistical classification and decision problem such that it becomes amenable to regression, discrimination and classification methods. Section 2.2 provides a short review of the methods used for secondary structure prediction and their achievements. Based on the secondary structure we define the four supersecondary classes which have to be predicted by the statistical methods we will introduce later. Section 2.3 introduces tertiary structure and fold prediction. Emphasis is given to the statistical content of the methods used previously by protein researchers. In Section 2.4 we will present standard and new regression and discriminant methods to be used competitively for prediction including the neural nets. These methods are applied in our case study of a data set of 268 protein sequences [Grassmann et al. (1998)] described in 2.5 which we investigated recently in collaboration with colleagues from the biophysiscs discipline in our research program. Selected results of the fitting of the models to these protein data will be given in Section 3 and discussed in Section 4.

## 2. Methods.

2.1. *Proteins and their classification.* From a chemical point of view, a protein is a polymer or polypeptide consisting of a long chain of amino acids linked by peptide bonds to a one-dimensional directed polypeptide chain, see Brandon and Tooze (1991) for an illustrative introduction. Important for the 3D geometry is that fact that each amino acid in the sequence contains a central carbon atom ($C_\alpha$ atom) and that each amino acid is characterized by its side chain (residue) attached to the $C_\alpha$ atom. Interatomic forces bend and twist the protein into a characteristic 3D folded state. The sequence of the $C_\alpha$ atoms represents the so-called backbone of the protein. Their three-dimensional coordinates represent the genuine 3D structure. For the local geometry and all other

details see Neumaier (1997).

Nature has provided twenty amino acids; see Brandon and Tooze (1991) or any standard monograph on molecular biology for a listing and characterization. A *protein of length* $N$ is then formally represented as an ordered sequence

$$P = (s_1, \ldots, s_N)$$

with elements $s_i$ from the finite set $A = \{A_1, \ldots, A_{20}\}$. The length of proteins varies considerably between the tens and a few thousands. An average sized protein has a length of 100-200 amino acids. For combinatorial reasons the number of possible proteins is therefore huge: Given an averaged sized protein of 150 amino acids, the number of possible sequences would be $20^{150} \approx 10^{200}$. At present, the number of proteins existing in living nature is not known. Based on a number of assumptions, Zhang (1997) estimated the number of human proteins roughly to $5 - 10 \times 10^5$. Similarities of the shape and functional similarities of proteins motivated researchers to define *structural classes* for proteins, say classes $C_1, \ldots, C_K$. The largest set of structural classes would be obtained by the complete description of a protein by all $C_\alpha$ atoms and their 3D coordinates. Although, the determination of the full set of 3D coordinates of the $C_\alpha$ atoms is an important task performed in cristallography [Zanotti (1992)] and NMR spectroscopy [Torda and van Gunsteren (1992)] broader classifications lead to structural families. Holm and Sander (1994) describe e.g. 270 fold classes for 838 families with a class occupancy ranging between 1 and 73. Using some type of sampling statistics and empirical data, Wang (1998) estimates a total number of 1150 protein superfamilies and about 650 protein folds to exist in nature. That means we are dealing roughly with a number of sequences of the order of magnitude of perhaps $10^5 - 10^6$ to be classified into about $10^2 - 10^3$ classes, given all proteins have been once sequenced.

The definition of structural classes of proteins started with the identification of local structure in the primary amino acid sequence. There are only three types of so-called *secondary structures*: *$\alpha$-helices*, *$\beta$-sheets* and *coils* ($\gamma$). For illustrative details see Branden and Tooze (1991) and Fetrow et al. (1997). Secondary elements can combine with each other to form **motifs** or so-called **super-secondary structures** which finally assemble globally and form the tertiary structure. Classification and prediction of secondary structure is considered also as intermediate step to tertiary structure [Stolorz et al. (1992)]. A very simple global classification is obtained by characterizing the protein by the presence or absence of $\alpha$-helices and $\beta$-sheets. This results in four *super-secondary classes (SSC)*: only $\alpha$, only $\beta$, one part $\alpha$ plus one part $\beta$ ($\alpha + \beta$), and $\alpha$ and $\beta$ alternating ( $\alpha/\beta$ ). Based on topological similarity of the backbone, a definition of 38 fold classes has been proposed by Pascarella and Argos (1992)

and recently enlarged to 42 fold classes by Reczko et al. (1994). That set of classes was later enlarged to 45 classes [Reczko and Bohr (1994)] and further to 49 classes by Reczko et al. (1997).

The raw information given by the protein sequence $P = (s_i, i = 1, \ldots, N)$ with elements from $A = \{A_1, \ldots, A_{20}\}$ is usually reduced and restructured for protein classification and prediction such that each protein is represented by an element of a suitable predictor space (feature space) $X$. A very simple example is the space of the frequency distributions of amino acids, the 20-dimensional unit cube $X = [0,1]^{20}$ where each protein is represented by a vector $(f_1, \ldots, f_{20})$ of the relative frequencies of the 20 amino acids in its sequence. Other feature spaces have been constructed by some type of 'reading' information. There, a moving window $x_j^{(a)} = (s_j, s_{j+1}, \ldots, s_{j+a-1})$ say of length $a$, is gliding along the protein sequence (Figure 2). The sample $x_j^{(a)}, j = 1, \ldots, N$ represents the sequence $P = (s_1, \ldots, s_N)$ and a suitable feature space $X$ is e.g. the space of the frequency distribution of all $a$-tuples with elements of $A$. Special cases are the dipeptides obtained for $a = 2$ which give raise to a 400 dimensional space of dipeptide frequencies. We consider in the following the complete protein as sampling unit. In some cases, sequences of sub-domains of proteins or motifs were treated as independent samples even if they originated from the same protein. Two *feature spaces* $X$ will be used in our analysis: Firstly, the space of the amino acid frequencies (AAF)

$$X_1 : x = (f_1, \ldots, f_n)$$

where $f_i$ denotes the relative frequency of the amino acid $A_i$ in the sequence. Secondly, the space of the dipeptide matrices (DPF)

$$X_2 : x = (f_{1,1}, \ldots, f_{1,20}, f_{2,1}, \ldots, f_{2,20}, \ldots, \ldots f_{20,1}, \ldots, f_{20,20})$$

where $f_{i,k}$ denotes the relative frequency of the amino acid pairs $(A_i, A_k)$ in an ordered sequence of residues (the case of a moving window of length $a = 2$).

A *classification / prediction rule* $R$ is a rule which maps the feature information $x \in X$ of each primary sequence $P = (s_1, \ldots, s_N)$ into a finite set of structural classes $C = C_1, \ldots, C_K$. For convenience, the structural classes $C_k$ are represented by the consecutive numbers $\{1, \ldots, K\}$. From a statistical point of view the protein structure prediction is nothing more than a prediction of an element of a finite set of structural classes based on information $xinX$ where $X$ is an Euclidean Space, e.g. a subset of $\Re^n, n > 1$. However, the space $X$ is not straightforwardly given in practice. There are many options to define $X$, see above, such that it represents relevant information and is still dimensionally tractable. When using moving windows $x_j^{(a)}$, the window size $a$ can be chosen

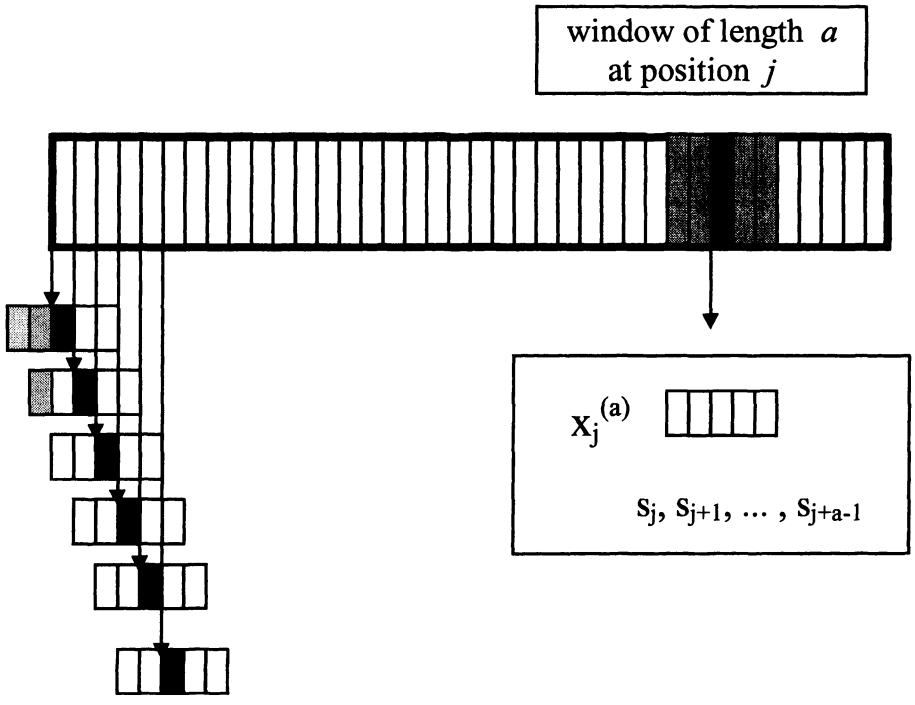


FIG. 2. *Moving window information read from the protein sequence. A moving window of length 5 glides in this example from the left to the right. At the start (and at the end, respectively) the segments have to be augmented by spacers.*

depending how much neighborhood information is thought to be useful. Given only a limited number of observations available in practice, one has to account for the sparseness of data in $X$.

2.2. *Secondary structure prediction.* Because of its direct connection with the SSC classification used below and because of the wealth of previously obtained results we address here secondary structure prediction for further illustration. The secondary structure of the amino acid sequence is defined as a local property and induces an one-to-one mapping

$$R_{SEC} : P = (s_1, \ldots, s_N) \longrightarrow Q = (r_1, \ldots, r_N)$$

from the set of all possible sequences $\{(s_1, \ldots, s_n) : s_i \in A, i = 1, \ldots, n, n = 1, 2, \ldots\}$ to the set $\{(r_1, \ldots, r_n) : r_i \in (\alpha, \beta, \gamma), i = 1, \ldots, n, n = 1, 2, \ldots\}$. Each element of $P$ is mapped onto exactly one element from $\{\alpha, \beta, \gamma\}$. The rule $R_{SEC}$ assigns to a protein its secondary structure $Q$ as an estimate $\hat{Q} = (\hat{r_1}, \ldots, \hat{r_N})$ depending on a sample of $n$ known pairs $(P_j, Q_j) j = 1, \ldots, n$, of proteins and structures.

Let us use this framework to consider the definition of a classification error. The naive *per protein error rate* is the ratio $w/N$ of the number $w$ of incorrectly assigned secondary classes of the residues of that protein $P$ of length $N$. This error is further differentiated with respect to the three secondary types. This yields a misclassification table or confusion matrix [Ripley (1996) Chap. 2.7]:

| | | True Class | | | |
|---|---|---|---|---|---|
| | | $\alpha$ | $\beta$ | $\gamma$ | |
| Assigned Class | $\alpha$ | $w_{\alpha\alpha}$ | $w_{\alpha\beta}$ | $w_{\alpha\gamma}$ | $\hat{N}_\alpha$ |
| | $\beta$ | $w_{\beta\alpha}$ | $w_{\beta\beta}$ | $w_{\beta\gamma}$ | $\hat{N}_\beta$ |
| | $\gamma$ | $w_{\gamma\alpha}$ | $w_{\gamma\beta}$ | $w_{\gamma\gamma}$ | $\hat{N}_\gamma$ |
| | | $N_\alpha$ | $N_\beta$ | $N_\gamma$ | $N$ |

The diagonal contains the number of correctly and the-off diagonal contains the numbers of incorrectly classified amino acids. Therefore,

$$e_i = (N_i - w_{ii})/N_i$$

is the *structure specific error rate*, $i = 1, 2, 3$, and

$$e = \left( N - \sum_{i=1}^{3} w_{ii} \right) /N$$

the *total error rate*, usually denoted by $Q_3$ in secondary structure prediction. If classification or prediction is performed for a set of proteins one calculates an *overall structure specific error rate* and *overall total error rate* by pooling all residues of all the proteins. Another measure of discordance is the so-called *Matthew correlation* [Matthew (1975)] defined as

$$MC_i = \frac{w_{ii} \left( \sum_{j,k \neq i} w_{jk} \right) - \left( \sum_{j \neq i} w_{ij} \right) \left( \sum_{j \neq i} w_{ji} \right)}{\hat{N}_i \left( \sum_{j \neq i} \hat{N}_j \right) N_i \left( \sum_{j \neq i} N_j \right)}$$

for $i = 1, 2, 3$. $MC_i$ is up to the factors $N_i$ the square root of the chi-square statistic for the classification into the $i$-th secondary structure category if the data are organized as a four fold table with the numbers $w_{ii}$ of correct classifications into category $i$ and the numbers $\sum_{j=k;j,k \neq i} w_{jk}$ of correct classification into the non-$i$ category.

Other error estimates are obtained by splitting the data into a training set and a test set and calculating the *training error rate* and the *test error rate*, or

TABLE 1
*Previous results on secondary structure prediction.*

The overall percentage of prediction (Q3 accuracy) is given and where available the sample sizes of the training and the test set in brackets $[n(training), n(test)]$. In some cases the method could not be described in short terms and is missing (−). If more than one result is published, only the best is reported. In two cases where only the individual predictions for the $\alpha$-helices and the $\beta$-sheet was published those are reported in parentheses. jack = jackknife procedure, $CV(x) = x$-fold cross validation. For references see *http://www.dkfz-heidelberg.de/biostatistics/protein/protlit.html.*

| AUTHOR | METHOD | % | |
|---|---|---|---|
| Lim (1974) | physico-chemical characteristics | 59 | |
| Chou & Fasman (1978) | preference index | 57 | |
| Garnier et al. (1978) | GOR, maximum likelihood | 56 | |
| Gibrat et al. (1987) | GOR II | 63 | |
| Zvelebil et al. (1987) | GOR + evolutionary conservation | 66.1 | [-,11] |
| Levin et al. (1986) | KNN + homology | 62.2 | |
| Biou et al. (1988) | GOR Combined, | 65.5 | [67,-; jack] |
| Levin & Garnier (1988) | KNN | 63.0 | |
| Holley & Karplus (1989) | ANN | 63.2 | [48,14] |
| Qian & Seinowski (1988) | ANN | 64.3 | [91,15] |
| Rooman & Wodack (1988) | - | 62 | |
| Kneller et al. (1990) | ANN + a priori information | 65 | |
| King & Sternberg (1990) | symbolic machine learning | 60 | [43,18] |
| Muskal & Kim (1992) | tandem ANN - $\alpha$ | 95.0 | [105,15] |
| | tandem ANN - $\beta$ | 95.4 | |
| Salzberg & Cost (1992) | machine learning | 71 | [100,28] |
| Stolorz P et al. (1992) | tandem FNN | 63.5 | [91,14] |
| Zhang et al. (1992) | ANN, nearest neighbor hybrid | 66.4 | [107,-] |
| Sasagawa & Tajima (1992) | ANN | 56.2 | [33,29] |
| Asai et al. (1993) | Hidden Markov methods | 66.0 | [120,-;jack] |
| Leng et al. (1993) | two level method | 69.3 | |
| Yi & Lander (1993) | KNN + ANN + scoring | 68 | |
| Rost & Sander (1993) | ANN + Jury | 70.8 | [130,-;CV(7)] |
| Rost & Sander (1994) | - | 72.5 | |
| Ellis & Milius (1994) | GOR | 62.2 | [239,-] |
| Wako & Blundell (1994) | - | 77 | [14,-] |
| Geourjon & Deleage (1994) | self-optimized, binary, similarity | 69 | |
| Solovyev & Salamov (1994) | LDA + multiple alignment | 68.2 | [126,-;jack] |
| Barlow (1995) | hierarchical mixture of ANNs | 63 | [91,14] |
| Salamov & Solovyev (1995) | KNN + scoring table | 72.2 | [126,-] |
| Chandonia & Karplus (1996) | FNN+ sequence profiles | 72.9 | [318,-;CV(32)] |
| DiFrancesco et al. (1996) | Logistic regression | 71.5 | [115,-;CV(7)] |
| Frishman & Argos (1997) | local pairwise alignment | 74.8 | [125,-;jack] |
| Ito et al. (1997) | 3D-1D compatibility/pseudo energy | 69.3 | [325,77] |
| Kawabata & Doi (1997) | BW-mod GOR + mult. alignement | 68.2 | [126,-:CV(7)] |
| Fiser et al. (1997) | Deleage method - $\alpha$ | 68.6 | [80,-] |
| | Deleage method - $\beta$ | 65.0 | |
| Levin (1997) | KNN (update of Levin & Garnier, 1988) | 72.8 | [372,111] |
| Rychelwski & Godzik (1997) | segmented similarity after alignment | 72.4 | [256,256] |

by calculating a *cross-validation error* (CV) or the *jackknife error* [Efron and Tibshirani (1993)], see also Grassmann et al. (1998).

Table 1 gives an overview of the development of the **secondary structure prediction**. Since its beginning, prediction accuracy has improved from less than 60% to more than 70%. Among the methods not using intrinsic biochemical information were mostly the nearest neighbor methods and artificial neural networks (ANNs). From a statistical point of view to mention is the quadratic logistic regression applied by DiFrancesco et al. (1996). They obtained a prediction error of 27.6% (with 7 fold CV). Interestingly, this rate was reduced to 21.5% when they incorporated techniques from bioinformatics as e.g. the relative frequencies of residues at each position after multiple alignment of homologue sequences, a variability score describing conserved residue patterns, or insertions and deletions. For details of the statistical modeling and the performance of the maximum likelihood estimation method see Di Francesco et al. (1996).

2.3. *Tertiary structure prediction.* Prediction of 3D structure is extremely complicated and has been confined to only a small number of shorter sequences. An illustrative view of the present state of tertiary structure prediction is obtained from reports on the recent contest of the Asilomar Conference in 1994 [Defay and Cohen (1995)]. Out of 33 proteins 14 were examined successfully in 12 laboratories. Fold prediction from primary sequence information was performed by 9 research groups performing 23 predictions on 11 sequences and obtaining 4 totally correct predictions. Hubbard and Park (1995) classify in another exercise 9 out of 27 sequences. They apply methods based on evolutionary information contained in multiple sequence alignments and hidden Markov models using various computer algorithms and alignment scores. In contrast to these predictions aiming at the 3D structure we will consider in our statistical classification two sets of classes:

a) the *four super-secondary classes (SSC)* defined above:

$\mathbf{C} = \{$only $\alpha$, only $\beta$, one part $\alpha$ plus one part $\beta$ $(\alpha + \beta)$, and $\alpha$ and $\beta$ alternating $(\alpha/\beta)$ $\}$

b) the 42 classes of Reczko et al. (1994) based on topological similarity of the backbone and presented by the numbering of the classes (www[2]):

$\mathbf{C} = \{1, 2, \ldots, 42\}$

The SSC were chosen for our study for reasons of convenience and because of the possibility of comparison with the work of Reczko et al. (1994). SSC has also been investigated by Geourjon and Deleage (1994) and Efimov (1994), also Barton (1995). Supersecondary structure beyond these four classes has been investigated by Sun and coworkers. They used a vector projection method [Sun

et al. (1996)] and later also a feed forward neural net with one hidden layer of about half of the number of input units [Sun et al. (1997)] to predict a set of 11 standard motifs in 56 non-redundant proteins selected from a set of 240 sequences from the PDB. The motifs were defined as potential building blocks for tertiary structure and are characterized by a well defined 3D structure related to the backbone [Sun and Jiang (1996)]. Prediction accuracy obtained with the neural net for the 11 super-secondary classes of motifs ranged between 68% and 80% [Sun et al. (1997)] and was similar to the accuracy obtained in secondary structure prediction (Table 1). The vector projection method [Sun et al. (1996)] was tailored to those motifs and yielded an accuracy between 83% and 96%. This result may encourage stepping from primary via secondary to tertiary structure prediction. Reczko and Bohr (1994) actually tried this approach to some extent by combining the 42 fold classes with the SSC and they could improve their previous accuracy of 71% for the SSC prediction to about 91%.

2.4. *Statistical methods.* The problem the classification and prediction of secondary or tertiary structure can be formulated statistically in the framework of statistical decision theory. For this purpose we refer to Chapter 2 in Ripley (1996). General statistical decision theory and especially the background of Bayesian methods are found in the monograph of Berger (1985). Stolorz et al. (1992) introduced and discussed Bayesian analysis for secondary structure prediction. Grassmann et al. (1998) identified statistical methods of classification and discrimination as possible tools for fold prediction and applied them straightforwardly. They distinguish two cases: (i) methods based on the *posteriori class probability* and (ii) methods based on the *class conditional probability*. In both cases, an input vector $x$ is assigned to its structural class $k$ by a decision rule $d(x)$. The input vector $x = (x_1, \ldots, x_p)$ is an element of the feature space $X$ associated with the sampling units (i. e. the protein sequences), the structural class $k$ is an element of $C$ associated with the protein fold classification. Random elements of $X$ and $C$ are denoted by X and C, respectively.

**(i) posteriori class probability**

Case (i) builds the decision on the posteriori class probability of class $k$ given $x$

$$\mathrm{p}(k|x) = \mathrm{P}(C = k|X = x)$$

A sequence $x$ is assigned to that class $k$ for which $p(k|x)$ is maximum. This assignment minimizes the total risk if a standard loss function is assumed [Ripley (1996) Chap. 2.1]. Such, the decision rule $d(x)$ is given as

$$d(x) = \{k \in C : p(k|x) = \max_j p(j|x)\}$$

This decision rule is directly related to regression which enables the application of logistic regression and in its sequel the use of the feed forward neural networks. Set

$$f_k(x) = \mathrm{E}[Y_k|x] = \mathrm{P}(C = k|x) = p(k|x) \tag{2.1}$$

where $Y_k, k = 1, \ldots, K$ are "dummy" variables coding the class variable $C$ as follows

$$Y_k = 1 \quad \text{if} \quad C = k \qquad \text{and} \qquad Y_k = 0 \quad \text{if} \quad C \neq k.$$

An example is the multiple logistic model

$$f_k(x) = p(k|x) = \frac{\exp\big(\eta_k(x)\big)}{\sum_{m=1}^{k} \exp\big(\eta_m(x)\big)} \tag{2.2}$$

with the linear predictor $\eta_k = \beta \cdot x$ [Ripley (1996) Chap. 3.5]. This generalizes the well known correspondence of Fisher's linear discrimination and linear regression. Maximum likelihood methods are used to fit the model and to estimate $f_k(x) = p(k|x)$. The multiple logistic regression (2.1) and (2.2) is equivalent to the single-layer feedforward neural network (FNN) which uses as input the feature vector $x = (x_1, \ldots, x_p) \in X$ and has the $K$ output units $Y_1, \ldots, Y_K$. $\eta_k(x) = {w_k}^T \cdot x$ represents the output function with weight vectors, see e.g. Grassmann and Edler (1996) and Schuhmacher et al. (1994). Below, we will apply this FFN and also the feedforward network FNN(H) with a layer of H hidden units [Ripley (1996) Chap. 5). To minimize the error between the current state net output and the target output we use as error function the Kullback-Leibler distance which is equivalent to the use of the likelihood function. Weight decay regularizes the FNN. Notice, the number of hidden units plays an important role for the structure of the non-linearity and the dimension of the parameter space. Two further regression methods based on the posterior class probability are applied below as described in Grassmann et al. (1998). These are the *additive model* of Hastie and Tibshirani (1990) known as the so-called BRUTO method [Hastie et al. (1994)], and the *projection pursuit regression* (PPR) of Friedman and Stützle (1981).

**(ii) class conditional probability** Case (ii) builds the decision on the *class conditional probability* of the feature $x$ given the class $k$

$$p(x|k) = P(X = x|C = k)$$

Using Bayes formula

$$p(k|x) = \frac{p(x|k) \cdot p(k)}{\sum_y p(x|k) \cdot p(k)} = \frac{p(x|k) \cdot p(k)}{p(x)}$$

the decision rule d(x) can obviously be rewritten as

$$d(x) = \{k \in C : p(x|k) \cdot p(k) = \max_j p(x|j) \cdot p(j)\} \tag{2.3}$$

The prior probabilities $p(k)$ are either assumed to be known or they are estimated by the relative class frequencies $\hat{p}(k)$. The conditional densities $p(x|k)$ can be estimated either by assuming a parametric model $p(x|k, \theta)$ or by nonparametric methods (kernel or nearest neighbor methods), see Ripley (1996; Chap. 2 and 6). Assuming for $p(x|k)$ the multidimensional normal density as a parametric family we obtain the Linear Discriminant Analysis (LDA). Assuming different variance-covariance matrices for the different classes yields the Quadratic Discriminant Analysis (QDA). If the number of classes is large and if the number of available sequences with fold structure information is limited, the full QDA requires the estimation of a too large number of unknown parameters. Therefore QDA is restricted to the so-called QDA-MONO where only the diagonal elements (variances) differ between the classes.

Finally we use the K-*Nearest Neighbor Classification* (KNN) which assigns an object with feature vector $x$ to the majority class of its neighbors. Decision rule 2.3 is applied with an estimate of the class conditional probability of the form

$$\hat{p}(x|k) = \frac{B_k}{n_k \cdot A(H, x)}$$

where $B_k$ denotes the number of the $K$ nearest neighbors of $x$ that belong to class $k$, $n_k$ is the total number of objects in class $k$, and $A(K, x)$ is the content of the smallest hypersphere containing the $K$ nearest points to $x$ [Ripley (1996) Chap. 6.2]. The methods described above were computationally realized by S-Plus software (e.g., *lda, qda, fda, knn*).

2.5. *Data.* The data used for illustration of the statistical classification and prediction described above originate form Reczko et al. (1994). They considered 268 proteins including a few sub-domains which had been classified into the four SSC and into 42 fold categories related to the Pascarella and Argos (1992) classification. Figure 3 exhibits the frequency of the SSC in the data set of the 268 protein sequences. This sample was subdivided into a training set of 143

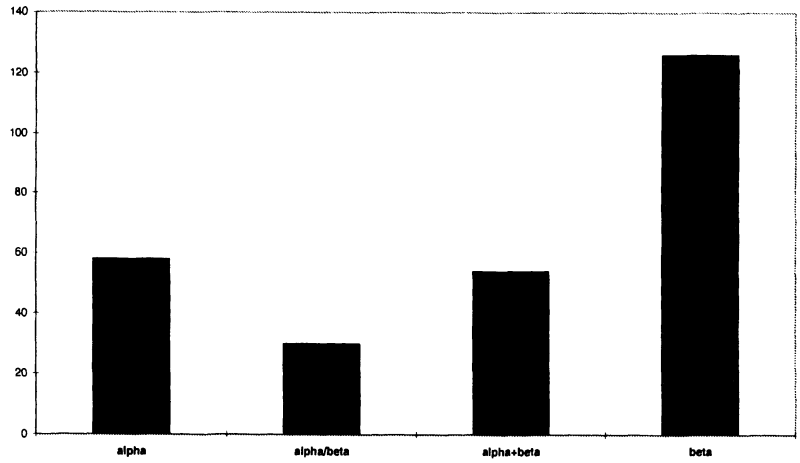


FIG. 3. *Occupation frequency of the four supersecondary classes (SSC) of the data set of 268 sequences.*

sequences and a test set of 125 sequences [same as used by Reczko et al. (1994)] by random sampling stratified according to the 42 classes such that each of the 42 fold categories was occupied at least by one sequence in the training set and in the test set and otherwise balanced at best, but putting 'excess sequences' into the training set, see www[2] for the set of sequences, the partition and the classification into the SSC and the 42 fold classes. Tentatively, we used a second partition with sizes of the training set and the test set in the ratio of about 2:1. In this case 90 proteins were randomly sampled into the test set and 178 remained in the training set.

**3. Results of the prediction analysis.** In this section we will illustrate the application of the methods described above through prediction based on primary sequence information. We used two simple feature spaces: the amino acid frequencies $X_1$ and the dipeptide matrices $X_2$. Since the number of 143 sequences in the training set was smaller than 400, the dimension of $X_2$, a principal component analysis (PCA) was applied in order to reduce the dimension. Deliberately, we set the cut off point to 90% explained variation and obtained so 74 remaining components. This defined a third feature space $X_3$ of dimension 74. For the use of PCA in protein classification see also Ferran et al. (1993). Table 2 outlines the classification task depending on the chosen feature informa-

TABLE 2
*Outline of the classification task.*

Classification was considered for two sets of classes: the supersecondary classes (SSC) and the 42 fold classes based on the backbone topology of Pascarella and Argos (1992). Feature information was available as the amino acid frequencies (AAF), the dipeptid frequencies (DPF), and the 74 first principal component values of the DPF: (DPF-74PC). The notation $p \to K$ informs on the dimension $p$ of the feature space and the number of classes $K$. The DPF column was not realised in this evaluation because of the too large dimension of the feature space $X$ in relation to the sample size.

| | Feature Information | | |
|---|---|---|---|
| Classification | AAF | DPF | DPF-74PC |
| SSC | $20 \to 4$ | $(400 \to 4)$ | $74 \to 4$ |
| 42-CAT | $20 \to 42$ | $(400 \to 42)$ | $74 \to 42$ |

tion (AAF or DPF) and the chosen classification (SSC or 42-CAT, the 42 fold classes). The error rate of the reclassification of the training sequences gives rise to the *apparent prediction error* (APE) which was determined to judge overfitting. As objective measures for the prediction error we calculated the *test prediction error rate* (TPE) and the *cross validation error rate* using a 10-fold cross validation (CV-10). We will focus here on prediction based on the AAFs of the SSCs ('$20 \to 4$') and of the 42 fold classes ('$20 \to 42$').

3.1. *Results for the case* $20 \to 4$. Figure 4 summarizes the error rates for the prediction of the SSCs through the AAFs ('$20 \to 4$' case). Table 3 provides the error rates numerically. The FNN reached perfect apparent prediction on the training set with 7 and 9 to 14 hidden units in our standard splitting of 143 training and 125 test sequences, see Table 3 upper part. The test error rate (TPE) and the cross validation error rate (CV) decreased with few minor exceptions when the number H of hidden units is increased from 0 (logistic regression) to 10. Increasing H forced the error rates to a plateau. Increasing H further lead to numerically unstable estimates. Best prediction accuracy was 77.6% (22.4% CV(10), 23.2% TPE) obtained with an FNN(10). Projection pursuit regression (PPR) could almost reach this accuracy, see Table 3 middle. With a comparable number of terms (H' = 12), PPR could even beat the FNN(10) in terms of TPE (22.0%). PPR became less predictive with a higher number of terms. The discriminant analysis methods performed reasonably good, except the full QDA, which is obviously over-parameterized. QDA showed a perfect prediction on the training set, but it became a disaster on the test set also in terms of the CV error. Remarkably, QDA-MONO yielded one of the best results in '$20 \to 4$' with 20.1% CV error. The additive model (BRUTO) performed almost identical to the LDA. It is also seen that the discriminant based methods show almost

no over-optimism in the apparent error rate (LDA: APE = 30.1 versus CV = 29.5, QDA-MONO : APE = 18.2 versus CV = 20.1). We investigated another splitting of the data into a larger training set and a smaller test set of a ratio of about 2:1 but we obtained worse results (see error rates in parentheses in Table 3). The better performance of QDA-MONO compared to LDA is exhibited in the discriminant plot in Figure 5. The three classes only $\alpha$ (1), $\alpha$ and $\beta$ alternating ($\alpha/\beta$) (2), one part $\alpha$ plus one part $\beta$ ($\alpha + \beta$) (3), and only $\beta$ (4) are further separated in the QDA-MONO discriminant plot and especially the two mixed types are better discriminated.

Because of the identity of the data sets we can now compare our results obtained by standard statistical methods directly with those of Reczko et al. (1994) who had used a cascade correlation network (a partially recurrent neural net allowing for varying topologies). Our accuracy is higher than theirs reported as 71% (TPE = 29%).

3.2. *Results for the case* 20 $\rightarrow$ 42. When we predicted the 42 fold classes from the 20 amino acid frequencies ('20 $\rightarrow$ 42'), see Table 4, the prediction accuracy became worse, as could be expected because of the larger number of classes. LDA exhibited almost as good results as the FNN(12) in terms of the test error rate. PPR showed results similar as the FNN as long as the number of hidden units H and number of terms H' was small. When increasing H and H', PPR became inferior to FNN. The result of QDA was comparable to that of LDA. The KNN was not calculated in this case.

3.3. *Results for the case* 400/74 $\rightarrow$ 4/42. The cases '400 $\rightarrow$ 4' and '74 $\rightarrow$ 4', described in detail in Grassmann et al. (1998), are summarized here for comparison with the cases above. The few successful technically working applications of the FFNs in the case '400 $\rightarrow$ 4' were not reliable because of the too high dimension of the feature space compared with the number of observations. Therefore, we restricted the analysis to the feature space $X_3$ of the first 74 principal components. In the case of '74 $\rightarrow$ 4' the FNN(10) provided in terms of CV the highest prediction rate of 77.6%. LDA gave 73.1% and QDA-MONO and BRUTO only 62.7% and 64.6%, respectively. In the case '74 $\rightarrow$ 42' the FNN(9) provided the highest prediction rate of 67.5%. LDA yielded an even better result of 69.8%; QDA-MONO 61.6% and BRUTO 64.6%. The prediction accuracy Reczko et al. (1994) obtained by a cascade-correlation network was about 73%. They could improve the classification by enlarging the set of classes to 45 [Reczko and Bohr (1994)] and to 49 [Reczko et al. (1997)]. They report then an accuracy of about 82% for the 42 classes when using another constraint network.

TABLE 3
*Prediction of supersecondary classes (SSC) from the amino acid frequencies (AAF).*

The apparent error (APE), test error (TPE) and cross-validation error are presented for the neural networks with varying numbers of hidden units: FNN(H) the projection pursuit regression with varying number of terms H' PPR(H', linear discriminant analysis: LDA, Quadratic discriminant analysis: QDA, quadratic discriminant analysis restricted to varying variances: QDA-MONO, generalized additive model: BRUTO, K-th nearest neighbor method: KNN. The set of 286 sequences were split between training and test set in two ways denoted by '143'/'125': (143 training, 125 test) and '178'/'90' : (178 training, 90 test).

| | Classification Error in % | | | | |
|---|---|---|---|---|---|
| | APE training set '143' ('178') | | TPE test set '125' ('90') | | CV-10 cross-validation |
| FNN: H hidden units | | | | | |
| 0 | 30.8 | (34.3) | 33.6 | (34.4) | 36.9 |
| 1 | 41.3 | (40.4) | 37.6 | (38.9) | 42.5 |
| 2 | 25.9 | (25.3) | 30.4 | (42.2) | 33.6 |
| 3 | 30.1 | (18.0) | 32.0 | (35.6) | 30.2 |
| 4 | 9.1 | (11.2) | 31.2 | (37.8) | 31.7 |
| 5 | 2.8 | (9.0) | 29.6 | (28.9) | 27.2 |
| 6 | 0.7 | (11.2) | 32.0 | (28.9) | 26.9 |
| 7 | 0.0 | (11.8) | 24.0 | (27.8) | 27.6 |
| 8 | 2.1 | (10.1) | 25.6 | (30.0) | 26.1 |
| 9 | 0.0 | (11.2) | 26.4 | (28.9) | 26.1 |
| 10 | 0.0 | (11.2) | 23.2 | (27.8) | 22.4 |
| 11 | 0.0 | (10.1) | 20.8 | (26.7) | 25.7 |
| 12 | 0.0 | (2.8) | 21.6 | (24.4) | 24.6 |
| 13 | 0.0 | (3.9) | 21.6 | (26.7) | 23.1 |
| 14 | 0.0 | (2.8) | 27.2 | (24.4) | 21.6 |
| 15 | 2.8 | (2.8) | 20.0 | (24.4) | 19.8 |
| PPR: H' terms | | | | | |
| 1 | 36.5 | | 42.4 | | |
| 2 | 25.2 | | 32.8 | | |
| 4 | 16.8 | | 29.6 | | |
| 8 | 6.3 | | 24.0 | | |
| 10 | 0.7 | | 32.8 | | 28.0 |
| 12 | 0.7 | | 22.0 | | |
| 13 | 0.7 | | 20.0 | | |
| 20 | 0.0 | | 32.0 | | 27.2 |
| LDA | 30.1 | (29.2) | 36.0 | (33.3) | 29.5 |
| QDA | 0.0 | (0.6) | 60.0 | (72.2) | 60.4 |
| QDA-MONO | 18.2 | (16.9) | 28.0 | (32.2) | 20.1 |
| BRUTO | 30.1 | (29.8) | 36.0 | (34.4) | 29.5 |
| KNN | 0.0 | (-) | 19.2 | (-) | 22.8 |

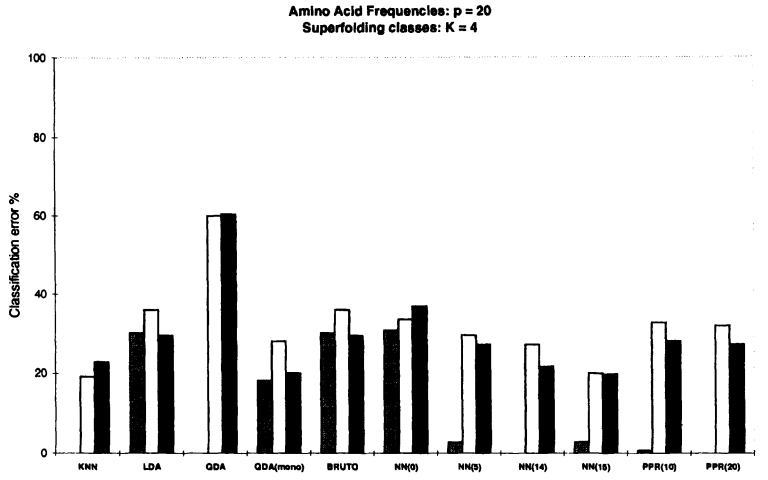


FIG. 4. *Histogram of the error rates (apparent, test, and cross-validation error rate) of the classification into the four SSCs using the AAF information.*

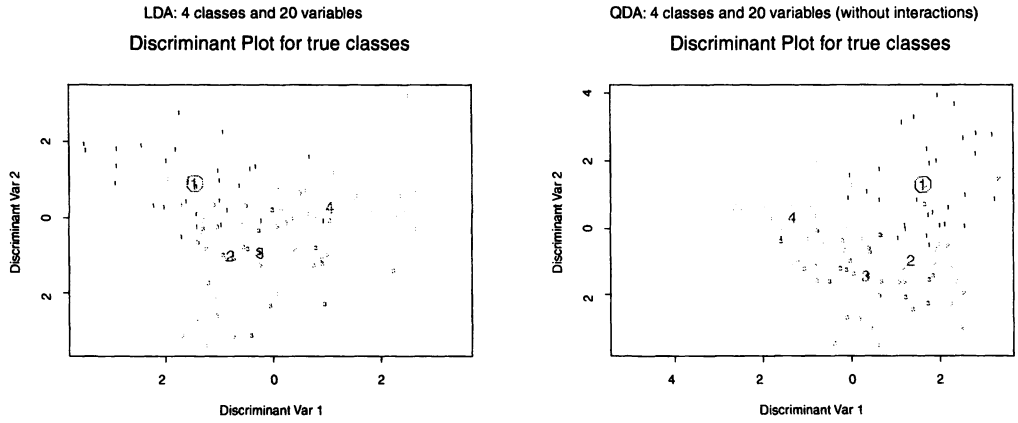


FIG. 5. *Discriminant plot of the Linear Discriminant Analysis (LDA), upper part, and the Quadratic Discriminant Analysis without varying interaction (off-diagonal) terms (QDA-MONO), lower part, for the classification* $20 \rightarrow 4$ *of 125 test sequences into the four supersecondary classes (SSC)* $\alpha$ *(1),* $\alpha$ *and* $\beta$ *alternating* $(\alpha/\beta)$ *(2), one part* $\alpha$ *plus one part* $\beta$ $(\alpha + \beta)$ *(3) and only* $\beta$ *(4) on the basis of the 20 amino acid frequencies (AFF)*

TABLE 4

*Prediction of the 42 fold classes based on backbone topology of Pascarella and Argos (1992) from the amino acid frequencies (AAF).*

The apparent error (APE), test error (TPE) are presented for the neural networks with varying numbers of hidden units: FNN(H) the projection pursuit regression with varying number of terms H': PPR(H'), the linear discriminant analysis: LDA, the quadratic discriminant analysis: QDA, the quadratic discriminant analysis restricted to varying variances: QDA-MONO, and the generalized additive model: BRUTO. The set of 286 sequences was split into 143 training sequences and 125 test sequences.

| | | Classification Error in % APE training set | TPE test set |
|---|---|---|---|
| FNN: | H hidden units | | |
| | 0: Log. Regr. | 0.0 | 46.4 |
| | 1 | 76.2 | 74.4 |
| | 2 | 59.4 | 61.6 |
| | 4 | 28.0 | 48.0 |
| | 6 | 16.8 | 32.8 |
| | 8 | 9.2 | 31.2 |
| | 12 | 0.7 | 27.2 |
| PPR: | H' terms | | |
| | 6 | 61.5 | 60.8 |
| | 10 | 48.3 | 56.0 |
| | 15 | 33.6 | 48.0 |
| | 17 | 27.3 | 44.0 |
| | 18 | 23.1 | 44.8 |
| LDA | | 2.8 | 27.5 |
| QDA | | 0.0 | 29.7 |
| QDA-MONO | | 0.0 | 71.4 |
| BRUTO | | 2.8 | 27.5 |

**4. Discussion.** The prediction of protein folds from their amino acid sequence is an impressively long-standing challenge in molecular biology and biophysics [Finkelstein (1997)]. After a few blind predictions in the seventies it was realized in the eighties that the efforts are 'not hopeless'. Two large scale blind predictions performed in 1994 [Moult et al. (1995), Prediction Center (1996)] exhibit the difficulty to predict 3D folds systematically if the proteins are not closely related to previously known ones, see Lemer et al. (1995), Defay and Cohen (1995) and Hubbard and Tramontano (1996). Prediction experience of the past years showed that the most successful tools are knowledge based systems in combination with experience and statistical methods [Rost and O'Donoghue (1997)]. To our knowledge, tertiary structure predictors have used almost always rather small data sets. Statistical procedures which exhibit their power on large sizes much better have not been applied systematically. Our results obtained with statistical classification methods show that their application is also 'not

hopeless' and that their combination with biophysical methodology may add quality to future prediction, which is interesting in face of the fastly increasing information from the protein data bases. A statistical approach to the problem, with careful attention to assumptions, variation, sampling, defensible precision estimates, and realistic estimates of error probabilities, could strengthen existing procedures as well as provide new ones. Compared with the present aims of protein prediction [Rost and O'Donoghue (1997)] our results above fall somehow short when considered for improving the direct protein prediction. This is not surprising given the simplicity of the feature spaces used in our analysis and the fact that no physical and chemical properties of the molecular level were included. Further research is needed when a richer - and then also more complex - feature space is used, see e.g. the proposal of class-directed structure where only representative members of classes will be fully structurally characterized [Terwilliger et al. (1997)]. Our investigation focussed in the role of standard and new statistical methods with the goal of identifying a possibly best statistical procedure. We considered especially the neural networks. FNNs are in fact non-linear regression methods subordinated under posterior class probability based classification. A further aim was the application of these methods on a larger data set than it has been used in most previous evaluations of automatic protein fold prediction. From our analyses we conclude that linear discriminant and nearest neighbor methods are potent competitors to the more flexible neural networks. The results obtained from the modern discriminant and a regression methods (e.g., BRUTO, PPR) were mixed. In some cases these methods competed very well in other cases the results we obtained so far were disappointing in the sense that they were not able to yield an improved prediction and had sometimes serious problems to cope with the ill-posedness of the classification problem. Typical for this was the failure of the QDA when the number of input variables became larger than the sample size. Search for more efficient regularization methods is needed to exploit the power of these new statistical methods. Previous predictions especially for secondary structure proved the usefulness of FNNs. This was corroborated in our investigation where the FNNs competed well with other methods. This is not surprising since the FNNs are non-linear regression methods and implement standard statistical tools. However, as universal approximators, neural nets are always in danger of overfitting, which we experienced in our analysi too. Therefore, the bias-variance trade off has to be considered carefully. Automatic smoothing and regularization are statistical methods to be investigated further. A clear disadvantage of all FNNs is the lack of interpretation of the weights and the fact that quite different weights and weight patterns can lead to the same prediction outcome.

The optimal method for assessing the validity of the prediction procedure

would be the use of an independent validation set sampled from the set of all proteins. We used a test set after dividing the sample of 268 sequences into 143 training and 125 test sequences to calculate the test error rates (TPE). Those may be still a little optimistic, but the bias is usually in the order of a few percentages, when compared with the CV error. Similar small biases have been observed by DiFrancesco et al. (1996) for secondary structure prediction. This corroborates the recommendation of using some sort of cross-validation error, at least as long as no independent validation data are available. We calculated the cross-validation error (CV-10 fold) mostly in addition to the TPE. In some cases as e.g., PPR, however, computing of the CV error became very time consuming and was not performed for each architecture. The fact that our error rates obtained by the statistical procedures are around 30% is not too disappointing given that no higher order information from the protein was taken into account. Without that it might be difficult to surpass that margin. Obviously, more protein sequence data and perhaps both, more informative feature spaces and functionally more realistic class definitions are needed. Inclusion of information on distant interaction in the protein sequence and its quantitative presentation could be as helpful as the use of physico-chemical properties of the amino acids. At present it seems that the number of classes and the classification itself is too much tailored to the existing information on 3D structure. The number of existing relevant structural families is estimated to about 200-500 but present fold class prediction is limited to much smaller numbers, as our analysis of 42 classes or the analysis of Sun et al. (1996) of 11 classes.

The prediction methods from above assume independent sampling of a complete and correct sequence. Only then are the statistical model estimates and the accuracy measures valid statistics. However, in practice occur errors during sequencing and in a number of cases interest focuses in parts of proteins only, as e.g. motifs. This creates for any procedure - not only the statistical ones - the errors-in-variables problem and the non-independent sampling problem which both need further investigation.

Usually, protein researchers distinguish two situations: (i) Presence of sequence similarity such that the investigated sequences is similar to a sequence of known 3D structure in the data base. (ii) Absence of sequence similarity such that no similar sequence exists in the 3D data base. Similarity is defined as sequence identity after alignment of at least 25% to 30%, which is at the same time considered sufficient to infer structural similarity [Rost and O'Donoghue (1997), Schneider et al. (1997)]. The distinction between (i) and (ii) established the prediction paradigm: If a protein has been newly sequenced, then search the 3D data base for at least one sufficiently similar sequence. If one is found structure and function is predicted from the knowledge available from that. If

none is found an automated structure prediction is tried by comparing the sequence information of the new sequence with the sequence information available for all proteins whose 3D structure has been clarified so far. In our analysis we did not account thoroughly for the effect of the similarity between the 268 sequences. Details on the sequence similarity is provided in www[2]. However, it is shown in Grassmann et al. (1998) how with our data the prediction accuracy decreases when the dissimilarity increases. Further research on the performance of the statistical procedures is needed on dependency of sequence similarity. The recently provided representative sample from the PDB data bank of 838 proteins of Holm and Sander (1996) could be an excellent data set. There, all sequences have less than 25% sequence identity and fall into 270 different fold classes with class occupancy ranging between 1 and 73, see *ftp://ftp.embl-heidelberg.de/databases/fssp/*.

**Acknowledgements** The authors thank two anonymous referees for their critical but also very constructive and helpful comments. We thank Sandor Suhai and Martin Reczko for providing us the with the data and some biophysical background and for John Crowley drawing once our attention to neural nets. For stimulating discussions on statistical classification and prediction and help in applying them the second author is very grateful to Trevor Hastie. Part of this research was supported by the German Academic Exchange Service (DAAD, Doktorandenstipendium HSP II/AUFE). Finally, the first author is grateful to AMS for support to attend the Summer Research Conference in Seattle.

## REFERENCES

ANFINSEN, C., HABER, E., SELA, M. and WHITE, F.J. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the USA* **47** 1309–1314.

BAIROCH, A. and APWEILER, R. (1995) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research* **24** 21–25.

BARTON, G.J. (1995) Protein secondary structure prediction. *Current Opinions in Structural Biology* **5** 372–376.

BERGER, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

BERNSTEIN, F.C., KOETZLE, T.F., WILLIMAS, G.J.B., MEYER, E., BRYCE, M.D., ROGERS, J.R., KENNARD, O., SHIKANOUCHI, T. and TASUMI, N. (1977) The Protein Data bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology* **112** 535–542.

BENSON, D.A., BOGUSKI, M.S., LIPMAN, D.J. and OSTELL, J. (1997) GenBank. *Nucleic Acids Research* **25** 1–6.

BRANDEN, C. and TOOZE, J. (1991) *Introduction to Protein Structure.* Garland Publ., New York.

CREIGHTON, T.E. (1984) *Proteins. Structure and Molecular principles.* Freeman, New York.

CREIGHTON, T.E. (1990) Understanding protein folding pathways and mechanisms. In *Protein Folding*, Gierasch, L. M., King, J. (eds), Amer. Assoc. Adv. Sci., Washington, 157–170.

DEFAY, T. and COHEN, F.E. (1995) Evaluation of current techniques for ab initio protein prediction. *Proteins* **23** 431–445.

DICKERSON, R.E. and GEIS, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology.* Benjamin Cummings, Menlo Park CA.

DIFRANCESCO, V., GARNIER, J. and MUNSON, P.J. (1996) Improving protein secondary structure prediction with aligned homogous sequences. *Protein Science* **5** 106–113.

DINUR, U. and HAGLER, A. T. (1991) New approaches to empirical force fields. In *Reviews in Computational Chemistry*, Vol. II, Lipkowitz, K. B., Boyd, D. B. (eds). VCH Pbl., New York, 99–164.

EFIMOV, A.V. (1994) Super-secondary structures in proteins. In *Protein Structure by Distance Analysis*, Bohr & Brunak (Eds) IOS Press Amsterdam, 187–200.

EFRON, B. and TIBSHIRANI. R.J. (1993) *An Introduction to the Bootstrap.* Chapman & Hall, Cambridge.

FASMAN, G. (1989) *Prediction of Protein Structure and the Principles of Protein Conformation.* Plenum, New York.

FETROW, J.S., PALUMBO, M.J. and BERG, G. (1997) Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* **27**, 249–271.

FERRAN, E.A., FERRARA, P. and PFLUGFELDER B. (1993) Protein classification using neural networks. In: ISMB-93. Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, Hunter, L,. Searls, D., Shavlik, J. (Eds). AAAI Press, Menlo Park, CA,127–135.

FINKELSTEIN, A.V. (1997) Protein structure: what is it possible to predict now? *Current Opinion in Structural Biology* **7** 60–71.

FRIEDMAN, J.H. and STÜTZLE, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association* **76** 817–823.

FRIEDRICHS, M.S., GOLDSTEIN, R.A. and WOLYNES, P.G. (1991) Generalized protein tertiary structure recognition using associative memory Hamiltonians. *Journal of Molecular Biology* **222** 1013–1034.

GEOURJON, C. and DELEAGE, G. (1994) SOPM: A self-optimized method for protein secondary structure prediction. *Protein Engineering* **7** 157–164.

GIERASCH, L.M. and KING, J. (1990) *Protein Folding: Deciphering the Second Half of the Genetic Code.* Amer. Assoc. Adv. Sci., Washington.

GRASSMANN, J. (1996) Artificial neural networks in regression and discrimination. In *Softstat '95, Advances in Statistical Software 51*, Faulbaum, F. and Bandilla, W. (Eds). Lucius & Lucius, Stuttgart, 399–406.

GRASSMANN, J. and EDLER, L. (1996) Statistical classification methods for protein fold class prediction. In *COMPSTAT. Proceedings in Computational Statistics*, Prat A. (ed). Physica-Verlag, Heidelberg, 277–282.

GRASSMANN, J., SUHAI, S., RECZKO, M. and EDLER,L. (1998) Protein Fold Class Prediction: Statistical Classification versus Artificial Neural Networks. Manuscript submitted.

HASTIE, T. and TIBSHIRANI, R.J. (1990) *Generalized Additive Models.* Chapman & Hall, Cambridge.

HASTIE, T., TIBSHIRANI, R.J. and BUJA, A. (1994) Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89** 1255–1270.

HOLLEY, L. and KARPLUS, M. (1989) Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the USA* **86** 152–156.

HOLM, L. and SANDER,C. (1994) Searching protein structure databases has come of age. *Proteins* **19** 165–173.

HOLM, L. and SANDER, C (1996) Alignment of three-dimensional protein structures: network server for database searching. *Methods in Enzymology* **266** 653–62.

HUBBARD, T. and TRAMONTANO, A. (1996) Update on protein structure prediction: results of the 1995 IRBM workshop. *Folding and Design* **1** R55–R63.

LAMBERT, M. and SCHERAGA H. (1989) Pattern recognition in the prediction of protein structure. I-III. *Journal of Computational Chemistry* **10** 770–831.

LEMER, C.M.R., ROOMAN, M.J. and WODAK, S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23** 337–355.

MATTHEW, B. W. (1975) Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta* **405** 442–451.

MICKLER, T.A. and LONGNECKER, D. E. (1992) The immunosuppressive aspects of blood transfusion. I. *Intensive Care Medicine* **7** 176–188.

MOULT, J., JUDSON, R., FIDELIS, K. and PEDERSEN, J.T. (Eds) (1995) Large scale experiment to assess protein structure prediction methods. *Proteins* **23** ii–iv.

NEUMAIER, A. (1997) Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Reviews* **39** 407–460.

PASCARELLA, S. and ARGOS, P. (1992) A data bank merging related protein structures and sequences. *Protein Engineering* **5** 121–137.

PERUTZ, M.F. (1978) Hemoglobin structure and respiratory transport. *Scientific American* **239** 92–125.

PREDICTION CENTER (1996) Protein structure prediction center on WWW URL http://Prediction center.llnl.gov/.

RECZKO, M. and BOHR, H. (1994) The DEF data base of sequence based protein fold class predictions. *Nucleic Acids Research* **22** 3616–3619.

RECZKO, M., KARRAS, D. and BOHR, H. (1997) An update of the DEF database of protein fold class prediction. *Nucleic Acids Research* **25** 235.

RECZKO, M., BOHR, H., SUBRAMAMIAM, S., PAMIDIGHANTAM, S. and HATZIGEORGIOU, A. (1994) Fold class prediction by neural networks. In *Protein Structure by Distance Analysis*, Bohr & Brunak (Eds). IOS Press Amsterdam, 277–285.

RICHARDS, F.M. (1991) The protein folding problem. *Scientific American* **264** 54–63.

RIPLEY, B.D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

ROST, B. and O'DONOGHUE, S. (1997) Sisyphus and prediction of protein structure. *CABIOS* **13** 345–356.

ROST, B. and SANDER, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins, Structure, Function and Genetics* **19** 55–72.

ROWEN, L., MAHAIRAS, G. and HOOD, L. (1997) Sequencing the human genome. *Science* **278** 605–607.

SCHULZ, G.E. (1988) A critical evaluation of methods for prediction of protein secondary structures. *Annual Review of Biophysics and Chemistry* **17** 1–21.

SCHUMACHER, M., ROSSNER, R. and VACH, W. (1994) Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis* **21** 661–682.

SCHUMACHER, M., ROSSNER, R. and VACH, W. (1994) Neural networks and logistic regression: Part II. *Computational Statistics and Data Analysis* **21** 683–701.

SCHNEIDER, R., de DARUVAR, A. and SANDER, C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research* **25** 226–239.

STERNBERG, M.J.E. (1996) *Protein Structure Prediction.* Oxford University press, Oxford.

STOLORZ P, LAPEDES A and XIA Y (1992) Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology* **225** 363–377.

SUN, Z. and JIANG, B. (1996) Patterns and conformations of commonly occurring supersecondary structures (basic motifs) in Protein Data Bank *Journal of Protein Chemistry* **15** 675–690.

SUN, Z., ZHANG, C.-T., WU, F.-H. and PENG, L.-W. (1996) A vector projection method for predicting supersecondary motifs. *Journal of Protein Chemistry* **15** 721–729.

SUN, Z., RAO, X.-Q., PENG, L.-W. and XU, D. (1997) Prediction of protein supersecondary structure based on the artificial neuural network. *Protein Engineering* **10** 763–769.

TAYLOR, W.R. (1992) *Patterns in Protein Sequence and Structure.* Springer, New York.

WANG, Z.X. (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering* **11** 621–626.

WWW[1]: *http://www.dkfz-heidelbberg.de/biostatistics/protein/protlit.html*
WWW[2]: *http://www.dkfz-heidelbberg.de/biostatistics/protein/gsme97.html*
ZHANG. C-T. (1997) Relations of the numbers of protein sequences, families and folds. *Protein Engineering* **10** 757–761.

GERMAN CANCER RESEARCH CENTER, HEIDELBERG
BIOSTATISTICS UNIT - R0700
PO BOX 10 19 49
69009 HEIDELBERG
GERMANY
EDLER@DKFZ-HEIDELBERG.DE

BRAIN[2]
LEITENWEG 8A
97286 WINTERHAUSEN
GERMANY
JANET.GRASSMANN@BRAINN.DE