PRED-CLASS: Cascading Neural Networks for Generalized Protein Classification and Genome-Wide Applications

Claude Pasquier, Vassilis J. Promponas, and Stavros J. Hamodrakas*
Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens, Greece

ABSTRACT **A cascading system of hierarchical, artificial neural networks (named PRED-CLASS) is presented for the generalized classification of proteins into four distinct classes—transmembrane, fibrous, globular, and mixed—from information solely encoded in their amino acid sequences. The architecture of the individual component networks is kept very simple, reducing the number of free parameters (network synaptic weights) for faster training, improved generalization, and the avoidance of data overfitting. Capturing information from as few as 50 protein sequences spread among the four target classes (6 transmembrane, 10 fibrous, 13 globular, and 17 mixed), PRED-CLASS was able to obtain 371 correct predictions out of a set of 387 proteins (success rate ~ 96%) unambiguously assigned into one of the target classes. The application of PRED-CLASS to several test sets and complete proteomes of several organisms demonstrates that such a method could serve as a valuable tool in the annotation of genomic open reading frames with no functional assignment or as a preliminary step in fold recognition and ab initio structure prediction methods. Detailed results obtained for various data sets and completed genomes, along with a web sever running the PRED-CLASS algorithm, can be accessed over the World Wide Web at http://o2.biol.uoa.gr/PRED-CLASS. Proteins 2001;44:361–369.** © 2001 Wiley-Liss, Inc.

Key words: protein classification; artificial neural network; transmembrane, fibrous, globular, and mixed proteins; genome annotation; genome-wide analysis

INTRODUCTION

The prediction of protein tertiary structures from information contained in amino acid sequences remains a challenging problem in structural molecular biology, even though great progress has been achieved during the last few years.[1–3] Over the last 3 decades, several computational methods have been developed for the prediction of one-dimensional (1D) structural features of proteins from their sequences alone. For example, secondary structure prediction schemes aim to propose (sometimes with noteworthy success[4–6]) the probable locations of secondary structure elements. In many cases, they are reported to be a fundamental initial stage (or a refining final stage) for ab initio calculations,[7] fold recognition,[8,9] or homology modeling techniques.[10] Machine learning techniques have been a common practice to mine the information hidden in the vast amount of protein sequences resulting from completed and ongoing genome projects, combined with available experimental functional or structural knowledge or information. In particular, different types of artificial neural network (NN) predictors have often served as a powerful tool for this.[11–14] However, algorithms to predict generalized topological features of protein molecular structures could prove to be useful tools as a preliminary step in protein structure prediction, functional class determination, or both.

We recently published a simple NN-based classification scheme (PRED-TMR2[15]) to distinguish between *transmembrane* (TM) and globular, water-soluble proteins, integrated with a previously reported method (PRED-TMR[16]) for the fast and accurate detection of TM segments. We now extend the capabilities of the system by further classifying nontransmembrane proteins into three classes: *fibrous* (FIBR; e.g., collagen and elastin), *globular* (GLOB; e.g., various types of enzymes), and a last group of proteins composed of both FIBR and GLOB domains, called hereinafter *mixed* (MIX) proteins (e.g., several intermediate filament proteins).

In this work, we illustrate the learning procedure followed (NN training) and the results obtained on several sets of known proteins for an estimation of the classification error rate. The method's predictive power combined with reasonable time performance on currently available complete proteome sets (e.g., the recently sequenced *Drosophila melanogaster* genome[17]) indicates that PRED-CLASS could be a precious source of information in the automatic or manual annotation of genomic orphan open reading frames (ORFs; ORFans[18]). The generalized classification obtained by the method suggests that PRED-CLASS could be useful as a starting point in fold recogni-

Abbreviations: 1D, one-dimensional; C, any of the FIBR, GLOB, MIX, and TM classes; FIBR, fibrous; FFT, fast Fourier transform; f_{nc}, number of proteins belonging to class C falsely classified in another class; GLOB, globular; MIX, mixed; NN, neural network; ORF, open reading frame; SCOP, structural classification of proteins; SEL, selectivity; SENS, sensitivity; SProt: SwissProt; TM, transmembrane; t_{pc}, number of proteins belonging to class C correctly classified in C.

Grant sponsor: European Commission—Training and Mobility Researchers; Grant number: ERBFMRXCT960019.

*Correspondence to: Stavros J. Hamodrakas, Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 15701, Greece. E-mail: shamodr@cc.uoa.gr

Received 17 November 2000; Accepted 23 April 2001

TABLE I. SProt Identification Codes for Sequences Composing the Training Set Used for NN2 and NN3

FIBR	GLOB	MIX
CA16_HUMAN	ADHA_UROHA	ABP2_HUMAN
CA1E_HUMAN	AZU1_METJ	AINX_RAT
CCC4[a]	C550_THIPA	DESM_CHICK
CH18_DROVI	CATA_YEAST	FILS_BOVIN
CH19_DROGR	CTR2_VESCR	GFAP_HUMAN
ELS_CHICK	DYRA_CITFR	IF3T_TORCA
FBOH_BOMMO	HBA1_ARCGA	IFE_BRALA
KR2A_SHEEP	IGJ_HUMAN	ION3_CARAU
KRB2_SHEEP	LYC1_CAPHI	LAM1_HUMAN
TPM1_CHICK	RNPH_BACSU	NEST_RAT
	RUB1_PSEOL	NF60_LOLPE
	THTR_AZOVI	NFM_MOUSE
	TRYP_ASTFL	PERI_RAT
		PLST_CARAU
		TANA_XENLA
		VIM4_XENLA
		XNIF_XENLA

[a]CCC4 has not been deposited in SProt. It is a structural protein found in the egg shell of the fruit fly *Ceratitis capitata.*[19]

tion or ab initio structure prediction methods, and, in combination with comparative studies on completed genomic sequences, it could give further insight into the evolution of protein structure and function.

MATERIALS AND METHODS

Information Gathering

A set of 11 proteins (6 TM, 2 FIBR, and 3 GLOB) with known structural characteristics was used for the training of the first network, as described in detail elsewhere.[15] Another set of 40 proteins (10 FIBR, 13 GLOB, and 17 MIX; see Table I) was selected for the learning process of the newly employed NNs. This made a total of only 50 protein sequences, if we consider that the sequence with corresponding SwissProt (SProt)[20] ID ELS_CHICK was used in the training of all three networks.

Another set of 387 protein sequences reliably assigned to the four target classes (147 TM, 73 FIBR, 55 GLOB, and 112 MIX) served as a test data set to evaluate the predictive power of the system.

The nonredundant set of 148 integral membrane proteins recently compiled by Möeller et al.[21] served as an ideal representative set of well-characterized prediction targets of the TM class. For classification purposes, no detailed information concerning the location or the orientation of membrane spanning segments was necessary. One protein used in the training process of the first component network (SProt ID: LECH_HUMAN) was present in this set and was, therefore, removed. This set, which was larger than the initial test set on which PRED-TMR2 was tested, was used to gain a more realistic approximation of the performance of the first network classifier.

The 55 representative GLOB proteins are a subset of the 65 well-known GLOB proteins collected by Levitt and Greer:[22] they are typical GLOB proteins varying in sequence and functional characteristics. All 10 sequences present in the training set were removed as well.

A set of 130 sequences belonging to the MIX class was extracted from SProt (release 35); we were looking for entries containing all of the following keywords in their feature (FT) fields: HEAD, ROD, and TAIL, which correspond to typical domain definitions in known MIX structures. All 17 sequences used for the training phase, as well as an incomplete sequence (SProt ID IFEA_HELPO), were also removed.

The collection of FIBR proteins was based on expert knowledge deposited in the existing literature and is composed of typical representatives of this class, such as different collagen types, noncytoskeletal keratins, elastins, and FIBR chorion proteins, all collected from SProt release 35.

Retrieving representative sets of nonhomologous FIBR and MIX protein sequences was not an easy task because of the low number of proteins in these classes deposited in the public databases. In addition, no safe automatic way exists to extract FIBR protein sequences from public databases, as there are no specific and selective keywords. As far as the MIX class is concerned, performing a search with the keywords HEAD, ROD, and TAIL within the FT fields in SProt release 39 yielded just 55 more protein sequences. For this reason, we decided to retain all protein sequences from SProt release 35 that, to our knowledge, could fit into these two classes so that the test could be as general as possible, not taking possible sequence similarity into account. As more data become available, further testing on larger data sets should be performed.

We also considered that another useful test would be to screen the classification results in the context of the SCOP (structural classification of proteins) classification scheme of protein domains.[23] On these grounds, we chose the ASTRAL[24] subset of SCOP (version 1.48) sequences with less than 40% pairwise similarity. This set is composed of 2619 sequences of protein domains classified (in the first level of the SCOP, version 1.48, hierarchy) into seven classes: 484 all-α, 602 all-β, 632 α/β, 568 $\alpha+\beta$, 48 multidomain, 55 membrane and cell surface proteins and peptides, and 230 small proteins.

Protein sequence data collected for the training and evaluation phases of the PRED-CLASS system are available over the World Wide Web at http:/o2.biol.uoa.gr/PRED-CLASS.

System Architecture and Component NN Topology

The overall classification system consists of three successive, multilayer, feedforward (acyclic) artificial NNs (Fig. 1), each one with a single hidden layer at which the computation takes place. Some common features shared by all three NNs are the following:

1. There is full connectivity, as every node in each network layer is connected to every other node in the adjacent forward layer.
2. There is a small number (two, three, and three, respectively) of nodes in the hidden layer responsible for the actual learning process carried out by each component network.

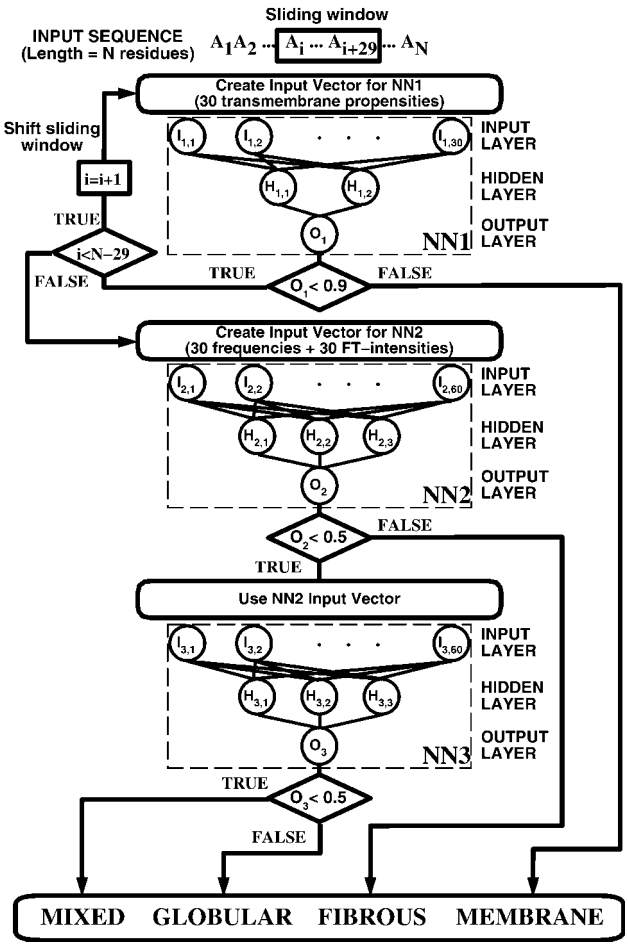

Fig. 1. PRED-CLASS architecture: individual component NNs and their layered structure. The input type for each subsystem, network connectivities, information flow, and decision scheme for the output layer of each NN are indicated.

3. The activation function on each node is a nonlinear, sigmoid logistic function[25] of the weighted sum of all synaptic weights (plus a constant bias not shown in Fig. 1). Consequently, output values are in the 0–1 range.

The NN used for the discrimination of TM and nontransmembrane proteins (NN1) has an input layer of 30 neurons. For each sequence, overlapping sliding windows of 30 residues are iteratively used to create input vectors (Fig. 1), whose components are the TM propensities of the corresponding residue types.[15,16] The desired output is 1 or 0 depending on the observation that this segment is related to a TM region or not. Output values larger than an empirically estimated threshold (0.9) prompt toward the classification in the membrane class, whereas lower values forward the protein in the next level of classification.

In addition, NN2 and NN3 have an input layer of 60 nodes, which means that 60 parameters for each sequence are required to feed the network during the learning or classification operation. Before the learning process, all network synaptic weights are initialized to small random values.

With backpropagation[26] as the learning algorithm for each NN, two passes of computation are required during the training stage: a forward pass, where all synaptic weights remain unaltered and network signals are computed on each neuron in an hierarchical order by the activation function. Relative entropy (or the Kulback–Liebler distance) between the target and the true distribution on the output nodes serves as an objective function, as it is the most appropriate in cases of dichotomous output. In fact, the relative entropy cost function corresponds to the negative log likelihood if the observed Boolean output value is a probabilistic function of the input instance. Thus, our effort to look for the set of model parameters (network synaptic weights) that maximize the likelihood of the data, given the model, is essentially identical to that for minimizing the relative entropy function. Some useful characteristics of the relative entropy function are the following:

1. Its minimum value for binary target response values is zero, and it occurs when the target and true distributions are identical.
2. It can be proved (data not shown) that it depends on the relative output errors and not on their absolute values, which means that it gives the same weight to small and large values.

The target of this phase is to minimize the relative entropy between the target and true distribution, with iterative adjustment of the network synaptic weights, according to a gradient descent strategy. Several sequential iterations with a randomized order of presentation of training examples might be necessary until the network is considered to have converged, stabilizing the error rate below a desired threshold, ideally leading to a global minimum of the objective function.

An important issue in the design of an NN classification system is the network's generalization, that is, its ability to give correct predictions when it is presented with unseen examples. With a small number of training samples and a relatively large number of synaptic weights, there is always the possibility that the network's free parameters will adapt to the special features of the training data (overfitting). A straightforward way to overcome this problem is to use a sufficient number of training examples (usually more than 30 times the number of adjustable network parameters). When this is not possible, a number of techniques have been proposed to control overfitting. In our application, we employed both nonconvergent (early stopping) and convergent (fixed penalty/weight decay) methods. In the case of the early stopping method, the training process is stopped before the optimization procedure has finished. We follow the common method, which is to withhold and use part of the training data (50, 20, and 20% for the three networks) as an internal validation set. Training is stopped at the point at which the classification error on the holdout subset begins to rise. To overcome the dependence of the exact point of stopping on the actual split of the initial data set, this task is completed in two

TABLE II. System Performance on a Test Set of 387 Protein Sequences

	Predicted					
	TM	FIBR	GLOB	MIX	Total observed	SEL (%)
Observed						
TM	139	0	8	0	147	94.5
FIBR	1	72	0	0	73	98.6
GLOB	0	1	54	0	55	98.2
MIX	3	3	0	106	112	94.6
Total predicted	143	76	62	106	387	
SENS (%)	97.2	94.7	87.1	100.0		

steps. In the first step, the data set is split several times (five times in our case), and each time a model is fitted and tested, the value of the error function at the point that stopping is invoked is stored. In the second step, the complete training set is used to train the final model; it is stopped when the error criterion reaches the mean value obtained in the previous step. The weight decay method assigns a penalty on large weights, forcing the network's regression surface to be smooth.

By keeping the number of synaptic weights as small as possible, the training process leads more quickly to convergence because fewer free parameters have to be readjusted for optimization in each backpropagation cycle. At the same time, the number of training examples required to achieve good generalization (small fraction of classification errors) can be reduced, as it is reported to be proportional to the number of free parameters.[27]

In the prediction phase, just like the forward pass in learning, network weights are globally fixed (those obtained after the convergence of the training process), and the NN is presented with an unknown example for classification. In the same hierarchical manner, the input signal propagates once in the forward direction, and the output value constitutes the network's decision based on the already studied training examples.

Training Parameters

In our study, those proteins classified as nonmembrane by NN1 are candidates for classification by NN2. After a careful inspection of the sequential and structural characteristics of several FIBR, GLOB, and MIX proteins (data not shown), the following 60 input parameters were chosen as appropriate input to NN2:

1. Thirty values corresponding to the composition of the examined sequence in all 20 residue types and 10 different groupings of residues sharing common structural and/or physicochemical properties. The amino acid groupings used in this study were the following: AVLIFWDEQMHK (α-helix formers),[28] VLIFWYTCQM (β-sheet formers),[28] GPDNSCKWYQTRE (β-turn formers),[28] CVILMFYWAP (hydrophobic), DEHKRSTNQ (polar), HRKDE (charged), HRK (positively charged), DE (negatively charged), HFWY (aromatic), and VLIA (aliphatic).
2. Thirty values corresponding to the highest intensity for periodicities detected for each residue or group type by a fast Fourier transform (FFT) algorithm.[29] The implementation of the FFT algorithm within this method is applied with the default parameters, encoding protein sequences in a numerical string of 0s and 1s to note the absence or presence, respectively, of any desired residue type in a specific position in the examined sequence. Higher intensities for a particular amino acid type (or group) suggest the existence of an underlying periodic pattern in which this residue type is involved.

These two types of parameters clearly reflect patterns of composition as well as relative distributions of amino acid types among sequences and repetitive elements. Although such an approach might seem rather simplified, NNs are capable of capturing subtle patterns in available data and succeed in recognizing weak underlying signals buried in the examined data.

The output neuron of NN2 is considered activated (or fired in the NN terminology) when its value (O_2) is greater than or equal to 0.5, indicating a positive case of an FIBR protein. In the opposite case ($O_2 < 0.5$), the sequence is further examined by NN3, which has exactly the same topology as NN2 and accepts parameters identical to those of NN2. The result $O_3 \geq 0.5$ in the NN3 output node indicates a case of a GLOB protein; otherwise, the protein is classified as MIX.

RESULTS AND DISCUSSION

Results on an Evaluation Test Set of 387 Proteins

To assess the performance of the system, we applied several tests. As described in a previous work,[15] NN1 demonstrated a perfect performance on a set of 101 nonhomologous TM proteins (101 correct predictions) and a very good performance in a subset of PDB_SELECT[30] composed of 995 GLOB proteins (97.7% correct assignments to the GLOB class). We created a new test set with a total of 387 well-characterized protein sequences from all four target classes (147 TM, 73 FIBR, 55 GLOB, and 112 MIX), as described in the Materials and Methods section, to evaluate the performance of the new integrated system.

The overall performance of the classification scheme is approximately 96% (371 correct assignments), and a summary of the results obtained by PRED-CLASS is presented in Table II. A first measure of the method's success is its selectivity (SEL) for each particular class C (FIBR, GLOB, MIX, and TM):

TABLE III. Distribution of PRED-CLASS Predictions Among the Seven Classes in the First Level of the SCOP Hierarchy Based on the SCOPlt40pc Set

	TM	FIBR	GLOB	MIX	Total
All-α	15	9	447	13	484
All-β	1	12	588	1	602
α/β	16	4	610	2	632
α + β	11	4	553	0	568
Multidomain	3	0	44	1	48
Membrane and cell surface	28	0	27	0	55
Small	0	43	187	0	230
Total	74	72	2456	17	2619

$$SEL_c = 100*t_{pc}/(t_{pc} + f_{nc}) \quad (1)$$

SEL_c represents the percentage of correct assignments in each class [true positive hits (t_{pc})] compared with the total members of the class [true positives plus false negatives ($t_{pc} + f_{nc}$)], where f_{nc} is the number of proteins belonging in class C but erroneously classified in another class. A higher selectivity of 98.6% was obtained for the FIBR class, and the lowest (94.5%) was obtained for the TM class, yielding a mean selectivity of 96.5%. The first classifier, which decided whether a protein sequence corresponded to a TM protein or not, was trained on as few as 11 sequences.

The sensitivity (SENS) of classification for each class C is also a useful evaluation criterion:

$$SENS_c = 100*t_{pc}/(t_{pc} + f_{pc}) \quad (2)$$

$SENS_c$ is the percentage of correct assignments in each class compared with the total number of assignments in this class [true positives plus false positives ($t_{pc} + f_{pc}$)], where f_{pc} is the number of erroneous assignments in class C. Along these lines, classification in the MIX class was 100% sensitive, yielding no false predictions in the MIX class, whereas the lowest sensitivity was obtained for the GLOB class (87.1%), with a mean sensitivity value of 94.7%. This result, however, could be rather misleading because the GLOB class apparently is the most abundant in the protein universe, whereas in our test set this class of proteins is somehow underrepresented and a more sensitive performance of the GLOB classifier should be expected. Additionally, because no MIX proteins were used as negative examples in the training phase of the first NN, it seems sensible that three out of the four false assignments in the TM class come from MIX proteins. A maximal collection of proteins in the FIBR, GLOB, and MIX classes, removing redundancy as in ref. 21, would certainly serve as a valuable set for training and evaluating classification systems such as PRED-CLASS.

Case by Case Study of Erroneous Classifications

Wrong classifications in this evaluation phase were examined manually for a better understanding of the method's performance. Half of the 16 erroneous predictions were for 8 TM proteins classified in the GLOB class (SProt IDs: COXK_BOVIN, COXD_BOVIN, TONB_ECOLI, VS10_ROTBN, GEF_ECOLI, FDOH_ECOLI, BCS1_YEAST, and NNTM_BOVIN). Half of them (COXK-_BOVIN, COXD_BOVIN, BCS1_YEAST, and NNTM_BOVIN) are bound to the mitochondrion membrane, whereas 2 of them (TONB_ECOLI and GEF_ECOLI) contain a signal membrane-anchored N-terminal sequence.

Another erroneous classification occurred for an FIBR collagen-type IV precursor (CA44_HUMAN) possessing a signal N-terminal sequence ranging from residue 1 to 38, incorrectly classified as TM. A careful inspection of the output of the first NN reveals that the segment responsible for the wrong classification is located exactly in this region. After this signal peptide is removed from the sequence, it is correctly classified in the FIBR class.

A total number of six false predictions for MIX proteins were obtained, three of them being type II cytoskeletal keratins (K2CA_HUMAN, K2CB_HUMAN, and K2CF_HUMAN) misclassified in the TM class. These sequences are almost identical (>98% pairwise identity), and the false classification is due to a long segment rich in glycine and hydrophobic residues. The remaining three wrongly predicted MIX proteins (K1CJ_BOVIN, K2M2_SHEEP, and NFH_MOUSE) were assigned to the FIBR class.

A subtle case was the only GLOB protein that was classified in the FIBR class: ferredoxin. Its characteristic cysteine-rich, iron-binding domains exhibit a periodicity of approximately 3.4 residues, and when three cysteines are artificially mutated into another residue type, ferredoxin is correctly classified in the FIBR class, indicating that a probable reason for this wrong assignment is both the overall composition and the detected periodic patterns.

Results in the Context of the SCOP Classification of Protein Domains

A rather qualitative test of the method was performed against the sequences of protein domains as classified in the SCOP database,[23] with the ASTRAL[24] subset (see the Materials and Methods section), which is freely accessible via the Internet, chosen as the most appropriate resource for this task. The chosen threshold of sequence similarity in this data set ensures that, on the one hand, most of the redundancy (in terms of sequence similarity) is removed, and, on the other hand, several sequence representatives exist in most of the classification levels. This data set

TABLE IV. PRED-CLASS Results on 30 Complete Proteome Sets[†]

Species	TM	FIBR	GLOB	MIX
Bacterial				
Aquifex aeolicus	21.9	0.1	72.8	5.2
Bacillus subtilis	26.1	0.2	68.9	4.7
Borrelia burgdorferi	25.6	0.2	70.5	3.7
Campylobacter jejuni	25.4	0.1	70.9	3.6
Chlamydia muridarum	26.7	0.3	67.2	5.7
Chlamydia pneumoniae	29.6	0.5	65.2	4.7
Chlamydia trachomatis	25.3	0.1	69.1	5.5
Deinococcus radiodurans	20.2	3.0	73.3	3.5
Escherichia coli	23.8	0.5	72.7	3.0
Haemophilus influenzae	21.3	0.2	74.6	3.8
Helicobacter pylori	22.6	0.1	71.5	5.7
Helicobacter pylori J99	23.2	0.3	71.0	5.4
Mycobacterium tuberculosis	21.3	4.6	71.9	2.1
Mycoplasma genitalium	27.1	0.0	68.7	4.1
Mycoplasma pneumoniae	22.8	0.0	72.0	5.1
Rickettsia prowazekii	29.7	0.1	66.8	3.3
Synecocystis sp. (strain PCC 6803)	26.7	0.5	69.4	3.4
Thermotoga maritima	23.8	0.3	70.4	5.5
Treponema pallidum	24.5	1.4	68.5	5.6
Ureaplasma parvum	24.4	0.2	73.1	2.3
Xylella fastidiosa	19.9	1.1	76.7	2.3
Mean	24.4	0.6	70.7	4.2
Standard deviation	2.7	1.1	2.7	1.2
Archaeal				
Aeropyrum pernix	18.3	18.2	61.2	2.3
Arcaeoglobus fulgidus	21.3	0.2	75.2	3.2
Methanobacterium thermoautotrophicum	20.9	0.3	73.6	5.1
Methanococcus jannaschii	20.5	0.1	76.7	2.7
Pyrococcus abyssi	23.3	0.3	72.8	3.6
Pyrococcus horikoshii	27.4	2.0	67.3	3.2
Mean	21.9	3.5	71.1	3.3
Standard deviation	2.8	6.6	5.3	0.9
Eukaryotic				
Caenorhabditis elegans	38.2	0.8	17.1	43.8
Drosophila melanogaster	24.6	3.8	53.0	18.6
Saccharomyces cerevisiae	28.7	0.5	55.2	15.6
Mean	30.5	1.7	41.8	26.0
Standard deviation	5.7	1.5	17.5	12.6
Overall mean	24.5	1.3	67.9	6.2
Overall standard deviation	3.8	3.3	10.8	7.8

[†]The percentage of proteins predicted to belong in the TM, FIBR, GLOB, and MIX classes are listed.

(SCOPlt40pc) contains 2619 sequences of protein domains, whereas the entire SCOP (version 1.48) contains 21,828 protein domains. The distribution of PRED-CLASS predictions among the seven classes of the first level of the SCOP hierarchy is presented in Table III. Although most sequences in this test set correspond to GLOB water-soluble protein domains, there are several points to make without consideration of further details:

1. Out of the 2286 protein domains classified in the all-α, all-β, α/β, and α+β classes in the SCOP classification scheme, 2198 (~96%) were predicted to belong to the GLOB class. Only 43 of them (1.9%) were predicted to be TM, 29 (1.3%) were predicted to be FIBR, and 16 were predicted to be MIX (0.7%).
2. Membrane or cell surface protein domains are predicted to belong either to the TM or GLOB classes (possible for membrane-anchored protein domains)
3. Most of the domains classified by PRED-CLASS in the FIBR class correspond to small protein domains, which are usually domains rich in disulfide bonds. No small proteins have been classified in the TM or the MIX class.

A thorough analysis of the PRED-CLASS predictions against all SCOP levels of hierarchy is currently in progress, with much attention paid to cases where PRED-CLASS predictions are not consistent within the same fold, family, or even superfamily.

Genome-Wide Application of PRED-CLASS

With the great impact of genomic sequences in public databases, the main goal of all efforts to devise reliable predictions, on aspects of protein structure and function, is the tractability of the developed methods to handle huge data sets with efficiency, with respect to computational time and resources. Our method's time performance is rather reasonable, and no external parameters need to be defined by the end user, which makes automatic genome-wide predictions with PRED-CLASS possible. For example, the recently sequenced genome of the fruit fly *D. melanogaster*,[17] composed of 13,604 protein sequences, was analyzed overnight on a Silicon Graphics O2 workstation with 128 MB of main memory and one 300-MHz R5000 processor.

Thirty completed, nonredundant proteome sets were obtained from the Proteome Analysis web site[31] at the European Bioinformatics Institute server (http://www.ebi.ac.uk/proteome/). The organisms whose proteomes have been analyzed ranged from simple archaea and bacteria to complex eukaryotic multicellular organisms, such as *D. melanogaster*, and the results are summarized in Table IV. A significant proportion of these sequences does not show clear sequence similarity to proteins of known structure or function and, consequently, remain uncharacterized.

Even if no detailed annotation can be obtained for these unique protein sequences, it is very interesting to answer the following questions: What is the (approximate) expected frequency of proteins in the TM, FIBR, GLOB, and MIX classes in a newly sequenced genome? How are these frequencies correlated to the taxonomy of the studied organism and/or specific environment factors or cellular processes? Previous studies raised the question of whether the proportion of membrane proteins encoded in a genome is correlated to its size or not[32,33] with contrasting conclusions.

According to the predictions obtained with PRED-CLASS, even when the frequencies of each class of proteomes of the same kingdom are averaged [Fig.

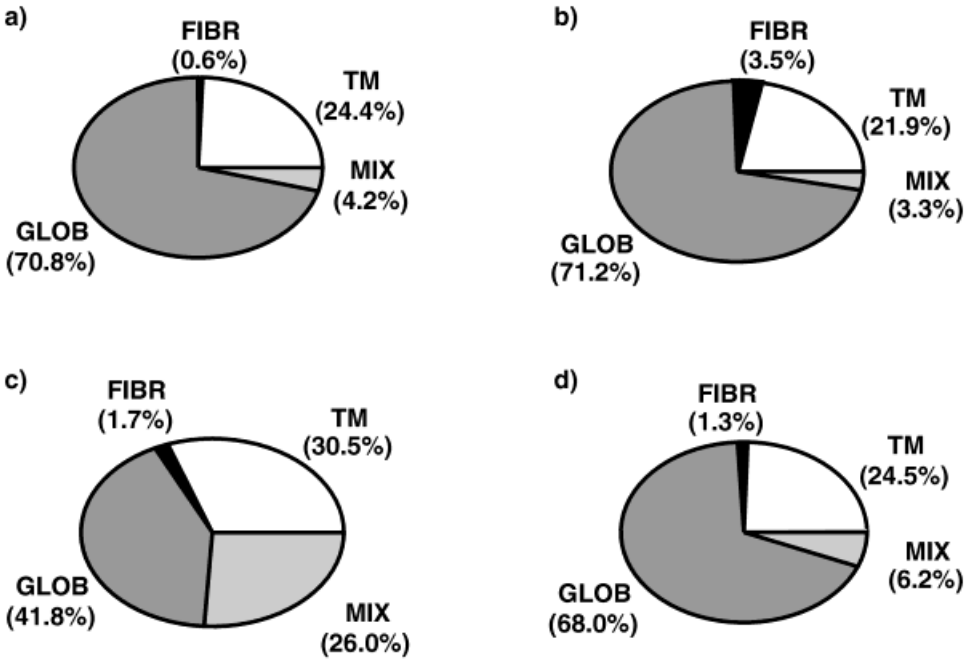

Fig. 2. Mean proportions of proteins encoded (a) in 21 complete bacterial genomes classified in the TM, FIBR, GLOB, and MIX classes as predicted by PRED-CLASS; (b) in 6 complete archaeal genomes classified in the TM, FIBR, GLOB, and MIX classes as predicted by PRED-CLASS; (c) in 3 complete eukaryotic genomes classified in the TM, FIBR, GLOB, and MIX classes as predicted by PRED-CLASS; and (d) in all 30 complete genomes analyzed in this work classified in the TM, FIBR, GLOB, and MIX classes as predicted by PRED-CLASS.

2(a–c)], the standard deviations calculated are rather high (see Table IV) to support a generalized hypothesis. This apparently leads to the conclusion that there is substantial variation even within the same domain of life, and further phylogenetic relations as well as organism-specific information should be considered before any hypothesis is suggested. For example, if we examine the frequency of predicted TM proteins in bacteria, where frequency values range between 19.9% (*Xylella fastidiosa*) and 29.7% (*Rickettsia prowazekii*), in agreement with ref. 32, the mean and median of the bacterial distribution are equal (both 24.4%) and close to the mean frequency of the complete set of proteomes (24.5%), whereas the extreme values lie approximately two standard deviations from the mean, indicating a normal-like distribution.

Most of the proteomes of archaeal and eukaryotic organisms studied in this work fall between the limits imposed by the bacterial distribution for TM proteins, with the only sound exception being *Caenorhabditis elegans*, in which about 38% of the genes are predicted to encode for TM proteins. This extremely high value cannot be counted as an artifact of our method because it is in agreement with previous studies[32,33] following different strategies for this characterization. Another strange finding for *C. elegans* is the fairly large number of proteins predicted to belong to the MIX class (43.8%) combined with a very small incidence of predicted GLOB proteins (17.1%). These values are the maximum and minimum frequencies in the respective classes among all genomes analyzed in this study. Apparently, a detailed study of the *C. elegans* genome is necessary to explain these findings, combining several types of computer-based predictions with experimental knowledge.

CONCLUSIONS

The classification of protein structure and function is a major goal in structural molecular biology, aimed at understanding the principles that govern the folding procedure when linear amino acid chains fold to a three-dimensional structure, adapting a preferable shape for their desired function. Here, we have demonstrated PRED-CLASS, a robust system of simple artificial NNs for the generalized classification of proteins with information encoded in single amino acid sequences. Although protein sequence information is submitted to the system in machine-friendly numerical representations, the underlying principles are based on well-known attributes of protein structure.

The high overall prediction accuracy (~96%) indicates that PRED-CLASS could be pertinent for single-sequence or genome-wide analysis, either as a stand-alone application or as a part of a more elaborate computational analysis. PRED-CLASS could be used as an assisting tool in genome functional annotation projects, where some potential functions for a protein sequence become more possible and others are excluded after a correct prediction in one class or another. Such

generalized predictions could prove to be valuable in the primary stages of threading methods or ab initio protein structure prediction. In the former, predictions could reduce the number of folds possibly compatible with an examined sequence, whereas in the latter case structural constraints (e.g., the existence of coiled-coil regions in FIBR or MIX proteins) are indicated, drastically improving the execution time and performance of such approaches.

As a stand-alone method, PRED-CLASS has been used to analyze several complete proteomes from all domains of life (archaea, eubacteria, and eukaryotes). It is possible that there exists a tendency in eukaryotic complete genomes to code a larger proportion of TM proteins, in agreement with a previously reported work.[32] This finding could be explained by the invention of new protein functions through evolutionary mechanisms to accommodate the need of multicellular organisms for communication with the cell environment. However, because of the insufficient complete genome information in this domain of life and the lack of appropriate experimental evidence, it may be too early to make such an assumption.

Plans for future work in the improvement and further extension of the proposed system are already in progress, resulting from some weaknesses of the method (described in the Results and Discussion section) and the emerging need for divergent accurate predictors. TM protein detection could apparently gain in sensitivity through the filtration of secretory signal peptides in a preprocessing stage. It is highly recommended that a computer program such as SignalP[34] be used to identify signal peptides so that they are removed from the sequence before PRED-CLASS classification. A PRED-CLASS run against all the nontransmembrane proteins of SProt release 39 with signal peptides revealed that improved sensitivity in TM protein detection can be achieved if this strategy is followed (data not shown). Selectivity could be improved by the retraining of NN1 with mitochondrial TM proteins as well.

An important class of outer membrane proteins,[35] in a great majority, are wrongly classified in the GLOB class. However, this prediction is essentially correct because quite a few of them are β-barrels.

A feature useful to end users would be a reliability index accompanying each prediction so that the quality of individual predictions could be known before further analysis steps are decided. The main problem with this task is the availability of large data sets of proteins with reliable characterization in all four classes.

Although the time performance of the system's current version is adequate, a parallel implementation of PRED-CLASS would make genome-wide predictions faster by an estimated factor of 1.5–2. A novel fold-class prediction method for globular, water-soluble proteins is currently in the final development and testing stages and is intended to be integrated into the PRED_CLASS system.

ACKNOWLEDGMENTS

The authors thank Professor E. Moudrianakis for a critical reading of the manuscript and the anonymous referees for their valuable comments.

REFERENCES

1. Finkelstein AV. Protein structure: what is it possible to predict now? Curr Opin Struct Biol 1997;7:60–71.
2. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins 1999;Suppl 3:88–103.
3. Koehl P, Levitt M. A brighter future for protein structure prediction. Nat Struct Biol 1999;6:108–111.
4. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O. Prediction of protein secondary structure at 80% accuracy. Proteins 2000;41:17–20.
5. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:192–202.
6. Rost B. Better 1D predictions by experts with machines. Proteins 1997;Suppl 1:192–197.
7. Pedersen JT, Moult J. Ab initio protein folding simulations with genetic algorithms: simulation on the complete sequence of small proteins. Proteins 1997;Suppl 1:179–184.
8. Murzin AG. Distant homology recognition using structural classification of proteins. Proteins 1997;Suppl 1:105–112.
9. Rice DW, Fischer D, Weiss R, Eisenberg D. Fold assignments for aminoacid sequences of the CASP2 experiment. Proteins 1997; Suppl 1:113–122.
10. Sanchez R, Šali A. Evaluation of comparative protein structure modeling by MODELLER 3. Proteins 1997;Suppl 1:50–58.
11. Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10:53–60.
12. Fariselli P, Casadio R. HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. Comput Appl Biosci 1998;12:41–48.
13. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins 1999;36:340–346.
14. Aloy P, Cedano J, Olivia B, Aviles X, Querol E. 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. Comput Appl Biosci 1997;13: 231–234.
15. Pasquier C, Hamodrakas SJ. An hierarchical neural network system for the classification of transmembrane proteins. Protein Eng 1999;12:631–634.
16. Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng 1999;12:381–385.
17. Adams MD, Celniker SE, et al. The genome sequence of *Drosophila melanogaster*. Science 2000;287:2185–2195.
18. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics 1999;15:759–762.
19. Vlahou D, Konsolaki M, Tolias P, Kafatos FC, Komitopoulou M. The autosomal chorion locus of the medfly *Ceratitis capitata*. I. Conserved synteny, amplification and tissue specificity but sequence divergence and altered temporal regulation. Genetics 1997;147:1829–1842.
20. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
21. Möeller S, Krivencheva EV, Apweiler R. A collection of well characterised integral membrane proteins. Bioinformatics 2000; 16:1159–1160.
22. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J Mol Biol 1977;114:181–239.
23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247: 536–540.
24. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for

protein structure and sequence analysis. Nucleic Acids Res 2000; 28:254–256.
25. Minsky M, Papert, S. Perceptrons: an introduction to computational geometry. Cambridge, MA: MIT Press; 1968.
26. Rumelhart DE, Hinton GE, Williams RJ. In: Rumelhart D, McClelland J, Group PR, editors. Parallel distributed processing: explorations in the microstructure of cognition. Volume 1, Foundations. Cambridge, MA: MIT Press; 1986. p 318–362.
27. Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 1999.
28. Hamodrakas SJ, Etmektzoglou T, Kafatos F. Amino acid periodicities and their structural implications for the evolutionarily conservative central domain of some silkmoth chorion proteins. J Mol Biol 1985;186:583–589.
29. Pasquier C, Promponas VJ, Varvayannis NJ, Hamodrakas SJ. A web server to locate periodicities in a sequence. Bioinformatics 1998;14:749–750.
30. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
31. Apweiler R, Biswas M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Oin T, Phan I, Zdobnov E. Proteome analysis: application of InterPro and CluSTr for the functional classification of proteins in whole genomes. In: Bornberg-Bauer E, Rost U, Stoye J, Vingron M, editors. Proceedings of the German Conference on Bioinformatics 2000. Berlin: Logos Verlag; 2000. p 149–157.
32. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7:1029–1038.
33. Stevens TJ, Arkin IT. Do more complex organisms have a greater proportion of membrane proteins in their genomes? Proteins 2000;39:417–420.
34. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
35. Schulz GE. β-barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.