WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Predicting experimental properties of integral membrane proteins by a naive Bayes approach

Antonio J. Martin-Galiano, Pawel Smialowski, and Dmitrij Frishman*

Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

ABSTRACT

Integral membrane proteins (iMPs) are challenging targets for structure determination because of the substantial experimental difficulties involved in their sample preparation. Accordingly, success rates of large-scale structural genomics consortia are much lower for this class of molecules compared to globular targets, underscoring the pressing need for predictive strategies to identify iMPs that are more likely to overcome laboratory bottlenecks. On the basis of the target status information available in the TargetDB repository, we describe the first large-scale analysis of experimental behavior of iMPs. Using information on recalcitrant and propagating iMP targets as negative and positive sets, respectively, we present naive Bayes classifiers capable of predicting, from sequence alone, those proteins that are more amenable to cloning, expression, and solubilization studies. Protein sequences are represented in the space of 72 features, including amino acid composition, occurrence of amino acid groups, ratios between residue groups, and hydrophobicity measures. Taking into account unequal representation of main taxonomic groups in the TargetDB, sequence database had a beneficial effect on the prediction results. The classifiers achieve accuracies of 70%, 63–70%, and 61% in predicting the amenability of iMPs for cloning, expression, and solubilization, respectively, thus making them useful tools in target selection for structure determination. Our assessment of prediction results clearly demonstrates that classifiers based on single features do not possess acceptable discriminative power and that the experimental behavior of iMPs is imprinted in their primary sequence through relationships between a restricted set of key properties. In most cases, sets of 10–20 protein features were found actually relevant, most notably, the content of isoleucine, valine, and positively-charged residues.

Proteins 2008; 70:1243–1256.

Key words: machine learning; protein expression; protein solubility; structural genomics; target selection.

INTRODUCTION

The availability of a high-resolution, three-dimensional structure of a protein is an important prerequisite for achieving a complete understanding of its function and, ideally, obtaining structural information for each and every protein fold would be highly desirable. However, experimental limitations have led to a marked bias toward solving easily tractable, small globular architectures, as reflected by the over-representation of these entities in structural databases (PDB,[1] SCOP,[2] CATH[3]). Large-scale structural genomics projects emerged in the end of the last decade as a worldwide concerted initiative aimed at revealing at least one representative structure for every fold existing in nature[4] and eventually at producing enough structures to cover the entire sequence space with structural models of sufficient quality.[5,6] At the onset of the global structural genomics effort, many consortia pursued those proteins more amenable to structural characterization.[7] However, as the "low hanging fruit" of structural biology gets depleted and at the same time available experimental techniques improve, the focus of many researchers is being shifted toward more challenging and under-represented "high-hanging fruit" targets. Arguably, the most genuine example of difficult-to-solve targets is constituted by membrane proteins (MP). Hence only a handful of high-resolution structures are available for this important class of biological molecules (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html) in spite of their unquestionable scientific interest.

The transmembrane (TM) domains of MPs are composed of either α-helix bundles or β-sheets, with the former architecture being by far the most common. α-Helical MPs can be additionally subdivided into peripheral, where a small TM domain mediates the attachment of one or more soluble domains to the membrane, and integral membrane proteins (iMPs) composed of

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: German Mycobacterium Tuberculosis Structural Genomics Consortium, BMBF

*Correspondence to: Dmitrij Frishman, Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany. E-mail: d.frishman@wzw.tum.de

Received 20 July 2006; Revised 18 April 2007; Accepted 3 May 2007
Published online 17 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21605

helix bundles embedded in the membrane, and exerting more elaborate functions than mere anchoring. MPs constitute around 30% of each proteome,[8] play a crucial role in many essential cellular processes, and represent half of all proteins targeted by commercially available drugs.[9]

Since the publication of the first MP structure,[10] the structure determination rate of new MP structures has been orders of magnitude lower than that of water-soluble proteins.[11] Molecular biologists encounter substantial difficulties in over-expressing iMPs to sufficient amounts and purity due to their unusual amino acid composition, cellular toxicity when present at high quantities, and saturation of the secretory machinery leading to MP aggregation and accumulation in the cell. Furthermore, owing to their strongly hydrophobic character, iMPs can only be extracted from the lipid bilayer and solubilized using detergents. However, the choice of the ideal concentration and detergent type that would keep an iMP in solution and preserve its native structure may be extremely difficult. Various strategies have been applied to overcome these obstacles, such as the utilization of slow-growing host strains,[12] low growth temperature, and low concentration of expression inducers.[13] Overall, significant progress in high-throughput production of MPs has been reported,[14] and there is no doubt that in the near future structure determination of MPs will experience explosive growth, repeating the history of structural research on globular proteins.[15] Structural genomics projects are now increasingly beginning to include MPs in their target lists along with globular proteins, and three experimental consortia exclusively devoted to MPs currently exist: the *Mycobacterium tuberculosis* MP project (http://www.membraneprotein.magnet.fsu.edu/)[16]; the ProAMP initiative in Germany (Proteomewide Analysis of Membrane Proteins: https://binfo.bio.wzw.tum.de/proj/proamp)[17]; and the Membrane Protein Network (http://www.mepnet.org).

Rational target selection is a major cost-saving factor in any large-scale structural genomics project. Any target list must satisfy at least two fundamental conditions. First, it has to be optimal with respect to the scientific agenda of a given consortium. For example, if the focus of the research is on proteins with novel folds, appropriate bioinformatics tools will be used to identify such targets based on structure prediction and sequence comparison. Second, a target list should ideally include proteins more amenable to crystallographic or NMR analysis. Factors determining experimental success of structure determination are poorly understood. However, under a given set of experimental conditions, protein amenability to high-throughput experimentation is an individual trait, ultimately determined by its amino acid sequence. The experimental progress of many structural genomics initiatives (22 at the time of writing) is being monitored by the TargetDB resource,[18] which dynamically tracks the status of each target protein. Based on these data, attempts have been made to correlate sequence information with available data on experimental behavior at different stages of structure determination by means of retrospective data mining.[19,20] Recent approaches have reached an overall prediction accuracy of 72% for protein solubility[21] and 70% for crystalizability.[22] However, this work has focused on globular proteins, for which a significant body of experimental evidence has been accumulated.

For iMPs, such data are only beginning to emerge in sufficient quantities and the factors determining the experimental fate of iMPs are even more enigmatic that those of globular proteins. For example, contradicting reports have been published regarding the impact of the number of TM regions on expression yields. While Dobrovetsky *et al.*[14] and Korepanova *et al.*[23] reported higher expression levels for proteins with relatively few TM helices, Eshaghi *et al.*[13] observed higher rates of successful expression for proteins with more than eight TM helices. Finally, the systematic study of Daley *et al.*[24] did not confirm any of these trends. Specific structural features of iMPs and their experimental behavior, which are vastly different from that of globular proteins, necessitate the development of data mining approaches specialized for this class of proteins. The selection of a suitable training dataset as well as the identification of relevant sequence attributes represents a significant challenge, especially in view of the virtual absence of truly high-throughput experimental pipelines designed to handle iMPs and the resulting scarcity of available data. Data analysis is further complicated by the fact that even highly sequence-similar proteins, both globular[25] and membrane,[26] can behave differently under the same experimental conditions. Therefore, selection of experimentally tractable MPs cannot be conducted at the level of entire protein families alone. In addition, classification algorithms need to be able to suggest the most promising candidates even from a collection of homologs, a major topic in structural genomics. Here we report the development of three machine-learning approaches specially tailored for the prediction of clonability, expressibility, and solubility of iMPs based on their primary sequence, using the entire body of experimental data currently available in TargetDB.

MATERIALS AND METHODS

Data on structural targets

Amino acid sequences and experimental progress information of structural targets pursued by 22 structural genomics consortia were downloaded from the TargetDB database (http://targetdb.pdb.org; March 2006 version).[18] For each protein, TargetDB lists its current experimental status which may be one of the following: Selected, Cloned, Expressed, Soluble, Purified, Crystallized, Diffraction-Quality Crystals, Diffraction (native diffraction-data or phasing diffraction-data), NMR Assigned, HSQC, Crystal Structure, NMR Structure, and In PDB. Targets with three or more predicted TM helices according to the TMHMM 2.0 algorithm[27] were considered iMPs.

DOI 10.1002/prot

Table I
The Number of Amenable and Recalcitrant Targets at Different Stages of Experimental Structure Determination

		Experimental status							
Behavior	Dataset	Cloned	Expressed	Soluble	Purified	Crystallized	Diffraction	Structure	In PDB
Amenable	Globular proteins	54,178	36,768	19,109	13,199	6,408	3,973	3,105	2,760
	iMPs	2,454	672	100	63	22	8	6	4
Recalcitrant	Globular proteins	22,968	13,668	11,233	2,750	3,821	1,264	431	0
	iMPs	969	1,725	523	25	29	9	2	0

Redundant targets sharing a given level of sequence similarity were detected based on BLAST alignments, covering at least 90% of the length in both proteins compared.

Datasets of experimentally recalcitrant and well-behaved proteins were constructed separately for globular proteins and for iMPs (see Table I and Fig. 1) based on

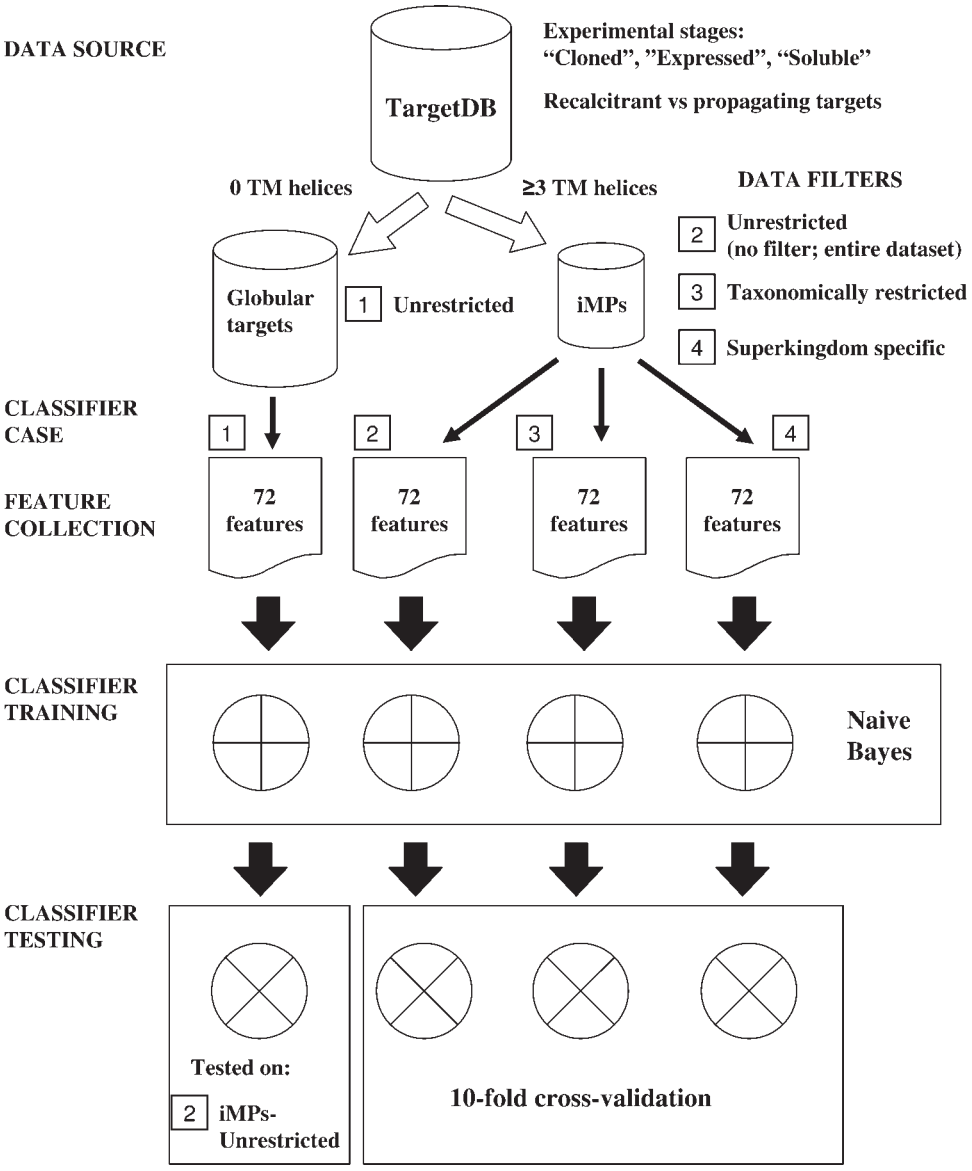

Figure 1
Flowchart of the study. See text for details.

Table II
Taxonomic Distribution, Numbers of Contributed Targets, and Success Rates of Selected Structural Genomics Consortia

		Amenable				Recalcitrant			
Status	Dataset	Archebacteria	Eubacteria	Eukaryota	Total	Archebacteria	Eubacteria	Eukaryota	Total
Cloned	Unrestricted	21	499	34	554	1	106	40	147
	Restricted	21	55	34	110	1	41	40	82
Expressed	Unrestricted	35	146	224	405	209	166	1344	1719
	Restricted	35	146	181	362	209	166	375	750
	Bacteria	—	146	—	—	—	166	—	312
	Eukaryota	—	—	224	—	—	—	1344	1568
Soluble	Unrestricted	16	27	18	61	19	114	190	323
	Restricted	16	27	18	61	19	114	133	266

the information obtained from TargetDB. Targets found at a certain experimental stage were considered amenable to all preceding stages as well, even when not stated explicitly. A target was classified as manifestly recalcitrant with respect to the next stage of structure determination if it remained in its current status for at least 9 months (from June 2005 to March 2006). As seen in Table I, an amount of data sufficient for statistical analyses of iMPs is available only for the first three experimental stages—cloned, expressed, and soluble—and thus only these three stages are considered here.

It should be noted that the information on target status provided by TargetDB, while being enormously useful, must be treated with caution. Differences between the 22 structural genomics consortia contributing to the database, in terms of their scientific objectives, experimental procedures applied, as well as the amount of funding available, have caused a considerable diversity in the size of respective target lists, taxonomic origin of target proteins, and success rate of structure determination. Moreover, the decision to report a certain experimental status for a given target may in part be based on subjective criteria. For example, the expression efficiency reported by the Joint Center of Structural Genomics consortium targets was 99.0% while only 1.5% of those were solubilized, presumably reflecting expression rates insufficient to undertake further studies. Thus, like most biological databases, TargetDB is unavoidably inhomogeneous and contains a sizeable amount of spurious and/or subjective assignments. In an attempt to partially alleviate this problem, we chose not to consider target data from the consortia reporting extreme (very low or very high) success rates as well as from those that did not contribute a sufficiently large number of targets to at least one experimental status considered here. Targets from the consortia showing success rates outside of the intervals 35–95% for cloning and 5–90% for expression and solubilization were not considered. Additionally, valid consortia for a given status were required to contribute at least five targets to that status and at least one target to both recalcitrant and amenable datasets. In total 6, 7 and 6 consortia satisfied these empirically derived criteria for cloned, expressed, and soluble targets, respectively.

Selection of positive and negative training datasets is further complicated by the fact that the amenability of a protein to experimental procedures in the course of structure determination may correlate with its taxonomic origin. Arguably, the best documented effect of this kind is the higher stability of proteins from thermophilic species[28,29] compared with their mesophilic counterparts. Eukaryotic proteins are generally more challenging structural targets due to their larger size, complex domain organization, and frequent occurrence of intrinsically disordered regions.[30] If not properly accounted for, organism-specific amino acid usage[31] will unavoidably introduce strong bias in any predictions based on sequences of taxonomically distant species.

To reduce taxonomic bias in our calculations, we created random subsets of TargetDB data such that each superkingdom contributed no more than 50% to both propagating and recalcitrant target sets at the three experimental stages considered (Fig. 1, Case 3). Specifically, at the cloning stage, the number of eubacterial targets was restricted to 55 and 41 for propagating and recalcitrant proteins, respectively. The number of eukaryotic targets was reduced to 181 and 375 for propagating and recalcitrant proteins at the expressed stage and to 133 for recalcitrant proteins at the soluble stage. Other entries in Table II did not require any adjustment as they already included less than 50% of targets from any single superkingdom. An obvious disadvantage of the approach described earlier is that it leads to a dramatic reduction in the total number of instances available for training. An alternative way to avoid taxonomic bias is to construct separate classifiers for different superkingdoms (Fig. 1, Case 4). Unfortunately, because of the paucity of available data this approach was only applicable to the expression status in bacterial and eukaryotic proteins.

Throughout this article we refer to the target status data for globular proteins, originally reported in the TargetDB, as well as to the data on iMPs, after adjusting the contribution of individual consortia (see earlier), as *unrestricted*. Furthermore, the dataset obtained by restricting the contribution of taxonomic groups is referred to as *restricted*. Finally, we call separate datasets

DOI 10.1002/prot

Table III
Summary of Protein Sequence Features

Attribute type	Number of features	Description
Global features	5	GRAVY index, length, molecular weight, net charge and pI
TM-specific features	8	Average soluble section length, fraction of the longest soluble section, fraction of residues in TM helices, length of the longest soluble section, number of TM helices, and occurrence of GxxG-like motifs, GxxxG-like motifs and GxxG-/GxxxG-like motifs per 100 TM residues
Residue composition	20	Occurrence of the 20 amino acids
Residue group composition	6	Acidic (D, E), aromatic-uncharged (F, W, Y), basic (H, K, R), nonpolar hydrophobic–aliphatic (A, L, I, M, V), polar–neutral (N, C, Q, S, T, Y), small residues able to contribute to helix–helix interaction trough GxxxG-like motifs (G, A, S)[32]. Proline (α-helix breaker)[33] is not considered here since it already is as single residue.
Residue group composition ratios	21	Relative occurrence of principal residue groups, including proline
Alternative residue group composition	12	Side-chains that are carbonyl-containing (D, E, N, Q), amide-containing (N, Q), branched (I, L, T, V), bulky (side chain volume $\geq$100 cubic Å)[34] (F, K, R, Y, W), charged (D, E, H, K, R), hydrophobic (hydrophaty index $>$0)[35] (A, C, F, I, L, M, V), hydrogen bonded (E, D, H, K, N, Q, R, S, T, W, Y),[36] straight chains containing hydroxyl (S, T), short straight chains (side chain length $<$4.5 Å)[37] (A, C, G, S, T), strongly basic (net charge = 1)[38] (K, R), strongly charged (net charge = 1 or $-$1)[38] (D, E, K, R), and tiny (side chain length $<$3 Å)[37] (A, G)
Total	72	

A detailed descriptor list with average values and standard deviations of each feature for recalcitrant and successful targets at different stages of experimental structure determination is available as supplementary material (S1).

for individual superkingdoms (and the associated classifiers) *superkingdom-specific*.

Protein sequence features

Two classes of sequence features were considered: global features that characterize a certain property of an amino acid sequence as a whole using a single number, and compositional features that give relative occurrence or occurrence ratios of different residues and residue groups (Table III). The GRAVY index (grand average of hydrophobicity) was defined as the average hydrophobicity of all amino acid residues according to the Kyte-Doolittle scale.[35] The net charge was calculated as the sum of all negative and positive charges in the sequence. TMHMM 2.0[27] was used to predict the number and location of TM helices. The average and longest loop length in each protein were calculated based on the distances between predicted TM segments. TM helices interact tightly due to the presence of GxxG- and GxxxG-like motifs (where G is any of the following three small residues: G, A, or S; and x is any residue).[32] The frequencies of these motifs per 100 TM residues were also considered as features, since the fold compactness could, in principle, influence the experimental behavior of the iMPs. Amino acid composition was either computed individually or for a set of principal groups of amino acids according to their physicochemical properties as defined by Taylor.[39] Ratios between the occurrences of residue groups were also taken into account, as those have been reported to be informative for detecting amenable targets.[19] In addition, another 12 alternative partitions, reflecting differences in residue charge and size as well as the presence of certain chemical groups such as amides, were included in the analysis, even though they overlapped partially with other groupings.

As seen in Table III, a broad set of features (64 in total) was derived from complete amino acid chains of both globular proteins and iMPs, while eight further features were calculated solely on the basis of TM portions. Naturally, TM-specific features were all assigned the value of zero for globular proteins.

Data mining for prediction of experimentally tractable sequences

On the basis of the features presented in Table III, we built Bayesian classifiers[40,41] for recalcitrant versus successful targets in three binary classification problems: cloning, expression, and solubilization (Fig. 1). Naive Bayes classification is based on the principle that each instance should be classified to the most probable class, with the probability distribution calculated for each class using training instances.[41] A distinctive characteristic of the method is that features are assumed to be independent, which allows estimating these probabilities separately for each feature. This simple, yet powerful technique, has been shown to perform well even when the "naive" independence assumption between attributes is violated.[40]

Sets of globular proteins and iMPs were prepared for the three experimental stages from the TargetDB repository considered here (see above). In addition, the iMP dataset was processed to prevent a strong taxonomical effect either by restricting disproportionate contributions of individual superkingdoms (Case 3), or by constructing superkingdom specific classifiers (Case 4). Sets of 72 features were extracted from every protein dataset and utilized to construct independent naive Bayes classifiers.

It is known that even a small unbalance in the population of different classes can be detrimental to the performance of any classifier that is trained to maximize the overall percentage of correctly classified samples. To compensate for the effects caused by uneven class population, a cost-sensitivity method[41] was imposed on the Bayesian classifier. A cost-sensitivity matrix was built according to the relative ratio of instances, so that misclassification in the less populated class incurred proportionally higher cost.

For each experimental stage (cloning, expression, and solubilization) we investigated two different learning problems. In Case 1 we trained our classifiers on data for globular proteins and then validated them on iMPs (see Fig. 1). In Cases 2, 3, and 4 the classifiers were trained and tested using 10 times stratified crossvalidation over the iMP dataset (Fig. 1). Crossvalidation allows evaluating classifier performance over all instances by repetitive training and evaluation coupled with walking across the dataset. In our application, data is randomized, stratified (for equal class representation), and divided into 10 parts. In each of the 10 training and evaluation cycles, one-tenth of the data is set aside for evaluation while the remaining nine-tenths are used for training. The overall performance (accuracy) was defined as the percentage of correctly classified examples:

$$\text{Overall performance} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})}$$

where TP, TN, FP, FN are the number of true positives, true negatives, false positives, and false negatives, respectively. In addition, we separately assessed the performance on the successful class using the true positive rate (percentage of correctly classified positive examples) and on the recalcitrant class using the true negative rate (percentage of correctly classified negative examples) defined respectively as

$$\text{TP rate} = \text{TP}/(\text{TP} + \text{FN})$$

and

$$\text{TN rate} = \text{TN}/(\text{TN} + \text{FP})$$

Given the imbalance between positive and negative targets in our datasets, additional estimations of the prediction quality were carried out by calculating the Matthews correlation coefficient (MCC)[42] defined as

$$\text{MCC} = \frac{\big((\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})\big)}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

We also computed the gain for the positive and negative sets:

$$\text{Gain (for positives)} \frac{= \big(\text{TP}/(\text{TP} + \text{FP})\big)}{\big((\text{TP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})\big)}$$

$$\text{Gain (for negatives)} \frac{= \big(\text{TN}/(\text{TN} + \text{FN})\big)}{\big((\text{TN} + \text{FP})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})\big)}$$

The MCC adopts values between -1 and 1, with negative values indicating that the number of correctly classified instances is below random expectation and low/high positive values being a sign of moderate/good performance, respectively. On the other hand, gain directly reflects the probability of correct assignment to either the positive or the negative class compared to the value expected if the null hypothesis, that is, the classifiers have no predictive power of the natural occurrence of the class, were right. For example, a gain of 1.6 for the positive set means that the ability of the classifier to discriminate correct positive instances is improved by 60% relative to random sampling. In contrast to MCC, gain is calculated independently for every class rather than for the complete classifier. The statistical significance of both MCC and gain estimators depends on the number of instances.

The low degree of sequence redundancy, with the probability of 10-fold crossvalidation between sequence-similar proteins below 1%, is due to the generally low coverage of MP sequence space by TargetDB. However, one should keep in mind that the redundancy will unavoidably increase with the growth of the number of MP sequences in future updates of TargetDB, necessitating more sophisticated methods of crossvalidation.

Selection of relevant features

We used the WEKA machine learning workbench[43] for feature selection as well as for classification. Features were selected using the wrapper approach as described by Kohavi and John.[44] A wrapper system consists of three main components: a classification algorithm (induction algorithm), whose performance is optimized by the feature selection process; a search engine (feature selection search), which determines how to progress through the configuration space of the absence/presence patterns of features; and an evaluation function (feature evaluation), which measures how well the classifier is performing using the currently selected features. The "best first" forward selection served as the search engine. The classification error estimated with up to fivefold crossvalidation was utilized to evaluate accuracy. Crossvalidation was aborted when the net progress of the accuracy fell below 1%. The forward selection process was terminated when accuracy did not increase by at least 0.1% after adding five consecutive features.

Availability

The method described in this contribution for the taxonomically restricted datasets has been made available online as part of the MEMEX server at http://webclu.bio.wzw.tum.de:8080/memex/. Query sequences in multiple FASTA format are evaluated and the predicted propensity for every class (propagating/recalcitrant) and status (cloned, expressed, soluble) is provided as output.

DOI 10.1002/prot

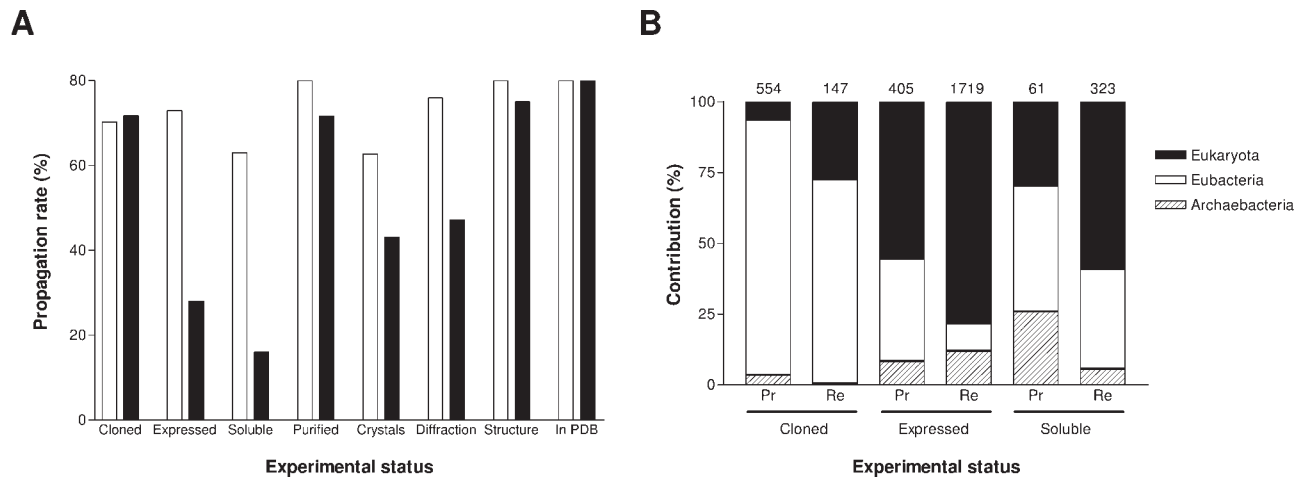

Figure 2

Success rates for globular and membrane proteins at different stages of structure determination. ***A:*** *Success rates for globular proteins (white bars) and iMPs (black bars) surviving a given experimental step were calculated considering all proteins that reached the previous step as 100%.* ***B:*** *Relative contribution to the propagating (Pr) and recalcitrant (Re) iMP curated datasets split by superkingdom. The total number of targets is given on the top of the respective bars.*

RESULTS AND DISCUSSION

Experimental rates of iMPs

We began this study by assessing the most difficult bottlenecks in structure determination of iMPs by comparing the progression of globular and membrane targets from one experimental state to the next. As seen in Figure 2(A), success rates at the expression and solubilization stages are markedly lower for iMPs, one-half and one-fourth of those for globular proteins, respectively. For these two stages, differences in the success rate between globular and iMPs were extremely significant (*P*-value $<2.2 \times 10^{-16}$) as deduced by chi square analysis. Although better expression rates for iMPs have been reported,[13] those were achieved in a medium-scale project focused on *Escherichia coli,* where several host strains, vectors, and conditions were tested for a limited number of targets.

Success rates at other stages did not show statistically significant differences (*P*-value >0.05), although given the small number of iMPs at advanced experimental stages, especially crystallization, this conclusion remains yet to be confirmed. Peripheral MPs, defined as those possessing one or two predicted TM helices, displayed intermediate behavior between iMP and globular proteins with respect to expression and solubilization (data not shown). iMPs thus generally appear to be much more challenging targets than globular proteins.

Data presented in Figure 2(A) underscore the pressing need for reliable prediction of iMPs that are more likely to overcome the expressibility and solubility bottlenecks. Similar success rates of cloning experiments—roughly 70%—have been achieved both for globular proteins and iMPs, still leaving substantial room for improvement. By contrast, it is currently not feasible to analyze more advanced experimental stages, as the amount of data available for training is too small for deriving statistically relevant conclusions. For instance, only 63 iMPs have survived the purification stage (Table I).

To conclude, our statistical survey demonstrates that available data on experimental progress of iMPs, while being much less abundant than those for globular proteins, is already sufficient to make an assessment of targets at three experimental stages: cloned, expressed, and soluble. Even after the filtering steps described in Materials and Methods, we were left with 2877 iMPs from all three kingdoms of life (and 69 individual species) targeted by nine structural genomics consortia (Table II) involving dozens of experimental groups. It is thus reasonable to believe that the dataset used here is not overly biased by taxonomic factors or individual characteristics of each lab (available skills, experimental procedures, personal preferences), allowing for meaningful generalizations.

Construction of classifiers to predict experimental success

Our approach involves associating regions of the feature space describing iMPs with their experimental properties in order to identify particularly promising targets for structural genomics. We applied Bayesian statistics[40] and cost-sensitivity analysis to the primary sequences of TargetDB proteins in order to construct classifiers capa-

Table IV
Single Features with the Highest Discriminatory Power

Status	Dataset	Feature	Performance rates (%) Overall	TP rate	TN rate
Cloned (without cost-sensitivity)	Unrestricted	% EDHKNQRSTWY	80.6 (0.260)[a]	97.3 (1.03)[b]	17.7 (3.02)[b]
		% polar	80.3 (0.248)	96.9 (1.03)	17.7 (2.88)
		% hydrophobic	80.0 (0.200)	98.6 (1.02)	10.2 (3.11)
		% ACFILMV	79.9 (0.199)	98.0 (1.02)	11.6 (2.89)
		Ratio small/hydrophobic	79.9 (0.175)	99.8 (1.01)	4.8 (4.17)
	Restricted	% polar	70.2 (0.388)	79.1 (1.26)	58.8 (1.57)
		Length	66.2 (0.313)	91.8 (1.13)	32.9 (1.74)
		Ratio aromatic/polar	64.1 (0.309)	55.4 (1.32)	75.3 (1.30)
		Mw	65.1 (0.287)	90.1 (1.12)	31.8 (1.67)
		Ratio hydrophobic/polar	62.0 (0.267)	53.6 (1.27)	72.9 (1.26)
Expressed (without cost-sensitivity)	Unrestricted	Ratio small/polar	81.2 (0.216)	16.8 (2.74)	96.4 (1.03)
		% V	81.5 (0.206)	13.3 (2.95)	97.6 (1.02)
		% A	80.7 (0.189)	15.6 (2.52)	96.0 (1.02)
		% AG	80.9 (0.186)	14.1 (2.60)	96.7 (1.02)
		% small	81.0 (0.158)	9.9 (2.66)	97.7 (1.01)
	Restricted	% K	69.7 (0.197)	15.2 (1.99)	96.0 (1.04)
		% V	69.1 (0.176)	15.7 (1.82)	94.8 (1.04)
		% F	69.2 (0.169)	9.9 (2.09)	97.7 (1.03)
		% small	68.5 (0.146)	11.9 (1.78)	95.9 (1.03)
		% I	68.5 (0.143)	11.0 (1.81)	96.3 (1.02)
Soluble (with cost-sensitivity)	Unrestricted	Ratio hydrophobic/polar	58.1 (0.221)	75.4 (1.51)	54.8 (1.10)
		% G	52.8 (0.197)	78.7 (1.40)	48.0 (1.10)
		% L	58.3 (0.184)	68.9 (1.44)	56.3 (1.08)
		% polar	52.1 (0.170)	75.4 (1.35)	47.7 (1.08)
		Ratio aromatic/polar	65.9 (0.160)	52.5 (1.50)	68.4 (1.05)
	Restricted	% I	73.4 (0.201)	41.0 (1.76)	80.8 (1.05)
		Ratio hydrophobic/polar	53.5 (0.198)	77.0 (1.36)	48.1 (1.11)
		% T	55.0 (0.161)	68.8 (1.32)	51.9 (1.08)
		% polar	48.6 (0.153)	77.0 (1.25)	42.1 (1.09)
		% AG	50.8 (0.152)	73.8 (1.27)	45.5 (1.09)

Performance values are sorted by MCC. Only the five features with the highest MCC values per dataset are shown.
[a]Values in parentheses in this column indicate MCCs.
[b]Values in parentheses in this column indicate gains.

ble of discerning manifestly recalcitrant targets from those amenable to cloning, expression, and solubilization (Fig. 1). In an initial test we built 72 separate classifiers for each of the three experimental stages, one for each protein feature listed in Table III, both with and without cost-sensitivity; the total number of classifiers was thus 3 stages $\times$ 72 features $\times$ 2 cost-sensitivity conditions (yes/no) = 432. We observed that noncost-sensitive classification resulted in most, if not all, instances being attributed to the most populated class, while the application of a cost matrix led to the opposite outcome, whereby the least populated class was preferred. Overall, classifiers built on single features showed unacceptable imbalance between the true positive and true negative rates (Table IV) and, consequently, rendered poor MCC values, indicating that single features only have a marginal discriminative capacity. These results are in line with a previous study carried out exclusively on *E. coli* MPs, where MP expression did not correlate significantly with any single protein property, including the number of TM helices.[24] Furthermore, single-feature classifiers may be extremely unstable with respect to certain attributes. For example, considering that all members in one large family may possess a closely overlapping range of feature values (e.g. similar numbers of TM helices) across both classes, it is highly possible that recognition thresholds based on single features will classify such entire families as positives or negatives, even in spite of the differences in the experimental behavior of individual family members.

The failure of the simplistic approach based on single features naturally calls for the application of powerful data mining tools that take advantage of a large set of features. The leading assumption of this analysis is that the experimental behavior of iMPs is imprinted in their primary sequence through relationships between a restricted set of key properties, as previously suggested for globular proteins.[20,45]

We experimented with four alternative training setups schematically depicted in Figure 1. The first learning problem involves predicting experimental properties of iMPs using a classifier trained on recalcitrant and amenable *globular* proteins. Three further learning tasks utilize unrestricted, restricted, and superkingdom-specific sets of iMPs (see Materials and Methods) for training. This

DOI 10.1002/prot

Table V
Performance of the Bayesian Classifier

Dataset	Proteins analysed	Performance (%)								
		Cloned			Expressed			Soluble		
		Accuracy	TP rate	TN rate	Accuracy	TP rate	TN rate	Accuracy	TP rate	TN rate
Unrestricted	Globular	77.7 (0.200)[a]	92.2 (1.04)[b]	23.1 (2.11)[b]	66.0 (0.038)[a]	29.6 (1.13)[b]	74.6 (1.01)[b]	77.6 (0.077)[a]	18.0 (1.47)[b]	88.9 (1.01)[b]
	All iMPs	73.2 (0.275)	79.1 (1.00)	51.0 (2.43)	70.0 (0.227)	52.6 (1.69)	74.1 (1.07)	60.7 (0.157)	60.7 (1.42)	60.7 (1.06)
Restricted	All iMPs	70.3 (0.397)	72.7 (1.31)	67.1 (1.50)	64.2 (0.233)	56.4 (1.41)	68.0 (1.13)	59.6 (0.157)	60.7 (1.37)	59.4 (1.07)
Super-kingdom-specific	Bacterial iMPs	—	—	—	68.9 (0.346)	66.3 (1.59)	70.1 (1.20)	—	—	—
	Eukaryotic iMPs	—	—	—	62.8 (0.150)	57.1 (1.46)	63.8 (1.05)	—	—	—

[a]Values in parentheses in this column indicate MCCs.
[b]Values in parentheses in this column indicate gains.

four-tier strategy was chosen in order to assess the relative relevance of different protein types as well as their taxonomic origin and sequence characteristics for predicting clonability, expressibility, and solubility.

Similar to single feature classifiers, those trained on globular proteins also rendered unbalanced predictions when applied to iMPs (Table V). The most populated class was strongly favored, with MCC and gain values being close to 0 and 1, respectively. It is apparent that globular proteins are poor predictors for iMPs due to their vastly different amino acid composition and general architecture [Fig. 2(A)].

Therefore, classifiers based exclusively on iMPs were constructed. Because of the intrinsic difficulties in structure determination of MPs, the iMP subset constitutes only a tiny fraction of TargetDB, making statistical inference based on these data being highly sensitive to any inherent bias. As illustrated in Figure 2(B), a prominent bias present in the taxonomically unrestricted iMP dataset is the vastly different representation of eubacterial, archaeal, and eukaryotic sequences at the cloned, expressed, and soluble stages. As described in the Methods section, we sought to circumvent the taxonomic bias by either restricting the number of contributing targets from a single superkingdom ($\leq$50% of the total number of either positives or negatives), or by constructing separate superkingdom-specific classifiers. Unfortunately, because of scarcity of data, the second approach was only feasible for one experimental status (Expression) upon merging the eubacterial and archaebacterial targets into one generic category "Bacteria." Both of these strategies unavoidably led to a reduced total number of instances available for training and allowed us to improve the results only in a few selected cases.

The overall accuracy of Bayesian classifiers trained on the sets of iMPs yielded notably better, more balanced predictions compared to the single-feature classifiers and those based on globular proteins. In particular, significantly improved MCC values and performance rates/gain values for the less populated class were observed while performance rate/gain for the more populated class changed only marginally. For example, the success rates achieved by the classifier trained on globular proteins for the Expressed state were 29.6% (gain 1.13) and 74.6% (gain 1.01) for the less and more populated classes, respectively, while the corresponding values were 52.6% (1.69) and 74.1% (1.07) for the taxonomically unrestricted, and 56.4% (1.41) and 68.0% (1.13) for the taxonomically restricted sets of iMPs (Table V).

Although the datasets used in this work contained between 5 and 20% of entries sharing more than 25% sequence identity, we did not remove homolog sequences, as it is known that even minor variations in amino acid sequences can lead to vastly different amenability of proteins to experimental procedures.[25,26] Our intention was thus to develop a classifier that would be instrumental in selecting more promising targets from sequence families. The influence of sequence redundancy on the classification accuracy was estimated at several allowed identity levels (see supplementary Fig. S3). When only nonredundant targets were considered, insignificant accuracy reduction (less than 2.5%, P-value >0.13 calculated by chi square analysis) was observed for two out of six classifiers tested, while for the remaining four classifiers it did not change.

A naive classifier based on best BLAST hit was strongly biased toward the most represented class, yielding MCC values around half of those obtained by the Bayesian approach (Supplementary Table S4). In addition, more than 35% of proteins could not be classified by BLAST at all as they did not share sequence similarity with any other structural targets in our dataset.

By design, superkingdom-specific classifiers explicitly take into account taxonomic information (Table V), while for the rest of the predictors the influence of the taxonomic bias had to be assessed a posteriori. Scrutiny revealed that the discriminative power of the classifiers built on unrestricted input data strongly depended on

Table VI
Performance of the Unrestricted and Restricted Classifiers for Individual Taxonomic Superkingdoms

		Performance (%)								
		Archaebacteria			Eubacteria			Eukaryota		
Status	Dataset	Accuracy	TP rate	TN rate	Accuracy	TP rate	TN rate	Accuracy	TP rate	TN rate
Cloned	Unrestricted	NC	NC	NC	74.5 (0.216)[a]	81.4 (1.05)[b]	42.5 (1.86)[b]	58.1 (0.143)[a]	38.2 (1.23)[b]	75.0 (1.09)[b]
	Restricted	NC	NC	NC	68.7 (0.362)	74.5 (1.27)	61.4 (1.48)	73.0 (0.469)	79.4 (1.47)	67.5 (1.47)
Expressed	Unrestricted	49.6 (0.038)[a]	57.1 (1.09)[b]	48.3 (1.02)[b]	54.2 (0.171)	90.4 (1.09)	22.3 (1.36)	76.3 (0.109)	27.2 (1.59)	84.5 (1.02)
	Restricted	65.6 (0.109)	45.7 (1.38)	68.9 (1.03)	66.7 (0.332)	64.4 (1.37)	68.8 (1.30)	62.9 (0.195)	51.9 (1.36)	68.3 (1.11)
Soluble	Unrestricted	45.7 (−0.021)	93.8 (0.99)	5.3 (0.92)	61.0 (0.208)	66.7 (1.47)	59.6 (1.09)	63.0 (−0.066)	22.2 (0.69)	66.8 (0.99)
	Restricted	68.6 (0.421)	87.5 (1.33)	52.6 (1.53)	63.1 (0.208)	63.0 (1.50)	63.2 (1.09)	54.3 (−0.063)	33.3 (0.80)	57.1 (0.98)

NC: Noncalculable. Only one target in the negative set.
[a]Values in parentheses in this column indicate MCCs.
[b]Values in parentheses in this column indicate gains.

the relative contribution of different superkingdoms to the positive and negative datasets (Table VI). For example, the success rate in predicting eubacterial negatives and eukaryotic positives was ≤27%, whereas for eubacterial positives and eukaryotic negatives it was ≥84%. Based on the taxonomically aware approach, a significant improvement in prediction accuracy was achieved for the expressability of eukaryotic proteins, a major experimental issue, which increased from 27.2% to 51.9% (Table VI) and 57.1% (Table V) for taxonomically restricted and eukaryote-specific classifiers, respectively. The boost in predictive power observed for positive targets was

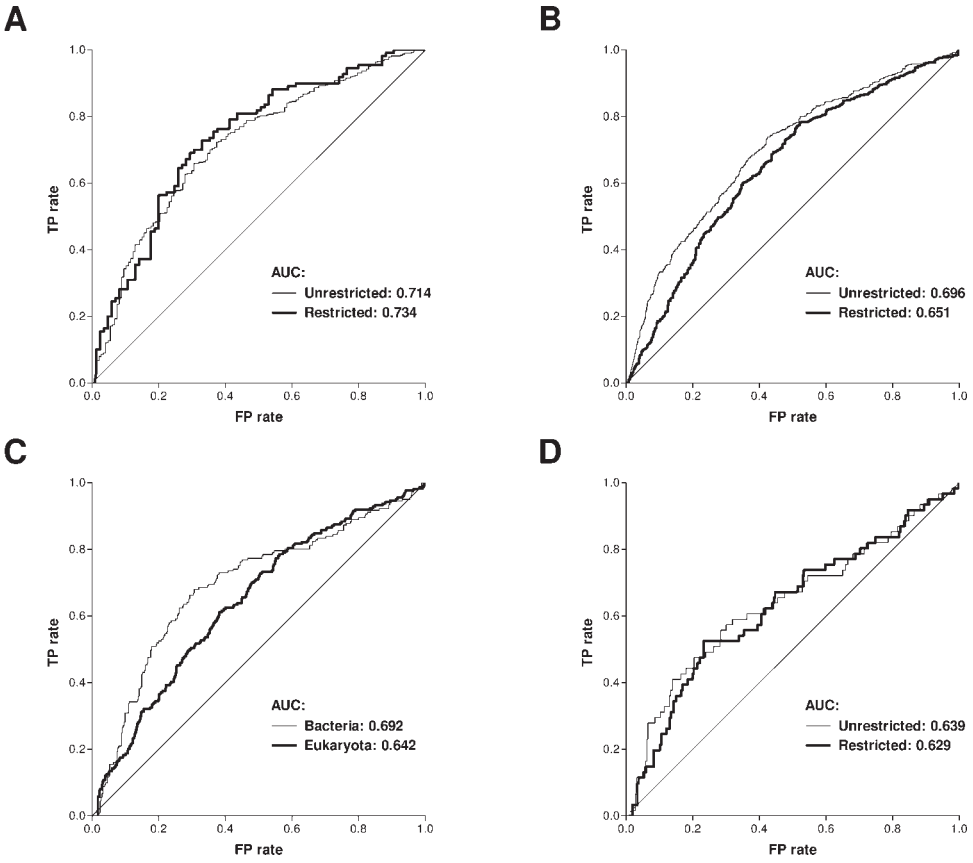

Figure 3
ROC curves of the classifiers. ROC curves for positives are displayed for both unrestricted (thin line) and restricted (thick line) datasets combined in the same graph. ***A:*** *cloned;* ***B:*** *expressed;* ***C:*** *expressed (bacteria vs eukaryotes);* ***D:*** *soluble. FP rate = FP/(FP+TN). AUC, area under the curve.*

DOI 10.1002/prot

Table VII
Features with the Highest Separation Power for Clonability, Expressability, and Solubilization

			Mean ± SD		
Dataset	Status	Properties	Recalcitrant	Amenable	P-*value*[a]
Cloned	Unrestricted	% V	8.1 ± 2.0	9.6 ± 2.5	1.0 E −14
		% polar	23.1 ± 4.4	20.0 ± 3.5	1.9 E −13
		% NQ	6.0 ± 2.0	4.8 ± 1.8	4.9 E −10
		Fraction longest loop length	0.31 ± 0.19	0.23 ± 0.17	3.2 E −06
		% K	4.1 ± 2.2	3.2 ± 2.0	3.5 E −06
	Restricted	% NQ	6.5 ± 2.0	4.9 ± 2.0	2.6 E −07
		Ratio aromatic/polar	0.49 ± 0.13	0.61 ± 0.20	4.1 E −07
		Longest loop length	146 ± 150	77.5 ± 72	1.6 E −04
		% V	7.9 ± 1.8	9.0 ± 2.6	2.4 E −04
		%ST	12.7 ± 2.7	11.5 ± 2.3	1.0 E −03
Expressed	Unrestricted	%I	8.7 ± 2.5	7.4 ± 2.6	6.5 E −18
		% A	7.3 ± 2.8	9.1 ± 3.9	1.8 E −16
		% AG	13.1 ± 4.7	16.0 ± 6.4	2.0 E −16
		Ratio small/polar	0.85 ± 0.33	1.08 ± 0.52	2.2 E −16
		% V	7.6 ± 1.9	8.7 ± 2.7	1.6 E −14
	Restricted	% I	8.8 ± 2.6	7.3 ± 2.7	2.6 E −17
		% R	4.2 ± 1.6	4.9 ± 2.0	5.2 E −08
		Ratio small/hydrophobic	0.54 ± 0.10	0.57 ± 0.11	6.3 E −08
		% V	8.1 ± 2.1	8.9 ± 2.7	1.0 E −06
		% F	6.8 ± 2.2	6.0 ± 2.5	1.8 E −06
	Bacteria	% K	3.8 ± 2.0	2.3 ± 2.0	1.8 E −14
		% I	9.1 ± 3.0	7.1 ± 3.2	8.9 E −12
		% V	8.6 ± 2.3	10.3 ± 2.8	1.9 E −11
		% R	4.2 ± 1.7	5.4 ± 2.1	4.4 E −10
		% F	6.1 ± 2.1	5.0 ± 2.2	3.0 E −08
	Eukaryota	% I	8.5 ± 2.3	7.6 ± 2.0	6.6 E −10
		Fraction of the longest loop length	0.26 ± 0.15	0.33 ± 0.17	1.2 E −07
		Ratio aromatic/acidic residues	2.2 ± 1.6	1.8 ± 1.0	2.5 E −07
		% DEKR	15.7 ± 3.9	17.4 ± 4.8	6.1 E −07
		% DEHKR	17.9 ± 4.2	19.6 ± 5.0	1.9 E −06
Soluble	Unrestricted	% I	7.2 ± 2.5	8.5 ± 3.5	5.8 E −03
		% G	6.9 ± 3.2	7.6 ± 2.3	4.3 E −02
	Restricted	% T	5.8 ± 1.6	4.9 ± 1.4	6.9 E −05
		% L	11.9 ± 2.9	13.0 ± 2.5	2.0 E −03
		% I	7.1 ± 2.5	8.5 ± 3.5	3.9 E −03

The mean values are percentages for those features expressed as percentages, and fractions for all other features.
[a]Up to five features with *P*-values $\leq 5 \times 10^{-2}$ are displayed sorted by significance. The *P*-values were calculated by the unpaired Student's *t*-test. The complete list of contributing features is presented in the supplementary material table S1.

achieved, in part, at the expense of decreased true negative rates, which went down from 84.5% to 68.3% (Table VI) and 63.8%, (Table V), respectively. The restriction of the superkingdom contribution also resulted in two to three times better MCC values for cloning and expression. Different success rates for positive and negative datasets are the direct consequence of the inequality in the number of positive and negative instances. Importantly, the eukaryote-specific predictor for expressable proteins yielded a gain of 1.46. Since the success rate in combination with the gain quantifies the actual utility of the classifier in detecting positives or negatives compared to random, high values of these parameters for the challenging task of detecting eukaryotic proteins that are more likely to be expressed are encouraging. On the other hand, prediction of solubility was improved by the taxonomically aware approach only for archaebacterial targets (Table VI), probably due to the data scarcity for this status. Notably, highly balanced predictions were obtained using multifeature classifiers ($\sim$ 61% both for positives and negatives).

A more detailed characterization of the classifier performance is given by receiver operating characteristic (ROC) curves (Fig. 3) whose shape reflects the classifier's ability to distinguish signal from noise[46] by graphically representing the relation between sensitivity (TP rate) and, indirectly, specificity (keeping in mind that FP rate = 1 − specificity). The diagonal line on the graph corresponds to a random classifier. Using ROC curves, the accuracy of a classifier can be conveniently estimated by the area under the curve (AUC). The unrestricted and restricted datasets show similar curves with comparable AUC values in the three experimental stages analyzed. The ROC curves for clonability and expressability [Fig. 3(A–C)] are characterized by gradual growth in the entire range of the FP rate values, which is typical for classifiers

equally specific and sensitive. On the other hand, the ROC curve for solubility goes up steeper in the first half of the chart [Fig. 3(D)] and then flattens out to some extent, a typical behavior of the classifiers that can discriminate relatively modest fractions of examples but with high degree of certainty. In the three experimental stages evaluated, we observed that the AUCs of the multiple-feature classifier were 4–9% higher than those of the best single-feature classifier (Table IV) (see Supplementary material S2).

Protein features determining experimental success

Data mining techniques are not only useful for practical classification purposes, but also for gaining an understanding of the subject under study. Specifically, we are interested in learning what MP properties distinguish difficult targets from easier ones. From the total of 72 features listed in Table III, sets of descriptors ranging from 9 to 20 in size were found informative for various classification tasks using the wrapper method (see Methods). The only extreme was found for the expressed status and the eukaryote-specific classifier (38 relevant features). The high complexity of the prediction task seems to be the reason for the unusually high number of contributing features in that case. Most relevant features showed highly significant differences between the positive and negative classes with *P*-values below 10^{-5} (Table VII). In general, these features tend to be frequencies of individual residues rather than residue group frequencies, group ratios, or TM features.

When comparing the features found informative by the unrestricted and taxonomically restricted classifiers, it immediately becomes obvious that we are facing the chicken-and-egg problem, as features determining the experimental success often strongly depend on the particular taxonomic composition of the dataset under study. The causal relationships between the taxonomic origin of a protein and its amino acid composition on the one hand, and its experimental behavior on the other hand, are not known. In order to eliminate any influence of taxonomy one would need to find those features that do not show taxonomic variability, but do correlate with experimental outcome.

For the unrestricted classification problem, increased occurrence of polar residues, asparagines, and glutamines as well as longer loops make proteins more recalcitrant to cloning while a higher fraction of valine residues correlates with increased amenability to cloning. Upon removing the taxonomic bias, the relative occurrence of amide-containing amino acids and valine were found to be the only feature still helpful for distinguishing recalcitrant and amenable proteins at the cloned stage.

Three out of five of the most relevant features for expression in the taxonomically unrestricted dataset involved small residues (%A, %AG, ratio small/polar residues). The differences between the values of small residue descriptors between bacterial and eukaryotic targets were extremely significant (*P*-values 10^{-88} to 10^{-136}; higher content in the bacterial set) but nearly indistinguishable for %I (*P*-value: 0.45), which strongly suggests that the former are related to the taxonomic origin of proteins, whereas the latter influence the experimental outcome through physicochemical and structural properties not directly related to taxonomy. The percentage of isoleucine was still significantly lower in expressed proteins when the taxonomic effect was accounted for by either contribution-restricted or taxonomic-specific datasets. As a general tendency in the restricted dataset, proteins with lower content of isoleucine and phenylalanine in combination with more arginine and valine constitute promising targets for expression studies.

At the soluble stage, the low number of instances available for training leads to markedly lower statistical significance of the findings. The protein length was not selected by the wrapper algorithm as a relevant feature, but it was smaller in solubilized (322 ± 213) than in nonsolubilized (421 ± 304) targets (*P*-value 2.6×10^{-3}) although the fraction of residues in TM helices remained basically unaltered.

Interestingly, no features explicitly related to predicted TM segments, such as the number of TM helices or the fraction of residues in TM helices, correlated with the experimental outcome at the cloned, expressed, and soluble stages, although in two cases the length of the hydrophilic loops was a differentiating factor. Apparently precise accounting for occurrence of individual amino acid residues captures the effect that hydrophobic TM segments have on protein experimental properties better than rough cumulative measures.

CONCLUSIONS

We have demonstrated that Bayesian classifiers trained on globular proteins are poor predictors of experimental properties of iMPs. Furthermore, no single feature of iMPs taken separately possesses sufficient predictive power to distinguish between experimentally tractable and recalcitrant iMPs. As revealed by the application of machine learning approaches, combinations of multiple features, commonly ranging between 10 and 20 in number, are required to yield satisfactory prediction performance. An overall prediction accuracy of 70%, 63–70%, and 61% was achieved for the cloning, expression, and solubilization stages of the structural genomics pipeline, respectively, with an acceptable balance between the performance for the positive and negative class. In particular, the accuracy levels for clonability and expressibility of iMPs are comparable to those reported for the solubility of globular proteins,[20,21] while the solubility of iMP the

 DOI 10.1002/prot

accuracy is lower, although still significant. Although the MCC values remain relatively modest, our predictors already produce results 37–59% better than random for the challenging cases (targets recalcitrant to cloning and those amenable to expression/solubilization) with *P*-values of 10^{-3} to 10^{-10} as judged by their respective success rates. Thus our classifiers are immediately useful for designing rational target selection strategies and may be instrumental in setting priorities on members of MP families according to their predicted experimental behavior.

Thorough investigation of the classifier performance highlights the necessity of applying taxonomic filters in data-mining studies on protein sequences, especially when dataset sizes are limited. Taxonomic differences affecting the variables under study can disguise the features actually responsible for the observed phenotype. Utilization of superkingdom-specific and taxonomically restricted classifiers turned out to be especially beneficial for handling particularly difficult classification tasks, exemplified by identifying eukaryotic proteins amenable to expression experiments. However, it is important to realize that, at all three experimental stages analyzed, the number of instances in positive and negative classes is unequal and some taxonomic bias is still present even in the restricted dataset (Table II). In general, we found that the classifiers trained on restricted datasets work reasonably well and we recommend their utilization through our web site.

Our study constitutes the first attempt to predict experimental behavior of MPs in the context of high-throughput structure determination. At present, the work is hindered by the severely limited amount of data available for training the classifiers. However, the dynamics of data growth in TargetDB suggests that more and more structural labs are starting to address this difficult class of molecules with ever-improving success rates. We therefore anticipate that in the not-so-distant future it will become possible to substantially enhance the accuracy of our predictors for the cloned, expressed, and soluble stages as well as to start developing specialized predictors for subsequent stages of structure determination. The quality of predictions is also bound to improve as additional, more detailed, and better curated information gets incorporated in TargetDB. We plan periodic updates of our predictors as more training data become available.

Finally, it is not unusual that a target that is well behaved with respect to one experimental stage will prove difficult at another stage. A well-documented example of such behavior are proteins from hyperthermophilic organisms which show higher crystallization rates but lower expression rates compared with their mesophilic counterparts.[28] Ideally, target selection should be aimed at detecting those proteins that have high chances to progress quickly through the entire structure determination pipeline. It would thus be desirable to construct integrated classifiers capable of detecting such overall promising candidates.

ACKNOWLEDGMENTS

A.J.M-G is supported by the BMBF competence network "Proteomics of membrane proteins." We thank Chad Davis and Juergen Cox for careful reading of the manuscript and useful comments.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
2. 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.
3. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
4. Sali A. 100,000 protein structures for the biologist. Nat Struct Biol 1998;5:1029–1032.
5. Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7(Suppl):967–969.
6. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
7. Stevens RC, Yokoyama S, Wilson IA. Global efforts in structural genomics. Science 2001;294:89–92.
8. Frishman D, Mewes HW. Protein structural classes in five complete genomes. Nat Struct Biol 1997;4:626–628.
9. Drews J. Drug discovery: a historical perspective. Science 2000;287:1960–1964.
10. Deisenhofer J, Epp O, Miki K, Huber R, Michel H. X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. J Mol Biol 1984;180:385–398.
11. White SH. The progress of membrane protein structure determination. Protein Sci 2004;13:1948–1949.
12. Miroux B, Walker JE. Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J Mol Biol 1996;260:289–298.
13. Eshaghi S, Hedren M, Nasser MI, Hammarberg T, Thornell A, Nordlund P. An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins. Protein Sci 2005;14:676–683.
14. Dobrovetsky E, Lu ML, Andorn-Broza R, Khutoreskaya G, Bray JE, Savchenko A, Arrowsmith CH, Edwards AM, Koth CM. High-throughput production of prokaryotic membrane proteins. J Struct Funct Genomics 2005;6:33–50.
15. Bowie JU. Are we destined to repeat history? Curr Opin Struct Biol 2000;10:435–437.
16. Walian P, Cross TA, Jap BK. Structural genomics of membrane proteins. Genome Biol 2004;5:215.
17. Essen L-O. Structural genomics of "non-standard" proteins: a chance for membrane proteins? Gene Funct Dis 2002;3:39–48.
18. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics 2004;20:2860–2862.
19. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. Nat Struct Biol 2000;7:903–909.

20. Goh CS, Lan N, Douglas SM, Wu B, Echols N, Smith A, Milburn D, Montelione GT, Zhao H, Gerstein M. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J Mol Biol 2004;336:115–130.
21. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on over-expression in *Escherichia coli*. Bioinformatics 2006;22:278–284.
22. Smialowski P, Schmidt T, Cox J, Kirschner A, Frishman D. Will my protein crystallize? A sequence-based predictor. Proteins 2006;62: 343–355.
23. Korepanova A, Gao FP, Hua Y, Qin H, Nakamoto RK, Cross TA. Cloning and expression of multiple integral membrane proteins from *Mycobacterium tuberculosis* in *Escherichia coli*. Protein Sci 2005;14:148–158.
24. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G. Global topology analysis of the *Escherichia coli* inner membrane proteome. Science 2005;308:1321–1323.
25. Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, Das R, Gerstein M, Arrowmith CH, Edwards AM. Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. Proteins 2003;50:392–399.
26. Surade S, Klein M, Stolt-Bergner PC, Muenke C, Roy A, Michel H. Comparative analysis and "expression space" coverage of the production of prokaryotic membrane proteins for structural genomics. Protein Sci 2006;15:2178–2189.
27. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–580.
28. Jaenicke R, Bohm G. The stability of proteins in extreme environments. Curr Opin Struct Biol 1998;8:738–748.
29. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 2000;1:76–88.
30. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic target selection for structural genomics on eukaryotes. Proteins 2004;56: 188–200.
31. Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des 1998;3:497–512.
32. Senes A, Engel DE, DeGrado WF. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr Opin Struct Biol 2004;14:465–479.
33. Reiersen H, Rees AR. The hunchback and its neighbours: proline as an environmental modulator. Trends Biochem Sci 2001;26:679–684.
34. Krigbaum WR, Komoriya A. Local interactions as a structure determinant for protein molecules: II. Biochim Biophys Acta 1979; 576:204–248.
35. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
36. Adamian L, Liang J. Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. Proteins 2002;47:209–218.
37. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 1988;32:269–278.
38. Klein P, Kanehisa M, DeLisi C. Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim Biophys Acta 1984;787:221–226.
39. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
40. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Besnard P, Hanks S, editors. Proceedings of the eleventh conference on uncertainty in artificial intelligence. San Mateo, CA: Morgan Kaufmann; 1995. pp 338–345.
41. Witten IH, Frank E. Data mining. Practical machine learning tools and techniques, 2nd ed. San Mateo, CA: Morgan Kaufmann; 2005. 525 pp.
42. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405:442–451.
43. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–2481.
44. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell J 1997;97:273–324.
45. Canaves JM, Page R, Wilson IA, Stevens RC. Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. J Mol Biol 2004;344:977–991.
46. Domingos P, Pazzani M.Beyond independence: conditions for the optimality of the simple Bayesian classifier. In: Saitta L, editor. Machine learning: proceedings of the thirteenth international conference. San Francisco, CA: Morgan Kaufmann; 1995. pp 105–112.

 DOI 10.1002/prot