Journal of Universal Computer Science, vol. 16, no. 5 (2010), 604-621
submitted: 26/6/09, accepted: 12/2/10, appeared: 1/3/10 © J.UCS

Investigating a Correlation between Subcellular Localization and Fold of Proteins

Johannes Aßfalg, Jing Gong, Hans-Peter Kriegel, Alexey Pryakhin, Tiandi Wei, Arthur Zimek
(Institute for Informatics, Ludwig-Maximilians-Universität München, Germany
www: `http://www.dbs.ifi.lmu.de`
{assfalg,gongj,kriegel,pryakhin,tiandi,zimek}@dbs.ifi.lmu.de)

Abstract: When considering the prediction of a structural class for a protein as a classification problem, usually a classifier is based on a feature vector $x \in \mathbb{R}^n$, where the features represent certain attributes of the primary sequence or derived properties (e.g., the predicted secondary structure) of a given protein. Since the structure of a protein (i.e., its native conformation) is stable only under specific environmental conditions, it is commonly accepted to assume proteins being evolutionarily adapted to specific subcellular localizations and according to their physicochemical environment. Our statistical evaluation shows a strong correlation between the subcellular localization of proteins and their structural class. The correlation is strong enough to allow for a classification of proteins into their structural class solely based on information regarding the subcellular localization. We conclude that knowledge regarding the subcellular localization of proteins can be useful as a feature for the structural classification of proteins.
Key Words: bioinformatics, protein subcellular localization, protein fold prediction
Category: J.3, I.2.6

1 Introduction

It is common opinion that the three-dimensional structure of a protein is already encoded in its amino acid sequence [Anfinsen (1973)]. Being denatured, a protein can regain its enzymatic activity on its own in its normal physiological milieu. These findings have motivated researchers since decades to learn to assign the three-dimensional structure to a protein given the sequence of amino acids. Many approaches to this problem are homology-based. For a target protein of unknown structure, a template protein of known structure governs the assignment of structural properties along the amino acid sequence. From a machine learning point of view, these methods could be called lazy learners, since no model is sought to explain the fold but only the similarity to other objects leads to assigning the same properties. While these approaches are fairly successful, deriving a good similarity measure in terms of good sequence alignments only does not necessarily lead to biologically new insights regarding the folding process. Since, as Godzik put it, *"most proteins fold on their own [...], without checking what the structure of their homologs is in databases but following physical laws governing their behavior"* [Godzik (2003)], it might be a more revealing approach to eagerly find a model explaining which properties of a sequence were responsible to fold the sequence into a certain structure. For this purpose, the primary sequence is often transformed into a feature vector $x \in \mathbb{R}^n$, where $x_i \in \mathbb{R}$ represents the numerical value for a certain property such as, e.g., the percentage of alanine in the complete sequence of amino acids.

Considering the fold recognition problem as a classification task is based on the assumption that the number of (naturally occurring) protein folds is limited (see for example [Chothia (1992)], [Govindarajan et al. (1999)], or [Wolf et al. (2000)]). Two forces may have played a role in the limitation of the actual variety of folds: divergent evolution of protein function (since all folds are derived from a relatively small group of shared common ancestors) and convergent evolution of protein structure (since certain folds are physicochemically much more favored and thus may have originated independently in many cases) [Govindarajan and Goldstein (1996)]. Thus, the space of protein structures C is assumed to be finite and discrete, and a classifier should learn a function $\mathbb{R}^n \rightarrow C$. Of course, this function, once discovered, is of paramount interest, since it would be an approximation of the forces of nature guiding the folding process. In the meantime, however, an appropriate feature space is still to be discovered. This classification approach seeking a model to explain the fold may yield findings that are possibly of interest for the most fundamental approach to the protein-fold problem, the family of *ab initio* methods.

In this paper, we are interested in the subcellular localization of a protein (which is also often predicted based on the amino acid sequence) as an attribute in a feature space that is related to the fold of a protein. This idea is motivated by the fact that many proteins are specialized to operate in a certain compartment or region of a cell. Since each cellular compartment maintained a specific physicochemical environment throughout evolution, and the native conformation of a protein is only stable in a certain environment, the localization of a protein might have a considerable impact on its final three-dimensional structure.

Andrade et al. [Andrade et al. (1998)] have shown the amino acid composition on the surface of proteins to carry information regarding the subcellular location of the protein, since the surface of proteins is directly exposed to the environment. Our reasoning is, *vice versa*, that the subcellular localization of a protein is a valuable information in predicting the structure of a protein. For this purpose, we analyze whether there exists a correlation between the known localization and the known structural class of a protein, based on a newly compiled, up-to-date data set. If so, a method predicting the structural class of a protein solely based on its known subcellular localization should perform better than a random classifier. Finally, we investigate whether a feature representing the subcellular localization of a protein is valuable in comparison to other well-established features and proves useful in combination with a well-known set of features.

Table 1: Covered subcellular localizations and corresponding keywords in SWISS-PROT.

ID	Subcellular localization	Keywords in SWISS-PROT	ID	Subcellular localization	Keywords in SWISS-PROT
1	Chloroplast	Chloroplast	8	Peroxisome	Peroxisome, Peroxisomal
2	Cytoplasm	Cytoplasm(ic)			Microsome, Microsomal
3	ER	Endoplasmic reticulum			Glyoxysome, Glyoxysomal
4	Golgi apparatus	Golgi			Glycosome, Glycosomal
5	Lysosome	Lysosome, Lysosomal	9	Extracellular	Extracellular
6	Mitochondrion	Mitochondrion, Mitochondrial			Secreted
7	Nucleus	Nucleus, Nuclear	10	Vacuole	Vacuole, Vacuolar

2 Material and Methods

2.1 Classes of Subcellular Localization and Fold Classes

We restricted our efforts to those subcellular localizations covered by many prediction methods. Since prediction methods for subcellular localizations have attracted constant attention in recent years and are probably going to reach reasonable accuracy in the upcoming years, a connection of localization prediction and structural classification of proteins may become practicable in the near future.

So far, existing prediction methods for subcellular localization differ widely in coverage of predicted localization as well as in reliability. As a reasonable subset of subcellular localizations, we consider the ten localization classes listed in Table 1. This selection is based on a broad range of well performing prediction methods using different approaches like amino acid composition (for example [Hua and Sun (2001)], [Park and Kanehisa (2003)], or [Yu et al. (2004)]), sorting signals [Bannai et al. (2002)], [Small et al. (2004)], homology search [Lu et al. (2004)], frequent subsequences (see [Gardy et al. (2005)]), and hybrid methods (i.e., using several of these approaches, see for example [Nakai and Horton (1999)], [Horton et al. (2006)], [Höglund et al. (2006)], [Bhasin and Raghava (2004)], or [Garg et al. (2005)]). By these and other methods (see [Aßfalg et al. (2009)] for a more detailed discussion), the localization classes listed in Table 1 are mostly covered.

As classification systems for fold classes, we use the hierarchical classification databases CATH [Orengo et al. (1997)] (release 3.0.0) and SCOP [Murzin et al. (1995)] (release 1.71).

2.1.1 CATH

The CATH database [Orengo et al. (1997)] is a hierarchical classification of protein domain structures. The classification is achieved via a semi-automatic procedure. The four main levels of classification are protein class (C), architecture (A), topology (T) and homologous superfamily (H). 'Class' is the top level essentially describing the secondary structure composition of each domain. Four major classes are recognized: mainly-alpha, mainly-beta, alpha-beta, and few-secondary-structure. In contrast, 'architecture' summarizes the shape revealed by the orientations of the secondary structure units but ignores the connectivity between the secondary structures. For example, barrels and sandwiches are certain architectures. At the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies. When structures belonging to the same T-level also have suitably high similarities combined with similar functions, the proteins are assumed to be evolutionarily related and are considered as belonging to the same homologous superfamily.

2.1.2 SCOP

The SCOP database (Structural Classification Of Proteins) [Murzin et al. (1995)] provides a detailed and comprehensive description of the structural and evolutionary relationships between proteins, but it is quantitatively of smaller coverage than CATH since SCOP has been constructed manually by visual inspection and comparison of structures. Like CATH there is also a hierarchy of four main levels of classification: class, fold, superfamily and family. The top level classification is 'class' grouping proteins w.r.t. their secondary structure composition. Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins sharing the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies. Superfamilies collect proteins that have low sequence identities but whose structural and functional features suggest that a common evolutionary origin is probable. Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide some evidence of a possible common descent despite the absence of high sequence identity.

2.2 Data Sets

Our experiments require a data set unifying localization information as annotated e.g. in SWISS-PROT [Wu et al. (2006)] and structural classification information as in CATH and SCOP. We assign localization classes according to SWISS-PROT keywords as given in Table 1.

Table 2: Numbers of domains and classes on each level in the different data sets.

Data Set	DS 1	DS 2	DS 3	Data Set	DS 1	DS 2	DS 3
CATH Domains	37,841	34,178	1,791	SCOP Domains	17,978	13,081	2,662
Class	4	4	4	Class	11	10	6
Architecture	34	27	10	Fold	580	125	28
Topology	545	213	24	Superfamily	849	158	29
Hom. Superfamily	919	309	34	Family	1,274	190	30

We distinguish localization annotations that are experimentally confirmed from those annotated as "by similarity", "potential", or "probable" in SWISS-PROT. The latter can be assumed to be less reliable. However, in the first steps, we consider the total data set. For evaluation of a feature-vector-based classification, we return to this distinction. The total amount of extracted proteins with localization annotation was 125,160. After filtering for entries with a unique localization annotation, our data set contains 80,668 entries. Both, SWISS-PROT entries, as well as SCOP or CATH entries, may correspond to entries in the Protein Data Bank (PDB) [Berman et al. (2000)]. Among the 80,668 entries in our data set, the total number of corresponding PDB entries is 12,332. The number of the distinct PDB entries is 11,013. One PDB entry may correspond to several entries in SCOP or CATH, since SCOP and CATH register domains rather than complete proteins. After all, the join between SWISS-PROT and domains in CATH or SCOP results in a data set for CATH containing 37,841 entries and a data set for SCOP containing 17,978 entries (see entries for data set 'DS 1' in Table 2).

For assessing the usability of the subcellular localization as a single feature allowing for fold classification, the data sets are reduced as follows: To ensure generalizability, each class should be represented by some examples in each fold. Note that, for the same reason, in a classification scenario it is required to have several similar examples to allow for learning a generalized concept. Thus, reducing the data set based on sequence similarity levels, as opposed to alignment-based methods, is not appropriate here. We therefore pruned all classes with less than 20 entries from the CATH and the SCOP data set. This results in a data set for CATH containing 34,178 entries, and a data set for SCOP of 13,081 entries. The number of classes in different levels is shown in the 'DS 2' column of Table 2 for CATH and SCOP.

To have classifiers trained on reliable data, we compiled a third data set. Entries with localization annotation not experimentally confirmed are pruned. Again, CATH classes with less than 20 members are canceled. This results in a data set for CATH containing 1,791 entries. For SCOP, each remaining class contains at least 40 entries. The data set finally consists of 2,662 entries. Further details for CATH as well as for SCOP can be found in the 'DS 3' column of Table 2.

2.3 Correlation Analysis

In a first step to evaluate a possible relationship between the subcellular localization and the structure of a protein, we consider the conditional (or posterior) probability $P(C_i|L_j)$ for a protein to belong to structural class C_i, given its subcellular localization L_j. For a set of structural classes

$$C = \{C_1, \ldots, C_n\}$$

and the set of ten localization classes

$$L = \{L_1, \ldots, L_{10}\},$$

the posterior probability of a protein of a given localization for a certain structural class is defined as

$$P(C_i|L_j) = \frac{|C_i \cap L_j|}{|L_j|}.$$

Thus, we have to count the number of protein domains for each location and in each location for each structural class separately. This correlation analysis is performed on the full data set 'DS 1'.

One may suspect the correlation analysis to be biased by predominant structural classes in the PDB. However, if there are predominant structural classes but the posterior probabilities are not biased to these classes, the results appear even stronger. Thus, we allow for predominant structural classes in this data set. Furthermore, this data set has already been demonstrated to be challenging for localization prediction methods (see [Aßfalg et al. (2008)]).

2.4 Posterior-Probability-Based Classification

Based on the posterior probability, we can evaluate a simple classifier predicting the structural class of a protein of known localization class. Let C and L be defined as above. The structural class of a protein located in L_j is then given by

$$\arg\max_{C_i \in C} \{P(C_i|L_j)\}.$$

This means, the localization yields ways to tell the fold of a protein. Although this assumption is by far too simple, this is nevertheless a working classifier with surprising results.

Since a reasonable generalization is required in a classification task, this classifier is evaluated by 10-fold cross-validation on data set 'DS 2'.

Table 3: Amino acid attributes and the mapping of amino acids onto sets of three groups [Dubchak et al. (1999)].

Property	Group 1	Group 2	Group 3
Hydrophobicity	Polar: {R,K,E,D,Q,N}	Neutral: {G,A,S,T,P,H,Y}	Hydrophobic: {C,V,L,I,M,F,W}
Normalized van der Waals volume	0–2.78: {G,A,S,C,T,P,D}	2.95–4.0: {N,V,E,Q,I,L}	4.43–8.08: {M,H,K,F,R,Y,W}
Polarity	4.9–6.2: {L,I,F,W,C,M,V,Y}	8.0–9.2: {P,A,T,G,S}	10.4–13.0: {H,Q,R,K,N,E,D}
Polarizability	0–0.108: {G,A,S,D,T}	0.128–0.186: {C,P,N,V,E,Q,I,L}	0.219–0.409: {K,M,H,F,R,Y,W}

2.5 Subcellular Localization as a Feature in Structural Classification

2.5.1 Established feature space

The by far best known set of features for protein fold recognition was proposed by Dubchak et al. [Dubchak et al. (1995)] and used by many other machine learning approaches to the problem (e.g. [Ding and Dubchak (2001)], [Shen and Chou (2006)]). This feature set describes the properties of a sequence with three different descriptors called composition, distribution, and transition, based on different partitionings of amino acids into different sets of groups. Therefore, the set of amino acids is mapped onto a set of groups. The groups used by [Ding and Dubchak (2001)] were hydrophobicity, normalized van der Waals volume, polarity, and polarizability. The mapping for these sets of groups is defined by Table 3. The set of amino acids is also a set of groups (of size 20) as well as the predicted states of secondary structure (helix, sheet, and coil).

The descriptors can be applied to each of the sets of groups. The *composition* describes the percentage of amino acids of the sequence per group. This results in a number between 0 and 1 for each group, where the sum is 1. The *distribution* consists of five numbers for each group: the fractions of the entire sequence where the first residue of the corresponding group occurred, and where 25%, 50%, 75%, and 100% of those are contained. For each pair of groups, *transition* counts the number of transitions from one group to the other or *vice versa* within a given sequence. For the sets of groups of three members like secondary structural states or the groups described above, the composition descriptor and the transition descriptor each provide three numbers, the distribution descriptor five times three. Thus, for each of these groups a feature vector of dimensionality 21 is provided using the descriptors. For the single amino acids, only the composition is computed which in combination with the length of a sequence also results in a feature vector of dimensionality 21. Distribution and transition for the single amino acids would result in very high dimensional feature spaces. The combination of the feature vectors results in a feature space of $6 \times 21 = 126$ dimensions.

2.5.2 Feature evaluation

Since we propose to utilize the localization information as a new feature, the next step is to show that the new feature is important for classification. As a quality baseline, we compare the localization feature with the 126 features of Dubchak et al., using three feature evaluation techniques.

First, we chose an approach which utilizes the Support Vector Machine method (SVM) [Guyon et al. (2002)]. Features are ranked by the square of the weight assigned by the SVM. Feature evaluation for multiclass problems is handled by ranking attributes for each class separately using a one-versus-all method and then using the top of each pile to give a final ranking. [Guyon et al. (2002)] demonstrated experimentally on cancer data that the features selected by this method yield better classification performance. This method eliminates feature redundancy automatically and yields better and more compact feature subsets.

Second, we chose the RELIEF-method discussed in [Kira and Rendell (1992)] and [Kononenko (1994)]. The key idea of this approach is to estimate features based on their values among objects that are near to each other.

Finally, we employ the gain ratio method [Quinlan (1986)] which evaluates the significance of the feature f by measuring the gain ratio with respect to the class label c. The gain ratio is based on the notion of the entropy impurity value H and can be calculated as follows:

$$GainRatio(c, f) = \frac{H(c) - H(c|f)}{H(f)}.$$

We use the WEKA [Witten and Frank (2005)] implementation of the feature evaluation methods mentioned above. For all methods, we perform a 10-fold cross-validation.

2.5.3 Working in Concert

Finally, we evaluate the localization class as a feature in addition to the feature space as described above [Dubchak et al. (1995)]. As we are not interested in tuning a specific classifier, we use three well-known classification approaches, namely the k-nearest-neighbor classifier ($k = 10$), decision trees (J48 [Quinlan (1993)]) and support vector machines (SMO [Platt (1998)]) using a linear kernel, as implemented in WEKA. Since the addition of a new feature may per se facilitate a better classification accuracy, we compare three feature spaces for each structural class level of SCOP and CATH: first, the feature space 'D' as proposed by Dubchak et al. [Dubchak et al. (1995)] (126 dimensions), second, the feature space 'D' with an additional random feature: 'D+R'

(127 dimensions), and third, the feature space 'D' with the localization class as an additional feature: 'D+L' (127 dimensions). This way, we can assess the contribution of the localization-based feature to a general feature-based classification approach. Again, all classifiers are evaluated by 10-fold cross-validation.

While we assumed the location alone to be of value for the prediction of the structural class of a protein in an earlier step (cf. Section 2.4), the location of a protein is now (perhaps more realistically) assumed to guide to the correct fold along with information that is directly based on the sequence of amino acids. For both, feature evaluation (cf. Section 2.5.2) and feature-vector-based classification, we rely only on experimentally confirmed localization annotations. Thus, these experiments are based on data set 'DS 3'.

3 Results and Discussion

3.1 Strong Correlation between Localization and the Structural Class of Proteins

For both classification systems, CATH as well as SCOP, we found a distinct correlation between most localization classes and the structural class of a protein. Generally, the correlation is stronger on higher levels in the hierarchies of CATH and SCOP.

3.1.1 CATH

The results for the class level of CATH are depicted in Figure 1(a). For localization 1, the posterior probability of the CATH class "Mixed Alpha-Beta" is well above 80%. For other localizations, the dominant class is not always that eminent. However, for several localizations, the posterior probability of the dominant class is around 50% or above. As could be expected, on deeper CATH levels, the correlations are not equally pronounced, but nevertheless strong in many cases — see the supplementary material at `http://www.dbs.ifi.lmu.de/research/locfold/dataset/download.html`.

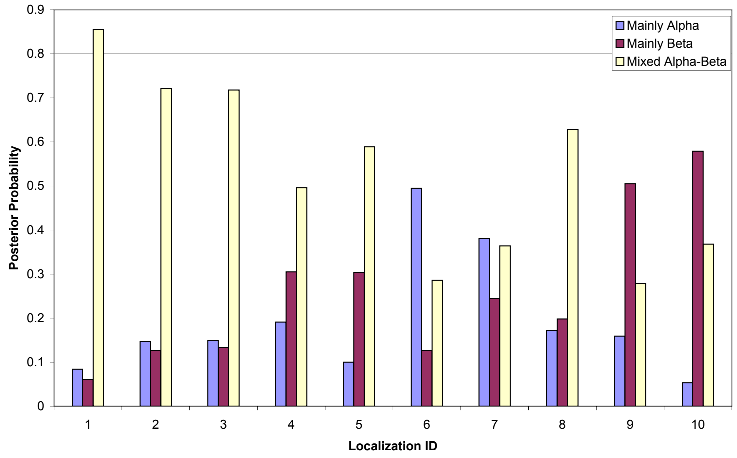
Exemplarily, the posterior probabilities for structural classes on the Architecture level of CATH are given in Table 4. The winner of the structural classes C_i in terms of posterior probability for each localization L_j (the index j reflecting the localization ID as defined in Table 1) is underlined in the table. The last column shows the prior probability for a given architecture, i.e., the percentage of entries of a certain class w.r.t. all entries. Here, the three highest entries are underlined.

Inspecting the values presented in this table is interesting, since one may suspect that the posterior probability

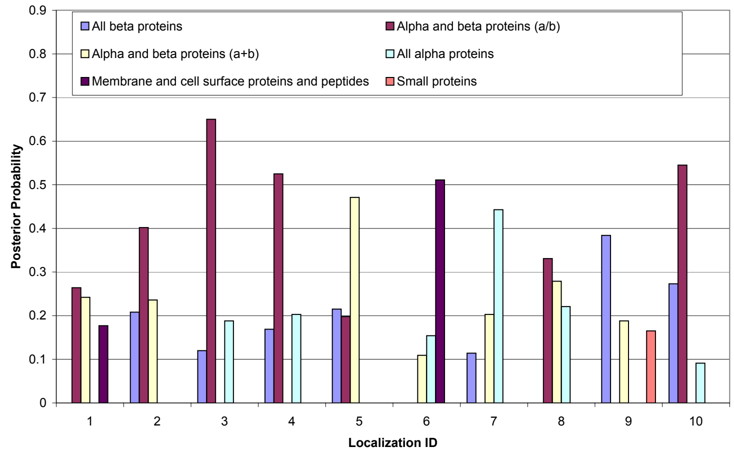
$$P(C_i|L_j) = \frac{|C_i \cap L_j|}{|L_j|}$$

Table 4: Posterior probabilities for CATH Architectures.

C_i (CATH Arch.)	$P(C_i\|L_1)$	$P(C_i\|L_2)$	$P(C_i\|L_3)$	$P(C_i\|L_4)$	$P(C_i\|L_5)$	$P(C_i\|L_6)$	$P(C_i\|L_7)$	$P(C_i\|L_8)$	$P(C_i\|L_9)$	$P(C_i\|L_{10})$	$\frac{\|C_i\|}{\sum_{j=1}^{34}\|C_j\|}$
C_1 1.10	0.0571	0.1065	0.0806	0.078	0.0778	0.2597	0.3341	0.0568	0.0877	0	0.1648
C_2 1.20	0.0098	0.0376	0.0363	0.1064	0	0.2119	0.0319	0.1154	0.0694	0.0263	0.0812
C_3 1.25	0	0.0027	0.0081	0	0.0222	0.0231	0.0147	0	0.0001	0.0263	0.0082
C_4 1.40	0.0054	0	0	0	0	0	0	0	0	0	0.0002
C_5 1.50	0.0116	0.0007	0.0242	0.0071	0	0	0	0	0.0022	0	0.0013
C_6 2.10	0	0.0087	0.004	0	0	0.0023	0.0012	0.0513	0.0356	0.0263	0.0132
C_7 2.20	0	0.0034	0	0	0	0	0.0876	0	0.0006	0	0.0143
C_8 2.30	0	0.0265	0	0	0	0	0.0186	0.0201	0.0121	0	0.0148
C_9 2.40	0.0259	0.0422	0.004	0	0.0704	0.0416	0.0778	0.0952	0.2774	0.5526	0.1071
C_{10} 2.60	0.0285	0.0316	0.121	0.2128	0.0667	0.0574	0.0275	0.011	0.1604	0	0.0695
C_{11} 2.70	0	0.0035	0	0.0922	0.1667	0	0.0005	0.0165	0.0042	0	0.0041
C_{12} 2.80	0	0	0.004	0	0	0	0.0002	0	0.0019	0	0.0005
C_{13} 2.90	0	0	0	0	0	0	0	0	0.0007	0	0.0002
C_{14} 2.100	0	0	0	0	0	0	0	0	0.0002	0	0.0001
C_{15} 2.102	0.0062	0	0	0	0	0.0256	0	0	0	0	0.0055
C_{16} 2.110	0	0	0	0	0	0	0	0	0.001	0	0.0002
C_{17} 2.120	0	0.0022	0	0	0	0	0	0	0.0034	0	0.0016
C_{18} 2.130	0	0.0037	0	0	0	0	0.0002	0	0.0005	0	0.0014
C_{19} 2.140	0	0	0	0	0	0	0.0011	0	0.0004	0	0.0003
C_{20} 2.150	0	0	0	0	0	0	0	0	0.0013	0	0.0003
C_{21} 2.160	0	0.003	0	0	0	0	0	0	0.0032	0	0.0018
C_{22} 2.170	0	0.0019	0	0	0	0	0.0307	0.0037	0.0016	0	0.0057
C_{23} 3.10	0.0437	0.0489	0	0.0355	0.0296	0.007	0.0089	0.1648	0.0795	0	0.0427
C_{24} 3.20	0.2239	0.0622	0	0.0922	0.0963	0.0047	0.0018	0.0952	0.0201	0.2237	0.0363
C_{25} 3.30	0.4362	0.1762	0.0484	0.0709	0.0741	0.1353	0.2111	0.1557	0.0739	0.0658	0.1526
C_{26} 3.40	0.1097	0.2893	0.2661	0.1631	0.0519	0.1211	0.0645	0.196	0.066	0.0526	0.1556
C_{27} 3.50	0	0.0318	0	0	0	0.0034	0	0	0.004	0	0.0122
C_{28} 3.55	0	0.0004	0	0	0	0	0	0	0	0	0.0001
C_{29} 3.60	0	0.0255	0	0	0.0148	0.0003	0.0018	0	0.0014	0	0.0092
C_{30} 3.65	0	0.0091	0	0	0	0	0	0	0	0	0.0030
C_{31} 3.70	0	0	0	0	0	0	0.0092	0	0	0	0.0014
C_{32} 3.80	0	0.0006	0	0	0	0	0.0007	0	0.0015	0	0.0007
C_{33} 3.90	0.0419	0.0771	0.4032	0.1348	0.3222	0.0142	0.0658	0.0165	0.0323	0.0263	0.0532
C_{34} 4.10	0	0.0047	0	0.0071	0.0074	0.0925	0.0103	0.0018	0.0571	0	0.0367



(a) Results for the three most likely CATH classes (classification level 'Class').



(b) Results for the three most likely SCOP classes (classification level 'Class'). Only 6 of the 11 SCOP classes can be found among the three most likely classes for each localization.

Figure 1: Posterior probabilities for different localizations (localization identifiers as in Table 1).

would reflect the prior probability for the same class C_i, given by

$$\frac{|C_i|}{\sum_{j=1}^{34} |C_j|}.$$

Hence, one may assume *prima facie* a dominant class C_i to dominate in posterior probability, too. However, the comparison clearly shows that this is not the case. The most

Table 5: Comparison between a random classifier and a Bayes classifier using localization information for different CATH and SCOP classification levels.

CATH level	**Class**		**Architecture**		**Topology**		**Hom. Superfamily**	
Classifier	**Bayes**	**Random**	**Bayes**	**Random**	**Bayes**	**Random**	**Bayes**	**Random**
Recall	0.602	0.252	0.425	0.049	0.729	0.004	0.786	0.007
Precision	0.576	0.252	0.376	0.039	0.277	0.004	0.246	0.007
Total Accuracy	0.580	0.251	0.306	0.039	0.177	0.004	0.128	0.004

SCOP level	**Class**		**Fold**		**Superfamily**		**Family**	
Classifier	**Bayes**	**Random**	**Bayes**	**Random**	**Bayes**	**Random**	**Bayes**	**Random**
Recall	0.515	0.247	0.787	0.008	0.749	0.006	0.826	0.006
Precision	0.531	0.247	0.397	0.008	0.330	0.007	0.340	0.006
Total Accuracy	0.464	0.243	0.202	0.008	0.120	0.006	0.118	0.005

dominant structural CATH architectural class is 1.10 (Orthogonal Bundle) comprising 16% of all proteins. The second and third largest architectural classes are 3.30 (2-Layer Sandwich) and 3.40 (3-Layer(aba) Sandwich). The remaining architectures are far less dominant. However, even the most predominant architectures do not conspicuously bias the posterior probability.

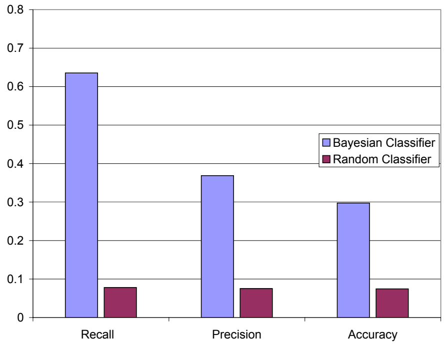
3.1.2 SCOP

The results for the SCOP class level are depicted in Figure 1(b). Similarly as for CATH, for most localizations one of the structural classes of SCOP is dominant. The most prominent example is localization 3, where the dominant class “Alpha and Beta (a/b)” has almost 65% posterior probability. Again, on deeper SCOP levels the correlations are less pronounced, but still strong in some localizations. Overall, the correlation between subcellular localization and structural classes seems weaker in the SCOP classification than in the CATH classification. Nevertheless, we again find the correlation not being dominated by the prior probability.

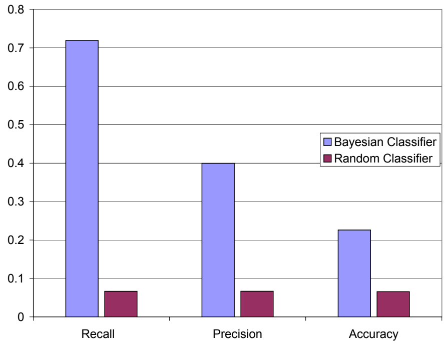
Additional tables for classifications on all hierarchical class-levels of both hierarchies, CATH and SCOP, are available online (see
`http://www.dbs.ifi.lmu.de/research/locfold/dataset/cath_sta.txt`
or
`http://www.dbs.ifi.lmu.de/research/locfold/dataset/scop_sta.txt`,
respectively). We observe the same tendency in all these remaining hierarchical class levels in SCOP and CATH there.

3.2 Well Performing Classifier Based on Posterior Probability Given the Subcellular Localization

The structural classification based on the posterior probability performs surprisingly well, considering the relatively simple assumption it is based on. We would not expect



(a) CATH



(b) SCOP

Figure 2: Comparison between a random classifier and a Bayes classifier using localization information. The results are averaged over the 4 CATH and SCOP classification levels.

the localization of a protein to completely determine its structural class. Otherwise, in a given cellular compartment, only proteins of a specific structure would occur. Clearly, this is not the case for most localizations. Although the absolute values for recall, precision, and accuracy on CATH structural classification based on the localization of a protein may seem not too impressive, the comparison to the expected value (given as random classification) shows a valuable contribution of the localization information to classification behavior. In Figure 2, the results for posterior probability classification and random classification are depicted averaged over all CATH levels. In Table 5, the values are given for each CATH level separately. The classification on SCOP structural classes shows a behavior similar to the classification according to CATH. In average over all SCOP levels, recall and precision are marginally weaker than for CATH, while the accuracy is better. In any case, the behavior is obviously superior to a random classification (see Figure 2). Table 5 presents recall, precision, and accuracy values for each SCOP level separately.

Classifiers for both classification systems, CATH and SCOP, are available online (see `http://www.dbs.ifi.lmu.de/research/locfold/`).

3.3 Subcellular Localization as a Feature Improves a Feature Space for Structural Classification

Finally, we report the experimental results to evaluate the subcellular localization as a feature in comparison to the established feature space as described above.

Table 6: Ranking of the localization feature by different feature evaluation methods (best ranking is 1, worst ranking is 127).

	CATH level				SCOP level			
Method	**Class**	**Arch.**	**Topol.**	**Hom.**	**Class**	**Fold**	**Superfam.**	**Family**
RELIEF	1	1	1	1	4	1	1	1
GainRatio	12	3	1	1	8	1	1	1
SVM-Rank	31	2	9	2	41	39	48	2

3.3.1 Competitive evaluation

RELIEF and GainRatio assign a very high rank on all levels of CATH and SCOP to the localization feature compared to the 126 established features (Table 6). At most levels, RELIEF and GainRatio both assign the top rank 1 to the localization feature. This means, if both methods were used to select just one feature out of all 127 features, the localization feature would have been chosen in all these cases before any other feature out of the already well-established feature space. The SVM-Rank, too, yields a position in the upper third of all features in all cases. This shows that a feature based on the subcellular localization of a protein is valuable even compared to a well-established feature space.

3.3.2 Performance in a combined feature space

Since the established feature space 'D' already allows for a good classification performance of well above 90% for our data sets, using the localization as an additional feature is not of great impact anymore. The results are reported in Table 7. However, we observe rather a slight increase in recall, precision, and accuracy than a deterioration. Generally, while the results seem rather inconclusive in the SCOP data set, in the CATH data set the positive impact of the localization feature is more conspicuous. Note that a difference of 0.001 in accuracy corresponds to approximately 1.8 or 2.7 differently classified proteins for CATH and SCOP, respectively.

On the class level, we would not expect any impact of a new feature, since the classes on class level, both in CATH and SCOP, represent the secondary structure composition only. Secondary structure information is equally included in all three feature sets. The localization feature improves the classification behavior more distinctly on deeper levels. Furthermore, the observed effect is more striking for the k-NN-classifier than for the decision tree and, in turn, than for the SVM classifier. On the one hand, SVM classifiers reach higher values for recall, precision, and accuracy anyway. So it is harder to increase values that already nearly reach 100%. But, on the other hand, the observation regarding the positive impact on the performance of the classifiers corresponds to the observations stated above (Section 3.3.1) for the feature evaluation methods. The RELIEF feature evaluation is related to the k-NN-approach, the GainRatio

Table 7: Comparison of classification results for different feature vectors.

	Features	**D**	**D+R**	**D+L**	**D**	**D+R**	**D+L**	**D**	**D+R**	**D+L**	**D**	**D+R**	**D+L**
	CATH level	**Class**			**Architecture**			**Topology**			**Hom. Superfamily**		
***k*-NN**	**Recall**	0.922	0.923	0.933	0.914	0.912	0.919	0.876	0.871	0.887	0.855	0.854	0.868
	Precision	0.981	0.983	0.982	0.926	0.923	0.930	0.920	0.918	0.930	0.876	0.878	0.892
	Accuracy	0.968	0.970	0.975	0.916	0.913	0.922	0.906	0.903	0.913	0.891	0.889	0.901
J48	**Recall**	0.957	0.957	0.957	0.933	0.929	0.945	0.913	0.911	0.929	0.906	0.906	0.925
	Precision	0.939	0.939	0.939	0.933	0.930	0.946	0.916	0.915	0.932	0.907	0.907	0.929
	Accuracy	0.976	0.976	0.976	0.927	0.928	0.943	0.939	0.938	0.944	0.931	0.931	0.943
SVM	**Recall**	0.956	0.956	0.956	0.946	0.946	0.954	0.958	0.958	0.959	0.947	0.946	0.954
	Precision	0.989	0.989	0.988	0.962	0.962	0.968	0.972	0.971	0.975	0.959	0.956	0.964
	Accuracy	0.984	0.984	0.984	0.955	0.955	0.962	0.967	0.966	0.970	0.962	0.960	0.967
	SCOP level	**Class**			**Fold**			**Superfamily**			**Family**		
***k*-NN**	**Recall**	0.968	0.967	0.964	0.948	0.949	0.950	0.948	0.944	0.951	0.946	0.948	0.951
	Precision	0.972	0.977	0.973	0.960	0.9621	0.961	0.958	0.956	0.957	0.953	0.955	0.957
	Accuracy	0.981	0.980	0.982	0.965	0.967	0.968	0.965	0.963	0.968	0.963	0.965	0.967
J48	**Recall**	0.966	0.966	0.966	0.959	0.959	0.958	0.956	0.956	0.960	0.957	0.955	0.958
	Precision	0.970	0.970	0.970	0.961	0.961	0.961	0.955	0.955	0.961	0.960	0.959	0.964
	Accuracy	0.984	0.984	0.984	0.968	0.968	0.968	0.965	0.965	0.969	0.968	0.967	0.968
SVM	**Recall**	0.982	0.985	0.982	0.988	0.989	0.987	0.989	0.989	0.986	0.990	0.990	0.987
	Precision	0.993	0.990	0.990	0.991	0.991	0.990	0.988	0.988	0.987	0.990	0.989	0.988
	Accuracy	0.992	0.992	0.991	0.991	0.992	0.991	0.991	0.992	0.991	0.993	0.993	0.991

Table 8: CATH: Depth of decision trees and usage of the localization feature.

CATH level	depth of tree	depth of localization feature	usage of localization feature
Class	10	–	0%
Architecture	13	3–12	8%
Topology	16	3–6	8.5%
Homol. Superfamily	12	3–7	6.2%

feature evaluation is related to the decision tree (and is, in fact, the method used in J48 to select a feature for a certain decision node), and the SVM-Rank is, obviously, related to SVM classification. Inspection of the constructed decision trees allows to conclude that the random feature contributed nothing to the classification, i.e., the random feature appears scarcely ever in a decision tree. In contrast, the localization feature appears almost always very early in the decision path (again corresponding to the high rank in GainRatio). To illustrate this observation, note the summary of decision trees for the CATH classification in Table 8. The value "usage of loc." is the frequency of the usage of the localization feature as split-attribute in all decision nodes in the corresponding tree.

4 Conclusion

Our line of reasoning was guided by the observation that many proteins are specialized to a certain compartment of a cell. Thus, a first, rather naive hypothesis to subsume our results may be stated as follows:

Assumption 1
The subcellular localization of a protein is sufficient to conclude the structural class of a protein.

This assumption can explain the observed correlation between both properties of a protein (Section 3.1). If this hypothesis were true, it could serve as a foundation for a classifier predicting the structural class solely based on the posterior probability given the subcellular localization of a protein (Section 3.2). This hypothesis may be true in some cases. For example, for the localization "nucleus", the most probable fold in SCOP is the "Histone-fold". The most likely family is "Nucleosome core histones". This makes perfect sense, since histones are typical proteins in the nucleus. However, the naive hypothesis is unlikely to hold in all cases, since this would mean that in a given cellular compartment only specific structural classes of proteins would occur. In fact, the absolute values of posterior probabilities suggest the opposite, since a probability of, e.g., 30% for a specific structural class (based on statistical evaluation) means that about 70% of all proteins in the very same localization exhibit a different structure.

These observations lead to a second, more refined hypothesis, according to our reasoning outlined in the introduction:

Assumption 2
In addition to the amino acid sequence of a protein, its subcellular localization considerably contributes to its finally adopted fold.

As the growing prosperity of localization prediction methods based on the amino acid composition or signal peptides suggests, the localization of a protein is itself coded in the amino acid sequence in many cases. In other cases, it may be encoded in the mRNA sequence, but it is then not present in the translated sequence anymore. However, the assumption is, that the localization of a protein may be an information of interest in assigning a structural class to an amino acid sequence. The refined hypothesis is backed by our results in Section 3.3.1 and biologically makes sense. All evaluation methods rank the localization feature in the upper third in a set of well-established, sequence-based features. Two of three methods even assign the first rank for several levels of structural classification. The next logical step is to use the localization feature indeed as one dimension among others in a feature space for classification. Here, our results remain somewhat inconclusive. While on the CATH data set, we note in many cases an improvement for classification performance, the results on the SCOP data set are not equally validating our hypothesis. However, the strong results of the statistical evaluation and the high ranking of the feature based on localization information may trigger further studies. Biologically, these findings make perfect sense since most proteins can be expected to be well adapted to the physicochemical properties of a specific localization.

References

[Andrade et al. (1998)] M. A. Andrade, S. I. O'Donoghue, and B. Rost. Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology*, 276:517–525, 1998.

[Anfinsen (1973)] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[Aßfalg et al. (2008)] J. Aßfalg, J. Gong, H.-P. Kriegel, A. Pryakhin, T. Wei, and A. Zimek. Supervised ensembles of prediction methods for subcellular localization. In *Proceedings of 6th Annual Asia Pacific Bioinformatics Conference (APBC), Kyoto, Japan*, 2008.

[Aßfalg et al. (2009)] J. Aßfalg, J. Gong, H.-P. Kriegel, A. Pryakhin, T. Wei, and A. Zimek. Supervised ensembles of prediction methods for subcellular localization. *Journal of Bioinformatics and Computational Biology*, 7:269–285, 2009.

[Bannai et al. (2002)] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.

[Berman et al. (2000)] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[Bhasin and Raghava (2004)] M. Bhasin and G. P. S. Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, 32(Web Server Issue):W414–W419, 2004.

[Chothia (1992)] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357: 543–544, 1992.

[Ding and Dubchak (2001)] C. H. Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

[Dubchak et al. (1995)] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 92:8700–8704, September 1995.

[Dubchak et al. (1999)] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. Kim. Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function, and Genetics*, 35:401–407, 1999.

[Gardy et al. (2005)] J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. L. Brinkman. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21(5):617–623, 2005.

[Garg et al. (2005)] A. Garg, M. Bhasin, and G. P. S. Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of Biological Chemistry*, 280(15):14427–14432, 2005.

[Godzik (2003)] A. Godzik. Fold recognition methods. In P. E. Bourne and H. Weissig, editors, *Structural Bioinformatics*, chapter 26, pages 525–546. John Wiley&Sons, 2003.

[Govindarajan and Goldstein (1996)] S. Govindarajan and R. A. Goldstein. Why are some protein structures so common? *Proceedings of the National Academy of Sciences of the United States of America*, 93:3341–3345, April 1996.

[Govindarajan et al. (1999)] S. Govindarajan, R. Recabarren, and R. A. Goldstein. Estimating the total number of protein folds. *Proteins: Structure, Function, and Genetics*, 35:408–414, 1999.

[Guyon et al. (2002)] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[Horton et al. (2006)] P. Horton, K.-J. Park, T. Obayashi, and K. Nakai. Protein subcellular localization prediction with WoLF PSORT. In *Proceedings of 4th Annual Asia Pacific Bioinformatics Conference (APBC), Taipei, Taiwan*, 2006.

[Hua and Sun (2001)] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[Höglund et al. (2006)] A. Höglund, P. Dönnes, T. Blum, H.-W. Adolph, and O. Kohlbacher. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, 22(10):1158–1165, 2006.

[Kira and Rendell (1992)] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning (ML 1992), Aberdeen, Scotland, UK*, 1992.

[Kononenko (1994)] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94, European Conference on Machine Learning, Catania, Italy, April 6-8, 1994, Proceedings*, 1994.

[Lu et al. (2004)] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–556, 2004.

[Murzin et al. (1995)] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[Nakai and Horton (1999)] K. Nakai and P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24(1):34–36, 1999.

[Orengo et al. (1997)] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[Park and Kanehisa (2003)] K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, 2003.

[Platt (1998)] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1998.

[Quinlan (1986)] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[Quinlan (1993)] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Shen and Chou (2006)] H.-B. Shen and K.-C. Chou. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22(14):1717–1722, 2006.

[Small et al. (2004)] I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for n-terminal targeting sequences. *Proteomics*, 4(6):1581–1590, 2004.

[Witten and Frank (2005)] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition, 2005.

[Wolf et al. (2000)] Y. I. Wolf, N. V. Grishin, and E. V. Koonin. Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, 299: 897–905, 2000.

[Wu et al. (2006)] C. Wu, R. Apweiler, A. Bairoch, D. Natale, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, 34:D187–D191, 2006.

[Yu et al. (2004)] C.-S. Yu, C.-J. Lin, and J.-K. Hwang. Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Science*, 13:1402–1406, 2004.