Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 21

Robust Remote Homology Detection by Feature Based Profile Hidden Markov Models

Thomas Plötz* Gernot A. Fink†

*Bielefeld University, Faculty of Technology, tploetz@techfak.uni-bielefeld.de

†Bielefeld University, Faculty of Technology, gernot@techfak.uni-bielefeld.de

Robust Remote Homology Detection by Feature Based Profile Hidden Markov Models*

Thomas Plötz and Gernot A. Fink

Abstract

The detection of remote homologies is of major importance for molecular biology applications like drug discovery. The problem is still very challenging even for state-of-the-art probabilistic models of protein families, namely Profile HMMs. In order to improve remote homology detection we propose feature based semi-continuous Profile HMMs. Based on a richer sequence representation consisting of features which capture the biochemical properties of residues in their local context, family specific semi-continuous models are estimated completely data-driven. Additionally, for substantially reducing the number of false predictions an explicit rejection model is estimated. Both the family specific semi-continuous Profile HMM and the non-target model are competitively evaluated.
In the experimental evaluation of superfamily based screening of the SCOP database we demonstrate that semi-continuous Profile HMMs significantly outperform their discrete counterparts. Using the rejection model the number of false positive predictions could be reduced substantially which is an important prerequisite for target identification applications.

KEYWORDS: Profile Hidden Markov Models (Profile HMMs), remote homology detection, protein sequence analysis, feature representation, target identification

---

*This work was supported by Boehringer Ingelheim and the Boehringer Ingelheim Pharma GmbH und Co. KG Genomics Group.

1 Introduction

Despite impressive improvements in both sensitivity and specificity obtained by the application of powerful probabilistic sequence analysis techniques, robust remote homology detection is still a challenging problem. Especially for target identification within drug discovery, the detection of new members of therapeutically relevant protein families is of fundamental scientific as well as commercial interest.

In the last decade, probabilistic modeling of protein families by means of Profile Hidden Markov Models became one of the dominating approaches for biological sequence analysis. They represent an important framework for obtaining fundamental biological insights by exploiting the results of major sequencing projects when aiming at general understanding of biological processes. The reason for the broad popularity of Profile HMMs is mainly given by the existence of efficient algorithms for both model estimation and evaluation. The basic technical challenge is the estimation of robust family models for highly diverging but related protein sequences with small amounts of training samples. In order to tackle this problem, several refinements of the basic Profile HMM estimation approach were proposed. By means of certain model regularization techniques, the parameters of Profile HMMs can be estimated using small sets of training samples. Additionally, during model evaluation target hits and misses are discriminated by means of specialized null models for log-odds scoring.

However, the incorporation of prior expert knowledge into the modeling process, e.g. by carefully designed Dirichlet mixtures [Brown et al., 1993], is critical regarding the detection of really new homologues. Models created by means of small sample sets containing data actually belonging to the appropriate target family, and larger amounts of potentially biased expert knowledge tend to focus on patterns already known. This means that their generalization abilities can be limited. In order to gain really new knowledge which is important for e.g. pharmaceutical purposes, alternative approaches need to be developed.

We present feature based semi-continuous Profile HMMs and application concepts for robust remote homology detection. The features extracted from a continuous signal-like protein data representation based on various biochemical properties and local residual context provide a richer sequence representation which is advantageous especially for remote homology detection. According to the critical incorporation of manually derived prior expert knowledge, the focus of our developments is on completely data driven techniques. Our principle goal is to estimate probabilistic models for protein families as un-biased as possible at every stage of the modeling process while keeping the robustness high even for small training sets. Compared to state-of-the-art discrete modeling approaches, feature based Profile HMMs presented in this article show superior performance for superfamily based remote homology detection tasks. Our new approach addresses the improvement of the general method of Profile HMM

based sequence comparison. Thus, also iterative model estimation processes (cf. [Karplus et al., 1998]) can benefit from our techniques.

Semi-continuous HMMs as introduced in [Huang and Jack, 1989] represent a modeling technique for effective exploitation of small training sets. Emissions are described by mixture densities based on a shared set of Gaussians which can be considered as a general representation of the feature space. The estimation of standard Profile HMMs, where both the emission space representation and the model structure are jointly optimized, requires rather large amounts of model specific training samples for data-driven model estimation. Contrary to this, the estimation of semi-continuous HMMs can be divided into training steps separately optimizing the emission space, and based on this the model structure. Only for the latter case model specific data is required. For the data-driven estimation of our feature space representation large amounts of un-annotated sequence data can be used, e.g. the complete SWISSPROT database. By applying different adaptation techniques, the general mixture representation of the emission space can be focused on particular protein families using small target specific training sets. Based on the new semi-continuous Profile HMMs robust remote homology detection becomes possible. For further reduction of the number of false predictions which are critical e.g. for drug discovery applications, we propose competitive model evaluation using a particular family model and an explicit rejection model estimated on general, i.e. un-annotated protein data.

This paper is organized as follows. Based on the analysis of current Profile HMMs (section 2) our new feature based sequence representation is presented in section 3. In section 4 the newly developed semi-continuous Profile HMMs are introduced which allow robust remote homology detection. Evaluation results based on SCOP superfamily experiments illustrating the improved performance of the new approach are presented in section 5.

2 Discrete Profile HMMs

Profile HMMs currently represent the most important statistical models used for probabilistic sequence analysis of biological data. The typical architecture of these models is shown in figure 1. Usually, the conserved parts of a multiple alignment of the sequences belonging to the protein family of interest are modeled by a linear sequence of match states M_i. A position in the alignment is considered conserved if some residue is present for the majority of sequences. In order to capture variations in sequence length insertions and deletions of residues are described by additional insert I_i and delete states D_i. Besides model estimation based on preceding separate multiple alignments, in the literature alternative approaches are described where models are created by iterative refinements using un-aligned training sequences [Krogh et al., 1994]. There are some extensions to the basic architecture with increased flexibility, e.g. in HMMER's Plan7 [Eddy, 2001].

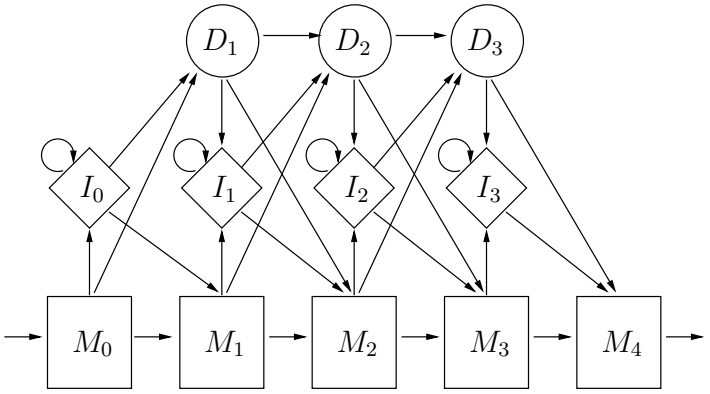


Figure 1: State-of-the-art Profile HMM

Currently, the emissions of Profile HMMs are modeled by state dependent discrete probability distributions over the set of 20 amino acids. Transition and emission probabilities are estimated using standard Baum-Welch or Viterbi training. For classification of sequence data the models are evaluated by computing the Forward or Viterbi scores, respectively. An excellent treatment of Profile HMMs can be found in [Durbin et al., 1998].

For detection tasks, the scores generated by aligning query sequences to the appropriate family models are evaluated regarding a threshold. Since these scores are depending on both the length of the sequences as well as on the length of the models, usually, they are considered regarding the scores generated by some background or null model. The resulting ratio of both scores is called the log-odds score and target hits are assumed for statistically significant values. The actual choice of the appropriate background model is rather crucial for the overall detection performance and target specific background models are widely used [Barrett et al., 1997].

Especially for remote homology detection tasks the number of training samples for estimating the target specific Profile HMM is usually rather small which is disadvantageous for *robust* model estimation. Thus, several so-called model regularization techniques were proposed which try to tackle this so-called sparse data problem. The currently most promising technique for obtaining statistically more “stable” amino acid distributions is based on the incorporation of prior knowledge by means of carefully designed Dirichlet distributions. For a review of model regularization techniques cf. e.g. [Karplus, 1995].

Common to all present Profile HMM approaches is the explicit and exclusive use of symbolic sequence representations. Furthermore, contextual information cannot be captured by the discrete probability distributions used.

3 Feature Based Profile HMMs

For most current Profile HMM based sequence analysis applications, raw amino acid data is processed. This seems obvious because it is usually the result of major sequencing projects and especially for target identification within the drug discovery pipeline often it is the only source of information. Consequently, discrete Profile HMMs are the methodology of choice for probabilistic sequence analysis.

Contrary to this, when applying HMMs to general pattern recognition problems like automatic speech recognition, usually, "natural" real-valued signals evolving in time are processed. Based on such signals (e.g. acoustic pressure for speech recognition), relevant features are extracted and used for (semi-)continuous HMMs. By means of the continuous emissions of such models, very flexible modeling of highly diverging data becomes possible usually significantly outperforming discrete HMMs.

The biological functions of proteins are principally caused by their biochemical properties which determine their three-dimensional structure. In fact, this spatial folding and thus the biochemical properties cause functional similarities of proteins motivating the definition of protein (super)families. Amino acid symbols "summarize" such properties of the residues which is usually implicitly respected by means of specific substitution matrices. However, this abstraction seems crucial because details investigated throughout the years in wet-lab research are neglected when processing symbolic data. The huge arsenal of powerful pattern recognition techniques cannot be used directly for protein sequences in their current representation.

In order to achieve better results for remote homology detection tasks we developed semi-continuous Profile HMMs. Their emissions are based on features extracted using pattern recognition techniques from a protein sequence representation which explicitly captures the underlying biochemical properties of the appropriate residues. Whereas the emissions are changed towards a continuous feature representation, the principle state-of-the-art Profile HMM architecture as shown in figure 1 remains the same. We first introduced the feature extraction approach described here in [Plötz and Fink, 2004].

In the following we will first explain our signal representation of the protein data (section 3.1). Subsequently, in section 3.2 the general feature extraction process as well as the actual modeling is described in detail.

3.1 Signal Representation of Proteins

The basic motivation for alternative representations of biological sequences is reasoned by the huge amount of pattern recognition techniques available. By means of such techniques, the application of Profile HMMs can benefit from uncovering possibly hidden characteristics of biochemical properties of protein

data. Explicitly exploiting such information can especially increase the performance of remote homology detection approaches.

Reconsidering the argumentation regarding the abstraction from biochemical properties when using raw amino acid data, the most promising signal representation approaches in fact rely on such properties. Kawashima et al. compiled a huge amount of so-called amino acid indices [Kawashima and Kanehisa, 2000]. Every index defines a mapping of amino acids to numerical values depending on its natural properties, e.g. hydrophobicity, or molecular weight. Once the amino acids are appropriately mapped to numerical values, a large variety of signal processing techniques can be applied.

In our approach we, basically, follow the idea of mapping residues to numerical values as defined in amino acid indices. However, limiting the sequence representation to an arbitrary but single index implies neglecting putative higher level relationships of the residuals. Furthermore, there is hardly any *exhaustive* prior knowledge, which property causes remote homologue sequences to belong to a distinct protein family – usually it cannot be specified exclusively. Therefore, we do not want to restrict the representation to a single biochemical property but to incorporate *multiple* properties of amino acids relevant for protein family affiliation.

We carefully selected 35 indices out of the huge pool of encoding schemes available, normalized each of them to the interval of $[-1, \ldots 1]$, and used the combination of them as a multi-channel signal representation which provides a rich characterization of the protein sequence analyzed. The actual selection of the indices followed biological considerations as explained in the following. On their website Kawashima and colleagues deliver a cluster map of the approximately 500 indices contained in the abovementioned database. The correlation coefficients of the particular indices are the basis for this statistical clustering. All biochemical properties considered are assigned to six coarse categories. When selecting the indices for our new signal-like protein sequence representation, we aimed at a broad coverage of these six categories in order to actually capture the most relevant biochemical properties of amino acids. Note that we do not focus on a completely redundancy-free selection of amino acid indices at this initial stage of feature extraction. Actually this is a very challenging problem which can hardly be solved manually. Instead, we perform an automatic redundancy reduction in the final stage of the overall feature extraction process (cf. section 3.2). In appendix A the 35 amino acid indices used for the signal-like representation of protein sequences are listed including the cluster map of Kawashima et al. where the selected indices are highlighted.

3.2 Feature Based Sequence Representation

The previously described sequence encoding method subsumes information from various sources in a multi-channel numerical representation. Generally, when

using Profile HMMs for protein families, two levels of residual context are considered. The classification result determining the decision regarding the probable affiliation of the sequence analyzed to a particular protein family is performed using the complete sequence. Thus, the global context is captured by the HMM. Contrary to this, for the estimation of emission probabilities no residual context is used at all.

In order to respect *local* signal characteristics already at the level of emission probabilities, in our feature extraction procedure local contexts of residues are considered. Consecutive samples of the 35 channel signals are analyzed using a sliding window approach (extracting *frames*). Starting from the first residue of a distinct sequence for each of the 35 channels 16 samples are used for short length signal analysis. The window size was heuristically determined in informal experiments. At the borders of the sequences the data is padded using prior probabilities of amino acids obtained from general protein data.

Basically, for remote homology detection the essentials of a particular protein family are of major interest. Thus, any putatively misguiding signal specialties relevant only for a minority of sequences belonging to the family of interest should be neglected. In summary, signal analysis should produce features enabling an abstract but meaningful view on the sequences representing the coarse shape. For extracting such features independently of the actual signal type, usually, a spectral analysis is performed. Transforming signals into a frequency based representation offers direct access to the desired shape approximation. Within the general pattern recognition domain a wide variety of such transformations has been developed, e.g. the Fourier or the Cosine transform. Once the transformation is estimated, distinct parts of frequencies of the signal can easily be removed by skipping single transformation coefficients. Thus, reducing the original signal to its coarse shape is straightforward.

While analyzing signals of protein sequences we found out that the standard spectral analysis approach using Fourier transform is not suitable for biological signals subsumed in the frames introduced above. This function transform assumes periodic signals of infinite length which is in no way the case for our data. Thus, we used a refined function transform technique which is more suitable for the short signals analyzed in our approach – the Discrete Wavelet Transform (DWT). Exemplary, in [Percival and Walden, 2000] a very detailed overview of general Wavelet based function analysis is provided whereas [Poularikas, 2000] gives a more general overview of function transforms including Wavelets. The basic advantage of the DWT for the analysis of the signal frames is the superb localization property in both time, i.e. the position of the residues, and frequency space. The coarse temporal signal structure of the protein sequences analyzed is determined channelwise, i.e. the DWT is applied to every channel of our multichannel signal representation individually.

In order to obtain the abovementioned coarse signals for every channel of a particular frame, we skip the upper five Wavelet detail coefficients. Informal

experiments on related data showed that the reconstruction of the original signals based on the first 11 Wavelet coefficients obtained after a two-stage multiscale analysis is suitable for the necessary abstraction from putatively misguiding details. Per frame the channel based feature vectors are concatenated to 385-dimensional vectors (35 channels $\times$ 11 DWT-coefficients).

The actual relevance of a single channel of the biological signal for the relationship to a distinct protein family can hardly be determined. Therefore, the combination of several biochemical properties needs to be considered. In order to improve remote homology detection, the goal of the proposed method is to avoid as many pitfalls in the early stages of sequence analysis as possible. Besides this, manually fixing the residues' properties tends to models whose generalization abilities regarding the detection of currently unknown homologues are limited. Thus, we propose a completely data driven approach. Potentially redundant information included in the concatenation of the DWT coefficients needs to be removed.

Generally, redundancy within the 385-dimensional feature vectors may originate from two different sources. First, even the most careful manual selection of relevant biochemical properties for the multi-channel signal-like representation of protein sequences (cf. section 3.1) does not necessarily guarantee that the particular amino acid indices do not provide completely complementary information. Second, the frames extracted using the sliding window approach for a particular protein sequence overlap substantially resulting in smoothed but (to some degree) redundant feature vectors.

The abovementioned redundancy is reduced using an automatic and data-driven approach, namely by finally performing a Principle Component Analysis (PCA) for the feature vectors of every frame. Usually, the (lower) dimension of the target feature space where the original feature vectors are projected to is determined by analyzing the spectrum of the eigenvalues of the appropriate covariance matrix with respect to the percentage of achievable reconstruction of the original data. In our case, a compact representation of the feature vectors in a 99-dimensional subspace is sufficient for more than 95% reconstruction. Thus, the dimensionality of the final feature vectors for protein sequence data is adjusted to 99.[1]

Figure 2 summarizes the feature extraction method described in this paper for an exemplary protein sequence (*Coxsackie virus and adenovirus receptor, domain 1 – Homo sapiens*; PDB-Id: 1f5w.). Starting from the plain protein sequence, frames of length 16 are extracted (upper part of the figure). For every frame the 35 channel encoding procedure is performed, resulting in the signal representation shown on the left side. The first 11 DWT coefficients of all chan-

[1]Note that due to the channel-wise linear DWT-based analysis of the sequence windows containing 16 residues at least 15 linear constraints exist. Choosing a full rank subset of the considered feature vectors would affect the definition of the principal components in a rather arbitrary way.

nels, illustrated on the right side, are concatenated to a single feature vector per frame. The concluding PCA (lower part) reduces the dimension to 99.

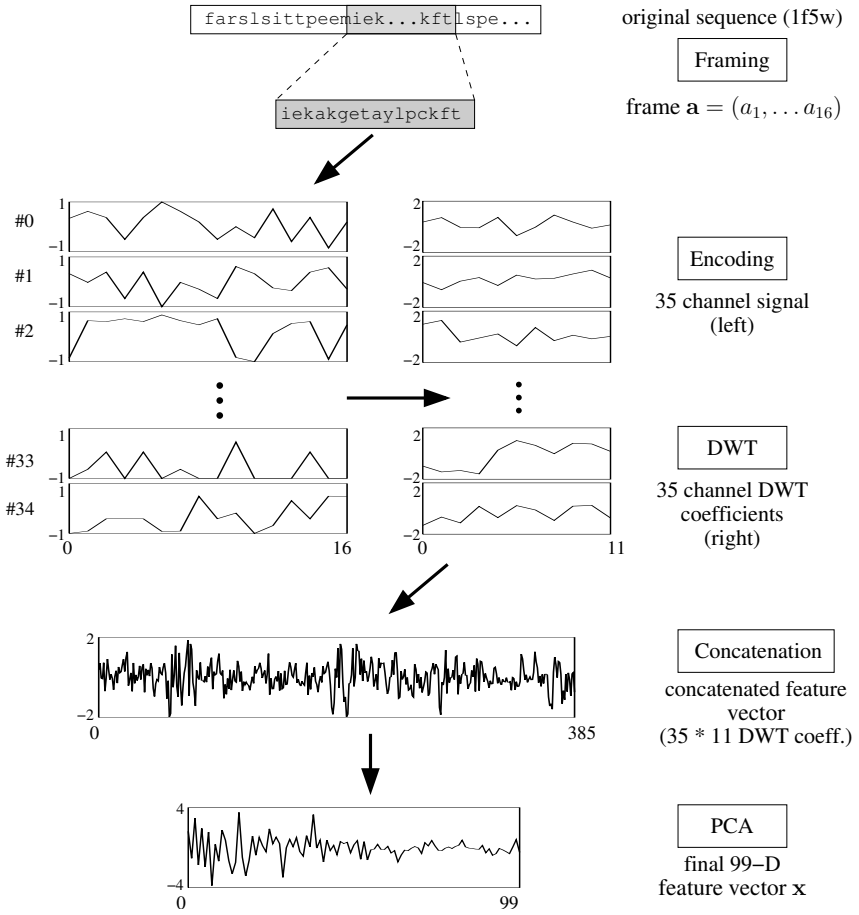


Figure 2: Feature extraction process for an exemplary protein sequence – *Coxsackie virus and adenovirus receptor, domain 1 (Homo sapiens)*; PDB-Id: 1f5w.

4 Robust Model Estimation and Remote Homology Detection

When processing feature vectors, generally continuous instead of discrete HMMs are used. Here, the continuous feature space is represented by means of mixture densities. For effective exploitation of training data, Huang and Jack proposed semi-continuous HMMs where all states share a common set of mixture densities which are weighted state-specifically [Huang and Jack, 1989]. Compared to continuous models only one global set of component densities needs to be estimated which is advantageous for small training sets. This shared set of densities can be considered as a general mixture representation of the feature space.

For a feature vector $\mathbf{x}$ corresponding to a frame of residues $\mathbf{a} = (a_1, \ldots, a_{16})$, the emissions $b_j(\mathbf{x})$ of HMM states j are defined as mixtures of K Gaussians $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k)$ with mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\mathbf{C}_k$ which are used for all HMM states but individually weighted by c_{jk}:

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) = \sum_{k=1}^{K} c_{jk} g_k(\mathbf{x}) \tag{1}$$

In our approach we replace the discrete emissions of state-of-the-art Profile HMMs with such semi-continuous emissions while keeping the original model topology as illustrated in figure 1.

By analyzing equation 1 it becomes clear that the estimation of semi-continuous HMMs can principally be divided into two separate steps. The model independent feature space representation, i.e. the Gaussian mixture component densities, can be obtained using general feature data. Subsequently, the model itself is optimized based on the resulting component densities and model specific training samples. We found that the separation of the estimation of a general feature space representation from position specific modeling is the basic advantage of semi-continuous modeling which can be exploited for robust estimation of protein families using small family specific sample sets.

In this section we present modeling techniques for data driven estimation of *robust* target specific semi-continuous Profile HMMs.

4.1 Robust Model Estimation

The parameters of the general mixture density based feature space representation are obtained by applying a modified k-means procedure to general protein data which is comparable to the Expectation-Maximization (EM) approach [Dempster et al., 1977]. The base for the un-supervised and completely data driven estimation of mixture densities are all sequences (approximately 90K) from the SWISSPROT database [Boeckmann et al., 2003] allowing the estimation of 1,024 Gaussians.

Technically, semi-continuous Profile HMMs are derived from the architecture of discrete models. Given the Profile structure, standard Viterbi training is performed using the component densities of the general feature space representation and small amounts of family specific data.

The mixture density representation of the feature space obtained from SWISSPROT captures the *global* properties of general protein data. In order to focus this representation to specific properties of proteins belonging to a particular target family, data driven mixture adaptation techniques are applied. Such transformations of the mixture parameters, i.e. mean vectors $\boldsymbol{\mu}_k$ and (not necessarily) covariance matrices $\mathbf{C}_k$, are used in order to optimize the coverage of general protein properties towards more family specific characteristics. Note that the

model structure, the transition probabilities as well as the state specific mixture weights c_{jk} remain unchanged during adaptation. Only the underlying mixture component densities are modified.

The initial estimation of the mixture densities using general protein data is, furthermore, the basis for the either probabilistic or deterministic assignment of target family specific samples to the particular Gaussians. Using these pre-trained parameters the actual mixture adaptation can be performed very robustly.

For the target family specific transformation of the densities, we investigated three different adaptation techniques which are described in the following. The number of family specific training samples, i.e. the amount of adaptation data which is usually very small, is denoted by T.[2]

Maximum Likelihood (ML) Estimation:

In the simplest case of target family based specialization of the feature space representation, the adaptation is performed by maximizing the likelihood of the mixture densities for the family specific sample sequences using EM up to convergence. Given the initial mixture representation of the feature space derived from SWISSPROT, the adaptation samples $\mathbf{x}_t$ are assigned probabilistically to all mixtures g in the iterative re-estimation of the parameters:

$$\hat{\boldsymbol{\mu}}_k^{m+1} = \frac{\sum_{t=1}^{T} \xi_t^m(k)\mathbf{x}_t}{\sum_{t=1}^{T} \xi_t^m(k)} \quad (2)$$

$$\hat{\mathbf{C}}_k^{m+1} = \frac{\sum_{t=1}^{T} \xi_t^m(k)\mathbf{x}_t\mathbf{x}_t^T}{\sum_{t=1}^{T} \xi_t^m(k)} - \hat{\boldsymbol{\mu}}_k^{m+1}(\hat{\boldsymbol{\mu}}_k^{m+1})^T \quad (3)$$

$$\hat{p}_k^{m+1} = \frac{1}{T}\sum_{t=1}^{T} \xi_t^m(k) \quad (4)$$

$$\xi_t^m(k) = P(g_t = k|\mathbf{x}_t, \hat{p}_k^m, \hat{\boldsymbol{\mu}}_k^m, \hat{\mathbf{C}}_k^m)$$

The probability of selecting the k-th Gaussian for the t-th adaptation sample ($g_t = k$) given the current estimates of the mixtures' parameters is denoted by $\xi_t^m(j,k)$, whereas $\hat{p}_k$ represents the prior probability of the k-th Gaussian. Since the parameters of all densities are re-estimated by the ML procedure, usually rather large sample sets are required for robust adaptation.

Maximum A-Posteriori (MAP) Adaptation:

Contrary to the ML approach, here, the iterative adaptation of the component densities is performed with respect to optimization of the posterior probability

[2]According to common experience of molecular biologists working in the field of drug discovery, typically (depending on the actual task) for remote homology detection only some dozens of sample sequences are available.

of the mixture parameters for the adaptation samples. Generally, prior parameter estimates $\hat{\boldsymbol{\mu}}_k$ and $\hat{\mathbf{C}}_k$ weighted by τ are combined with the re-estimation based on the family specific data by changing equations 2 and 3 to:

$$\hat{\boldsymbol{\mu}}_k^{m+1} = \frac{\tau\hat{\boldsymbol{\mu}}_k^m + \sum_{t=1}^{T}\xi_t^m(k)\mathbf{x}_t}{\tau + \sum_{t=1}^{T}\xi_t^m(k)} \tag{5}$$

$$\hat{\mathbf{C}}_k^{m+1} = \frac{\tau(\hat{\mathbf{C}}_k^m + \hat{\boldsymbol{\mu}}_k^m(\hat{\boldsymbol{\mu}}_k^m)^T) + \frac{\sum_{t=1}^{T}\xi_t^m(k)\mathbf{x}_t\mathbf{x}_t^T}{\sum_{t=1}^{T}\xi_t^m(k)}}{\tau + \sum_{t=1}^{T}\xi_t^m(k)} - \hat{\boldsymbol{\mu}}_k^{m+1}(\hat{\boldsymbol{\mu}}_k^{m+1})^T \tag{6}$$

Initial parameter estimates are obtained by applying the (modified) k-means algorithm to SWISSPROT data. The advantage of MAP adaptation is the balanced incorporation of prior information extracted from the larger set of un-labeled sequences depending on the actual amount of adaptation data. The more adaptation samples available, the stronger the influence of them and vice versa, the smaller the amount of target specific data, the higher the influence of the background estimation. We adjusted τ to the number of samples assigned to the particular mixture components as accumulated during the previous estimation steps which allows robust mixture adaptation even for small training sets.

Maximum Likelihood Linear Regression (MLLR):

For the third kind of adaptation, deterministic assignments of feature vectors $\mathbf{x}_t$ to mixture components are assumed. Originally developed for speaker adaptation of automatic speech recognition systems, Leggetter and Woodland proposed the modification of the mixtures' mean vectors only using affine transformations $\mathbf{W}_k$ [Leggetter and Woodland, 1995]. These transformations represent rotations and translations of the feature space estimated on small adaptation sets. They can be defined with respect to augmented D-dimensional mean vectors $\tilde{\boldsymbol{\mu}}_k = (1, \mu_{k_1}, \ldots, \mu_{k_D})^T$, where D in our case is 99:

$$\hat{\boldsymbol{\mu}}_k = \mathbf{W}_k\tilde{\boldsymbol{\mu}}_k \tag{7}$$

The transformations are generalized to groups of mixture components including densities not covered by the adaptation set via linear regression. Fischer and Stahl developed a simplified adaptation procedure by using a single regression class. This implies a global transformation matrix $\mathbf{W}$ which is applied to all augmented mean vectors $\tilde{\boldsymbol{\mu}}_k$ [Fischer and Stahl, 1999]:

$$\mathbf{W} = \left\{\sum_{t=1}^{T}\mathbf{x}_t\tilde{\boldsymbol{\mu}}_t^T\right\}\left\{\sum_{t=1}^{T}\tilde{\boldsymbol{\mu}}_t\tilde{\boldsymbol{\mu}}_t^T\right\}^{-1} \tag{8}$$

Contrary to ML and MAP adaptation, here instead of statistically re-estimating the mixtures' parameters, the densities themselves are transformed. The transformation estimated for mixtures actually covered by a small adaptation set is generalized to the complete feature space.

By means of all adaptation techniques described here the general feature space representation is focussed on particular target families in a completely data-driven way. For both ML and MAP adaptation all mixture parameters are re-estimated, in the latter case in combination with prior estimates of the mixture parameters. For MLLR only the single transformation matrix $\mathbf{W}$ needs to be estimated which requires considerably smaller amounts of target family specific data. Therefore, MLLR is especially attractive for remote homology detection tasks as addressed in this paper.

4.2 Robust Remote Homology Detection

The major difficulty for detection tasks is the discrimination between target hits and misses which is usually realized by threshold comparison of the scores. For independence regarding the actual length of a query sequence and for robust separation of sequences belonging to the target model and those who are not, discrete Profile HMM evaluation is based on more or less sophisticated null models for log-odds scores. When applying our feature based semi-continuous Profile HMMs to homology detection, we use a null model based on the prior probabilities of the mixture components estimated during model building.

In order to reduce the overall number of false detections, we furthermore apply a technique which is principally known from general detection tasks. Considering e.g. the problem of automatic speaker detection, usually an additional non-target model is estimated which explicitly covers all data *not* belonging to the target class. According to Reynolds such a model is called *Universal Background Model (UBM)* [Reynolds, 1997]. As an enhancement of the general UBM approach, our definition of the background model captures structural information using a left-right topology as outlined in figure 3. We evaluate both the UBM and the particular target model in a competitive manner which is combined with the log-odds scoring method described above. The UBM itself, consisting of $L_U = 30$ states, was estimated on the set of general SWISSPROT data by Baum-Welch training. The actual model length was determined heuristically in informal experiments.

4.3 System Overview

In figure 4 our approach for estimating semi-continuous Profile HMMs and an explicit UBM for robust remote homology detection is summarized graphically. Based on the feature representation of general protein data obtained using our new extraction method, a mixture representation of the general feature

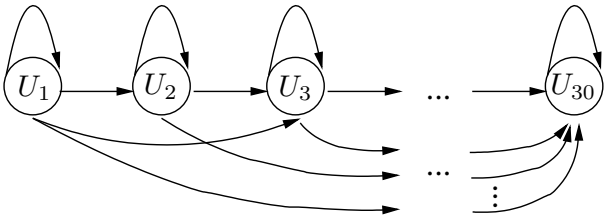


Figure 3: Left-Right topology of the UBM (sketch)

space is estimated using k-means (upper-left part). For semi-continuous modeling, the separate optimization of the emission space representation using large amounts of general protein data and the family specific training of the model structure is possible. By means of standard discrete models λ_D estimated on family specific training samples (upper-right), and the general feature space representation, semi-continuous Profile HMMs λ_G are obtained via Viterbi training (middle-right). Then, the mixture representation is optimized for the target families by applying adaptation techniques resulting in family specific models λ_S (lower-right). Finally, on SWISSPROT data the UBM is estimated (lower-left).

Compared to the state-of-the-art in probabilistic protein family modeling, namely discrete Profile HMMs as summarized in section 2 or treated in detail in e.g. [Durbin et al., 1998], semi-continuous feature based Profile HMMs as proposed in this article basically differ in their state specific emissions. The three-state model structure, and the probabilistic state transitions remain unchanged, though. In table 1 the properties of both modeling approaches are summarized.[3]

By means of an un-supervised estimation technique the 1,024 mixture components are obtained using sufficient amounts of general (un-annotated) protein data. The corresponding state-specific mixture weights (c_{jk} – cf. equation 1) are estimated during model training using small amounts of family specific training samples. Since usually only little training data is available, certainly not all of the 1,024 mixture components will contribute substantially to the state specific emission probabilities. Consequently, several mixture weights will be set to some small floor probability.

However, practical experience in alternative pattern recognition domains where (semi-) continuous HMMs are applied confirms that the resulting *empirical* probability distribution for mixture weights is suitable for robust modeling. Actually, model regularization of semi-continuous HMMs is only discussed at the level of density representations in the literature (cf. e.g. [Huang et al., 2001]). As the underlying protein feature space shows strong locality, i.e. for a particular target family only a small set of relevant Gaussians need to be considered per state. Thus, using the approaches presented in this paper the estimation of model parameters is possible even when only little family specific data is available.

[3]Since the detailed description of the particular protein family models including *all* parameters is very complex the reader interested in these technical details is referred to our website [Plötz, 2004].

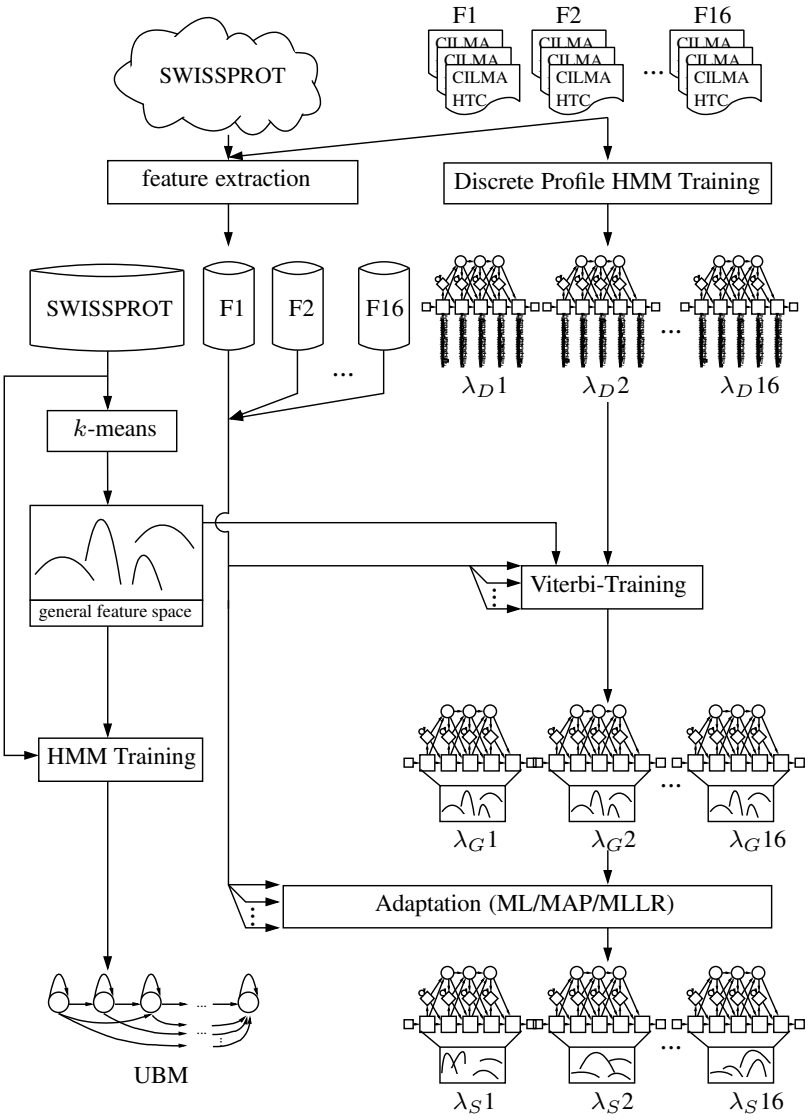


Figure 4: Overview of the model estimation procedure

5 Evaluation

The basic motivation for the developments described in this article is the improvement of remote homology detection methods for e.g. drug discovery tasks. In order to evaluate our new approaches of applying feature based semi-continuous Profile HMMs to this challenging task, we performed several experiments based on public data.

We concentrated on target identification tasks typical for early stages of the drug discovery process. Therefore, homologue sequences for single protein families are searched by database screening. Here, all query sequences are aligned to

Discrete Profile HMMs	SCFB Profile HMMs
• Three-state Profile architecture • Probabilistic state transitions (a_{ij})	
• Discrete state-specific emission probabilities $b_j(o)$ based on 20 standard amino acids – discrete distributions estimated on family specific training data and optionally regularized using background distributions	• State-specific emission probability vectors $b_j(\mathbf{x})$ based on 1,024 Gaussians which are weighted state specifically (c_{jk}) – empiric probability distribution estimated using family specific training data • Robust estimation of mixture components using general protein data and modified k-means • Family related specialization of mixture components using adaptation techniques

Table 1: Properties of probabilistic protein family models (SCFB: Semi-Continuous Feature Based). Only the emission parameters differ and for the larger number of parameters within the SCFB models robust estimation techniques are applied.

the appropriate model and depending on the scores generated the classification regarding target hit or miss is performed. Usually, the performance of detection techniques is measured as a function of the number of false negative predictions vs. the number of false positives which is summarized in ROC-curves [Baldi et al., 2000].

We compared our new approach to standard discrete Profile HMMs estimated using the SAM package v3.3.1 [Hughey and Krogh, 1996]. These models were created and evaluated using default parameters which e.g. implies Dirichlet model regularization, model training from un-aligned sequences, and Smith-Waterman like evaluation. According to the manual of SAM this configuration is reasonable for remote homology analysis using discrete Profile HMMs. Based on these models we created feature based semi-continuous HMMs as described in section 4 by means of our own general HMM framework ESMERALDA [Fink, 1999]. The topology of Profile HMMs (cf. figure 1) was kept fixed. As already noted, in this article the performance of the basic procedure is evaluated. Iterative model estimation approaches can benefit from our new approach, too.

5.1 Datasets

In order to evaluate the detection performance of the approaches presented in this paper, we applied the new Profile HMMs to the task of remote homology detection for superfamilies. Therefore, we used the SUPERFAMILY (cf. [Gough et al., 2001]) based hierarchy of the SCOP database [Murzin et al., 1995]. Sequences belonging to a distinct superfamily must not have similarity values above 95%. In fact, the data for every protein family covers almost the whole range of possible similarities. Thus, the performance for remote homology detection can actually be evaluated.

Generally, for the complex Profile HMM architecture (cf. figure 1) a certain amount of training material is required. Additionally, samples not used for training need to be available for performance assessment. Therefore, 16 superfamilies fulfilling these constraints were selected. Every superfamily contains at least 66 sequences and two thirds of the appropriate material was used for estimating the Profile HMMs. For the assessment of the models' detection performance approximately 34 sequences were used on average for every superfamily. Details regarding the sample sets can be found in appendix B. For the experiments, homologue sequences were searched for every superfamily considered analyzing the complete SCOP database (version 1.63) consisting of approximately 8,000 sequences.

5.2 Results

The remote homology detection experiments were performed individually for every superfamily. We compared the detection performance of discrete Profile HMMs with our feature based semi-continuous models. Additionally to the comparison with the baseline results of SAM, we evaluated the effectiveness of the three adaptation approaches described in section 4.1. The overall performance for superfamily based remote homology detection could be improved significantly which will be illustrated in the following.

The results of the detection experiments are illustrated by ROC-curves given in figures 5 and 6, respectively. The numbers of false negatives (x-axes) are compared to the corresponding numbers of false positives (y-axes). In addition to the complete ROC curves, individual "working areas" are highlighted. These areas shaded gray contain those parts of the plots which are most important for molecular biology research because the number of false positive predictions is reasonably limited. For the overall rating of our approaches, in figure 5 the results of all experiments are summarized in a single diagram. In order to demonstrate the effectiveness of the new techniques in more detail, in the remaining figures 6(a) and 6(b) the detection results for two exemplary superfamilies, namely *"Winged helix" DNA-binding domains (SCOP: a.4.5)*, and *Nucleic Acid Binding proteins (SCOP: b.40.4)*, are presented individually. Note that the actual restriction of the detailed presentation to two superfamilies was for clarity reasons. The results

are typical for the whole corpus and complete evaluation data can be found on our website [Plötz, 2004].

When inspecting the ROC-curves it becomes clear that feature based semi-continuous Profile HMMs significantly outperform their discrete counterparts for the task of remote homology detection. From the overall results shown in figure 5 it can be seen that the number of false negative predictions can generally be decreased while reducing the number of false positives significantly for all models. The ROC-curves corresponding to all variants of feature based semi-continuous Profile HMMs lie significantly below the reference curve of discrete models for the whole diagram (with an exception outside the working area for ML based adaptation which is due to insufficient training data). The outcome of the comparison of the performance of the particular adaptation techniques described in section 4.1 is that the MLLR approach is best.

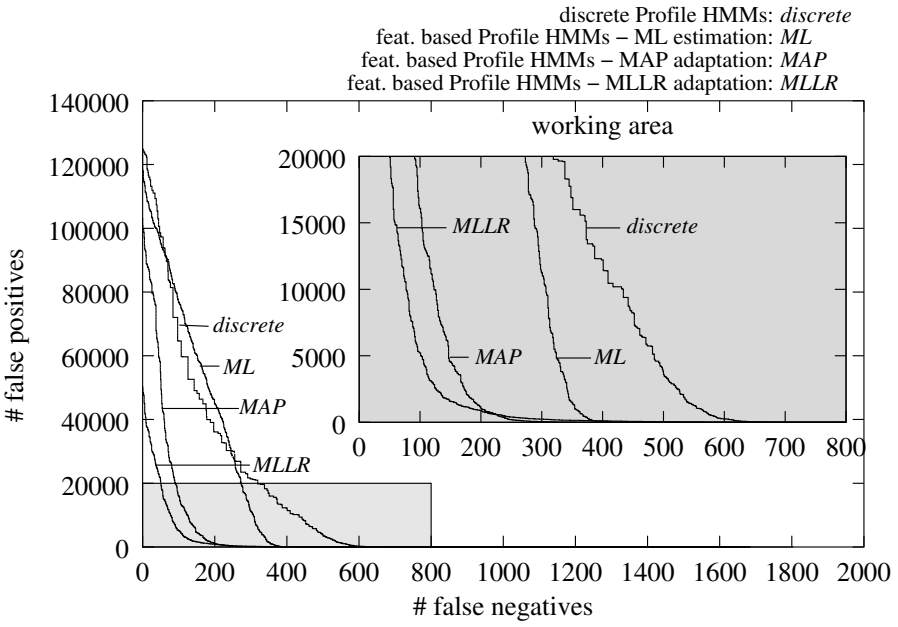
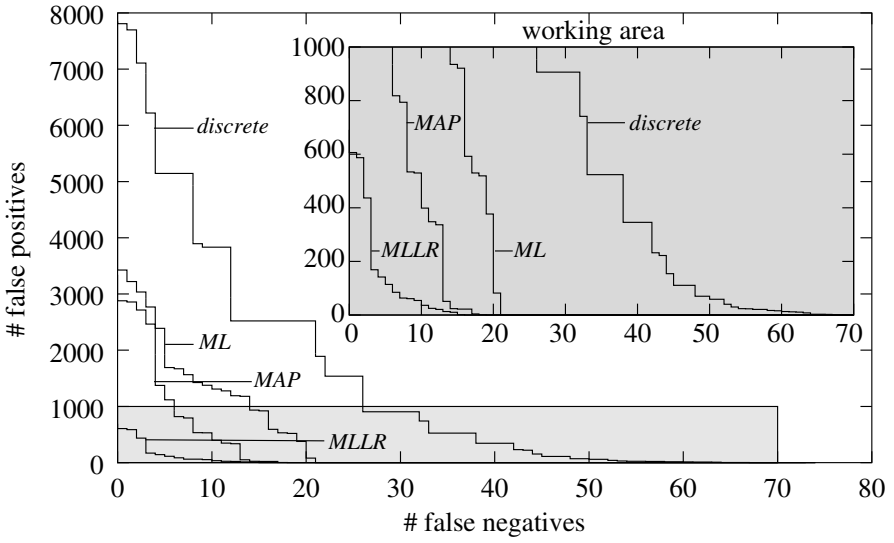


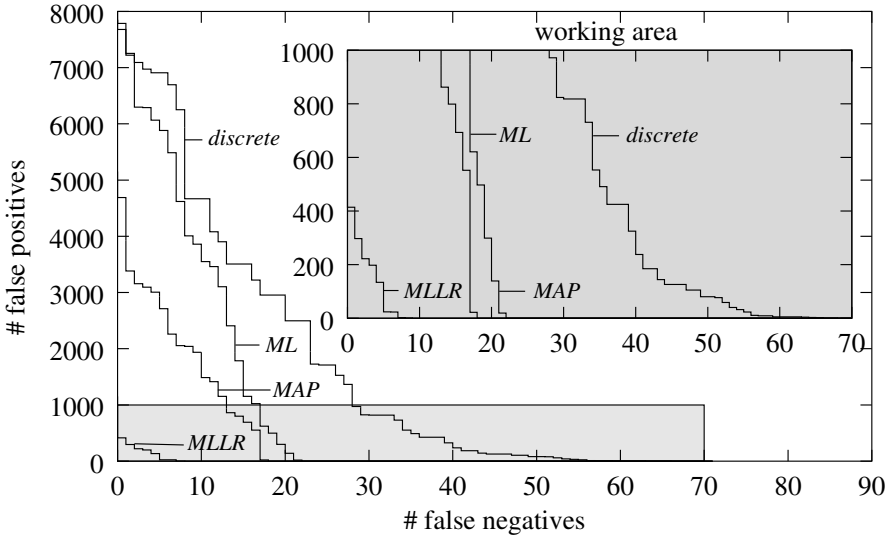
Figure 5: Summary of detection results for all superfamilies analyzed

The effectiveness of the competitive evaluation of both UBM and target models (cf. section 4.2) can be assessed by the maximum number of false positive predictions. Due to our explicit rejection model estimated on general protein data this number is dramatically reduced by almost 66 percent for all superfamilies and more than 80 percent for the exemplary superfamilies whose corresponding ROC-curves are presented individually.

To summarize, the results of the experimental evaluation presented in this section demonstrate the superior performance of feature based semi-continuous



(a) Individual ROC-curves for SCOP superfamily a.4.5



(b) Individual ROC-curves for SCOP superfamily b.40.4

Figure 6: ROC-curves representing the results of superfamily based remote homology detection utilizing SCOP

Profile HMMs compared to their discrete counterparts for the task of remote homology detection. By means of an explicit rejection model capturing proteins not explicitly belonging to the appropriate target family, the number of false positives could be reduced significantly. The number of positive predictions *generally* corresponds to the number of e.g. drug candidates. Since this data needs to be analyzed in further e.g. wet-lab examinations which are usually

very time-consuming and expensive, the reduction of false positives is of major importance.

6 Conclusion

In order to improve the performance of remote homology detection, in this paper feature based semi-continuous Profile HMMs were presented. We developed a multi-channel signal representation of fixed length sequence frames representing the biochemical properties of residues in their local neighborhood. The emissions of semi-continuous Profile HMMs are based on a mixture density representation of the feature space. In order to focus this representation on the properties of a particular protein family, we applied mixture density adaptation techniques with MLLR providing the best results. For reducing the number of false positive predictions, explicit rejection models were introduced which are evaluated in a competitive manner parallel to the particular target models.

The experimental evaluation of the new approach was performed by superfamily based screening of the SCOP database. It was shown that feature based semi-continuous Profile HMMs significantly outperform their discrete counterparts for remote homology detection for all superfamilies considered. The number of false predictions was substantially reduced which is an important prerequisite for e.g. target identification applications in the drug discovery area.

Compared to state-of-the-art discrete models our new approach of semi-continuous Profile HMMs represents a radical change of HMM based protein family modeling. The richer sequence representation using features which capture biochemical properties of residues in their local context is effective for emission modeling of HMMs. The estimation of background distributions can be performed completely data-driven on general protein data. The feature space representation can be specialized by adaptation techniques enabling the estimation of optimized models for particular target families using small amounts of specific training data.

7 Colophon

Thomas Plötz received the diploma in technical computer science from the University of Cooperative Education Mosbach, Germany, in 1998.
He received the diploma and a PhD degree (Dr.-Ing.) in computer science from the University of Bielefeld, Germany, in 2001 and 2005, respectively.

He joined the research group for Applied Computer Science (Angewandte Informatik) at the University of Bielefeld, Germany, in 2001. There he is work-

ing within the GRAS^{2}P project aiming at the detection of new members of therapeutically relevant protein families using statistical models.

He is interested in general aspects of pattern recognition with special focus on statistical techniques applied to various domains like bioinformatics, speech-processing, or automatic recognition of handwritten script.

Dr. Plötz is a member of the International Society for Computational Biology (ISCB).

Gernot A. Fink received the diploma in computer science from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1991 and the Ph.D. degree (Dr.-Ing.) also in computer science from Bielefeld University, Germany, in 1995. In 2002 he received the venia legendi (Habilitation) in applied computer science from the Faculty of Technology of Bielefeld University.

In 1991 he joined the Applied Computer Science Group (Angewandte Informatik) at the Faculty of Technology of Bielefeld University. Since 1997 he is assistant professor in the Applied Computer Science Group.

His fields of research are speech and handwriting recognition, spoken language understanding, man machine interaction, industrial image processing, and statistical methods for the analysis of genomic data. He has published various papers in these fields, and is author of a book on the integration of speech recognition and understanding and another on Hidden-Markov models for pattern recognition.

Dr. Fink is Member of the Institute of Electrical and Electronics Engineers (IEEE).

A Amino Acid Indices used

Table 2 provides information about the 35 biochemical amino-acid properties, selected for the signal based protein sequence encoding. These indices were selected from the compilation provided by [Kawashima and Kanehisa, 2000].

Channel Index	Description	Accession Key
0	Average flexibility indices	BHAR880101
1	Residue volume	BIGC670101
2	Transfer free energy to surface	BULH740101
3	Steric parameter	CHAM810101
4	Polarizability parameter	CHAM820101
5	A parameter of charge transfer capability	CHAM830107
6	A parameter of charge transfer donor capability	CHAM830108
7	Normalized average hydrophobicity scales	CIDH920105
8	Size	DAWD720101
9	Relative mutability	DAYM780201
10	Solvation free energy	EISD860101
11	Molecular weight	FASG760101
12	Melting point	FASG760102
13	pK-N	FASG760104
14	pK-C	FASG760105
15	Graph shape index	FAUJ880101
16	Normalized van der Waals volume	FAUJ880103
17	Positive charge	FAUJ880111
18	Negative charge	FAUJ880112
19	pK-a (RCOOH)	FAUJ880113
20	Hydrophilicity value	HOPT810101
21	Average accessible surface area	JANJ780101
22	Average number of surrounding residues	PONP800108
23	Mean polarity	RADA880108
24	Side chain hydropathy, corrected for solvation	ROSM880102
25	Bitterness	VENT840101
26	Bulkiness	ZIMJ680102
27	Isoelectric point	ZIMJ680104
28	Composition of amino-acids in extracellular proteins	CEDJ970101
29	Composition of amino-acids in anchored proteins	CEDJ970102
30	Composition of amino-acids in membrane proteins	CEDJ970103
31	Composition of amino-acids in intracellular proteins	CEDJ970104
32	Composition of amino-acids in nuclear proteins	CEDJ970105
33	Amphiphilicity index	MITS020101
34	Electron-ion interaction potential values	COSI940101

Table 2: Biochemical properties selected for sequence representation.

At the website of the authors most indices of the database are clustered with respect to six coarse categories:

A. Alpha and turn propensities,

B. Beta propensity,

C. Composition,

H. Hydrophobicity,

P. Physiochemical properties, and

O. Other properties.

In figure 7 an overview of the indices clustering is given. The indices used which were assigned by the authors to any of the categories are highlighted.

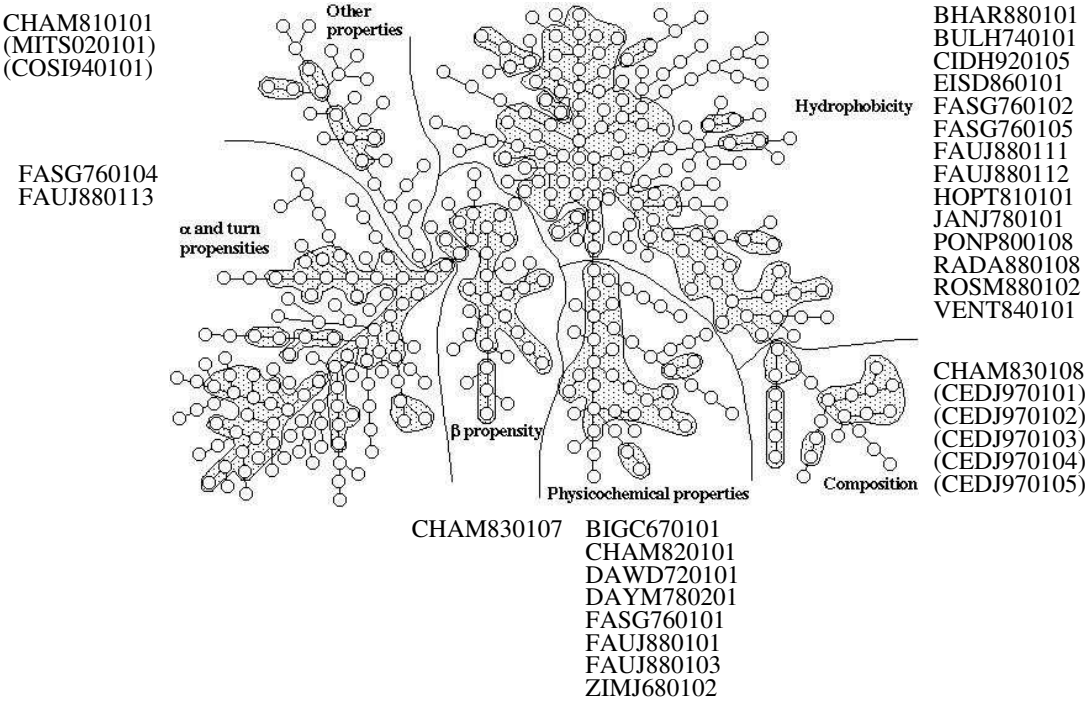


Figure 7: Clustering of the amino acid index database in [Kawashima and Kanehisa, 2000] as provided by the authors. The amino acid indices used for the approaches described in this article are marked explicitly. Note that for unknown reasons the cluster map does not include all indices used. These indices are manually assigned to the existing cluster map and written in parentheses.

B Overview of the sample set's characteristics

In the following details concerning the sample set used for the experimental evaluation of the new feature based semi-continuous Profile HMMs are given. In order to obtain suitable training and test sets the SCOP database was analyzed. Here, protein domains are classified regarding their affiliation to certain families, superfamilies, classes and folds. We used the SUPERFAMILY classification regarding sequence identities and limited it to 95%, i.e. all sequences contained in the sets must not contain similarity values of more thean 95%. In figure 8 the actual distribution of sequence similarities is illustrated. It can be seen, that similarities are almost uniformly distributed.

At the level of superfamilies, domains were selected which contain at least 66 members. This number is motivated by the fact that for both training and test a non-trivial number of samples is required for meaningful assessment. Limiting the minimum number of samples to 66, 16 superfamilies remain which can be used for the evaluation. For all superfamilies analyzed, the sequence sets are randomly divided into disjoint training and test sets (2/3 and 1/3, respectively). Table 3 provides detailed information about the sample set used.

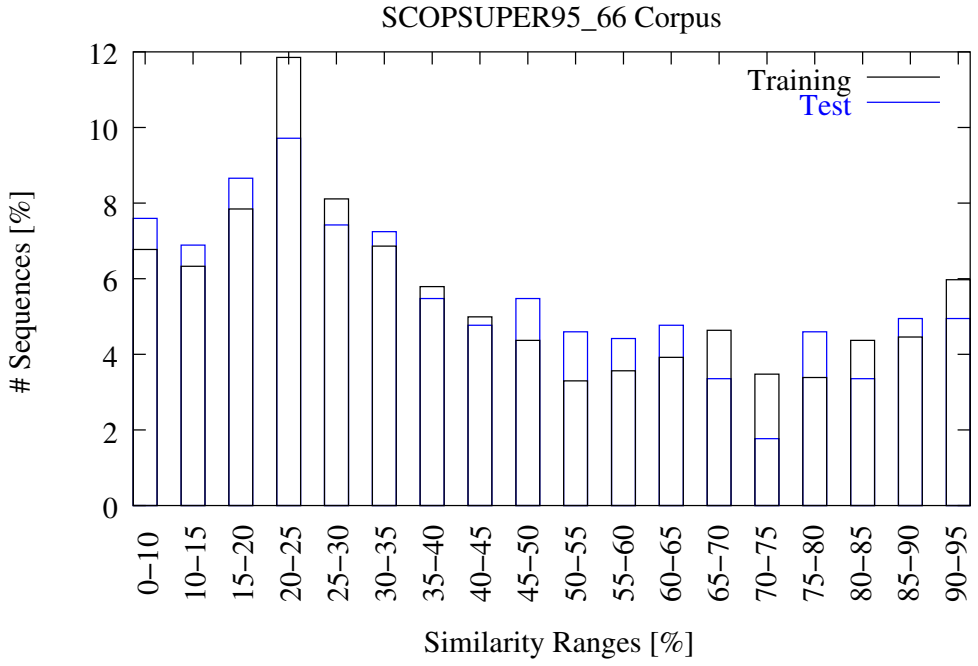


Figure 8: Histogram of similarity ranges for the SCOPSUPER95_66 corpus averaged over all 16 superfamilies involved (black: Training / blue: Test) illustrating the almost uniform distribution of similarities all over the whole range with one exception at 20-25%.

SCOP Id	SCOP superfamily name	# samples		length (mean/std.-derivation)	
		training	test	training	test
a.1.1	Globin-like	60	30	150.3 (13.6)	151.6 (11.1)
a.3.1	Cytochrome c	44	22	102.6 (24.1)	118.4 (32.6)
a.39.1	EF-hand	49	25	138.1 (48.0)	122.0 (39.3)
a.4.5	”Winged helix” DNA-binding domain	49	25	93.8 (26.6)	92.9 (23.1)
b.1.1	Immunoglobulin	207	104	108.9 (15.3)	106.7 (12.3)
b.10.1	Viral coat and capsid proteins	64	32	278.0 (92.9)	262.1 (85.2)
b.29.1	Concanavalin A-like lectins/glucanases	52	27	221.2 (51.2)	220.8 (72.9)
b.40.4	Nucleic acid-binding proteins	47	24	113.1 (36.6)	111.5 (47.2)
b.47.1	Trypsin-like serine proteases	55	28	231.4 (29.5)	226.0 (30.1)
b.6.1	Cupredoxins	50	26	143.9 (34.6)	139.0 (31.5)
c.1.8	(Trans)glycosidases	62	31	376.5 (76.4)	397.8 (84.0)
c.2.1	NAD(P)-binding Rossmann-fold domains	102	51	204.3 (58.9)	211.5 (75.1)
c.3.1	FAD/NAD(P)-binding domain	45	23	226.1 (93.3)	223.3 (86.3)
c.37.1	P-loop containing nucleotide triphosphate hydrolases	127	64	259.3 (120.4)	253.4 (85.6)
c.47.1	Thioredoxin-like	56	28	111.6 (38.2)	105.6 (35.3)
c.69.1	Alpha/Beta-Hydrolases	51	26	350.1 (103.7)	323.7 (25.0)
	Total:	1,120	566		

Table 3: Overview of the SCOPSUPER95_66 corpus used for the comparison of the effectiveness of both state-of-the-art discrete Profile HMM and the new feature based semi-continuous Profile HMMs. For every superfamily the numerical SCOP Id as well as their real name as defined in the database is given.

References

[Baldi et al., 2000] Baldi, P. et al. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16:412–424.

[Barrett et al., 1997] Barrett, C., Hughey, R., and Karplus, K. (1997). Scoring Hidden Markov Models. *Computer Applications in the Bioscience*, 13(2):191–199.

[Boeckmann et al., 2003] Boeckmann, B. et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370.

[Brown et al., 1993] Brown, M. P., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., and Haussler, D. (1993). Using Dirichlet mixture priors to derive Hidden Markov Models for protein families. In *Proc. Int. Conf. Intelligent Systems for Molecular Biology*, pages 47–55.

[Dempster et al., 1977] Dempster, A. et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

[Durbin et al., 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press.

[Eddy, 2001] Eddy, S. R. (2001). HMMER: Profile Hidden Markov Models for biological sequence analysis. *http://hmmer.wustl.edu/*.

[Fink, 1999] Fink, G. A. (1999). Developing HMM-based recognizers with ESMERALDA. In *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence, pages 229–234. Springer.

[Fischer and Stahl, 1999] Fischer, A. and Stahl, V. (1999). Database and online adaptation for improved speech recognition in car environments. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*.

[Gough et al., 2001] Gough, J. et al. (2001). Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *J. Molecular Biology*, 313:903–919.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

[Huang and Jack, 1989] Huang, X. D. and Jack, M. A. (1989). Semi-continuous hidden markov models for speech signals. *Computer Speech & Language*, 3:239–251.

[Hughey and Krogh, 1996] Hughey, R. and Krogh, A. (1996). Hidden Markov Models for sequence analysis: Extension and analysis of the basic method. *Computer Applications in the Bioscience*, 12(2):95–108.

[Karplus, 1995] Karplus, K. (1995). Evaluating Regularizers for Estimating Distributions of Amino Acids. In *Proc. Int. Conf. Intelligent Systems for Molecular Biology*, pages 188–196.

[Karplus et al., 1998] Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov Models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856.

[Kawashima and Kanehisa, 2000] Kawashima, S. and Kanehisa, M. (2000). AAindex: Amino acid index database. *Nucleic Acids Research*, 28(1):374.

[Krogh et al., 1994] Krogh, A. et al. (1994). Hidden Markov Models in computational biology: Applications to protein modeling. *J. Molecular Biology*, 235:1501–1531.

[Leggetter and Woodland, 1995] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, pages 171–185.

[Murzin et al., 1995] Murzin, A. G. et al. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Molecular Biology*, 247:536–540.

[Percival and Walden, 2000] Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistical Mathematics. Cambridge University Press.

[Plötz, 2004] Plötz, T. (2004). The GRAS^{2}P project. `www.techfak.uni-bielefeld.de/ags/ai/projects/GRASSP/`.

[Plötz and Fink, 2004] Plötz, T. and Fink, G. A. (2004). Feature extraction for improved Profile HMM based biological sequence analysis. In *Proc. Int. Conf. on Pattern Recognition*.

[Poularikas, 2000] Poularikas, A. D., editor (2000). *The Transforms and Applications Handbook*. CRC Press LLC, 2nd edition.

[Reynolds, 1997] Reynolds, D. A. (1997). Comparison of background normalization methods for text-independent speaker verification. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 963–966.