Pattern Recognition Methods for Advanced Stochastic Protein Sequence Analysis using HMMs

Thomas Plötz * and Gernot A. Fink

Faculty of Technology, Bielefeld University

P.O. Box 100 131, 33 501 Bielefeld, Germany

Tel/Fax: +49 521-106 2953 / +49 521-106 2992

{*tploetz,gernot*}*@techfak.uni-bielefeld.de*

Abstract

Currently, Profile Hidden Markov Models (Profile HMMs) are the methodology of choice for probabilistic protein family modeling. Unfortunately, despite substantial progress the general problem of remote homology analysis is still far from being solved. In this article we propose new approaches for robust protein family modeling by consequently exploiting general pattern recognition techniques. A new feature based representation of amino acid sequences serves as the basis for semi-continuous protein family HMMs. Due to this paradigm shift in processing biological sequences the complexity of family models can be reduced substantially resulting in less parameters which need to be trained. This is especially favorable when only little training data is available as in most current tasks of molecular biology research. In various experiments we prove the superior performance of advanced stochastic protein family modeling for remote homology analysis which is especially relevant for e.g. drug discovery applications.

Key words: Protein Sequence Analysis, Probabilistic Protein Family Modeling, HMM

* To whom correspondence should be addressed.

1 Introduction

One fundamental goal of research in molecular biology is the determination of proteins' functions. This is especially relevant for almost all variants of life sciences research, e.g. within the so-called drug discovery pipeline. Principally, the genomic data is exploited for the development of new therapies against severe illnesses like e.g. cancer.

The biological function of proteins is determined by their three-dimensional structure which depends on biochemical properties of the particular residues. One foundation of molecular biology states that similar functions of proteins are caused by similar structures. Furthermore, the three-dimensional structure of proteins is mainly determined by biochemical properties of the underlying primary structure, i.e. the linear sequence of amino acids. Thus, once the function of a particular protein could have been solved, related proteins can be obtained by sequence comparison which stands for one principle of molecular biology research. Since protein sequences are usually considered as strings of an alphabet consisting of the 20 standard amino acids, *computational* sequence comparison methods are predestinated for protein analysis at the beginning of the drug discovery process, namely target identification and target verification.

Currently, one very promising approach for protein family related analysis of amino acid sequences is the application of so-called Profile Hidden Markov Models (Profile HMMs) as probabilistic target family models. Given a training set of protein data, discrete HMMs are estimated. These models are then evaluated for unknown query sequences which are aligned to the explicit protein family models. Such explicit target family models are favorable for sequence analysis since family specific data is incorporated into the analysis.

Despite the substantial progress for remote homology analysis when applying Profile HMMs as described above, the general problem is, unfortunately, still far from being

solved. Although smart model regularization techniques have been developed, the robustness of protein family models is rather insufficient especially when only small training sets are available. For further breakthroughs in molecular biology research for both fundamental research and for commercial issues with respect to the pharmaceutical industry, improved methods for the basic step of protein sequence analysis are required. Since conventional methods including the application of even the most sophisticated discrete Profile HMMs apparently have reached their limits, new concepts are required.

In order to generally improve the computational analysis of protein sequences, in this article we present new concepts for HMM-based protein family modeling approaches. Due to the interpretation of protein sequence analysis as a pattern recognition problem, the general application of HMMs for bioinformatics purposes has become possible. Consequently, in our new approaches presented here, we generalize this idea by treating amino acid sequences as *signals* in their original meaning, i.e. representing some kind of biochemical measures depending on the particular position within sequences.

Based on this new protein sequence representation powerful features are extracted serving as the basis for all further processing. By means of these features, advanced protein family models become possible. We developed semi-continuous feature based Profile HMMs as direct replacements of the abovementioned discrete Profile HMMs. Due to the explicit consideration of the particular residues' biochemical properties, covered by the new feature representation, and robust model estimation and evaluation techniques applying general pattern recognition methods, the new semi-continuous Profile HMMs significantly outperform their discrete counterparts. Using the new feature representation protein family HMMs with reduced model complexity become possible. We developed the so-called *Bounded Left-Right* HMM model architecture containing a substantially smaller number of parameters that need to be trained for robust modeling which is especially relevant when only little training data is available.

Based on the general HMM framework ESMERALDA [1], the new techniques were integrated into a ready-to-use protein sequence analysis system which will be applied in the context of industrial molecular biology research. The focus of this article is on the presentation of the complete system as an innovative integrated framework for protein sequence analysis. We applied it for the experimental evaluation of the proposed concepts. Therefore, different corpora were defined exploiting the SUPERFAMILY hierarchy [2] of the SCOP database [3] at the level of superfamilies. It turned out that the new, more general pattern recognition oriented approaches for feature based protein sequence analysis using HMMs substantially outperform state-of-the-art techniques. The superior results are very promising for related tasks within e.g. the field of drug discovery indicating that the proposed techniques are favorable for generally improved remote homology analysis.

In the subsequent section we briefly review the state-of-the-art in probabilistic protein sequence analysis, namely Profile HMMs. Following this, section 3 in detail discusses the proposed advanced protein family modeling approaches using HMMs. As already mentioned before, an evaluation using considerable amounts of practical experiments was performed in order to judge the relevance of the new techniques. In section 4 the corresponding results are presented and discussed. Finally, we will give a summary.

2 State-of-the-art in Probabilistic Sequence Analysis using HMMs

Proteins can be interpreted as words over an alphabet with fixed lexicon, namely the set of amino acids. Consequently, most sequence analysis techniques are based on some kind of string processing algorithms. Usually, dynamic programming techniques are applied creating sequence alignments by matching or substituting, inserting, or deleting residues including the calculation of appropriate alignment scores which can be used for classification. Very popular implementations of traditional alignment techniques are e.g. BLAST

[4], or FASTA [5]. Basically the foundations of such so-called pairwise sequence comparison techniques remain the same, namely the direct sequence-to-sequence alignment.

Especially for the analysis of remote homologies the abovementioned pairwise alignment techniques are usually outperformed by explicit stochastic models of the protein families the sequences are belonging to. Here, remote homologies designate protein sequences which although sharing a common biological function contain only weak sequential similarities. The statistical properties of multiple protein family members are covered by a probabilistic model and query sequences are aligned to such models. Due to the explicit consideration of protein *family* information, such techniques are usually superior compared to the pairwise techniques mentioned above.

Currently, Profile HMMs are the most promising kind of stochastic protein family models. Originally introduced by Haussler and colleagues [6], Krogh et al. [7], and Baldi and coworkers [8], they are interpreted as probabilistic representation of a multiple alignment of sequences belonging to the same family. Usually, Profile HMMs are estimated for a particular protein family using training sequences for which the family affiliation is known in advance. Database search is performed by aligning query sequences to the model and calculating the appropriate score which is the base for family affiliation classification.

In this section we will briefly review the state-of-the-art in protein sequence analysis techniques using Profile HMMs In 2.1 the formal definition of (general) HMMs is summarized whereas the focus of sub-section 2.2 is on specific HMMs for bioinformatics purposes.

2.1 Definition of General Hidden Markov Models

Currently, Hidden Markov Models represent the predominating concept for the classification of general signals evolving in time covering both length and content variance. They

are used in various application fields, e.g. the automatic recognition of speech or handwritten script. The foundations of Hidden Markov Models are shortly but clearly presented in [9] and in more detail described in e.g. [10, chap.8].

Formally an HMM describes a two stage stochastic process. According to the so-called Markov Property in the first stage one state s_t of a finite set of states $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ is chosen depending only on a limited number of predecessors. For the most relevant first order models, the memory is bounded to the immediate predecessor s_{t-1}. Transitions between states are modelled with the probability $P(s_t|s_{t-1})$ and subsumed in a matrix

$$\mathbf{A} = [a_{ij}] = [P(s_t = S_j|s_{t-1} = S_i)], \quad \sum_j a_{ij} = 1, \quad 1 \leq i, j \leq N.$$

The initialization of the stochastic process is defined by the vector of initial probabilities $\boldsymbol{\pi} := [\pi_i] = [P(s_1 = S_i)]$. Depending on s_t in the second stage an observation symbol o_t is produced with probability $P(o_t|s_t)$. For bioinformatics applications these symbols are usually modeled as discrete emissions. Analogous to the transition matrix the emission probabilities can be summarized in a matrix

$$\mathbf{B} = [b_i(o_l)] = [P(o_t = O_l|s_t = S_i)], \quad 1 \leq i \leq N, 1 \leq l \leq M.$$

A linear sequence of emission symbols $\mathbf{o}$ represents the data to be handled. Formally a Hidden Markov Model is defined as a tuple $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$.

Pattern classification with HMMs requires the robust estimation of their parameters, namely the transition and emission probabilities as well as the model structure. Using representative samples the parameter values are usually estimated by means of variants of the well-known EM algorithm [11], most notably Baum-Welch, and Viterbi training.

Once HMMs are established serving as statistical models for distinct pattern families they have to be evaluated when classifying sequences of observations. Using the Forward algorithm the general probability $P(\mathbf{o}\,|\lambda_k)$ of an HMM λ_k producing the given sequence $\mathbf{o}$

can be calculated. More common is the deployment of the Viterbi algorithm which additionally decodes the most probable state-sequence chosen producing the observations:

$$P(\mathbf{s}\,|\mathbf{o},\lambda) = \frac{P(\mathbf{o},\mathbf{s}\,|\lambda)}{P(\mathbf{o}\,|\lambda)} \Rightarrow P(\mathbf{o},\mathbf{s}^*|\lambda) = \max_{\mathbf{s}\in S^T} P(\mathbf{o},\mathbf{s}\,|\lambda).$$

By means of the techniques discussed above, Hidden Markov Models can be estimated and evaluated very efficiently for various applications in different domains.

2.2 Profile Hidden Markov Models for Remote Homology Analysis

By abstraction from time HMMs have also been proven to be very effective for alternative domains. For bioinformatics purposes time dependency is substituted with residues' locations. In the following the foundations of current protein family HMMs are summarized.

Pairwise sequence comparison is often not suitable for the analysis of remote homologue protein sequences. Contrary to this, when exploiting multiple sequence alignments (MSAs), more information about the target *family* of interest is incorporated into sequence analysis. During alignment, for every column of a particular MSA probability distributions of amino acids are considered instead of single residues as for the pairwise case. A good overview of general sequence alignment algorithms is given in e.g. [12]. Usually the abovementioned position specific amino acid probability distributions are subsumed in so-called *Profiles* [13]. The generalization of Profile analysis refers to Profile Hidden Markov Models whose typical architecture is shown in figure 1.

The conserved parts of a multiple alignment of the sequences belonging to a target family 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. There are some extensions to the basic architecture with increased flexibility, e.g. in HMMERs Plan7 [14]. An excellent treatment of Profile HMMs can be found in [15].

Currently, the emissions of Profile HMMs are modeled by 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.

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 the length of the sequences, usually, they are considered with respect to 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 [16].

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* estimation. Thus, several model regularization techniques were proposed which try to tackle this so-called sparse data problem (cf. e.g. [17]). The currently most promising technique for obtaining statistically more “stable” amino acid distributions is based on the incorporation of prior knowledge using carefully designed Dirichlet distributions.

3 Advanced Stochastic Protein Sequence Analysis

Common practical experience of molecular biologists, especially in the research field of drug discovery, leads to the conclusion that even the most sophisticated probabilistic pro-

tein family modeling techniques (cf. the previous section) have reached their limits. Therefore, in this article we present new, and more general approaches representing some kind of paradigm shift. Currently, the development of the most promising approaches is very goal oriented, which means that several concepts are almost exclusively influenced by the actual biological task (cf. the Profile HMM architecture directly reflecting the matching/substitution, deletion, and insertion of amino acid residues).

Protein sequence analysis can generally be understood as a pattern recognition problem where more or less modified occurrences of patterns need to be assigned to the correct classes. Consequently, our approaches address the advanced incorporation and adoption of general pattern recognition techniques into the bioinformatics domain. Due to a more abstract view at the protein sequence analysis task , enhanced probabilistic models for protein families become possible which are described in this section. We will first discuss the central idea of feature based protein sequence analysis using semi-continuous Profile HMMs including robust model estimation and application schemes (sections 3.1 - 3.3, cf. also [18]). When using the proposed rich sequence representation the complex Profile HMM architecture is no longer needed. Thus, model topologies with reduced complexity can be used which is favorable for the robust estimation of protein family HMMs when only little training data is available. In section 3.4 the generalization to feature based protein family HMMs with a Bounded Left-Right topology is presented. An overview of the complete system for advanced stochastic protein sequence analysis is given in 3.5.

3.1 Feature Extraction from Protein Sequences

State-of-the-art HMM-based protein sequence analysis approaches are applied directly to symbolic primary structure sequence data. This data which is the more or less direct result of sequencing (after gene prediction, transcription, and translation) seems to be the

"natural" choice for most appropriate practical applications. The reason for this is that any further higher-level information about e.g. the three-dimensional structure of proteins is usually not available. However, according to the foundations of molecular biology proteins' functions are determined by their spatial conformation which directly depends on the biochemical properties of the underlying residues. The sequence of amino acid symbols represents some kind of summary of these properties only. In order to improve the capabilities of protein family HMMs we developed a richer protein sequence representation explicitly covering the abovementioned biochemical properties.

Once alternative representations for biological sequences are available the huge amount of powerful signal processing techniques already existing can be applied to the bioinformatics domain. Applications using protein family HMMs can thus benefit from uncovering possibly hidden characteristics of protein data. Explicitly exploiting such information can especially increase the performance of remote homology analysis approaches.

Currently, in the field of sequence analysis only little research is devoted to signal based representations. The most common approaches either use a mapping to some vector space [19,20], or to biochemical properties which is the basis for spectral analysis [21]. The most promising signal representations rely on biochemical sequence properties. Kawashima et al. compiled a huge amount of so-called amino acid indices [22]. Every index defines a mapping of amino acids to numerical values depending on biochemical properties.

In our new feature based representation of protein sequence data we are aiming at the explicit consideration of residues' relevant biochemical properties. Especially the local neighborhood of amino acids determines the spatial conformation of proteins and, therefore, its biological function. Basically, we follow the idea of mapping residues to numerical values as defined by amino acid indices. However, limiting the representation to an arbitrary but single index implies neglecting putative higher level relationships of residu-

als. 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 carefully selected 35 indices out of the pool of indices available for a multi-channel signal representation (cf. also [23]). The indices actually selected are listed in table 1 and [24, p. 203], respectively. The combination of multiple biochemical properties provides a rich characterization of protein sequences.

In order to respect *local* signal characteristics for HMM states' emissions, in our feature extraction procedure 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 time signal analysis. The window is moved along the whole sequence using single residue steps resulting in stepwise overlaps of 15 samples for each channel.

Basically, for remote homology analysis the signal analysis should produce features enabling a more abstract view on the actual sequences representing the coarse shape and neglecting detail. In order to extract 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. In our approach we use the Discrete Wavelet Transform (DWT) for the analysis of the coarse temporal signal structure (cf. e.g. [25]). For every channel of the signal representation of a particular protein sequence the Wavelet coefficients are determined using standard Daubechies filters of length 4. For the analysis of remote homologies we skip the upper five coefficients containing detail information, resulting in 11-dimensional feature vectors per channel which are concatenated to 385-dimensional vectors.

After combining the DWT coefficients of all channels into a single feature vector, poten-

tially redundant information needs to be removed. Thus, we finally perform a Principle Component Analysis (PCA) for the feature vectors of every frame whose components are normalized to the interval of $[-1 \ldots 1]$. The PCA-matrix itself was estimated beforehand using large amounts of general protein data (approximately 90K SWISSPROT sequences). Inspecting the eigenvalue spectrum of the data, it becomes clear, that a compact representation in a 99-dimensional subspace is sufficient for more than 95% reconstruction.

Figure 2 schematically summarizes the feature extraction method described in this article.

3.2 Robust Estimation of Feature Based Profile HMMs

Compared to the 20 discrete amino acid symbols, the new feature representation of protein data corresponds to a 99-D feature space. When processing feature vectors, generally discrete HMMs are not suitable for modeling. Instead, continuous modeling is usually the methodology of choice where the feature space is represented by state-specific mixtures. However, pure continuous HMMs seem problematic especially for remote homology analysis tasks since often only little training data is available. The smaller the amount of training data, the smaller the number of Gaussians which can be estimated robustly.

For effective exploitation of training data, in [26] semi-continuous HMMs were proposed where all states share a common set of mixture densities weighted state-specifically. 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 $\mathbf{C}_k$ used for all HMM states but individually weighted by c_{jk}:

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

By analyzing equation 1 it becomes clear that the model estimation can principally be divided into two separate steps. The model independent feature space representation 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. This can be exploited for robust estimation of protein family HMMs using small family specific sample sets.

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. The base for the unsupervised and completely data driven estimation of mixture densities is a huge pool of general protein data, i.e. sequences not explicitly assigned to the target protein family of interest. We used all sequences (approximately 90K) from the SWISSPROT database [27] allowing the estimation of a sufficient feature space representation, namely 1 024 Gaussians.

In our first approach addressing improved protein family modeling we developed semi-continuous Profile HMMs. This means that the discrete emissions of state-of-the-art protein family HMMs are replaced by the abovementioned semi-continuous emissions while keeping the original model topology as illustrated in figure 1. 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 the representation to specific properties of proteins belonging to a particular target family, data driven mixture

adaptation techniques are applied. Using such transformations of the mixture parameters, i.e. mean vectors $\boldsymbol{\mu}_k$ and (not necessarily) covariance matrices $\mathbf{C}_k$, the coverage of general protein properties is optimized towards more family specific characteristics. Note that the model structure, the transition probabilities a_{ij} as well as the state specific mixture weights c_{jk} remain unchanged during adaptation. Only the underlying mixture component densities are changed during adaptation.

We investigated two different adaptation techniques which are described in the following where the number of family specific training samples, i.e. the amount of adaptation data which is usually very small, is denoted by T.

Maximum A-Posteriori (MAP) Adaptation: The simplest case of target family based specialization of the feature space representation is the re-estimation of mixture parameters. Therefore, a maximum likelihood (ML) optimization, i.e. EM up to convergence, is performed applying the small family specific set of adaptation data. Since the parameters of all densities are re-estimated by ML, unfortunately, rather large sample sets are required for robust adaptation. Contrary to the ML approach, MAP adaptation of the component densities is performed with respect to optimization of the posterior probability 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:

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

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

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

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

Here $\xi_t^m(k)$ designates the probability of selecting a particular Gaussian for a given sample, and $\hat{p}_k$ represents the prior probability of the k-th Gaussian. Initial parameter estimates are obtained by applying a (modified) k-means algorithm [28] to SWISSPROT data. The advantage of MAP adaptation is the balanced incorporation of prior information extracted from the larger set of unlabeled 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 mixtures as accumulated during the previous estimation steps which allows robust mixture adaptation even for small training sets.

Maximum Likelihood Linear Regression (MLLR): For the second kind of adaptation, deterministic assignments of feature vectors $\mathbf{x}_t$ to mixtures are assumed. Originally developed for speaker adaptation of automatic speech recognition systems, in [29] the modification of the mixtures' mean vectors only using affine transformations $\mathbf{W}_k$ was proposed. These transformations represent rotations and translations of the feature space estimated on small adaptation sets. They can be defined with respect to augmented $D+1$-dimensional mean vectors $\tilde{\boldsymbol{\mu}}_k = (1, \mu_{k_1}, \ldots, \mu_{k_D})^T$, where D in our case is 99:

$$\hat{\boldsymbol{\mu}}_k = \mathbf{W}_k \tilde{\boldsymbol{\mu}}_k. \tag{5}$$

The transformations are generalized to groups of Gaussians including densities not covered by the adaptation set via linear regression. Fischer and Stahl developed a simplified procedure by using a *single* regression class. This implies a global transformation matrix $\mathbf{W}$ which is defined as follows [30]:

$$\mathbf{W} = \left\{ \sum_{t=1}^{T} \mathbf{x}_t \tilde{\boldsymbol{\mu}}_t^T \right\} \left\{ \sum_{t=1}^{T} \tilde{\boldsymbol{\mu}}_t \tilde{\boldsymbol{\mu}}_t^T \right\}^{-1}. \tag{6}$$

Contrary to MAP adaptation, here instead of statistically re-estimating the mixtures' parameters, the densities themselves are transformed. The transformation itself which is estimated for mixtures actually covered by a small adaptation set is generalized to the complete feature space. Since only the single transformation matrix $\mathbf{W}$ needs to be estimated which requires considerably smaller amounts of target family specific data, MLLR is especially attractive for remote homology analysis.

3.3 Robust Remote Homology Detection

In order to perform remote homology *detection* normally log-odds scores are used for a threshold based decision regarding target hit or miss. For the feature-based protein family HMMs we use a null model based on the prior probabilities of the mixture components estimated during model building.

We enhanced this procedure by furthermore applying a technique principally known from general detection tasks. Additionally, a non-target model which explicitly covers all data *not* belonging to the protein family of interest is competitively evaluated to a target model. According to [31] such a model, which was originally proposed for the task of automatic speaker detection, is called *Universal Background Model (UBM)*. As an enhancement of the general UBM approach, our definition of the background model used for Profile models optionally captures structural information using a standard left-right topology (otherwise we use the original single-state definition of UBMs). This 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.

In figure 3 our approach for estimating semi-continuous Profile HMMs and an explicit UBM for robust remote homology detection is summarized. Based on the new feature representation of general protein data a mixture representation of the general feature space is

estimated using k-means (upper-left). 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).

3.4 Protein Family HMMs with Reduced Complexity

The state-of-the-art in probabilistic protein sequence analysis techniques refers to stochastic representations of multiple sequence alignments. Therefore, Profile HMMs were developed explicitly respecting the dynamic programming "roots" of sequence alignment techniques for model creation. The topology of Profile HMMs is based on three different kinds of states, namely match, insert, and delete which correspond to the standard operations of dynamic programming techniques. Consequently, current protein family HMMs require a rather complex topology (figure 1). However, the basic drawback of such modeling techniques is the enormous amount of model parameters which need to be estimated thus requiring large training sets. Since usually only little training data is available often model regularization techniques need to be applied which is, however, problematic. Even with the most sophisticated model regularization techniques *robust* data-driven model estimation can hardly be realized. Data-driven techniques are especially favorable for remote homology analysis since putatively biased model regularization seems critical for detecting really new protein family members.

Based on the new feature based sequence processing we now focus on the reduction of the models' complexity in order to reduce the number of parameters which need to be trained while keeping the flexibility of the particular protein family models. Compared to

the feature based Profile HMMs which consist of substantially more parameters explicitly trained, especially for small training sets the new models with reduced complexity are favorable. According to the literature, currently there is only little research dedicated to the explicit reduction of the number of model parameters. Grundy and colleagues proposed the MEME system which heuristically combines rather simple motif HMMs to protein family models [32]. In [33] we presented feature based protein family modeling techniques using so-called sub-protein units (SPUs) which are obtained in a data-driven and un-supervised manner. However, for global protein family models state-of-the-art Profile HMMs currently still outperform SPU-based models.

When using the previously described feature representation, emissions of protein family HMMs are now based on a mixture density representation of the new feature space. The resulting continuous emission probability distributions are much broader than the discrete amino acid distributions of current Profile HMMs while keeping the specificity necessary for sequence classification. If features properly match the emission probability distributions of a particular state, the resulting contribution to the overall classification score is rather high which corresponds to the match case of dynamic programming. Contrary to this, if the features do not match the states' probability distribution, the local score will be small which corresponds to an insertion. Thus, the *explicit* discrimination between insert and match states is not needed any longer because it is implicitly performed already on the emission level. Furthermore, explicit deletes are only conceptual and can be replaced by jumps skipping direct neighbors which results in standard left-right topologies where every state is connected to all states adjacent to the right.

However, if arbitrary jumps within a protein family model are allowed, as defined for plain left-right topologies, especially for models covering larger protein families the number of parameters to be trained is still rather high. The number of transition probabilities N_t for a model containing L states is defined as follows:

$$N_t = \sum_{i=1}^{L} i + 1 = \frac{L}{2}(L+1) + 1.$$

For an exemplary protein family model consisting of 100 states, the number of transition probabilities is approximately 5 000.

Even for extremely diverging sequences belonging to a particular protein family it is rather unrealistic to assume *arbitrary* alignment paths through the appropriate protein family model which are allowed when using the plain left-right topology. Thus, a variant of standard left-right models is developed for protein family modeling – so-called *bounded left-right (BLR)* models. State transitions are restricted to the local context of a particular state resulting in substantially less transition parameters to be trained. The number of state transitions depends on the length of the underlying protein family model and it is defined by the ratio of the median of the training sequences' length to their minimum length. Using this heuristic, it is guaranteed that all training sequences can be aligned to the model when initializing it's length to the median of the length of the training data.

For local alignments optionally every state can serve as model entrance and exit. The corresponding transition probabilities are fixed by assuming uniform distributions which is reasonable according to [15, p. 113ff]. The length of BLR models is determined as the median of the lengths of the training data and the semi-continuous BLR models are initialized and trained using standard HMM algorithms. In figure 4 the BLR architecture of protein family models is illustrated.

Compared to the approximately 5 000 transitions for the complete left-right model architecture of the exemplary protein family given above, the number of parameters to be trained for the BLR topology is decreased to approximately 500 when assuming a median length of 100 and a minimum length of 20. For a corresponding three-state Profile HMM architecture the number of *transition* parameters for the given example is approximately

2 700. Additionally, the number of emitting states in BLR models is halved compared to standard three-state Profile HMMs. Note that due to respecting local amino acid contexts already at the level of emissions, usually feature based BLR models are significantly shorter than Profile HMMs.

3.5 System Overview

The focus of this article is on the development of a complete system for advanced stochastic protein sequence analysis which can be used for supporting molecular biology research at a larger scale. Based on the ESMERALDA framework for arbitrary HMM recognizers [1] the new techniques were implemented. In addition to the new approaches presented in this article, the system consists of powerful tools for both robust estimation and efficient evaluation of mixture densities, semi-continuous HMMs, Markov chains etc. Using this framework, high-throughput sequence analysis pipelines can be realized allowing for advanced and efficient analysis of remote homologue protein sequences.

The general application scheme of the described system is graphically summarized for the exemplary use of a BLR target model in figure 5. In the first frame (block 1), the general preprocessing steps necessary for the application of the new models are illustrated. These are the estimation of the general mixture density feature space representation, UBM training, and feature extraction for the sequences contained in the database which will be searched for remote homologues. These time-consuming steps are necessary to be performed only once. The target model estimation procedure is shown in frame 2. Based on the features extracted from a small sample set, a general semi-continuous feature based protein family HMM is estimated which is further specialized using either MAP, or MLLR adaptation. The resulting target model is competitively evaluated to the UBM. The actual remote homology detection process is illustrated in the third block.

4 Experimental Evaluation

The basic motivation for the developments of the approaches presented in this article is the improvement of remote homology analysis as usually performed in large scale for e.g. drug discovery tasks. We emphasized advanced stochastic techniques which are *generally* applicable to the task of protein sequence analysis. In order to judge their capabilities we performed an experimental evaluation thereby avoiding putative biases to particular protein families using public data of high quality. Since the major application of probabilistic protein family models within the drug discovery pipeline refers to target identification we concentrated on *detection* of remote homologue sequences of certain protein families by database screening. Query sequences are aligned to target models and depending on the scores generated the classification regarding hit or miss is performed.

Usually, the capabilities of detection techniques are measured as a function of the number of false negative predictions vs. the number of false positives which is summarized in ROC-curves [34]. Furthermore, especially for industrial use certain working points within the particular ROC-curves are relevant. Therefore, the percentage of allowed e.g. false negative predictions is fixed at 5% and the corresponding percentage of, here, false positive predictions is considered for judgment.

We compared the new techniques (referred to as SCFB – semi-continuous feature based – models) to discrete Profile HMMs estimated using the state-of-the-art SAM package v3.3.1 [35]. These models were created and evaluated using default parameters which e.g. implies Dirichlet model regularization. SCFB protein family HMMs were obtained as described in section 3 using our own general HMM framework ESMERALDA [1]. In this article the performance of the basic procedure is evaluated, i.e. iterative model estimation approaches, which are not addressed here, can also benefit from our new approaches.

4.1 Datasets

Generally, it is rather difficult to assess the power of protein family HMMs by means of unknown data. The actual family affiliation of the data processed needs to be known in advance. We, therefore, used the SUPERFAMILY [2] based hierarchy of the manually annotated, high-quality SCOP database [3] – v1.63 – at the superfamily level where sequences belonging to a distinct family must not have similarity values larger than 95%. The data for every family covers almost uniformly the whole range of possible similarities, i.e. the performance for *remote* homology analysis can actually be evaluated.

First we evaluated the general detection capabilities of feature based Profile HMMs in comparison to their discrete counterparts, i.e. the state-of-the-art. Therefore, "sufficient" amounts of training samples were assumed. Additionally, (annotated) samples not used for training were available for performance assessment. Within the database used 16 SCOP superfamilies fulfill these constraints and were thus selected for evaluation. Every superfamily contains at least 66 sequences and two thirds of the appropriate material was randomly chosen for estimating the Profile HMMs (on average 70 training samples for every superfamily). The detection experiments were then performed based on the approximately 8 000 sequences of the abovementioned database which among others contains the particular sequences. For further argumentation the corpus consisting of the corresponding training and test sets is referred to as *SCOPSUPER95_66*.

In the second set of experiments we evaluated remote homology detection using protein family HMM variants estimated when only smaller training sets were available. Therefore, the amount of sample sequences was successively reduced beginning from 44 samples which results in 44 sub-corpora. Note that for every sub-corpus the number of training sequences is equal for all superfamilies. Due to the enormous number of permutations possible, the (statistically correct) evaluation using leave-N out tests for $N = 1, \ldots, 43$ is

computationally not feasible. However, since we selected the particular sequences which were removed from the original training sets *randomly*, in fact reliable conclusions can be drawn when performing single detection experiments for every sub-corpus. The data sets used for the second kind of experiments are referred to as *SCOPSUPER95_44f*.

Due to space limitations a quantitative description of the corpora used is skipped *here*. In the following the results of the particular experimental evaluations are summarized. Exhaustive descriptions of the corpora and detailed evaluation results can be found on our website (www.techfak.uni-bielefeld.de/ags/ai/projects/GRASSP) or in [24], respectively.

4.2 Results

The results of the general evaluation of the capabilities of feature based protein family HMMs are presented in figure 6. The numbers of false negatives (x-axis) are compared to the corresponding numbers of false positives (y-axis). In addition to the complete ROC-curves, a "working area" is highlighted containing those parts of the curves which are most important for molecular biology research because the number of false positive predictions is reasonably limited. Analyzing the plots it becomes clear that both variants of feature based protein family HMMs substantially outperform state-of-the-art discrete Profile HMMs. It can be seen that the number of false negative predictions can generally be decreased while reducing the number of false positives. The ROC-curves corresponding to SCFB models lie significantly below the reference curve of discrete models for the whole diagram. The effectiveness of the competitive evaluation of UBM and target models can be assessed by the maximum number of false positive predictions (which are especially critical since they usually correspond to subsequent irrelevant but expensive wet-lab investigations). Due to our explicit rejection model this number is dramatically reduced by almost 66 percent for all superfamilies (cf. the particular maxima on the y-axis). Note that

the sensitivity of the UBM is not perfect resulting in a small amount of false rejections. Since hard decisions are delivered by the abovementioned competitive model evaluation, some ROC-curves of SCFB models do not cross the y-axis (marked with '+'). In table 2 the characteristic values for the 5% working points within the ROC-curves are given proving the superior performance of SCFB HMMs.

The second sets of experiments was directed to the assessment of the capabilities of protein family HMMs obtained when only little training data is available which is critical for robust estimation even when using sophisticated model regularization techniques. For an overview of the models' detection capabilities evaluated for the SCOPSUPER95_44f corpus, in figures 7 and 8 the ROC-curves corresponding to three exemplary sub-corpora are presented. In the first figure detection results for models estimated using 44 sequences per superfamily (upper diagram), and when using 30 samples each are shown (lower diagram). Figure 8 contains the results for models estimated using only 20 sequences per target. Analyzing the particular ROC-curves it can be seen that the new feature based protein family HMMs also substantially outperform state-of-the-art when the number of training samples is reduced. It becomes obvious that the smaller the training sets, the more favorable are especially the SCFB models with reduced complexity (BLR). The substantially reduced amount of model parameters can be robustly estimated even when only very little training data is available. Nevertheless, the flexibility of the models is sufficient for remote homology analysis.

In table 3 the corresponding characteristic values are given again proving the substantial progress for remote homology analysis when using the new approaches. In some of the experiments the characteristic values were not met. For prematurely ending ROC-curves (caused by UBM's false rejections) resulting in unreached working points in table 3 the appropriate global maxima at the endpoints of the ROC-curves are given in parentheses.

5 Summary

Due to the progress of genome sequencing projects providing huge amounts of biological data, the computational analysis of protein sequences has become more and more important for molecular biology research. Especially probabilistic models of protein families, namely Profile HMMs, are very promising for the classification of unknown data. Unfortunately, the general problem of *remote* homology analysis is still far from being solved, even when applying the most sophisticated probabilistic techniques.

We developed advanced probabilistic protein family models which represent a paradigm shift in protein sequence analysis addressing generally improved remote homology detection. Therefore, the bioinformatics problem was treated from a consequent pattern recognition point of view. By means of a rich feature based sequence representation semi-continuous protein family HMMs were developed, consisting of either Profile or a less-complex Bounded Left-Right (BLR) model architecture. The underlying general feature space representation is estimated using non-target specific sample sequences from SWISSPROT. For further model specialization mixture density adaptation techniques are applied, namely MAP or MLLR. Especially BLR models containing substantially smaller amounts of parameters are favorable for robust model estimation when only little training data is available. In combination with explicit background models their superior performance was demonstrated in various experiments based on the SCOP database at the level of superfamilies. The newly developed techniques were integrated into a ready-to-use sequence analysis system which can be used for remote homology analysis at a larger scale.

To conclude, the new approaches proposed in this article represent major improvements for remote homology analysis tasks. Furthermore, they can serve as the foundation for further developments which is very promising for general molecular biology research.

6 Acknowledgements

This work was supported by Boehringer Ingelheim and the Boehringer Ingelheim Pharma GmbH und Co. KG Genomics Group. The authors would especially like to thank Dr. Andreas Weith, Dr. Karsten Quast, Dr. Andreas Köhler, and Ogsen Gabrielyan for numerous fruitful discussions and their enthusiastic support.

References

[1] G. A. Fink, Developing HMM-based recognizers with ESMERALDA, in: Text, Speech and Dialogue, Vol. 1692 of Lecture Notes in Artificial Intelligence, Springer, 1999, pp. 229–234.

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

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

[4] S. F. Altschul, et al., Basic local alignment search tool, J. Mol. Biology 215 (3) (1990) 403–410.

[5] W. R. Pearson, D. J. Lipman, Improved tools for biological sequence comparison, in: Proc. Nat. Academy of Sciences USA – Biochemistry, Vol. 85, 1988, pp. 2444–2448.

[6] D. Haussler, et al., Protein modeling using Hidden Markov Models: Analysis of Globins, in: Proc. 26th Ann. Hawaii Int. Conf. System Sciences, Vol. 1, 1993, pp. 792–802.

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

[8] P. Baldi, et al., Hidden Markov Models of biological primary sequence information, in: Proc. Nat. Academy of Sciences USA – Biochemistry, Vol. 91, 1994, pp. 1059–1063.

[9] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

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

[11] A. Dempster, L. N.M., D. Rubin, Maximum likelihood from incomplete data via the *EM* algorithm, Journal of the Royal Statistical Society 39 (1977) 1–38, series B (methodological).

[12] N. C. Jones, P. A. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press Inc., 2004.

[13] M. Gribskov, et al., Profile analysis: Detection of distantly related proteins, in: Proc. Nat. Academy of Science USA – Biochemistry, Vol. 84, 1987, pp. 4355–4358.

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

[15] R. Durbin, et al., Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998.

[16] C. Barrett, et al., Scoring Hidden Markov Models, Computer Applications in the Bioscience 13 (2) (1997) 191–199.

[17] K. Karplus, Evaluating regularizers for estimating distributions of amino acids, in: Proc. Int. Conf. Intelligent Systems for Molecular Biology, 1995, pp. 188–196.

[18] T. Plötz, G. A. Fink, Robust remote homology detection by feature based Profile Hidden Markov Models, Statistical Applications in Genetics and Molecular Biology 4 (1).

[19] D. Anastassiou, Genomic signal processing, IEEE Signal Processing Magazine 18 (4).

[20] P. D. Cristea, Genetic signals: An emerging concept, in: Proc. Int. Workshop on System, Signals and Image Processing IWSSIP 2001, Bucharest, Romania, 2001.

[21] I. Cosic, The Resonant Recognition Model of Macromolecular Bioactivity – Theory and Applications, Birkhäuser Verlag, Basel, 1997.

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

[23] T. Plötz, G. A. Fink, Feature extraction for improved Profile HMM based biological sequence analysis, in: Proc. Int. Conf. on Pattern Recognition, 2004.

[24] T. Plötz, Advanced stochastic protein sequence analysis, Ph.D. thesis, Bielefeld University (Jun. 2005).

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

[26] X. D. Huang, M. A. Jack, Semi-continuous Hidden markov Models for speech signals, Computer Speech & Language 3 (1989) 239–251.

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

[28] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1967, pp. 281–296.

[29] C. J. Leggetter, et al., Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models, Computer Speech & Language (1995) 171–185.

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

[31] D. A. Reynolds, Comparison of background normalization methods for text-independent speaker verification, in: Proc. European Conf. on Speech Communication and Technology, Vol. 1, Rhodes, Greece, 1997, pp. 963–966.

[32] W. N. Grundy, et al., Meta-MEME: Motif-based Hidden Markov Models of protein families, Computer Applications in the Bioscience 13 (4) (1997) 397–406.

[33] T. Plötz, G. A. Fink, A new approach for HMM based protein sequence modeling and its application to remote homology classification, in: Proc. Workshop Statistical Signal Processing, IEEE, Bordeaux, France, 2005.

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

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

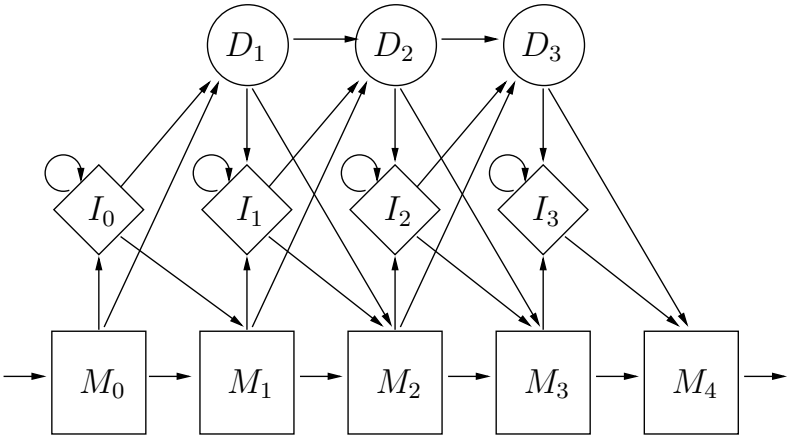

Fig. 1. State-of-the-art discrete Profile HMM architecture.

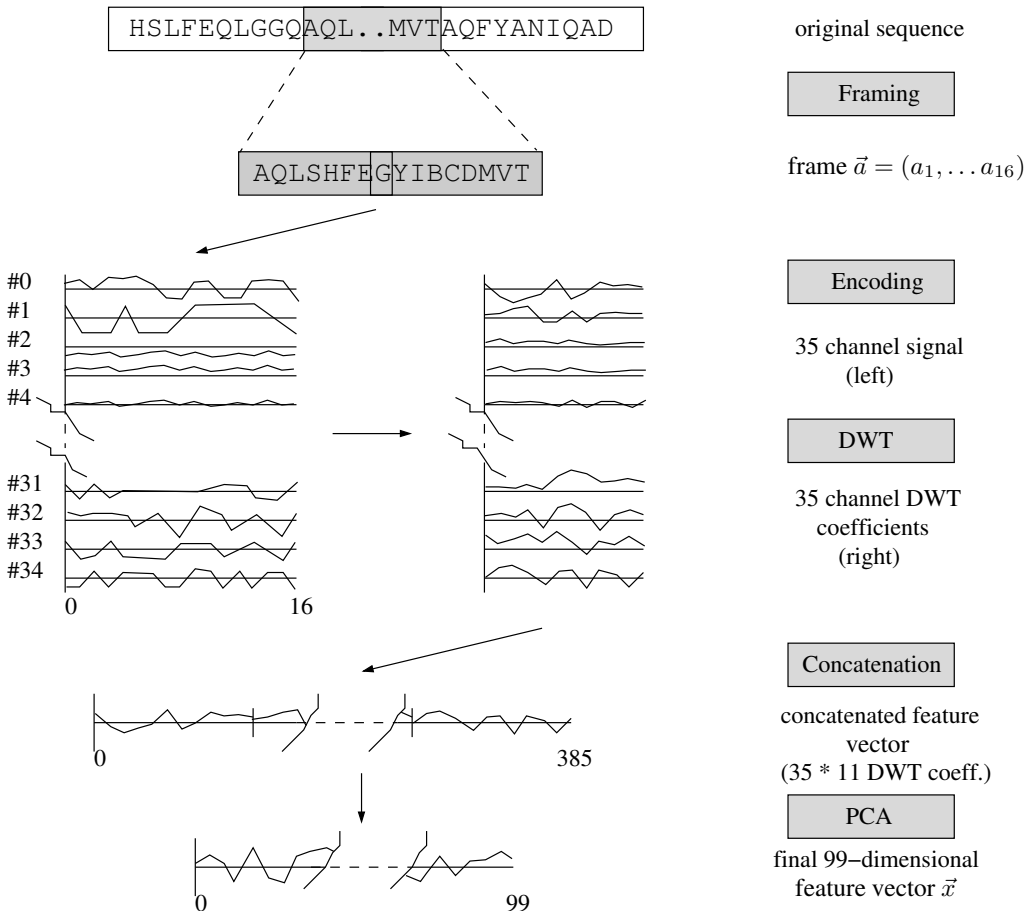

Fig. 2. Overview of the feature extraction method for protein sequences (cf. [23]).

Channel Index	Description
0	Average flexibility indices
1	Residue volume
2	Transfer free energy to surface
3	Steric parameter
4	Polarizability parameter
5	A parameter of charge transfer capability
6	A parameter of charge transfer donor capability
7	Normalized average hydrophobicity scales
8	Size
9	Relative mutability
10	Solvation free energy
11	Molecular weight
12	Melting point
13	pK-N
14	pK-C
15	Graph shape index
16	Normalized van der Waals volume
17	Positive charge
18	Negative charge
19	pK-a (RCOOH)
20	Hydrophilicity value
21	Average accessible surface area
22	Average number of surrounding residues
23	Mean polarity
24	Side chain hydropathy, corrected for solvation
25	Bitterness
26	Bulkiness
27	Isoelectric point
28	Composition of amino-acids in extracellular proteins
29	Composition of amino-acids in anchored proteins
30	Composition of amino-acids in membrane proteins
31	Composition of amino-acids in intracellular proteins
32	Composition of amino-acids in nuclear proteins
33	Amphiphilicity index
34	Electron-ion interaction potential values

Table 1
Biochemical properties selected for sequence representation.

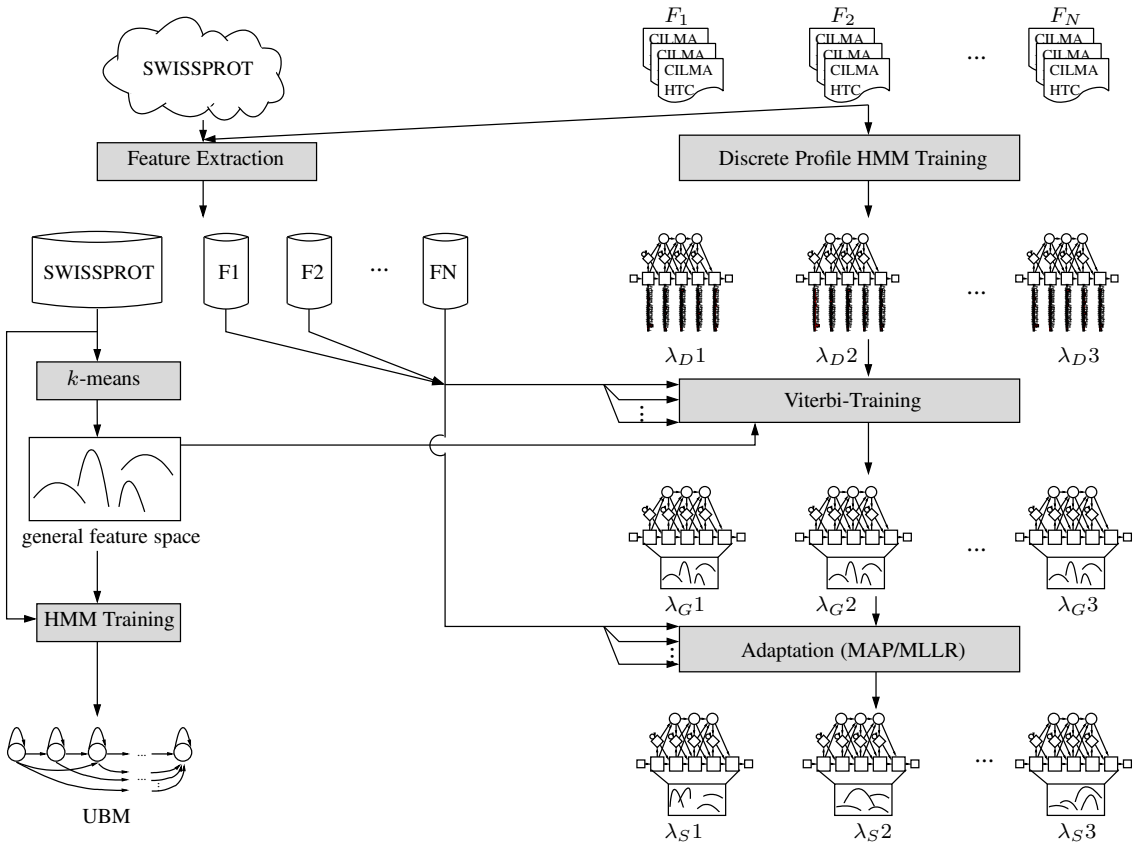

Fig. 3. Overview of the estimation procedure for feature based Profile HMMs.

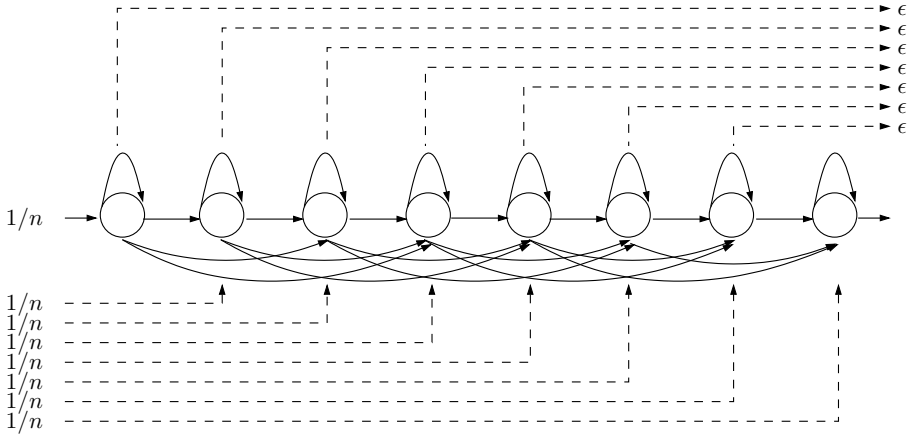

Fig. 4. Protein family model with bounded left-right topology.

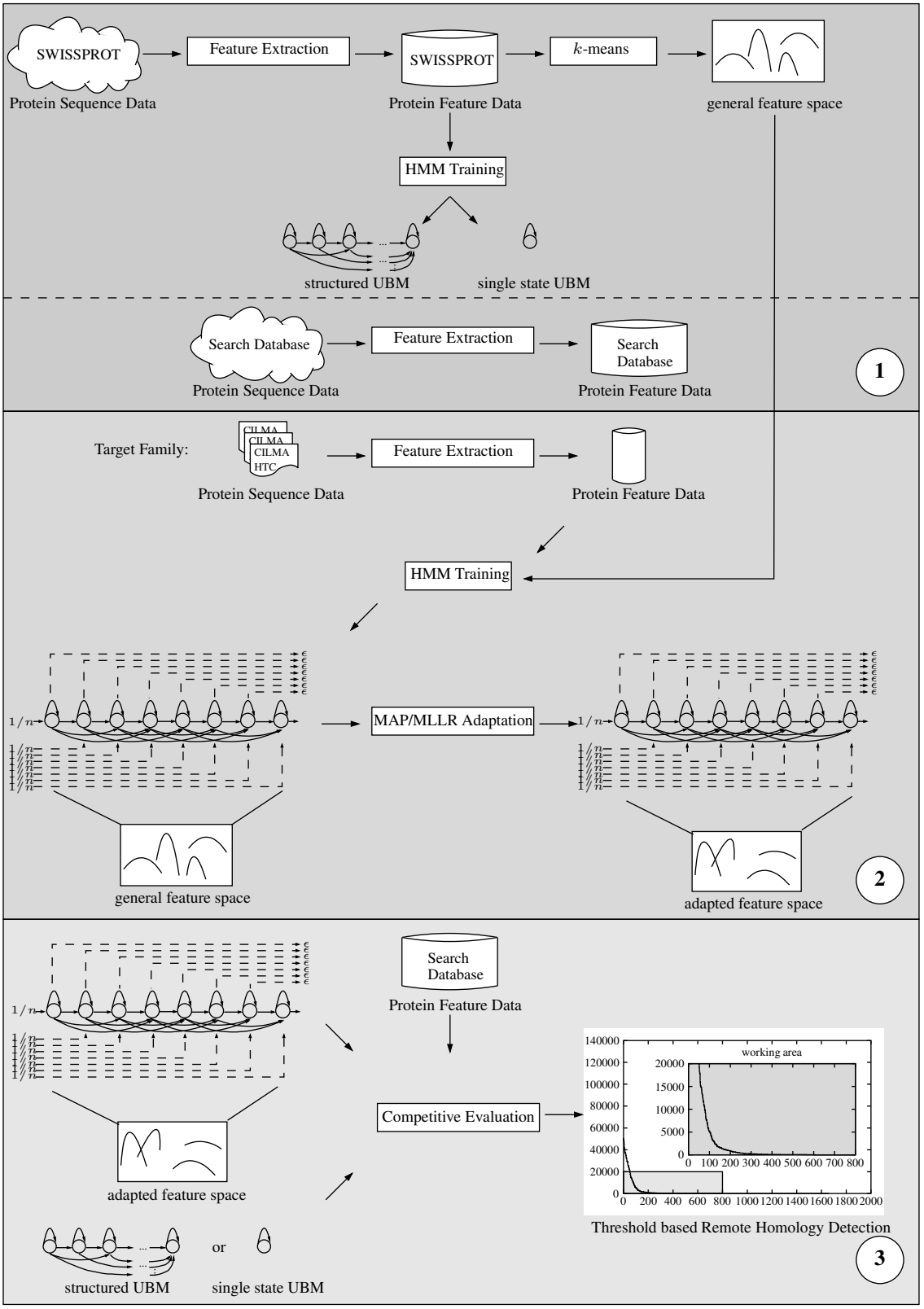

Fig. 5. Application scheme overview of semi-continuous protein family HMMs.

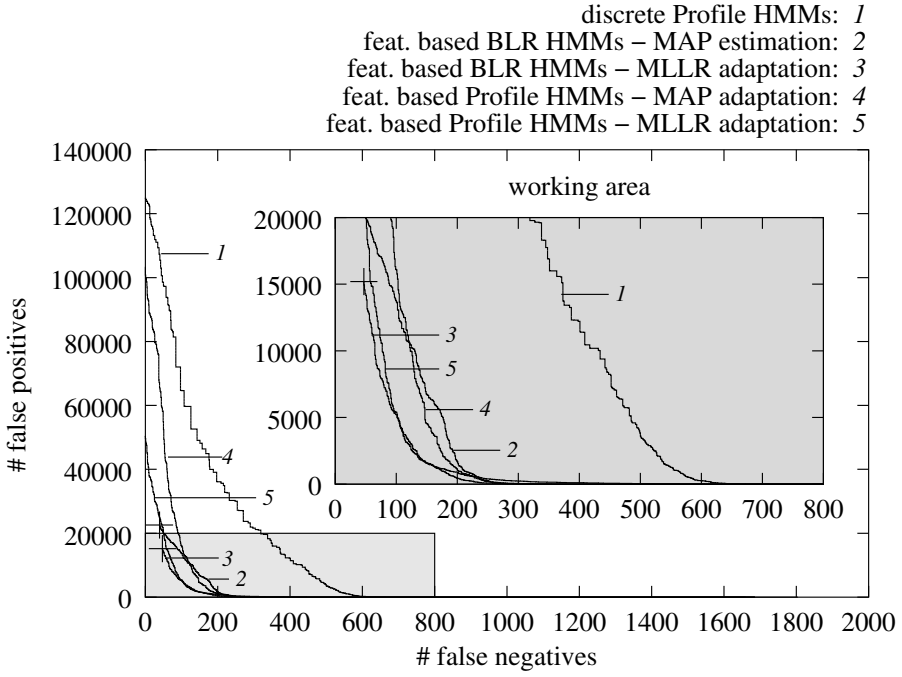

Fig. 6. ROC-curves for the experimental evaluation using the SCOPSUPER95_66 corpus.

HMM Variant	False Negative Predictions [%] for 5 % False Positives	False Positive Predictions [%] for 5 % False Negatives
Discrete Profile	26.1	57.6
SCFB Profile (MAP)	7.9	16.0
SCFB Profile (MLLR)	5.1	5.5
SCFB BLR (MAP)	8.9	11.9
SCFB BLR (MLLR)	4.9	4.7

Table 2

Characteristic values for SCOPSUPER95_66 detection experiments.

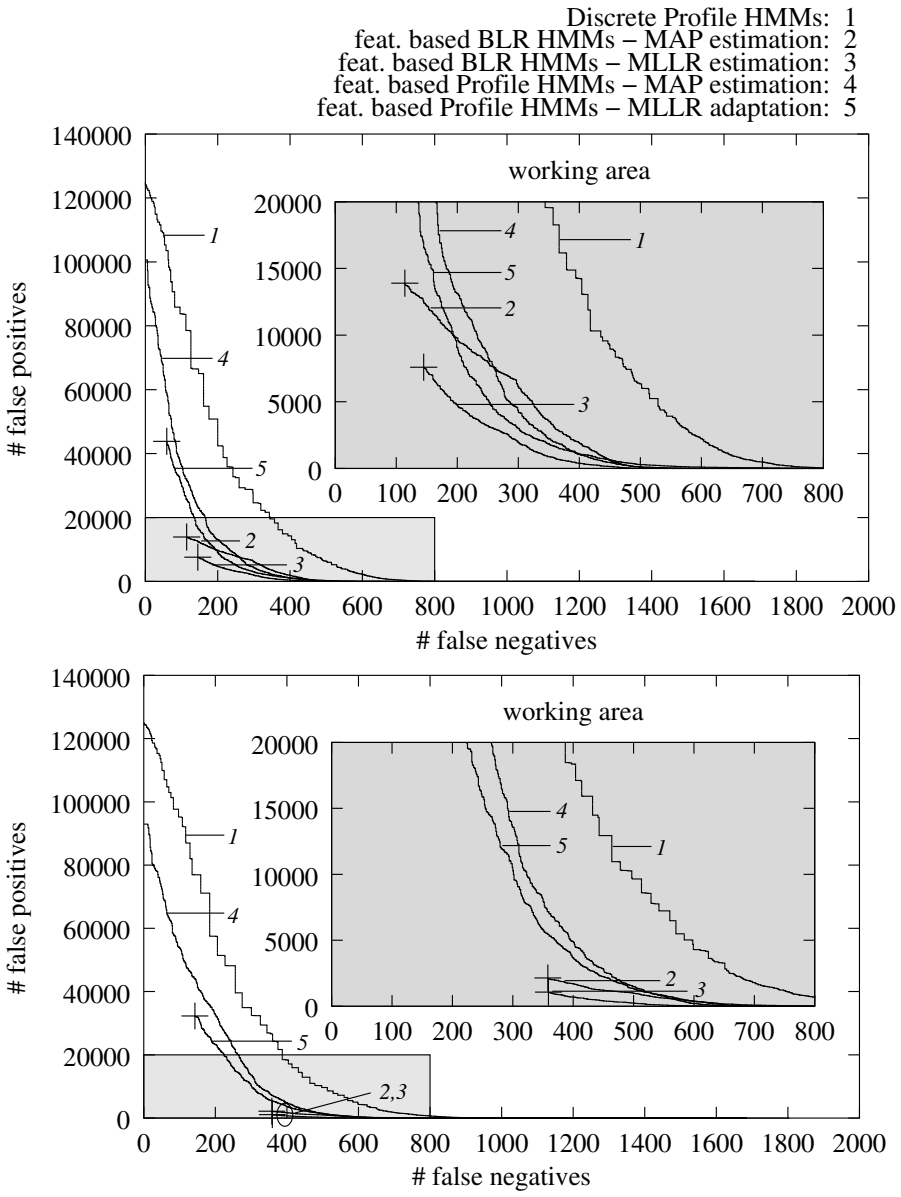

Fig. 7. ROC-curves for the experimental evaluation using the SCOPSUPER95_44f corpus; upper diagram: 44 training sequences / lower diagram: 30 training sequences.

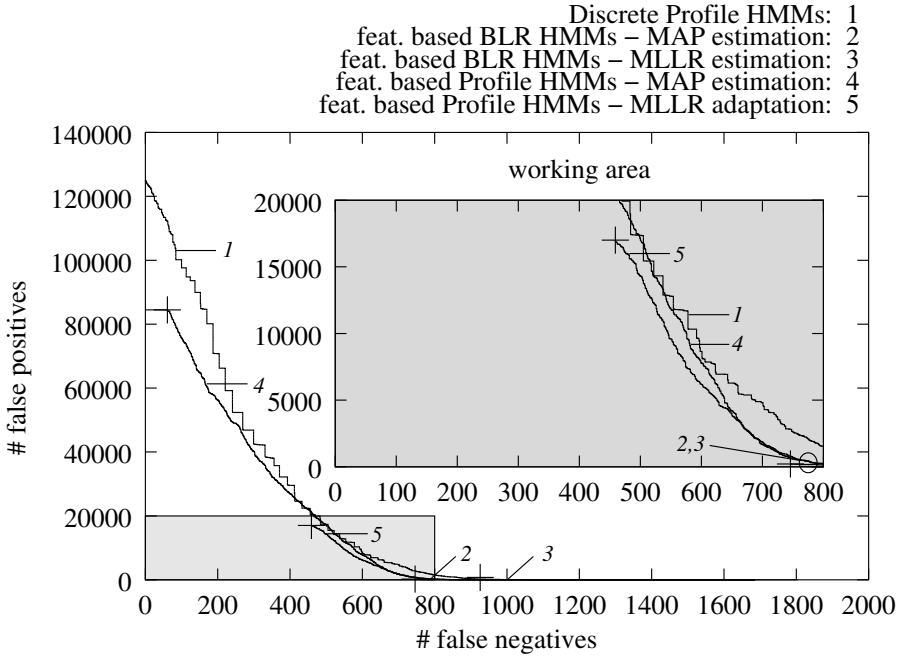

Fig. 8. Summary of detection results for SCOPSUPER95_44f (20 training sequences).

# Training Samples	HMM Variant	False Negative Predictions [%] for 5 % False Positives	False Positive Predictions [%] for 5 % False Negatives
44	Discrete Profile	27.8	68.7
	SCFB Profile (MAP)	15.1	31.4
	SCFB Profile (MLLR)	12.9	26.6
	SCFB BLR (MAP)	16.6	0.0 (11.1)
	SCFB BLR (MLLR)	9.3	0.0 (6.1)
30	Discrete Profile	31.6	78.2
	SCFB Profile (MAP)	20.9	45.0
	SCFB Profile (MLLR)	19.0	0.0 (25.9)
	SCFB BLR (MAP)	0.0 (19.8)	0.0 (1.7)
	SCFB BLR (MLLR)	0.0 (19.9)	0.0 (0.9)
20	Discrete Profile	36.3	80.2
	SCFB Profile (MAP)	34.5	62.2
	SCFB Profile (MLLR)	33.1	0.0 (13.6)
	SCFB BLR (MAP)	0.0 (41.4)	0.0 (0.2)
	SCFB BLR (MLLR)	0.0 (51.3)	0.0 (0.6)

Table 3
Characteristic values for SCOPSUPER95_44f experiments (44/30/20 training sequences).