In **Proc. Workshop Statistical Signal Processing** , Bordeaux, France, 2005.

# A NEW APPROACH FOR HMM BASED PROTEIN SEQUENCE FAMILY MODELING AND ITS APPLICATION TO REMOTE HOMOLOGY CLASSIFICATION

*Thomas Plötz and Gernot A. Fink*

Faculty of Technology
Bielefeld University, Germany
{tploetz,gernot}@techfak.uni-bielefeld.de

## ABSTRACT

Currently probabilistic models of protein families, namely HMMs, are the methodology of choice for remote homology analysis. Unfortunately, the topology of such so-called Profile HMMs is rather complex which, despite sophisticated regularization techniques, is problematic for robust model estimation when only little training data is available.

We propose a new HMM based protein family modeling method using building blocks which capture the essentials of particular targets only. They are estimated in a fully data-driven and unsupervised procedure. Contrary to current motif detection procedures we use a feature based protein sequence representation we developed earlier. Such small building blocks are automatically combined to global protein family HMMs which can be applied to remote homology analysis tasks.

The results of an experimental evaluation on a challenging task of remote homology classification prove that robust models containing substantially smaller amounts of parameters can be estimated using the new modeling approach. The smaller the number of parameters to be trained, the smaller the number of training samples required which is of major importance for e.g. drug discovery tasks.

## 1. INTRODUCTION

The correct classification of biological sequence data regarding its protein family membership is of fundamental scientific as well as commercial interest. In protein families biologically related amino acid sequences are grouped according to the function they encode. Although sharing a common biological function these sequences can be highly divergent at the residue level. Very frequently, sequence similarities of less than 40 percent occur. The analysis of such so-called remote homologies is relevant especially for drug design applications, namely for target identification and verification.

Currently, the most promising approaches for remote homology classification are based on probabilistic models for particular protein families, namely Profile HMMs. Protein data classification is performed by directly aligning the sequences of interest to the stochastic models (cf. e.g. [1] for an excellent treatment of probabilistic protein family modeling). Contrary to general pattern recognition applications like automatic speech recognition, the modeling base for Profile HMM approaches is mostly rather large. Even when estimating models for the functionally smallest units, the protein domains, usually very large models consisting of several hundred states are required. Generally, in order to train models including enormous amounts of parameters large training sets are required. However, especially for pharmaceutical applications usually only small amounts of sample data are available. The common model regularization by incorporating prior expert knowledge is critical especially for remote homology treatment since the resulting models tend to be biased towards facts already known before. Reducing the complexity of the models which directly corresponds to limiting training sets required for robust model estimation seems more promising.

In order to tackle the problem of remote homology classification, in this paper we propose a new signal-processing based approach for probabilistic protein family modeling using HMMs. Using our previously developed signal-like protein sequence representation which directly covers biochemical properties of residues in their local neighborhood, and features derived from it [2], building blocks for protein families are determined in a completely un-supervised and data-driven manner. In analogy to sub-word models in automatic speech recognition applications, we call these building blocks *Sub-Protein Units (SPUs)*. Such SPUs, which cover only those parts of a protein family relevant for successful sequence classification, are modeled using standard HMMs with less complex model architectures. The models can be estimated robustly using a variant of the EM algorithm. For this approach significantly less training samples are sufficient. By means of the SPUs most frequently occurring within the training sets, automatically protein family

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

models are derived by concatenation of the building blocks. Based on a representative superfamily classification task, we demonstrate the improved performance of our new approach compared to Profile HMMs while substantially reducing the number of model parameters required.

This paper is organized as follows. The state-of-the-art for HMM based remote homology classification is briefly summarized in section 2. Following this we present the new modeling approach and the results of the experimental evaluation (sections 3 and 4).

## 2. STATE-OF-THE-ART HMMS FOR REMOTE HOMOLOGY CLASSIFICATION

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 unaligned training sequences [3]. There are some extensions to the basic architecture with increased flexibility, e.g. HMMERs Plan7 [4].

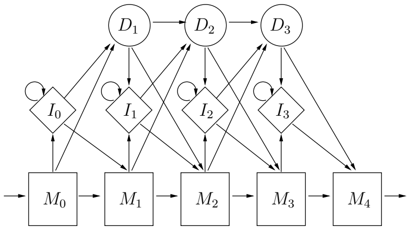

**Fig. 1**. State-of-the-art Profile HMM

Mostly, 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 Baum-Welch or Viterbi training. For classification of sequence data the models are evaluated by computing the Forward or Viterbi scores, respectively.

Compared to general pattern recognition applications, there is one major difference in using HMMs for protein sequence classification. Usually, rather large parts of proteins are modeled using global Profile HMMs. Even when modeling the smallest functional protein units, protein domains, large amounts of model parameters need to be trained. For local alignments where parts of the model match to parts of the sequence of interest the complex model architecture is required. The general concept of building blocks for protein family models is not generally used within the bioinformatics domain. Only very few approaches exist where global protein family models are created by concatenating smaller building blocks. Since protein domains represent the smallest functional protein unit, conservation based blocks, so-called motifs, serve as the base for protein family models [5]. However, although many motif detection techniques were developed, there is hardly any literature concerning motif based modeling of complete protein families.

## 3. MODELING PROTEIN FAMILIES USING SUB-PROTEIN UNITS (SPUS)

The basic goal of our new protein family modeling approach is to reduce the complexity of the models and thus the number of parameters required for robust remote homology classification. Therefore, the protein families (domains) which are currently modeled using Profile HMMs are analyzed regarding some kind of low-level building blocks using signal processing techniques. These building blocks are principally defined at the sequence level and represent "interesting" or dominant parts of proteins. Instead of analyzing the sequence of amino acids directly, which corresponds to a traditional motif detection approach, we extract the building blocks from features derived from the signal-like representation of the biochemical properties of residues in their local neighborhood (cf. [2] for a detailed description of feature based protein data classification). Since biochemical properties of protein data are explicitly considered, the resulting building blocks do not necessarily correspond to motifs.

In analogy to sub-word units in automatic speech recognition applications we will call the basic building blocks *Sub-Protein Units (SPUs)*. In our approach we emphasize the strictly data-driven determination of the SPUs in order to avoid potentially misguiding impacts of manual model regularization. Contrary to the usual global protein family models which cover the complete e.g. protein domains, the new protein family models now consist only of the concatenation of SPUs. Due to the feature representation and the smaller length variance of SPUs they will be modeled using standard HMM architectures with reduced complexity. The modeling itself is limited to classification relevant parts of a particular protein family. Sequence parts which are not belonging to SPUs are not explicitly modeled. Instead of this, they are covered by a protein family specific *General (G)* model. Thus, the overall number of states required for robust estimation can be reduced significantly which is favorable especially for pharmaceutical applications.

The overall process of modeling protein families using SPU based HMMs can be divided into three parts which are described in the following. In figure 3 these steps are graphically summarized including the SPU based annotation of an exemplary *Immunoglobulin* (d1f5wa_).

### 1. SPU Candidate Selection

The feature extraction method developed in [2] provides a richer sequence representation which allows better remote homology classification when using Profile HMMs. The selection of SPU candidates is directly based on the 99-dimensional feature vectors. In the first step, general SPU candidates need to be extracted from protein sequences. The SPU based annotation of the sample data will be used for SPU-model training and protein family creation. In the left part of figure 3 this step corresponds to the upper row where potential SPUs are marked red. For clarity the amino acid representation is shown. However, the selection is performed using the 99-dimensional feature vectors.

Various criteria for classifying parts of the overall feature representation of protein sequences as SPU or non-SPU (so-called *General Parts G*) are imaginable. The approach presented in this paper represents a general framework for protein family modeling based on building blocks which are conceptually below the level of direct biological functions. In the version shown here we define SPUs as high-energy parts of the protein sequence. Note that alternative approaches for SPU candidate selection can be used equivalently within our framework. All parts of the original training sample whose feature vectors' energy is below the average energy of the particular protein sequence are treated as General parts G.

The actual discrimination method based on the feature vectors' energy becomes reasonable when analyzing the feature extraction method in more detail. In order to extract reasonable features, a discrete wavelet transformation (using standard Daubechies filters of length 4, cf. [2]) is applied to the signal-like representation of biochemical properties of residues in their local neighborhood. Following this, the approximation and some detail coefficients are used as the base for further analysis. One fundamental property of the wavelet transformation is the concentration of signal energy in the upper coefficients. Thus, high feature vector energy is a reasonable indicator for relevance. For robust SPU candidate selection the energy signal of the feature vectors corresponding to a particular protein sequence is post-processed using DWT based smoothing, i.e. wavelet analysis of the original energy signal followed by re-construction using only a subset of the wavelet coefficients (approximately 66% obtained by skipping the remaining detail coefficients).

By means of these techniques, protein sequences are principally sub-divided into SPUs and General parts G which can be seen in the right part of figure 3 for an exemplary *Immunoglobulin*. SPUs are extracted from the energy signal of the protein sequence (solid line) where the average protein energy (dashed line) is below the actual feature vector energy. By means of post-processing two SPU candidates (dotted rectangles) are selected.

### 2. SPU Modeling

In the first step of the new protein family modeling approach protein sequences are annotated with respect to the SPU candidates or General decision. Following this, corresponding SPUs need to be identified in order to train HMMs for a non-redundant set of SPUs relevant for the particular protein family.

The SPUs estimated for the protein family model, and the General model which is unique for every protein family, are modeled using linear, semi-continuous HMMs (second row in the left part of figure 3). Once the training set is finally annotated using the non-redundant set of SPUs, these models are trained with the standard Baum-Welch algorithm.

In the approach presented here, the final set of SPUs relevant for a particular protein family is obtained by applying a variant of the EM algorithm for agglomerative clustering of the initial (unique) SPU candidates. Therefore, model evaluation and training of SPU-HMMs is alternated up to convergence. Here, convergence means a “stable” SPU based annotation of the training set, i.e. no differences between the annotations obtained in two succeeding iteration steps. During the iterative SPU determination unique models for corresponding SPUs are estimated since redundant models will not be hypothesized. The set of effective SPU candidates is stepwise reduced and the most frequent SPUs are used for the final annotation of the training set. The procedure which is comparable to the k-means clustering for HMMs proposed in [6], is summarized in figure 2.

### 3. Protein Family Modeling

Given the non-redundant set of SPUs relevant for the particular protein family, finally the global protein family model is created. The protein family itself consists of variants of SPU concatenations obtained during training (third row in the left part of figure 3). The $N$ variants which are most frequently occurring within the annotation of the particular training sets, are extracted for the conceptual family definition. Note that the actual value of $N$ is subject of further optimization and as a working version we currently use all variants which were observed more than once during training. Here, optional parts (marked with '?' in figure 3) as well as looped occurrences are possible. For actual protein sequence classification, all variants are evaluated in parallel and determine the final classification decision.

When limiting the modeling process to classification relevant parts of a particular protein family the overall number

**1. Initialization:**
Obtain initial set $S_0$ of SPU candidates by e.g. energy based annotation of training sequences.

**2. Training**
Perform Baum-Welch training for SPU candidate models (linear HMMs) using the current annotation of the training set.

**3. Annotation**
Use updated SPU models for obtaining new annotation of the training set – recognition phase.

**4. Termination**
Terminate if two subsequent annotations of the training set do not (substantially) differ, i.e. convergence, continue with step 2 otherwise.

**5. SPU Candidate List Reduction**
Reduce set of SPU candidates by discarding those elements which are not included in the final annotation of the training sequences. Perform final annotation of the training set using the remaining list of SPU candidates.

**6. Final Training**
Perform steps 2-4 until convergence and train SPU models using the final annotation of the training set.

**Fig. 2**. Algorithm for obtaining a non-redundant set of SPUs which are used for final protein family modeling.

of HMM states required can be significantly reduced compared to standard Profile HMM approaches. Since SPUs are determined using a rich feature based protein data representation, the protein family modeling will not be limited to motifs estimated using plain amino acid sequences which is favorable for remote homology classification.

## 4. EXPERIMENTAL EVALUATION

In order to prove the effectiveness of our new protein family modeling approach using Sub-Protein Units, we performed an experimental evaluation. We created disjoint datasets for training Profile HMMs as well as for testing. Using the SUPERFAMILY [7] based hierarchy of the SCOP database [8], protein family models were established for the classification of protein sequences at the superfamily level. Here, sequences belonging to a distinct superfamily must not have similarity values above 95%. This means, even sequences having sequence identities of only a few percent may belong to these superfamilies[1]. The training corpora as well as the evaluation data for every protein family cover almost the whole range of possible similarity values. Thus, the performance for remote homology classification can actually be evaluated. The datasets contain sequences for 16 superfamilies, whereas the training sets have an average size of 70 samples and about 36 sequences are used at an average for the evaluation.

The discrete Profile HMMs used as reference were estimated by applying the state-of-the-art Profile HMM framework SAM v3.3.1 [9]. Therefore, the `buildmodel` was used in default parameterization which implies model regularization by incorporating prior knowledge modeled via mixtures of Dirichlet distributions and local alignments obtained via Forward-Backward evaluation.

The proposed modeling approach was implemented using our own HMM framework ESMERALDA[10]. In order to prove the effectiveness of the new protein family models we compared the classification error $CE$ produced using the SPU based approach with state-of-the-art discrete Profile HMMs. Besides the classification error, the number of states contained in the final protein family model is of major importance. The smaller the number, the smaller is the number of training samples required for robust model estimation. In table 4 the classification errors are presented together with the appropriate number of states. Three different versions of the SPU modeling approach were evaluated (SPU based v1/v2/v3). These versions differ in the number of most frequent SPUs used for the final annotation of the sample sequences (cf. the description of SPU modeling in the previous section) resulting in different numbers of HMM states. It can be seen that the classification error obtained when applying the new SPU based protein family models significantly decreases compared to the classical modeling using discrete Profile HMMs. The number of states required for this improvement could be decreased significantly: Only approximately 40 percent of the original number of states are required while decreasing the classification error by almost 30 % (v3). Thus, the number of training sequences required, which is directly dependent on the number of states used, can be decreased substantially. Note that for the experiments described the EM algorithm was assumed converged if *no* differences between two subsequent annotations of the training sets were encountered.

[1] Note that *most* corpora used for biological sequence analysis contain a large amount of sequences with identities of 95% and above.

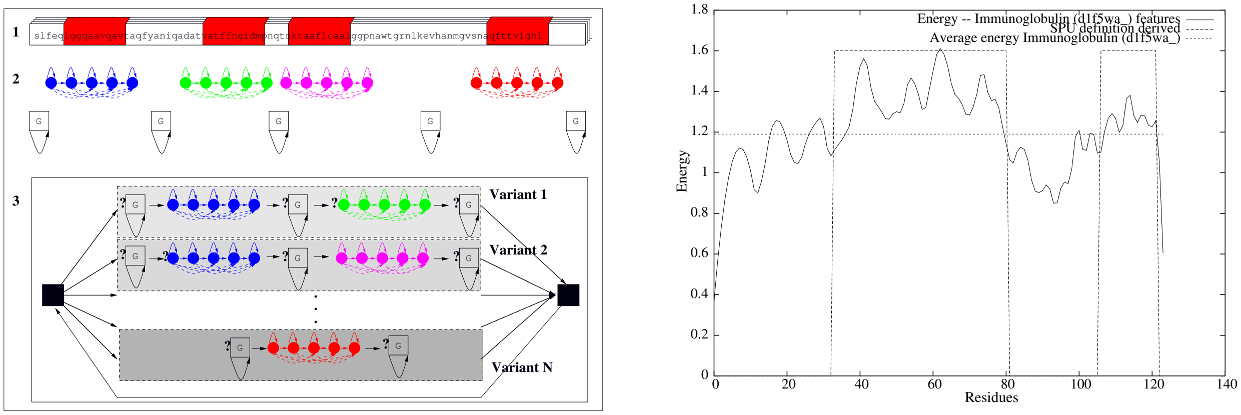

**Fig. 3**. Overview of the SPU based protein family modeling process (left) and result of the SPU-determination for an exemplary *Immunoglobulin* (d1f5wa_) protein (right)

| Modeling Type | # States $X$ | $\Delta X$<br>(vs. Profile HMMs) | # SPUs | $CE$ [%] | $\Delta CE$ [%]<br>(vs. discrete Profile HMMs) |
|---|---|---|---|---|---|
| Discrete Profile HMMs | 8726 | – | – | 32,9 | – |
| SPU based v1 | 2128 | -75,6 | 155 | 30,2 | -9.4 |
| SPU based v2 | 3398 | -61,1 | 169 | 24,9 | -24,2 |
| SPU based v3 | 3564 | -59,2 | 159 | 23,5 | -28,5 |

**Table 1**. Summary of the experimental evaluation ($CE$: Classification Error)

## 5. SUMMARY

In this paper we presented a new approach for HMM based protein family modeling using building blocks, namely Sub-Protein Units, estimated by analyzing a feature based sequence representation. Contrary to current Profile HMM based approaches, only the classification relevant parts of protein families are modeled. For a representative task of remote homology classification it could be shown that the number of parameters required for robust models and thus the amount of training data can be decreased substantially. This is especially important for pharmaceutical applications.

## 6. REFERENCES

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

[2] Thomas Plötz and Gernot A. Fink, "Feature extraction for improved Profile HMM based biological sequence analysis," in *Proc. Int. Conf. on Pattern Recognition*, 2004.

[3] Anders Krogh et al., "Hidden Markov Models in computational biology: Applications to protein modeling," *J. Molecular Biology*, vol. 235, pp. 1501–1531, 1994.

[4] Sean R. Eddy, "HMMER: Profile Hidden Markov Models for biological sequence analysis," *http://hmmer.wustl.edu/*, 2001.

[5] William N. Grundy, Timothy L. Bailey, Charles P. Elkan, and Michael E. Baker, "Meta-MEME: Motif-based Hidden Markov Models of protein families," *Computer Applications in the Bioscience*, vol. 13, no. 4, pp. 397–406, 1997.

[6] Michael P. Perrone and Scott D. Connell, "K-means clustering for Hidden Markov Models," in *Proc. Int. Workshop on Frontiers in Handwriting Recogntion*, L.R.B. Schomaker and L.G. Vuurpijl, Eds., Sept. 2000, pp. 229–238.

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

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

[9] R. Hughey and A. Krogh, "Hidden Markov Models for sequence analysis: Extension and analysis of the basic method," *Computer Applications in the Bioscience*, vol. 12, no. 2, pp. 95–108, 1996.

[10] Gernot A. Fink, "Developing HMM-based recognizers with ESMERALDA," in *Lecture Notes in Artificial Intelligence*, Václav Matoušek et al., Eds. 1999, pp. 229–234, Springer.