Feature extraction for improved Profile HMM based biological sequence analysis

Thomas Plötz and Gernot A. Fink
Faculty of Technology, Bielefeld University,
P.O. Box 100 131, 33 501 Bielefeld, Germany
{tploetz,gernot}@techfak.uni-bielefeld.de

Abstract

State-of-the-art systems for biological sequence analysis employ statistical modeling techniques, most notably so-called Profile HMMs. However, all approaches still rely on a purely symbolic sequence representation, which severely limits their capabilities in describing weak similarities between remotely homologue members of sequence families.

Therefore, we propose a multi-channel signal-like sequence representation based on a combination of several numerically encoded biochemical properties of the individual residues. From this representation features are extracted capturing relevant local sequence properties by applying wavelet and principal component analysis.

Evaluation results on a challenging task of sequence family classification prove that Profile HMMs trained on the feature-based sequence representation significantly outperform discrete models.

1. Introduction

In recent sequencing projects of complete organisms a vast amount of data has been produced which needs to be analyzed in order to obtain competent insights of biological processes. Following the paradigm, that similar function is caused by similar structure, especially for drug design tasks the classification of protein sequences regarding the family they are belonging to is of fundamental interest.

Traditional approaches for the classification of biological data, i.e. the primary structure of proteins, are based on alignment techniques. Using sophisticated dynamic programming methods, e.g. BLAST [1], two or more sequences are mutually aligned and scores are calculated providing measures for the similarity of the samples inspected. By means of these scores a familywise classification becomes possible. For close homologies such deterministic techniques are very powerful and widely used.

Besides highly conserved sequences, protein families can contain remote homologue members with weak mutual sequence similarities. Usually, these data is of great interest for the understanding of metabolism processes. Due to the importance of remote homologies, in the last years several probabilistic approaches have been developed for uncovering sequence relationships at higher levels. Basically, a statistical model for e.g. the protein family of interest is established which is subsequently used for classification and search tasks, respectively. Understanding the protein data as sequential streams the use of Hidden Markov Models became very popular. For the consensus of a distinct protein family a so-called Profile HMM [14] is established covering the alignment information in a probabilistic framework. Furthermore, structural information is incorporated by a specialized model architecture that distinguishes between match, insert, and delete states. The sequence of amino acids represents the emissions of the hidden states. Presently, Profile HMMs are the methodology of choice for a large variety of applications, e.g. [3, 14].

Although probabilistic approaches outperform traditional sequence analysis techniques, the overall situation is still not satisfying at all. Using state of the art Profile HMMs the classification error for remote homologies remains quite high. Since the fundamental understanding of higher level processes in complex metabolisms is explicitly depending on the previous analysis of primary structure data, improved techniques are strongly demanded.

The power of Profile HMMs is based on *statistic* modeling, represented by probabilistic state transitions and emissions. Thus, contrary to strongly deterministic techniques, a more fuzzy classification becomes possible. Nevertheless, consequently spoken the emissions itself are still discrete symbols – the 20 amino acids of the protein sequence.

In order to decrease the classification error of Profile HMMs for remote homologies we present an approach introducing a continuous emission space for protein sequences. Based on a multi-channel signal representation of biological data using the combination of 35 different amino acid indices, we create feature vectors substituting the basic amino acid symbols of present (discrete) Profile HMMs. We introduce the analysis of local relationships into the *emission space*. For every channel a short time analysis is performed on a fixed length frame using the Dis-

In Proc. Int. Conf. on Pattern Recognition, number 2, pages 315-318. IEEE, 2004.

crete Wavelet Transform (DWT). The concatenation of only some of the initial coefficients of all channels, neglecting detail information of the DWT, provides a coarse multi-channel spectrum, abstracting from potentially misguiding detail. By means of a principal component analysis (PCA) of the concatenated DWT coefficients, local characteristics are represented in a 99-dimensional feature space. Features, created using our data driven extraction process can be to some extend identified with residues in a special context.

Subsequently to a short description of present sequence analysis techniques using Profile HMMs in section 2, we describe the signal representation of the protein data used for the proposed approach. Section 4 in detail describes the proposed feature extraction method. In order to prove the effectiveness of the new method we performed an experimental evaluation on a set of remote homologue sequences for 16 protein superfamilies. Compared to a reference experiment using the state of the art Profile HMM framework SAM [12] impressing reductions of the classification error could be achieved. Section 5 describes the datasets used as well as the performed evaluation itself.

2. Biological sequence analysis using 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.

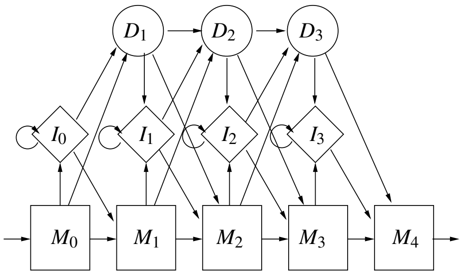

Figure 1. State of the art Profile HMM

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. There are some extensions to this basic architecture with increased flexibility, e.g. in HMMERs Plan7 [9].

Presently, the emissions of Profile HMMs are modeled by state dependent discrete probability distributions over the set of 20 amino acids. The model parameters, i.e. 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 [8].

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. Signal representation of protein sequences

The basic motivation for developing alternative representations for biological sequences is reasoned by the huge amount of powerful signal processing techniques already available. By means of such techniques, the application of Profile HMMs can benefit from uncovering possibly hidden characteristics of protein data. Thus, especially for remote homologies new biological insights can be obtained yielding decreased classification errors.

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 [2, 7] or to biochemical properties [6]. Spectral representations can then be derived from the encoded sequences. For classification purposes cross spectral energies are used as sequence similarity measures.

The most promising signal representation approaches rely on biochemical sequence properties. Kawashima et al. compiled a huge amount of such so-called indices [13]. Every index defines a mapping of amino acids to numerical values depending on biochemical properties, e.g. the hydrophobicity, molecular weight, or electron-ion interaction potential. Once the amino acids are appropriately mapped to numerical values, a large variety of signal processing techniques can be applied.

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, in our approach we do not want to restrict the representation to a single biochemical property. We carefully selected 35 indices out of the huge pool of encoding schemes available and used the combination of them as a multi-channel signal representation. The combination of multiple biochemical properties provides a rich characterization of the protein sequence analyzed.

4. Feature extraction for Profile HMMs

The previously described sequence encoding method subsumes information from various sources in a multi-channel numerical representation. In order to respect *local* signal characteristics, in our feature extraction procedure consecutive samples of the 35 channel signals are analyzed using a sliding window approach (extracting *frames* of the data). Starting from the first residue of a distinct sequence for each of the 35 channels 16 samples are used for the short time signal analysis. The window length was heuristically determined in preceding experiments. At the borders of the sequences the data is padded using prior knowledge about general amino acid distributions. Since Profile HMMs are used for sequence analysis tasks with a granularity of single amino acids, the window is moved along the whole sequence using single residue steps. Thus, at every step we have an overlap of 15 samples for each channel.

Basically, for remote homology detection 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. Principally, transforming signals into a frequency based representation offers direct access to the desired shape approximation. Within the general signal processing domain a wide variety of such transformations has been developed, e.g. the Fourier or the Cosine transform. Once the transformation is established, 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.

In the proposed approach we use the powerful Discrete Wavelet Transform (DWT) for the analysis of the coarse temporal signal structure, cf. e.g. [4]. Per channel of the signal representation of the appropriate protein sequence the Wavelet coefficients are determined using the standard Daubechies filter of length 4. For the detection of remote homologies we skip the upper five coefficients containing detail information, resulting in 35 11-dimensional feature vectors. Per frame the channel based feature vectors are concatenated to single 385-dimensional vectors.

As described in section 3, 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 the detection of remote homologies, the goal of the proposed method is to avoid as many pitfalls in the early stages of sequence analysis as possible. Thus, a completely data driven approach was developed. After combining the DWT coefficients of all channels into a single feature vector, potentially redundant information needs to be removed. Thus, we finally perform a Principle Component Analysis (PCA) for the feature vectors of every frame. Inspecting the eigenvalue spectrum of the data, it becomes clear, that a compact representation in a 99-dimensional subspace is sufficient.

Figure 2 schematically summarizes the feature extraction method described in this paper. Starting from the plain protein sequence, frames of length 16 are extracted (upper part of the sketch). 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 channels, illustrated on the right side, are concatenated to a single feature vector per frame. The concluding PCA (lower part) reduces the dimension to 99.

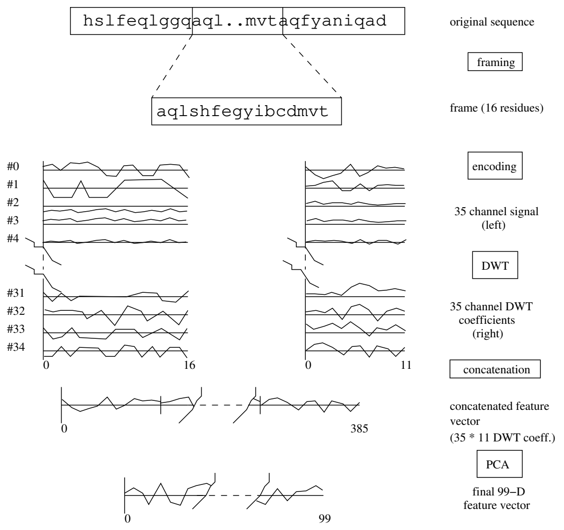

Figure 2. Feature extraction process (sketch)

By means of the proposed feature extraction method a powerful substitution scheme for the emissions of Profile HMMs has been developed. Using the extracted features of the training set for the appropriate model a vector quantization step is performed using standard techniques like k-means etc. The mixtures' weights of the resulting quantization of the emission space are used as emissions for the match as well as for the insert states of Profile HMMs.

5. Experimental Evaluation

In order to prove the effectiveness of substituting the standard emissions of Profile HMMs by features representing the coarse temporal characteristics in a finite context, we performed an experimental evaluation. We created disjoint datasets for training Profile HMMs as well as for testing. Using the SUPERFAMILY [11] based hierarchy of the SCOP database [5], profile HMMs 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

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

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 detection 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.

In order to compare our new feature based evaluation of Profile HMMs, initially we performed baseline experiments using the state of the art framework SAM [12] version 3.3.1. Using SAM, the classification is performed by means of log-odd scores as described in e.g. [8]. The baseline classification error is given by the results provided by using the E-value scores. The classification error of SAM based experiments here is approximately 32,9%.

The proposed feature extraction method was implemented using our own HMM framework ESMERALDA[10]. In order to prove the effectiveness of the new emission representation, we used exactly the same Profile HMM architecture as SAM, cf. figure 1. The classification error could significantly be reduced by 25,8% relative to about 24,4%.

Table 1 summarizes the substantial improvements for the SUPERFAMILY task. Since the corpora are based on real data (no simulation) a superior performance of our feature based sequence representation can also be expected for related tasks.

	CE	Δ CE
baseline – discrete Profile HMM (symbol based)	32,86%	–
multi-channel feature based Profile HMM (cont. emissions)	24,38%	-25,8%

Table 1. Experimental evaluation results

6. Conclusion

Profile HMMs presently are the methodology of choice for biological sequence analysis especially for close homologies. The detection of remote homologies, i.e. sequences belonging to a distinct protein family but containing only weakly conserved primary structures, is of fundamental interest for competent biological insights. Here, present Profile HMM approaches perform suboptimally.

In this paper we proposed an approach for the substitution of present (discrete) emissions of Profile HMMs by local features. We developed a multi-channel signal representation of fixed length sequence frames and extracted a coarse signal representation by calculating the first 11 DWT coefficients for every channel. Finally, feature vectors are created by removing redundancies from the concatenations of the multi-channel DWT features using PCA. The resulting 99-dimensional feature vectors are the base for the emissions of the states of Profile HMMs.

The experimental evaluation of the proposed method, based on classification tasks using the SUPERFAMILY corpus, proved the effectivity of the new technique. Comparing to the baseline result using the state of the art Profile HMM framework SAM, the classification error could significantly be decreased by more than 25% relative.

Beyond the demonstrated improved classification performance, a fundamental step towards an alternative sequence representation was performed. The proposed feature representation opens the Profile HMM domain for various signal processing approaches.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Molecular Biology*, 215(3):403–410, 1990.

[2] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing magazine*, 18(4), 2001.

[3] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov Models of biological primary sequence information. In *Proc. Nat. Academy of Science USA – Biochemistry*, volume 91, pages 1059–1063, Feb. 1994.

[4] W. Bäni. *Wavelets*. Oldenbourg, Munich Vienna, 2002.

[5] L. L. Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Scop database in 2002: refinements accomodate structural genomics. *Nucleic Acids Research*, 30(1):164–267, 2002.

[6] I. Cosic. *The Resonant Recognition Model of Macromolecular Bioactivity*. Birkhäuser Verlag, Basel, 1997.

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

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

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

[10] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek et al., editors, *Lecture Notes in Artificial Intelligence*, pages 229–234. Springer, 1999.

[11] J. Gough, K. Karplus, R. Hughey, and C. Chothia. 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, 2001.

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

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

[14] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden Markov Models in computational biology: Applications to protein modeling. *J. Molecular Biology*, 235:1501–1531, 1994.