Sequence Motif Identification and Protein Family Classification Using Probabilistic Trees

Florencia Leonardi and Antonio Galves

Instituto de Matemática e Estatística, Universidade de São Paulo

Abstract. Efficient family classification of newly discovered protein sequences is a central problem in bioinformatics. We present a new algorithm, using *Probabilistic Suffix Trees*, which identifies equivalences between the amino acids in different positions of a motif for each family. We also show that better classification can be achieved identifying representative fingerprints in the amino acid chains.

1 Introduction

A central problem in genomics is to determine the function of a new discovered protein using the information contained in its amino acid sequence [1]. Nowadays, the most popular methods to generate a hypothesis about the function of a protein are BLAST and Hidden Markov Models (HMM).

Probabilistic Suffix Trees (PST) were first introduced in [2] as a universal model for data compression. A major advantage of PST is its capacity of extracting structural information from the sequences under analysis. Recently, an implementation of PST has been successfully used in protein classification [3, 4], even though its performance decreases with less conserved families. Better results have been obtained using PST models for sparse sequences [5, 6]. A major drawback of these algorithms is their high complexity, which makes problematic their application in very large databases.

We present a new algorithm to estimate *Sparse Probabilistic Trees* (S-PT). We also show that the identification of sub-sequences of maximal mean probabilities (*fingerprints*) increases the classification rates of the PST algorithm. This is the basis of our F-PST algorithm.

2 The SPST and the F-SPST algorithms

It was suggested in the literature to use PST models to fit protein families. A PST is a stochastic chain $(X_0, X_1, \ldots)$ taking values on a finite alphabet $\mathcal{A}$ and characterized by two elements. The first element is the set of all contexts. A context $X_{n-\ell}, \ldots, X_{n-1}$ is the finite portion of the past $X_0, \ldots, X_{n-1}$ for each time which is relevant to predict the next symbol X_n. Observed that the length ℓ of the context depends on the past. The second element is a family of probability transitions associated to the set of the contexts. Given a context, its

associated probability transition gives the distribution of ocurrence of the next symbol inmediatly after the context.

In a PST the set of contexts has the *suffix property*: looking from the present to the past no context is a suffix of another context. This makes it possible to define without ambiguity the probability distribuition of the next symbol. The suffix property enable to represent the set of contexts as a tree. In this tree, each context $c = c_{-k}, \ldots, c_{-1}$ is represented by a complete branch, in which the first node on top is c_{-1} and so on until the last element c_{-k} which is represented by the terminal node of the branch.

In a PST model for a protein family, the alphabet $\mathcal{A}$ represents the set of twenty amino acids and the stichstic chains $(X_0, X_1, \ldots)$ are the sequences of amino acids belonging to the family.

caracterized by two elements. The first element is the set of all relevant contexts, which the set of all contexts models More precisely, let $(X_0, X_1, \ldots)$ be a symbolic sequence taking values on the finite alphabet $\mathcal{A}$ representing the set of the twenty amino acids. A PST model for this sequence PTprobabilistic tree basis of a PST model is the identification of the sub-sequences $X_{n-\ell}, \ldots, X_{n-1}$, called *contexts*, which are relevant to predict the next symbol X_n. The length of each context depends on $X_0, \ldots, X_{n-1}$, and the set of all contexts can be represented as a tree. In this tree, each context $c = c_{-k}, \ldots, c_{-1}$ is represented by a branch, whose sub-branch on top is determined by c_{-1}, the next subbranch is determined by c_{-2} and so on.

A *Sparse Probabilistic Tree* (SPT) is a PST in which some contexts are grouped together in an equivalence class. More precisely, the contexts of a SPT model are sequences of the form $A_{n-\ell}, \ldots, A_{n-1}$, where $A_i \subset \mathcal{A}$ for each i.

The S-PT algorithm works as follows. It starts with a tree consisting of a single root node. At each step, for every terminal node t with depth less than L and for every symbol x, the leaf x is added to t, if the sequence xt appears in the training sequences at least $N_{\min}$ times. For every pair of new leaves of a node, we test their *equivalence* using a log-likelihood ratio test. If the test accepts their equivalence, the leaves are merged together in a single leaf. The procedure is iterated with the new set of leaves. It stops when no more leaves can be merged.

To conclude the construction of the SPT we assigned to each leaf a transition probability estimated by the usual maximum likelihood procedure.

The *Fingerprint-PST* algorithm estimates the context tree and the transition probabilities in the same way as the PST. However, to classify a new sequence of amino acids, F-PST starts by identifying fingerprints defined as follows. Given a new sequence of amino acids, we look for the sub-sequence of length M with maximal probability, where M is the median of the already classified sequences in the protein family. If this maximum is bigger than a pre-defined threshold, the protein is identified as a member of the family.

3 Results and Discussion

We trained both S-PT and F-PST with a subset of families with more than 1000 sequences in the Pfam-A database, release 15.0 [7]. Then we applied the resulting models to classify all the sequences in the set Pfamseq. To establish the family membership threshold, we used the **equivalence number criterion** [8]. The quality of the model is measured by the number of true positives detected relative to the total number of proteins in the family.

Table 1 summarizes the classification rates obtained with our S-PT and F-PST together with the results produced by the PST implementation presented in [4]. It is clear that F-PST improves PST classification rates in a significant way. In the case of the S-PT algorithm, these are preliminary results as no attempt was made to optimize the choice of the parameters. However, it can be seen that in almost all families with high PST classification rate, the performance was improved with S-PT.

Figure 2 shows a context tree estimated with the S-PT algorithm for the AAA family. It is interesting to compare the equivalence classes in the tree with the classes obtained by grouping the amino acids by their physical and chemical properties, presented in Fig. 1. The coincidence of many classes obtained with the two procedures gives an indication of the S-PT ability to successfully retrieve biological information out of the amino acids sequences.

The preliminary results presented in this paper strongly suggest that these new algorithms can improve in a significant way the classification rates obtained with the implementation of the PST algorithm, presented in [4]. We are presently applying our algorithms to more families in the Pfam database to confirm this initial encouraging results.

References

1. Karp, R.M.: Mathematical challenges from genomics and molecular biology. Notices Amer. Math. Soc. **49** (2002) 544–553
2. Rissanen, J.: A universal data compression system. IEEE Trans. Inform. Theory **29** (1983) 656–664
3. Bejerano, G., Yona, G.: Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. Bioinformatics **17** (2001) 23–43
4. Bejerano, G.: Algorithms for variable length Markov chain modeling. Bioinformatics **20** (2004) 788–789
5. Eskin, E., Grundy, W.N., Singer, Y.: Protein family classification using sparse markov transducers. In: Proc. Int'l Conf. Intell. Syst. Mol. Biol. Volume 8. (2000) 134–145
6. Bourguignon, P.Y., Robelin, D.: Modèles de Markov parcimonieux: sélection de modèle et estimation. manuscript (2004)
7. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., Eddy, S.R.: The Pfam protein families database. Nucl. Acids Res. **32** (2004) D138–141

Table 1. Performance comparison between PST, F-PST and S-PT. PST and F-PST parameters: $L = 10$, $P_{\min} = 0.0001$, $\alpha = 0$, $\gamma_{\min} = 0.001$ and $r = 1.05$. S-PT parameters: $L = 10$, $N_{\min} = 2$, $\gamma_{\min} = 0.001$ and $r_{\max} = 3.841459$.

Family	Size	% True Pos. PST	% True Pos. SPST	% True Pos. F-SPST
7tm_1	515	93.0%	96.3%	97.7%
7tm_2	36	94.4%	97.2%	100.0%
7tm_3	12	83.3%	100.0%	100.0%
AAA	66	87.9%	90.9%	93.9%
ABC_tran	269	83.6%	85.9%	89.29%
actin	142	97.2%	97.2%	99.3%
adh_short	180	88.9%	89.4%	92.8%
adh_zinc	129	95.3%	91.5%	95.3%
aldedh	69	87.0%	89.9%	92.8%
alpha-amylase	114	87.7%	91.2%	94.7%
aminotran	63	88.9%	88.9%	90.5%
ank	83	88.0%	86.8%	86.6%
arf	43	90.7%	93.0%	93.0%
asp	72	83.3%	90.3%	91.7%
ATP-synt_A	79	92.4%	94.9%	97.5%

8. Pearson, W.R.: Comparison of methods for searching protein sequence databases. Protein Sci **4** (1995) 1145–1160
9. Bejerano, G.: Automata learning and stochastic modeling for biosequence analysis. PhD thesis, Hebrew University (2003)

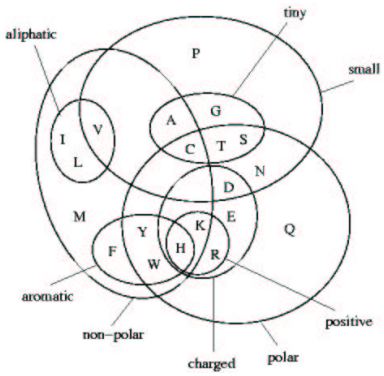

Fig. 1. A Venn diagram conveying several properties of the different amino acids (extracted from [9])

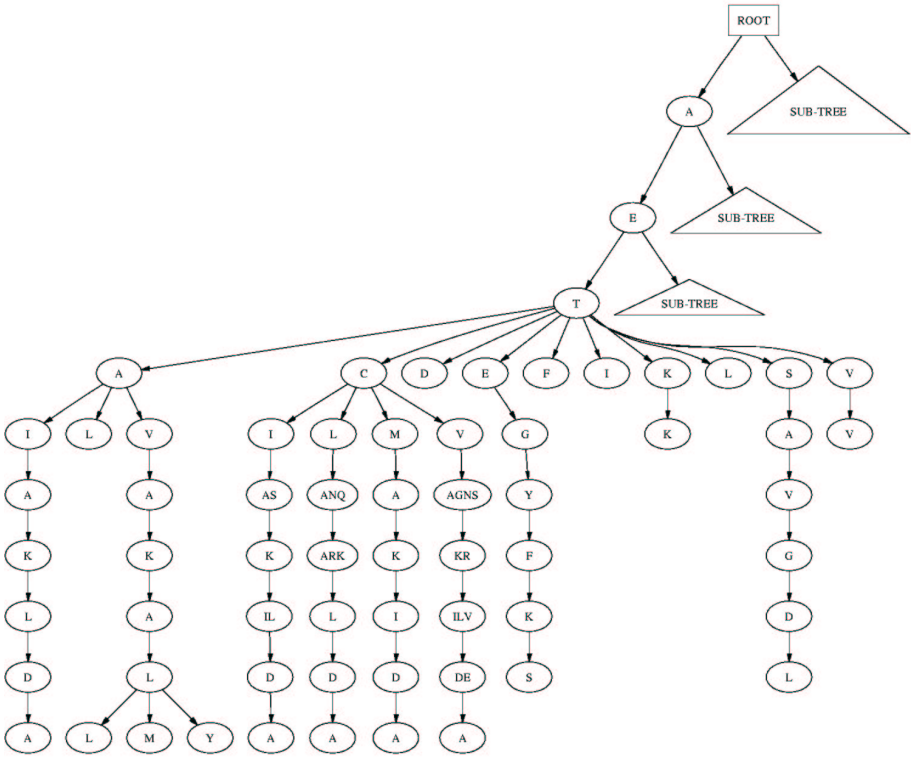

Fig. 2. The S-PT tree for the AAA family