BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 927–934

Markovian domain fingerprinting: statistical segmentation of protein sequences

Gill Bejerano [1,], Yevgeny Seldin [1], Hanah Margalit [2] and Naftali Tishby [1,*]*

[1]School of Computer Science & Engineering, The Hebrew University, Jerusalem 91904, Israel and [2]Department of Molecular Genetics & Biotechnology, Hadassah Medical School, The Hebrew University, POB 12272, Jerusalem 91120, Israel

Received on April 20, 2001; revised and accepted on July 9, 2001

ABSTRACT

Motivation: Characterization of a protein family by its distinct sequence domains is crucial for functional annotation and correct classification of newly discovered proteins. Conventional Multiple Sequence Alignment (MSA) based methods find difficulties when faced with heterogeneous groups of proteins. However, even many families of proteins that do share a common domain contain instances of several other domains, without any common underlying linear ordering. Ignoring this modularity may lead to poor or even false classification results. An automated method that can analyze a group of proteins into the sequence domains it contains is therefore highly desirable.

Results: We apply a novel method to the problem of protein domain detection. The method takes as input an unaligned group of protein sequences. It segments them and clusters the segments into groups sharing the same underlying statistics. A Variable Memory Markov (VMM) model is built using a Prediction Suffix Tree (PST) data structure for each group of segments. Refinement is achieved by letting the PSTs compete over the segments, and a deterministic annealing framework infers the number of underlying PST models while avoiding many inferior solutions. We show that regions of similar statistics correlate well with protein sequence domains, by matching a unique signature to each domain. This is done in a fully automated manner, and does not require or attempt an MSA. Several representative cases are analyzed. We identify a protein fusion event, refine an HMM superfamily classification into the underlying families the HMM cannot separate, and detect all 12 instances of a short domain in a group of 396 sequences.

Contact: jill@cs.huji.ac.il; tishby@cs.huji.ac.il

*To whom correspondence should be addressed.

1 INTRODUCTION

Numerous proteins exhibit a modular architecture, consisting of several sequence domains that often carry specific biological functions (reviewed in Bork, 1992; Bork and Koonin, 1996). For proteins whose structure has been solved, it can be shown in many cases that the characterized sequence domains are associated with autonomous structural domains (e.g. the C_2H_2 zinc finger domain). Characterization of a protein family by its distinct sequence domains (also termed 'modules') either directly or through the use of domain 'motifs', or 'signatures', is crucial for functional annotation and correct classification of newly discovered proteins. In many cases the underlying genes underwent shuffling events that have led to a change in the order of modules in related proteins. In other cases a certain module appears in many proteins, adjacent to different modules. A global alignment that ignores the modular organization of proteins may fail to associate a protein with other proteins that carry a similar functional module but in a different relative sequence location. Also, ignoring the modularity of proteins may lead to clustering of non-related proteins through false transitive associations[†]. Thus, ideally, clustering of proteins into distinct families should be based on characterization of a common sequence domain or a common signature and not on the entire sequence, allowing a single sequence to be clustered into several groups. For this, an unsupervised method for identification of the domains that compose a protein sequence is essential.

Many methods have been proposed for classification of proteins based on their sequence characteristics. Most of them are based on a seed Multiple Sequence Alignment (MSA) of proteins that are known to be related. The MSA

[†]For example, assume that proteins *A* and *B* have distinct single domains, and that protein *C* contains both domains. An algorithm may falsely deduce, since *A* and *B* are homologous to *C*, that *A* and *B* are homologous to each other.

can then be used to characterize the family in various ways: by defining characteristic motifs of the functional sites (as in Prosite, Hofmann *et al.*, 1999), by providing a fingerprint that may consist of several motifs (PRINTS-S, Attwood *et al.*, 2000), by describing a multiple alignment of a domain using a Hidden Markov Model (HMM) (Pfam, Bateman *et al.*, 2000), or by a position specific scoring matrix (BLOCKS, Henikoff *et al.*, 2000). All the above techniques, however, rely strongly on the initial selection of the related protein segments for the MSA, usually hand crafted by experts, and on the quality of the MSA itself. Besides being in general computationally intractable, when remote sequences are included in a group of related proteins, establishment of a good MSA ceases to be an easy task and delineation of the domain boundaries proves even harder. This becomes nearly impossible for heterogeneous groups where the shared motifs are not necessarily abundant, nor in linear ordering. It is highly desirable to complement these methods with efficient automatic generation of sequence signatures which can guide the classification and further analysis of the sequences. This need is especially emphasized in view of the large-scale sequencing projects, generating a vast amount of sequences that require annotation.

Unsupervised segmentation of sequences, on the other hand, has become a fundamental problem with many important applications such as analysis of texts, handwriting and speech, neural spike trains and indeed bio-molecular sequences. The most common statistical approach to this problem is currently the HMM. HMMs are predefined parametric models and their success crucially depends on the correct choice of the state model. In the common application of HMMs the architecture and topology of the model are predetermined and the memory is limited to first order. It is rather difficult to generalize these models to hierarchical structures with unknown *a-priori* state-topology (for an attempt see Fine *et al.*, 1998).

An interesting alternative to the HMM was proposed in Ron *et al.* (1996) in the form of a sub-class of *probabilistic finite automata*, the Variable Memory Markov (VMM) sources. While these models can be weaker as generative models, they have several important advantages: (i) they capture longer correlations and higher order statistics of the sequence; (ii) they can be learned in a provably optimal sense using a construction called *Prediction Suffix Tree* (PST); (Ron *et al.*, 1996; Buhlmann and Wyner, 1999); (iii) they can be learned very efficiently by linear time algorithms (Apostolico and Bejerano, 2000); (iv) their topology and complexity are determined by the data; and, specifically in our context (v) their ability to model protein families has been demonstrated (Bejerano and Yona, 2001).

In this work we apply a powerful extension of the VMM model and the PST algorithm, recently developed for stochastic mixtures of such models (Seldin *et al.*, 2001), that are learned in a hierarchical way using a Deterministic Annealing (DA) approach (Rose, 1998). Our model can in fact be viewed as an HMM with a VMM attached to each state, but the learning algorithm allows a completely adaptive structure and topology both for each state and for the whole model. The approach we take is information theoretic in nature. The goal is to enable a short description of the data by a (soft) mixture of VMM models, when the complexity of each model is controlled by the data via the *Minimum Description Length* (MDL) principle (see Barron *et al.*, 1998, for a review).

In effect we cluster regions of the input sequences into groups sharing coherent statistics. We grow a PST model for each group of segments, as complex as the group is statistically rich. We then refine this division by letting the PSTs compete over the segments. Embedding the competitive learning in a DA framework allows us to try and infer the correct number of underlying sources, and avoid many local minima. The output of our algorithm is a set of PST models, each of which has specialized in recognizing a certain protein region. The models can then be used to detect these regions in any protein.

In Seldin *et al.* (2001) we tested the algorithm on a mixture of interchanged running texts in five different European languages. The model was able to identify both the correct number of languages and the segmentation of the text sequence between the languages to within a few letters precision. Note that the segmentation there was not based on conserved regions (say, a few sentences, each repeating several times with minor variations), but rather based on the *conserved statistics* of running text segments in each language. In this paper we turn to observe statistical conservation in the context of protein sequences.

There are clear advantages to our approach compared to the common methods used for protein sequence segmentation. The method is automatic, there is no need for an alignment, the motifs themselves need not be few, abundant, or in linear ordering. When a signature is identified in a protein, its statistical significance can be quantitatively evaluated through the likelihood the model assigns to it. Given a group of related sequences the computational scheme we propose facilitates the segmentation of these sequences into domains through the use of the resulting statistical signatures, at times surpassing the susceptibility of single whole-domain HMMs. By characterizing protein families using these modular signatures it is possible to assign functional annotations to proteins that contain these modules, independent of their order in the protein. The detection of functional domains can then be used to define families and super-family hierarchies.

In Section 2 we outline the algorithm (a detailed description can be found in Seldin *et al.*, 2001). We then

turn in Section 3 to analyze promising results obtained for three exemplary diverse protein families (Pax, Type II DNA Topoisomerases and GST) and compare these with an alignment based approach. We conclude in Section 4 with a discussion and some directions for future work.

2 ALGORITHM OUTLINE

Several works precede the approach we follow in this paper. Learning a single VMM from a group of sequences using a PST model is defined in Ron *et al.* (1996). Strong theoretical results backing this approach when the underlying source exhibits Markovian-like properties are given in Ron *et al.* (1996) and Buhlmann and Wyner (1999). Equivalent algorithms of optimal linear time and space complexity for PST learning and prediction are proven in Apostolico and Bejerano (2000). In Bejerano and Yona (2001) partial groups of unaligned sequences from diverse protein families are each used as training sets. Resulting PSTs are shown to distinguish between previously unseen family members and unrelated proteins, in sensitivity matching that of an HMM trained on an MSA of the input sequences, while being much faster. Also noted there (see Figures 5 and 6 of Bejerano and Yona, 2001), when plotting the prediction along every residue of a protein sequence, is a correlation between protein domains and regions the family PST recognizes best *within* family members. That observation motivated the current work.

The algorithmic approach we take extends PST learning from single source modeling to several competing models, each specializing in regions of coherent statistics. Due to scope limitations we only outline it below.

2.1 Prediction suffix tree modeling

Consider a statistical model T assigning a probability to a protein sequence $\bar{x} = x_1 \ldots x_l, \forall x_j \in \Sigma$ the alphabet of amino acids. The higher $P_T(\bar{x})$ is the more we are confident that $\bar{x}$ belongs to the family of proteins T models. Treating $x_1, \ldots, x_l$ as a sequence of dependent random variables, PST modeling is built around the Markovian approximation $P_T(\bar{x}) = \prod_{j=1}^{l} P_T(x_j|x_1 \ldots x_{j-1}) \simeq \prod_{j=1}^{l} P_T(x_j|\,\mathrm{suf}_T(x_1 \ldots x_{j-1}))$ where the equality follows from applying the chain rule, and $\mathrm{suf}_T(x_1 \ldots x_{j-1})$ is the longest suffix of $x_1 \ldots x_{j-1}$ memorized by T during training. A PST T (Figure 1) is thus a data structure holding a set of short context specific probability vectors of the form $P_T(X_j|x_{j-d} \ldots x_{j-1})$. These short patterns of arbitrary lengths are collected from the training sequences *regardless* of relative sequence positions of the different instances of each pattern.

In Seldin *et al.* (2001) we define an MDL based variant of PST learning which is non-parametric and self-regularizing. It allows the PST to grow to complexity proportional to the statistical richness in the sequences it models. As an input it takes a collection of protein sequences $\{\bar{x}_1, \ldots, \bar{x}_n\}$, and a set of weights vectors $\{\bar{w}_1, \ldots, \bar{w}_n\}$, where the jth entry of $\bar{w}_i$, denoted $0 \leqslant w_{ij} \leqslant 1$, measures the degree of relatedness we currently assign between the jth element of x_i, x_{ij}, and the model we wish to train. For example, in order to train a PST only on specific regions in the proteins assign $w_{ij} = 1$ to those regions and $w_{ij} = 0$ elsewhere.

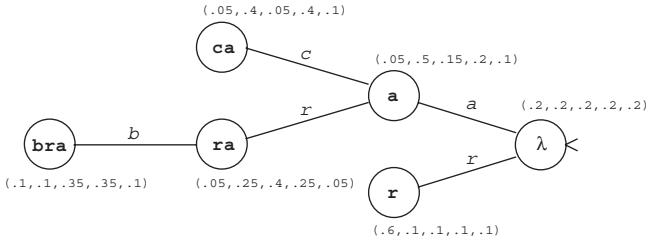

Fig. 1. An example of a PST over the alphabet $\Sigma = \{a, b, c, d, r\}$. The string inside each node is a memorized suffix and the adjacent vector is its probability distribution over the next symbol. (e.g. since $\mathrm{suf}_T(bacara) = ra$, the probabilities $P_T(?|bacara) = \{0.05, 0.25, 0.4, 0.25, 0.05\}$ for $\{a, b, c, d, r\}$ respectively).

2.2 Protein sequence segmentation

The relatedness between a PST model and a sequence segment is defined as the probability the model assigns to the segment (how well it predicts it). In order to partition the sequence between $k = 1, \ldots, m$ known PST models, we assign sequence segments from $\{\bar{x}_i\}$ to models in proportion to the relatedness between a segment and each of the competing models. The nm resulting vectors $\{\bar{w}_i^k\}_{i,k}$ constitute a soft partitioning of $\{\bar{x}_i\}$ between the models $(\forall i, j : \Sigma_k\, w_{ij}^k = 1)$. We may then retrain each model k with its new set of weights $\{\bar{w}_i^k\}_i$. This *soft clustering* (data repartitioning followed by model retraining) can be iterated until convergence to a set of PSTs, each modeling a distinct group of sequence segments[‡].

Clearly the quality of the solution we converge to depends on the number of models, and their initial settings. Both issues are solved by *iterative refinement*. We begin with a single model T_0 trained over the whole set $\{\bar{x}_i\}$ (with $\forall i, j : w_{ij}^0 = 1$). We then split T_0 into two identical replicas T_1, T_2 and randomly perturb both, to differ slightly. We then iterate repartitioning and training, splitting again when the models converge. Models that lose all grip on the data $(\Sigma_{i,j}\, w_{ij}^k = 0)$ are eliminated.

Finally, a resolution parameter $\beta > 0$ is introduced and is gradually increased from a low initial value. The

[‡] A similar iterative loop is used in soft clustering of points in R^n to k Gaussians.

Input: the set of unaligned sequences $\{\bar{x}_i\}$

1. Set $\forall i, j: w_{ij}^0 = 1$ and train PST T_0
2. Set: $\beta = \beta_0$, $m_{prev} = 0$, $\mathcal{T} = \{T_0\}$
3. Repeat until $\beta = \beta_{fin}$:
   (a) While $|\mathcal{T}| > m_{prev}$
      i. Set: $m_{prev} = |\mathcal{T}|$
      ii. Split in two and perturb all PSTs in $\mathcal{T}$
      iii. Repeat until convergence:
         A. Repartition the data into $\{\bar{w}_i^k\}_{i,k=1...|\mathcal{T}|}$ according to the PSTs predictions
         B. Retrain a new set of PSTs $\mathcal{T}$ using the new partitioning
      iv. Remove all empty models from $\mathcal{T}$
   (b) Increase β

Output: the final set of PST models $\mathcal{T}$

Fig. 2. The segmentation algorithm.

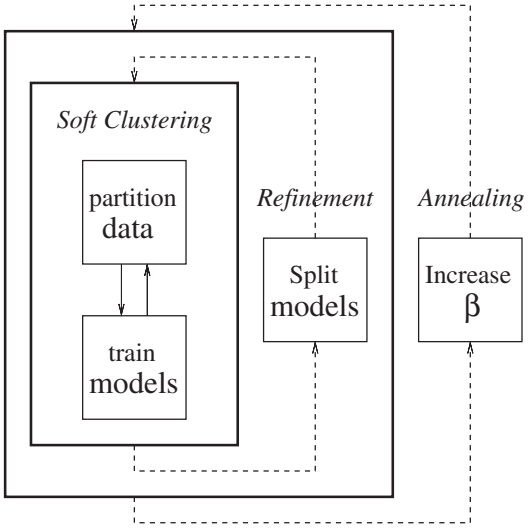

Fig. 3. Schematic description of the algorithm.

parameter β controls the hardness of the soft partition of sequence segments between the models. As β increases, segments separate more and more into distinct models.

Formally we set

$$w_{ij}^k = \frac{P(T_k)\, e^{\beta S_{T_k}(x_{ij})}}{\sum_{\alpha=1}^{m} P(T_\alpha)\, e^{\beta S_{T_\alpha}(x_{ij})}}$$

where $S_{T_k}(x_{ij}) \leqslant 0$ is a log-likelihood measure of relatedness between model k and symbol x_{ij}, and $P(T_k)$ corresponds to the relative amount of data assigned to model k in the previous segmentation. As β increases it induces a sharper distinction between the highest scoring $S_{T_k}(x_{ij})$ and the other models, for each x_{ij}. This *DA* procedure can avoid many local minima and generally yields better solutions than some other optimization algorithms (see Rose, 1998). A high-level pseudocode and a schematic description of the algorithm are given in Figures 2 and 3 respectively. See Seldin *et al.* (2001) for more details.

3 RESULTS

Recall from Figure 2 that the input to the segmentation algorithm is a group of unaligned sequences in which we search for regions of one or more types of conserved statistics. The different training sets were constructed using the Pfam (release 5.4) and Swissprot (release 38, Bairoch and Apweiler, 2000) databases. Various sequence domain families were collected from Pfam. In each Pfam family all members share a domain. An HMM detector is built for that domain based on an MSA of a seed subset of the family domain regions. The HMM is then verified to detect that domain in the remaining family members. Multi-domain proteins therefore belong to as many Pfam families as there are different characterized domains within them. In order to build realistic, more heterogeneous sets, we collected from Swissprot the *complete sequences* of all chosen Pfam families. Each set now contains a certain domain in all its members, and possibly various other domains appearing anywhere within some members.

Given such a set of unaligned sequences our algorithm returns as output several PST models (Figure 2). The number of models returned is determined by the algorithm itself. Each such PST has 'survived' repeated competitions by outperforming the other PSTs on some sequence regions. In practice two types of PSTs emerge for protein sequence data: models that significantly outperform others on relatively short regions (and generally perform poorly on most other regions)—these we call detectors. And models that perform averagely over all sequence regions—these are 'noise' (baseline) models and we can discard them automatically. We now turn to analyze the detectors—in which sequences do they outperform all other models and what is the correlation between detected regions and protein domains?

Several interesting results can come out of the analysis: First and foremost, a signature for the common domain or domains. Signatures for other domains that appear only in some proteins, may also appear. A signature may exactly cover a domain, revealing its boundaries. And when the Pfam HMM detector cannot model below the superfamily level, we may try to outperform it and subdivide into the underlying biological families.

Three of the Pfam-based sets we ran experiments on have been chosen to demonstrate examples covering all

the above cases. The three, very different, domain families are the Pax proteins, the type II DNA Topoisomerases and the glutathione S-transferases. At the end of the section we also compare our results to an MSA-based approach.

Ten independent runs of the (stochastic) segmentation algorithm, implemented in C++, were carried out per family. On a Pentium III 600 MHz Linux machine clear segmentation was usually apparent within an hour or two of run time.

Recall that each PST detector we examine is run over *all* complete sequences in the set it was grown on in order to determine its nature. In our experiments the signature left by each PST was the same between different runs, and between different proteins sharing the same domain(s). We therefore present only the output of all detector PSTs on representative sequences in a particular run.

3.1 The Pax family

Pax proteins (reviewed in Stuart *et al.*, 1994) are eukaryotic transcriptional regulators that play critical roles in mammalian development and in oncogenesis. All of them contain a conserved domain of 128 amino acids called the paired or paired box domain (named after the *Drosophila* paired gene which is a member of the family). Some contain an additional homeobox domain that succeeds the paired domain. Pfam nomenclature names the paired domain 'PAX'.

The Pax proteins show a high degree of sequence conservation. One hundred and sixteen family members were used as a training set for the segmentation algorithm, as described above. In Figure 4 we superimpose the prediction of all resulting PST detectors over one representative family member. This Pax6 SS protein contains both the paired and homeobox domains. Both have matching signatures. This also serves as an example where the signatures exactly overlap the domains. The graph of family members not having the homeobox domain contains only the paired domain signature. Note that only about half the proteins contain the homeobox domain and yet its signature is very clear.

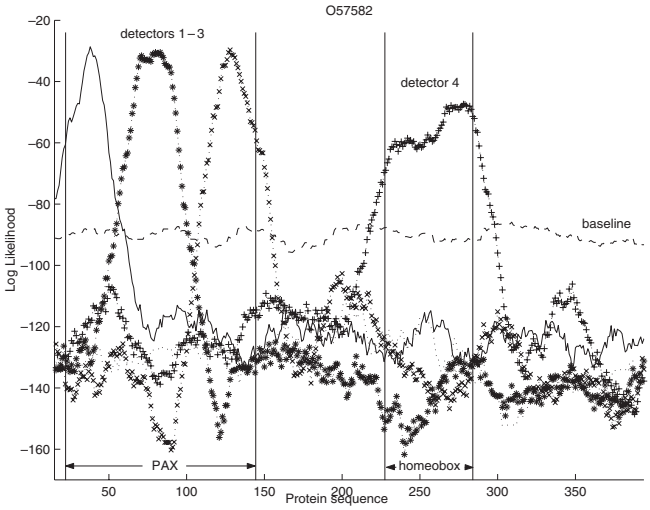

Fig. 4. Paired/PAX + homeobox signatures. We superimpose the log likelihood predictions $\log P_T(\bar{x})$ (Section 2.1) of all four detector PSTs generated by the segmentation algorithm, and an exemplary baseline model (dashed), against the sequence of the PAX6 SS protein. The title holds the protein accession number. At the bottom we denote in Pfam nomenclature the location of the two experimentally verified domains. These are in near perfect match here with the high scoring sequence segments.

3.2 DNA topoisomerase II

Type II DNA topoisomerases are essential and highly conserved in all living organisms (see Roca, 1995, for a review). They catalyze the interconversion of topological isomers of DNA and are involved in a number of mechanisms, such as supercoiling and relaxation, knotting and unknotting, and catenation and decatenation. In prokaryotes the enzyme is represented by the *Escherichia coli* gyrase, which is encoded by two genes, gyrase A and gyrase B. The enzyme is a tetramer composed of two gyrA and two gyrB polypeptide chains. In eukaryotes the enzyme acts as a dimer, where in each monomer two distinct domains are observed. The N-terminal domain is similar in sequence to gyrase B and the C-terminal domain is similar in sequence to gyraseA (Figure 8). In Pfam 5.4 terminology gyrB and the N-terminal domain belong to the 'DNA_topoisoII' family[§], while gyrA and the C-terminal domain belong to the 'DNA_topoisoIV' family[¶]. Here we term the pairs gyrB/topoII and gyrA/topoIV.

For the analysis we used a group of 164 sequences that included both eukaryotic topoisomerase II sequences and bacterial gyrase A and B sequences (gathered from the union of the DNA_topoisoII and DNA_topoisoIV Pfam 5.4 families). We successfully differentiate them into sub-classes. Figure 5 describes a representative of the eukaryotic topoisomerase II sequences and shows the signatures for both domains, gyrB/topoII and gyrA/topoIV. Figures 6 and 7 demonstrate the results for representatives of the bacterial gyrase B and gyrase A proteins, respectively. The *same* two signatures are found in all three sequences, at the appropriate locations. Interestingly, in Figure 6 in addition to the signature of the gyrB/topoII domain *another* signature appears at the C-terminal region of the sequence. This signature is compatible with a known conserved region at the C-terminus of gyrase B[‖],

[§] Apparently this family has been sub-divided in Pfam 6 releases.
[¶] The name should not be confused with the special type of topoisomerase II found in bacteria, that is also termed topoisomerase IV, and plays a role in chromosome segregation.
[‖] Corresponding to the Pfam 'DNA_gyraseB_C' family.

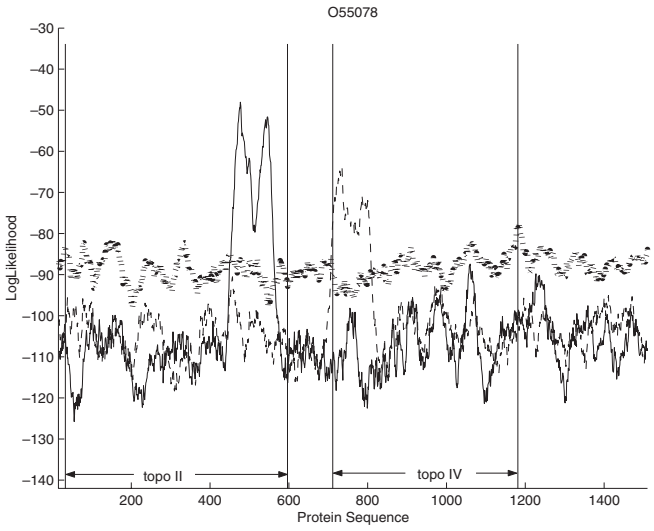

Fig. 5. Eukaryotic topoisomerase II signature. The legends in Figures 5–7 are equivalent to that of Figure 4, plotting the predictions of all detectors and a single baseline model.

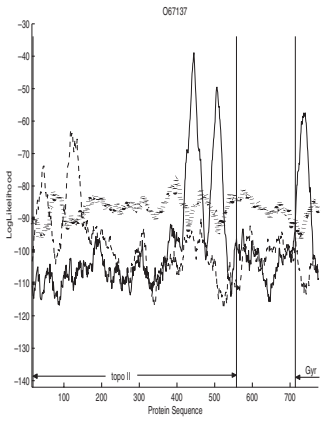

Fig. 6. Bacterial GyrB/topoII signature.

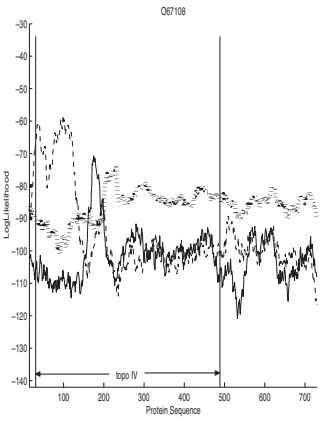

Fig. 7. Bacterial GyrA/topoIV signature.

that is involved in the interaction with the gyrase A molecule.

The relationship between the *E.coli* proteins gyrA and gyrB and the yeast topoisomerase II (Figure 8) provides a prototypical example of a fusion event of two proteins that form a complex in one organism into one protein that carries a similar function in another organism. Such examples have led to the idea that identification of such similarities may suggest the relationship between the first two proteins, either by physical interaction or by their involvement in a common pathway (Marcotte *et al.*, 1999; Enright *et al.*, 1999). The computational scheme we present can be useful in search for these relationships.

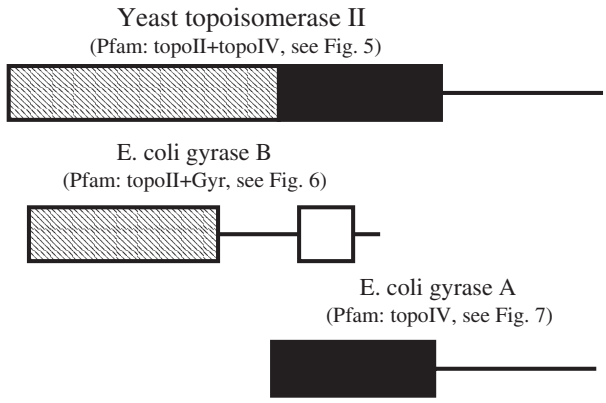

Fig. 8. Fusion event Illustration. Adapted from Marcotte *et al.* (1999). The Pfam domain names are added in brackets, together with a reference to our results on a representative homolog. Compare the PST signatures in Figures 5–7 with the schematic drawing above. It is clear that the eukaryotic signature is indeed composed of the two prokaryotic ones, in the correct order, omitting the C-terminus signature of gyrase B (short termed here as 'Gyr').

3.3 The glutathione S-transferases

The Glutathione S-Transferases (GST) represent a major group of detoxification enzymes (reviewed in Hayes and Pulford, 1995). There is evidence that the level of expression of GST is a crucial factor in determining the sensitivity of cells to a broad spectrum of toxic chemicals. All eukaryotic species possess multiple cytosolic GST isoenzymes, each of which displays distinct binding properties. A large number of cytosolic GST isoenzymes have been purified from rat and human organs and, on the basis of their sequences they have been clustered into five separate classes designated class alpha, mu, pi, sigma, and theta GST. The hypothesis that these classes represent separate families of GST is supported by the distinct structure of their genes and their chromosomal location. The class terminology is deliberately global, attempting to include as many GSTs as possible. However, it is possible that there are sub-classes that are specific to a given organism or a group of organisms. In those sub-classes the proteins may share more than 90% sequence identity, but these relationships are masked by their inclusion in the more 'global' class. Also, the classification of a GST protein with weak similarity to one of these classes is sometimes a difficult task. In particular the definition of the sigma and theta classes is imprecise. Indeed in the PRINTS database only the three classes, alpha, pi, and mu have been defined by distinct sequence signatures, while in Pfam all GSTs are clustered together, for lack of sequence dissimilarity.

Three hundred and ninety six Pfam family members were segmented jointly by our algorithm, and the results

were compared to those of PRINTS (as Pfam classifies all as GSTs). Five distinct signatures were found (not shown due to space limitations): (1) A typical weak signature common to many GST proteins that contain no sub-class annotation. (2) A sharp peak after the end of the GST domain appearing exactly in all 12 out of 396 (3%) proteins where the Elongation Factor 1 Gamma (EF1G) domain succeeds the GST domain. (3) A clear signature common to almost all PRINTS annotated alpha and most pi GSTs. The last two signatures require more knowledge of the GST superfamily. (4) The theta and sigma classes are abundant in nonvertebrates. As more and more of these proteins are identified it is expected that additional classes will be defined. The first evidence for a separate sigma class was obtained by sequence alignments of S-crystallins from mollusc lens. Although these refractory proteins in the lens probably do not have a catalytic activity they show a degree of sequence similarity to the GSTs that justifies their inclusion in this family and their classification as a separate class of sigma (Buetler and Eaton, 1992). This class, defined in PRINTS as S-crystallin, was almost entirely identified by the fourth distinct signature. (5) Interestingly, the last distinct signature, is composed of two detector models, one from each of the previous two signatures (alpha + pi and S-crystallin). Most of these two dozens proteins come from insects, and of these most are annotated to belong to the theta class. Note that many of the GSTs in insects are known to be only very distantly related to the five mammalian classes. This putative theta sub-class, the previous signatures and the undetected PRINTS mu sub-class are all currently further investigated.

3.4 Comparative results

In order to evaluate our findings we have performed three unsupervised alignment driven experiments using the same sets described above: an MSA was computed for each set using Clustal X (Linux version 1.81, Jeanmougin *et al.*, 1998). We let Clustal X compare the level of conservation between individual sequences and the computed MSA profile in each set. Qualitatively these graphs resemble ours, apart from the fact that they do not offer separation into distinct models.

As expected this straightforward approach yields less. We briefly recount some results (showing but one graph due to space limitations): the Pax alignment did not clearly elucidate the homeobox domain existing in about half the sequences. As a result, when we plot the graph comparing the same PAX6 SS protein we used in Figure 4 against the new MSA in Figure 9, the homeobox signal is lost in the noise. For type II topoisomerases the picture is slightly better. The Gyrase B C-terminus unit from Figure 6 can be discerned from the main unit, but with a much lower peak. However, the clear sum of two signatures we obtained for

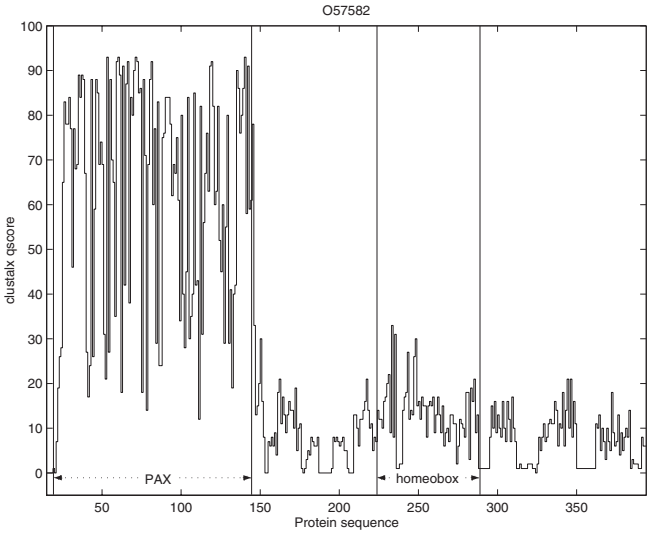

Fig. 9. Pax MSA profile conservation. We plot the Clustal X conservation score of the PAX6 SS protein against an MSA of all Pax proteins. While the predominant paired/PAX domain is discerned, the homeobox domain (appearing in about half the sequences) is lost in the background noise. Compare with Figure 4 where the same training set and plotted sequence are used.

the eukaryotic sequences (Figure 5) is lost here. In the last and hardest case the MSA approach tells us nothing. All GST domain graphs look nearly identical precluding any possible subdivision. And the 12 (out of 396) instances of the EF1G domain are completely lost at the alignment phase.

4 DISCUSSION

In this paper we have described a novel algorithm for detecting regions of conserved statistics within a group of sequences. We employed competitive learning to model the data using a mixture of PST models, governed by MDL considerations. Model refinement was achieved through a DA framework. We then demonstrated the capabilities of the algorithm in the proteins realm, by analyzing its output on three diverse protein groups.

We briefly recount the advantages of the proposed method: It is fully automated; it does not require or attempt an MSA of the input sequences; it handles heterogeneous groups well and locates domains appearing only a few times in the data; by nature it is not confused by different module orderings within the input sequences; it appears to seldom generate false positives; and it is shown to surpass HMM clustering in at least one hard instance.

Obviously no tool is without limitations. Highly similar statistical sources we wish to separate, domains we wish to detect that hardly appear in the data or very short domains may prove hard to segment. Indeed a few

sequences in nearly all groups presented above went unnoticed, as well as a GST family characterized in PRINTS. It seems that our segmentation algorithm can best be used in conjuncture with current alignment-based methods. The segmentation can first be applied to separate heterogeneous groups of proteins into groups sharing similarities. Those groups can then be profiled by HMMs or similar tools, using our signatures as guides to the alignment and domain boundaries.

In our opinion this tool may suggest a new perspective on protein sequence organization at large. Statistical conservation is *unlike* conventional sequence conservation. Regions may be statistically identical while completely dissimilar from an alignment point of view (running text in natural language is a good example, as we have demonstrated in Seldin *et al.*, 2001). We hope that this new, much more flexible notion of sequence conservation will eventually help better understand the constraints shaping the world of known proteins.

With this in mind, we are currently examining our results, striving to improve our understanding of the limitations inherent in segmenting protein data. Intriguing further applications of this new tool in the context of proteins include trying to refine existing classifications, looking for fusion events and composing a comprehensive library of detectors. Applications to other bio-sequences, DNA in particular, are also forthcoming.

ACKNOWLEDGEMENTS

The authors wish to thank Yael Altuvia for critically reading an early manuscript, and Arne Elofsson for sharing data. G.B. is supported by a grant from the Ministry of Science, Israel.

REFERENCES

Apostolico,A. and Bejerano,G. (2000) Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space. *J. Comput. Biol.*, **7**, 381–393.

Attwood,T., Croning,M., Flower,D., Lewis,A., Mabey,J., Scordis,P., Selley,J. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Barron,A., Rissanen,J. and Yu,B. (1998) The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theor.*, **44**, 2743–2760.

Bateman,A., Birney,E., Durbin,R., Eddy,S., Howe,K. and Sonnhammer,E. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Bejerano,G. and Yona,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **17**, 23–43.

Bork,P. (1992) Mobile modules and motifs. *Curr. Opin. Struct. Biol.*, **2**, 413–421.

Bork,P. and Koonin,E. (1996) Protein sequence motifs. *Curr. Opin. Struct. Biol.*, **6**, 366–376.

Buetler,T. and Eaton,D. (1992) Glutathione S-transferases: amino acid sequence comparison, classification and phylogentic relationship. *Environ. Carcinogen. Ecotoxicol. Rev.*, C **10**, 181–203.

Buhlmann,P. and Wyner,A. (1999) Variable length Markov chains. *Ann. Stat.*, **27**, 480–513.

Enright,A., Iliopoulos,I., Kyrpides,N. and Ouzounis,C. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Fine,S., Singer,Y. and Tishby,N. (1998) The hierarchical hidden Markov model: analysis and applications. *Mach. Learn.*, **32**, 41–62.

Hayes,J. and Pulford,D. (1995) The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 445–600.

Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the Blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Jeanmougin,F., Thompson,J., Gouy,M., Higgins,D. and Gibson,T. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.*, **23**, 403–405.

Marcotte,E., Pellegrini,M., Ng,H., Rice,D., Yeates,T. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Roca,J. (1995) The mechanisms of DNA topoisomerases. *Trends Biol. Chem.*, **20**, 156–160.

Ron,D., Singer,Y. and Tishby,N. (1996) The power of amnesia: learning probabilistic automata with variable memory length. *Mach. Learn.*, **25**, 117–149.

Rose,K. (1998) Deterministic annealing for clustering, compression, classification, regression and related optimization problems. *IEEE Trans. Inf. Theor.*, **80**, 2210–2239.

Seldin,Y., Bejerano,G. and Tishby,N. (2001) Unsupervised sequence segmentation by a mixture of switching variable memory Markov sources. *Proc. 18th Intl. Conf. Mach. Learn. (ICML)*. Morgan Kaufmann, San Francisco, CA, pp. 513–520.

Stuart,E.T., Kioussi,C. and Gruss,P. (1994) Mammalian Pax genes. *Annu. Rev. Genet.*, **28**, 219–236.