A new type of Hidden Markov Models to predict complex domain architecture in protein sequences

Raluca Uricaru, Laurent Brehelin, Eric Rivals

A new type of Hidden Markov Models to predict complex domain architecture in protein sequences

Abstract: *Profile Hidden Markov Models (pHMMs) represent sequence regions, called domains or motifs, that are conserved among the proteins of a family. They are routinely used either i/ to recognize the presence of a domain in a protein and thereby to test its membership of a known family, or ii/ to tag the precise position of a domain in the sequence. However, a majority of proteins are composed of several domains, and during evolution, events such as rearrangements or duplications may create different domain architectures in proteins of the same family. Due to their intrinsic linear structure, pHMMs cannot model several distinct domains whose number and relative order may be variable in a family. We lack efficient tools to perform recognition and tagging in the case of complex domain architectures. Here, we propose a generalized HMM to solve exactly this. In our solution, called cyclic profile HMM (cpHMM), specific transitions can model the repetition of units, as well as different relative orders of domains. In a cpHMM, complete domains are modeled by nested pHMMs. We provide a program for the construction of cpHMM that takes as input pHMMs, thereby allowing the user to capitalized on already developed pHMMs (PFAM). We adapted recognition and tagging algorithms to cpHMMs and test them on both the family of PentatricoPeptide Repeats proteins (PPR) and on the superfamily of saposins. Our results demonstrate that cpHMMs improve on pHMMs for the recognition and tagging of proteins with complex domains architectures, while keeping their efficiency. The architecture of PPR proteins has been manually annotated for a subfamilly in arabidopsis, however only the recognition with the PFAM PPR motif has been previously performed for the rice and poplar tree. Comparing our results with the annotations of arabidopsis PPR, we show that more than* 88% *of the motifs are precisely recognized by the cpHMM. Moreover, we completed the recognition of PPR, as well as the determination of their architecture, for both rice and poplar tree proteomes.*

Keywords: Motif, domain, profile HMM, domain architecture, tagging, recognition, cyclic permutation, duplication.

1 Introduction

Profile Hidden Markov Models (pHMMs) [8] are probabilistic models specially designed for the modelling of protein and domain families. They are widely used to align new protein sequences on the already known proteins of a given family, or to recognize new members of a protein family. For example, PFAM [3], the well-known base of protein families, makes intensive use of pHMMs (via the HMMER [6] software) for building, updating, and searching the database. pHMMs are probabilistic automata and as such depicted as a graph in where nodes represent the state and arrows transitions between states. The core of a profile HMM (see the insert in Fig. 1) is a linear sequence of match (M) states, one for each conserved position (consensus column) of a multiple alignment. Each M state emits (aligns to) a single residue, with a probability that is determined by the frequency of observed residues in the corresponding alignment column. In addition, states D and I model the delete and

insert gaps of the alignment. Two dummy states, B and E, represent the beginning and end of the sequence. Given a pHMM and a protein, one can compute the probability this protein being generated by the pHMM using the *Viterbi* algorithm [12]. The *recognition* problem, *i.e.*, inferring to which family a new protein belongs to, is solved by computing its probability for all pHMMs, representing all potential families, and to classify the protein in the family whose pHMM yields the highest probability (if the latter is above a given threshold).

A profile HMM can adequately model a sequence region conserved among the proteins of a family. However, the majority of proteins are multi-domains, *i.e.*, composed of several domains. It is known that the relative order of these units may be altered by domain swaping, circular permutations (*e.g.*, in DNA methyltransferases, lectins, or saposins), or other rearrangements [1,4,14]. Moreover, the number of domains may vary among related proteins through domain duplications [4]. Rearrangements and duplications are illustrated in the Supplementary Material. Profile HMMs are inapropriate to model such cases of variable domain architecture (*i.e.*, organisation), since they have a linear, cycle-free structure. Typical examples of families with complex multi-domain architectures are that of *PentatricoPeptide Repeat* proteins [9], and the Saposin superfamily [11].

Here we propose a generalization of pHMMs, termed *cyclic profile HMMs*, that enables the modelling of proteins families with both variable number of repeated units (or groups of units) and variable relative order of these units. While a similar idea of domain context was used to improve recognition compared to pHMMs [5], this generalization has never been applied to model complex domain architectures. We develop adequate algorithms for cpHMMs (Section 2), test them on the two above mentioned families (PPR and Saposins, see Section 3), and demonstrate their practical efficiency for both the recognition and tagging[1] tasks (Section 4). Compared to existing ad-hoc solutions, this work emphasizes the ability of cpHMMs to perform recognition and tagging of multi-domains architecture by optimizing a global criterion.

2 Cyclic profile HMMs

Cyclic profile HMMs are probabilistic models made up of several profile HMMs (one for each different unit) linked by transitions and some additional emitting and non-emitting states. Fig. 1 shows an example of cpHMM that can be used to model the PPR family. Each *2-circle state* is not a simple state, but a complete nested profile HMM (their names correspond to those of PPR motifs). These pHMMs adopt the same structure as those built with HMMER [6] (see the insert). In the cpHMM, *diamond states* correspond to emitting states (similar to those in pHMMs). They are associated with an amino acid probability distribution and are used to model the regions located between motifs or domains (regions not modeled by pHMMs). Circle states are non-emitting states, which are used to group transitions common to several states and to avoid an explosion of the number of transitions. As for pHMMs, two non-emitting dummy states, denoted by S and F in Fig. 1, serve to begin and end a sequence.

The structure of cpHMM can incorporate prior knowledge about the domain architecture of the family. In the example of Fig. 1, the *cyclic part* allows the repetition of 3 PPR motifs in any order, while the *linear part* reflects the fact that in N-terminal part of PPR proteins some motifs occur in a fixed linear order or are missing. Generally, any regular expression[2] that describes a domain architecture can be translated in a cpHMM in a straightforward way. Note that a cpHMM can also be built without prior knowledge about the architecture: it suffices to incorporate all domains/motifs

[1] We term "Tagging" the task of finding the nature and positions of each unit in the protein sequence.
[2] Regular expression are PROSITE like patterns where the symbols represent the domains or motifs

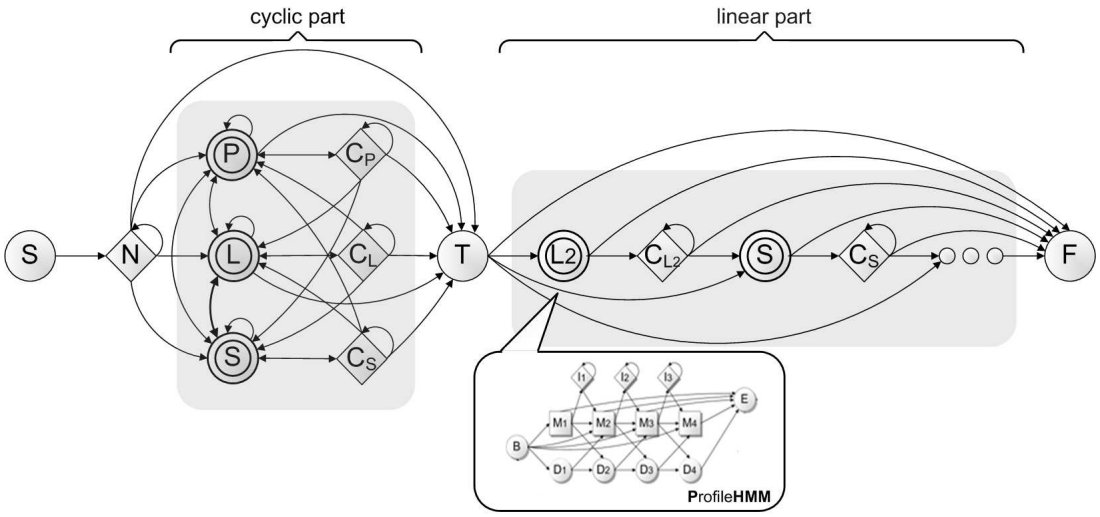

Figure 1. A cyclic profile HMM (cpHMM) modelling the PPR motif architecture. The insert represents a profile HMMs as built by HMMER and used in PFAM. In the cpHMM structure, double circle represent nested profile HMMs, while simple circle are normal states; arrows represent transitions between states. In this cpHMM, a cyclic part allowing any relative order between motifs, precedes a linear part in which motif must appear in fixed order. In general, the linear part may be empty.

in the cyclic part and none in the linear part. Concerning the amino acid distribution associated with diamond states, any prior knowledge about the amino acid composition of the inter units sequence can be incorporated, like amino-acid distribution of the organism or of a reference protein database.

2.1 Recognition and tagging

cpHMMs generalize profile HMMs, but are still Hidden Markov Models. Thus, the algorithmic toolbox developed for the general class of HMMs is available for cpHMMs. Recognition and tagging are performed with Forward and Viterbi algorithms mentioned above [12], which we summarize now. Let us consider an amino acid sequence $O = o_1 \dots o_T$ and H a cpHMM.

The recognition problem—does the protein belong to the family modeled by H?—requires to compute the probability, $P(O|H)$, that O is generated by H. If the probability lies above a given threshold then the protein is considered to belong to the family. As a sequence O may be generated following several paths between states S and F in H, to obtain $P(O|H)$ one needs to compute and sum up the probability of O for all these generating paths in H. The exponential number of potential paths makes a naive exploration impractical. Fortunately, this computation can be done in polynomial time by a dynamic programming algorithm known as the *forward* algorithm [12].

Concerning the tagging problem, the classical approach in HMMs involves to find out the state sequence (from S to F), called the *Viterbi path*, that has the highest probability to generate O. This path is then used to tag the sequence in a straightforward way: every portion of sequence generated in the Viterbi path by one of the pHMM, say H', that composes H is considered to be a motif/domain associated with H'. As for the recognition problem, the computation of the Viterbi path can be done in polynomial time by a dynamic programming algorithm called the *Viterbi algorithm* [12].

The number of states of the model asks for a lot of attention when implementing the forward and Viterbi algorithm. Taking advantage from the free cycle structure of the profile HMMs, HMMER [6] manages to obtain a time complexity that is linear in the number of states N, *i.e.*, $O(NT)$. For cpHMMs, the cycles introduce a quadratic term in the complexity, which does not depend on the total number of states in H the cpHMM, but solely on the number of nested pHMMs. Actually, if N_B denotes the maximal branching factor of the cpHMM structure (in Fig. 1 $N_B = 4$), then the complexity of the algorithms is $O(NT + N_B^2 T)$. Note that N_B is usually small. Hence, in practice the time complexity can still be considered as linear in N and leads to low computing times. In the experiments of Section 3, applying the cpHMM model for PPR proteins to the set of $40,000$ *arabidopsis* proteins for both recognition and tagging takes less than 20 minutes on a classical desktop computer.

3 Results

We perform analysis on a large plant specific protein family, the Pentatrico-Peptide Repeat protein family (PPR) [9], and on a more evolutionary divergent Saposin superfamily. In both case most proteins contains several domains that can be repeated and whose order may be altered.

3.1 Analysis of the Pentatrico-Peptide Repeat protein family

The Pentatrico-Peptide Repeats (PPR) family of proteins comprises 466 proteins in *Arabidopsis thaliana*, making it one of the largest plant specific family. PPR are involved in specific RNA editing in plant organels as well as cytoplasmic male fertility restoration [7,9]. PPR proteins contain tandem repeats of PPR motifs (named P, L, S), as well as other non PPR motifs (E, E+, Dyw). Half of the PPR proteins form the PPRP subfamily, while the other is called the PCMP subfamily [2]. PPR architecture can be described using regular expressions (where letters represent motifs): (P*-S*)* for PPRP subfamily and (P-L-S*)*-[E-[E+ -[Dyw]]] for PCMP [9] (see illustration of Fig. 5 in Supplementary Material).

	identical	improved prediction	slightly different	different
#proteins	98	30	24	45

Table 1. Comparison of the cpHMM motif annotation with the expert manual annotation of Bruyère and Lecharny (IBP, Paris Orsay) on the PCMP subfamily of the PPR proteins.

PPR recognition and tagging in other species. A member of the PPR family is characterized by the occurrence of at least one PPR motif (PFAM PF01535). A fine classification of PPR motif led to the definition of mainly 3 subtypes of motifs denoted P, L, or S for which pHMMs have been constructed from arabidopsis sequences [9]. As PPR motifs are exclusive of the family members, the PFAM domain as well as the P, L, or S pHMMs are able to recognize all arabidopsis PPR proteins. However, precise annotation of the domain architecture remains tricky because of the inherent resemblance of the 3 types of motifs (all of them come from an ancestral motif), of possible overlap in annotation, and of the divergence at the amino-acid level. We compared the automatic annotation derived from our cpHMM to the manual expertise (Table 1). For each protein, we aligned the two domain architectures with a global alignment algorithm and classify them into 4 classes according to the differences

observed. In Class 1, annotations are identical; in Class 2 also except that some motifs considered as being in the twilight zone in the manual annotation and marked as unsure, where predicted by the cpHMM. In Class 3, the annotations differ in only one motif, while in class 4 several positions may be different. Only 45 proteins of the family have more than one motif difference between the annotations. In this class, these represent only 11% of the motifs (ie, less than 2 motifs per protein over an average length of 15 motifs). Moreover, the differences in classes 3 and 4 are majoritarily located at the $N-$ and $C-$ termini of the proteins. Overall, the automatic annotation by the cpHMM is valid in more than 88% of the motifs. The automatic annotation of the 197 proteins is carried out in less than a few minutes with a reasonable precision, while the manual expertise required months.

In the light of these results, we used the cpHMM to detect PPR in the rice and poplar tree proteomes. The number of PPR detected and their distribution between PPRP, PCMP, and in classes of PCMP are given in Supplementary Material. For the rice, the current annotation contains 522 PPR proteins, while we could detect 562 (and 629 for poplar). Altogether the distribution in subclasses of PCMP is similar to the one in *arabidopsis*, showing the validity of the approach and confirming the overall conservation of the family. These results demand validation by cross-checking with other annotations.

3.2 Saposin superfamily recognition and tagging

Saposins are small proteins that activate various lipid-degrading enzymes in the lysosomes [10]. Many proteins in this superfamily are formed of several type of Saposin domains, differ by their domain architecture but share common structural features [10,1]. Saposins perform different functions and are widely spread in eukaryotes (plants, amibes, animals).

We extracted from PFAM [3] the hmm for the Saposin type A, Saposin like type B region 1 and 2 domains (Id and Accession numbers: SapA PF02199, SapB_1 PF05184, SapB_2 PF03489) and built a cpHMM from them. We consider the set of proteins representative from different domain architectures containing at least one SapB-1 domain (40 proteins instead of 41 since one protein id does not exist anymore in Swissprot/Trembl, as given in PFAM entry PF05184). HMMER [6] recognizes 38 out of 40 proteins with E-value threshold at $< e-2$, while our cpHMM identifies all of the 40 proteins with E-value $< e-5$. Moreover, when considering the 5 proteins in the Saposin entry of Sisyphus (AL00047861) [1], which consists in structurally related proteins, HMMER [6] detects only 4 of them at E-value $< e-3$ and the fifth proteins at $3e-2$, while all are recognized by cpHMM with E-value $< e-6$. Sisyphus database gathers alignment of proteins whose relationships "identification using the existing computational tools still remains difficult or impossible" and as such represents an interesting test case. In both experiments, the expected architecture of the 3 Sap domains were readily annotated by the cpHMM.

4 Discussion

By adapting the Viterbi and Forward algorithm to cpHMMs, we provide a useful tool to recognize or tag complex domain architecture in protein sequences. Experiments of PPR proteins recognition in the whole rice or poplar proteomes demonstrate the practical effiency of our solution. Moreover, cpHMMs can reuse already developed profile HMMs and thus take advantage of databases like PFAM [3]. In comparison, previously developed alignment algorithms to detect circular permutations in proteins [15,13] represent a complementary approach to identify single rearrangement in closely related sequences (alignment sensitivity is far less than that of pHMMs). In another line of research, the proposal of [5] provides fully integrated domain tagging of all domains for a given protein, using both

sequence similarity and domain context. However, it can neither build a specific model for a given protein family, nor recognize whether a protein belongs or not to a certain family.

In conclusion, a cpHMM can automatically perform both tagging and recognition, and can be adjusted and parameterized for a specific protein family, as in the case of PPR proteins. As it can process the entire protein sequence, it yields a globally optimal "multiple tagging" of several distinct domains, and a measure of its statistical significance. It gives a global E-value, computed as in HMMER [6] and enables the recognition, for a fixed level of confidence, of the proteins of a particular family. For the purpose of automatic annotation, the validation obtained for the PCMP subfamily of the PPR proteins in *arabidopsis*, more than 88% agreement for motif tagging, shows the ability of cpHMMs for systematic annotation of new genomes. cpHMM are versatile tools, which can easily adapt to new situations like identifying PPR proteins in of other species, or even model other protein families with complex domain architectures (e.g., leucine rich repeats, kelch motif repeats).

References

[1] A. Andreeva, A. Prlic, T. Hubbard, and A.G. Murzin. SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nuc Acids Res*, 35(S1):D253–259, 2007.

[2] S. Aubourg, N. Boudet, M. Kreis, and A. Lecharny. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol. Biol.*, 42(4):603–613, 2000.

[3] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S.R. Eddy. The PFAM protein families database. *Nuc Acids Res*, 32(S1):D138–141, 2004.

[4] E. Bornberg-Bauer, F. Beaussart, S. K. Kummerfeld, S. A. Teichmann, and J. Weiner. The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences*, 62(4):435–445, 2005.

[5] L. Coin, A. Bateman, and R. Durbin. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc Natl Acad Sci*, 100(8):4516–4520, 2003.

[6] SR Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[7] E. Kotera, M. Tasaka, and T. Shikanai. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature*, 433(7023):326–30, 2005.

[8] A. Krogh. An introduction to hidden Markov models for biological sequences. *Computational Methods in Molecular Biology*, 1998.

[9] C. Lurin, C. Andres, and S. Aubourg et al. Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. *Plant Cell*, 16(8):2089–2103, 2004.

[10] RS Munford, PO Sheppard, and PJ O'Hara. Saposin-like proteins (SAPLIP) carry out diverse functions on a common backbone structure. *J. Lipid Res.*, 36(8):1653–1663, 1995.

[11] A.G. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[12] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 2(77):257–286, 1989.

[13] S. Uliel, A. Fliess, A. Amir, and R. Unger. A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, 15(11):930–936, 1999.

[14] S. Uliel, A. Fliess, and R. Unger. Naturally occurring circular permutations in proteins. *Prot. Eng.*, 14(8):533–542, 2001.

[15] J. Weiner, T. Geraint, and E. Bornberg-Bauer. Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, 21(7):932–937, 2005.