Hidden Markov Models for the Detection of Motif Repeats in Protein Sequences

Raluca Uricaru, Laurent Bréhélin, Eric Rivals
uricaru@lirmm.fr, brehelin@lirmm.fr, rivals@lirmm.fr
LIRMM, Université Montpellier 2

12 Octobre 2006

Proteins are composed of structural units, called domains, which can be detected using computational methods. In many families, proteins contain several domains, or several conserved sequence parts, called motifs. During the evolution, rearrangements (swaps or circular permutations) may alter the relative order of these units (domains or motifs) [Bornberg05], while tandem duplications can change their number. In a family, it results in a variable organization of the units along the chain.

Profile HMMs are the prefered models to represent motifs or domains. They serve to recognize new members of a protein family. Profile HMMs have a linear structure which can model a single unit, and if iterated they can detect repeats of this unit. To perform a "multiple tagging", that is to identify the sequence regions corresponding to each possible motif, one needs to combine the results of each HMM. This was done following heuristic criteria (e.g., choosing the best scoring motif when the detections overlap). Up to now, we lack a method to perform a multiple tagging automatically and optimally according to a global criterion.

To solve this problem, we generalize Profile HMMs to a novel structure called Cyclic Profile HMMs, which can predict the most probable motif organization of a protein. It can cope with repeated units whose number varies, and with changes in the relative order of the units. We used Cyclic HMMs to tag multiple motifs in the PPR (**P**entatrico **P**eptide **R**epeats) protein family of plants.

PPR proteins contain tandem repeats of PPR motifs; there are 3 such motifs, named **P**, **L**, **S**. Roughly half of the PPR proteins form the PPRP subfamily (the proteins from this subfamily usually have only P motif repeats), while the other half is represented by the PCMP subfamily. The PPRPs can be defined using the following regular expression (like in PROSITE, but in our case, letters represent motifs) : (P*-S*)* ; the more complex structure of the PCMPs can also be summarized using a regular expression : (P-L-S*)*-[E-[E+ -[Dyw]]] [Lurin04].

Having as input the profile HMMs representing each PPR motif, helped solving the problem accurately. These HMMs were used to build a Cyclic Profile HMM that can process the entire motif sequence. It allows to obtain the globally optimal "multiple tagging" solution and gives us the means to measure its statistical significance. As a result, we obtained an automatic, "multiple tagging" tool.

Thanks to the manual motif annotation for the PCMPs, from the *Arabidopsis thaliana* we could validate our results. We found less than 10% discordance and we were able to retrieve the expected PCMP classes distribution [Rivals06]. After this validation, we used our tool to annotate the PPRP subfamily in *Arabidopsis* and to perform the first motif annotation of the complete PPR family

proteins of rice. This allows to compare the PPR motif organization between the model plants of monocotyledons and dicotyledons.

In conclusion, we prove the capacity of the Cyclic Profile HMM structure to solve the problem of "multiple tagging". We verified the feasibility and the efficiency of the method in terms of execution time, for a large number of sequences and a large number of motifs. Cyclic Profile HMMs are a versatile structure that adapts to new situations, like the PPR family of other species, or even to other protein families with complex unit organization (leucine rich repeats, kelch motif repeats). In future work, we will concentrate on the discovery of similarities for the subfamilies of a species and on the fine identification of new subfamilies.

[Bornberg05] Bornberg-Bauer E. et al., CMLS Cell. Mol. Life Sci., 62 : 435 - 445 (2005) ;

[Lurin04] Lurin C., et al., Plant Cell, nb. 8, 16 : 2089 - 2103 (2004) ;

[Rivals06] Rivals E. et al., Plant Physiology, 141 (2006).