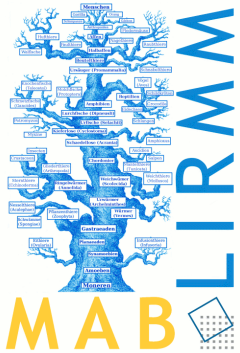
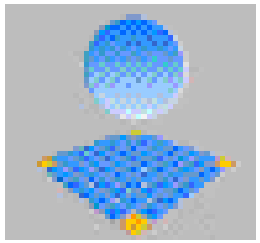
A new type of Hidden Markov Models to predict complex domain architectures in protein sequences

Raluca Uricaru, Laurent Bréhélin and Eric Rivals

LIRMM, CNRS Université de Montpellier 2





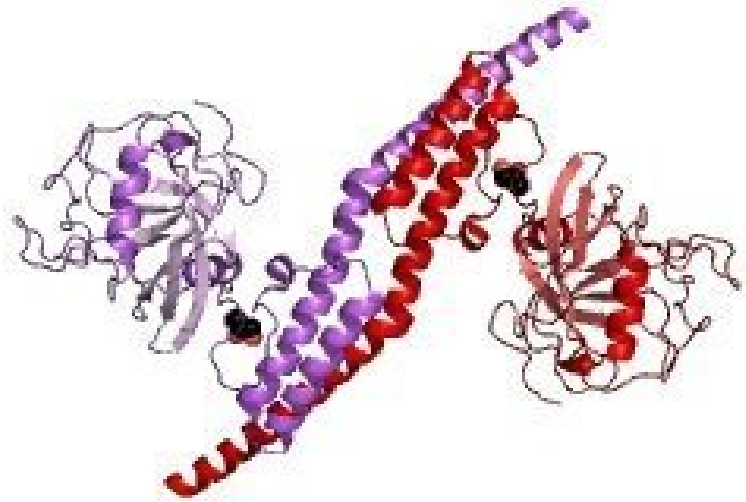
Summary

Summary

General Context

Motifs, Domains, Pfam

- **Motifs** = sequence regions conserved in a protein family
- **Domains** = conserved sequence regions with structural properties



TPR domains
(Pfam, NAR 2004)

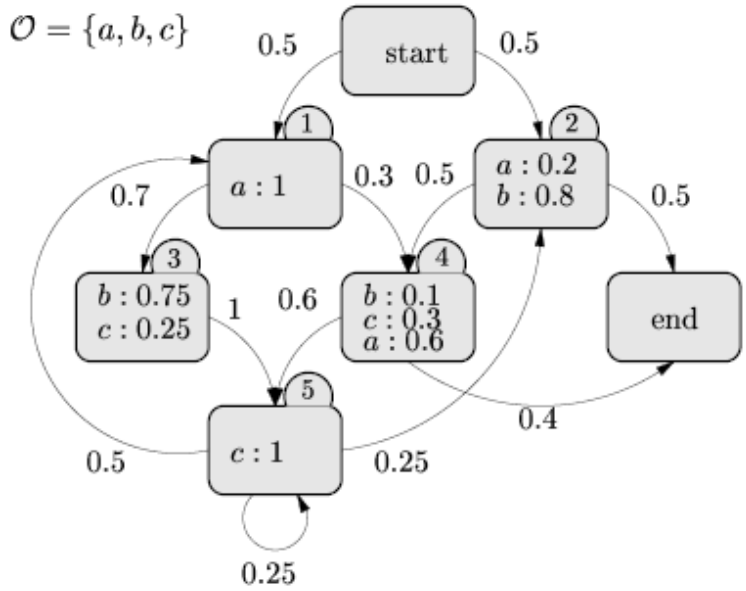
many proteins are composed of several domains

- **Pfam** collection of protein domains & families
  uses pHMMs (via HMMER software) to model domains

Profile Hidden Markov Models (pHMMs)

Hidden Markov Models (HMM)

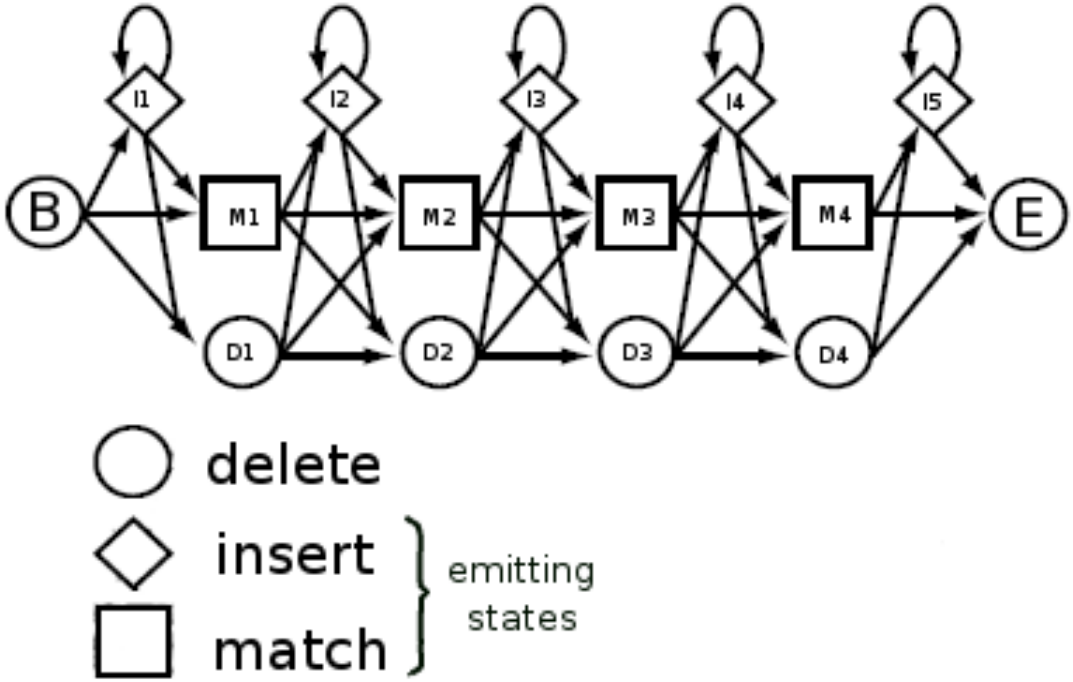
- close to probabilistic automata
- define stochastic, non-deterministic processes
- hidden, markovian generation process



Profile HMMs

- particular HMMs designed to represent domains/motifs
- strongly linear structure, left-right models

Profile Hidden Markov Models (pHMMs) (2)



Profile HMM : 4 match states example (S.R. Eddy, 1996)

Profile Hidden Markov Models (pHMMs) (3)

Problems

- test a protein's membership of a known family
  $\longrightarrow$ recognition
  compute the probability that the protein sequence is generated by the model

  Forward algorithm

- tag the precise position of a domain in the sequence
  $\longrightarrow$ tagging
  find the state sequence (Viterbi path) with the highest probability

  Viterbi algorithm

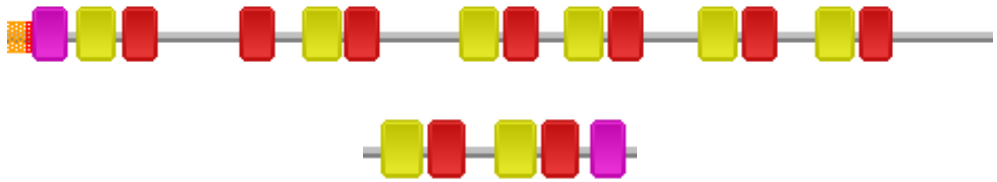
Profile Hidden Markov Models (pHMMs) (4)

Difficulties

- proteins are usually composed of several different domains
- rearrangements & duplications
  $\longrightarrow$ different domain **architectures** in the same family

Q3YMU5_DROSI :

Q5R848_PONPY :



Distinct domain architectures for Saposins (Pfam)

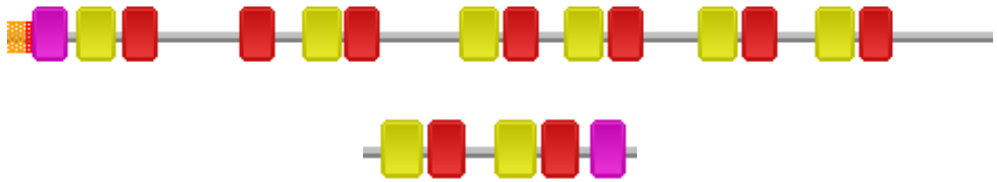
Profile Hidden Markov Models (pHMMs) (4)

Difficulties

- proteins are usually composed of several different domains
- rearrangements & duplications
  $\longrightarrow$ different domain **architectures** in the same family

Q3YMU5_DROSI :

Q5R848_PONPY :



Distinct domain architectures for Saposins (Pfam)

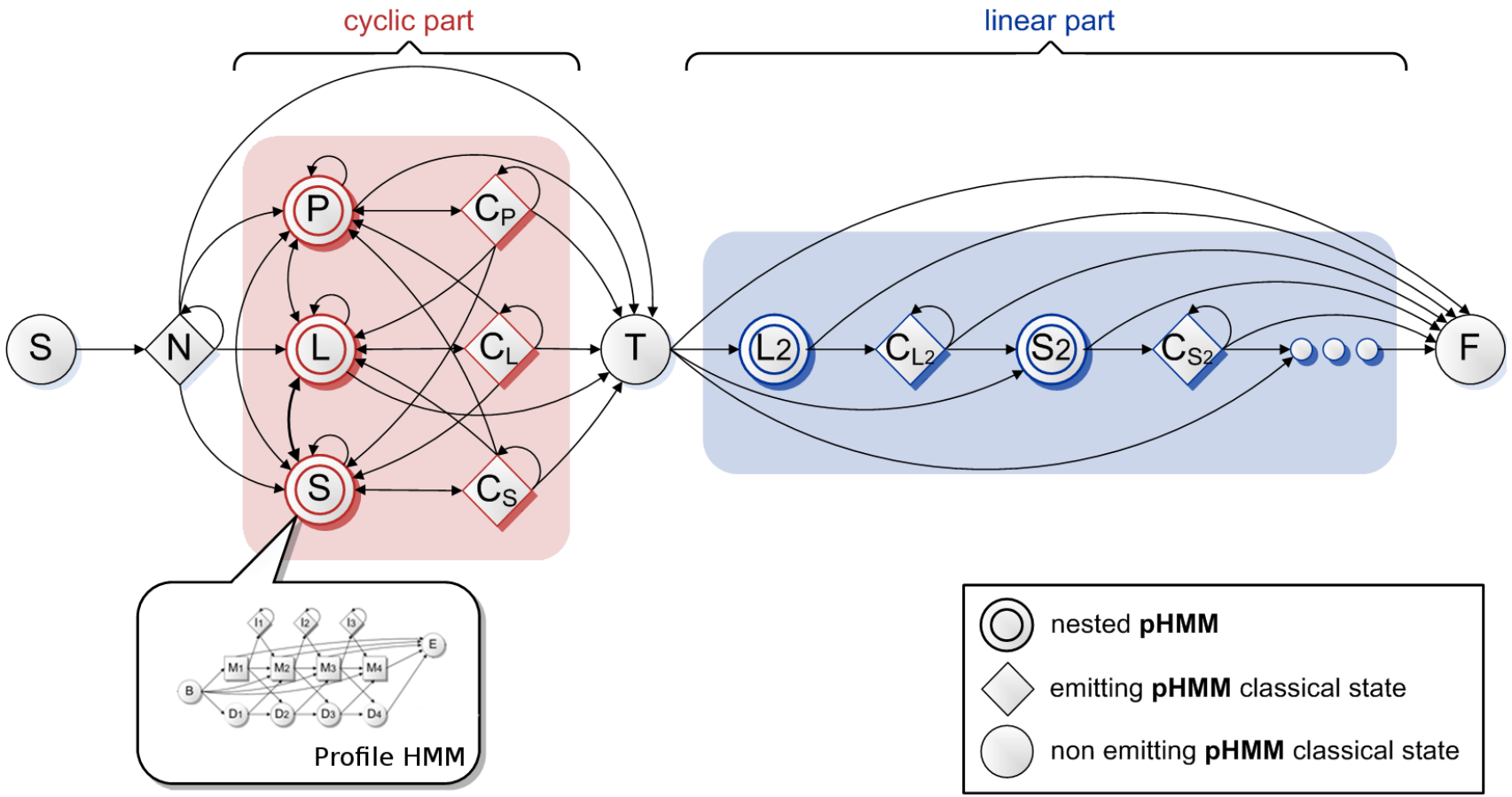
pHMM : has a linear structure,
models a single domain

$\longrightarrow$ unable to deal with complex domain architectures

Summary

Solution : Cyclic profile HMM (CpHMM)

generalized pHMM - *made up of nested pHMMs*



CpHMM : advantages

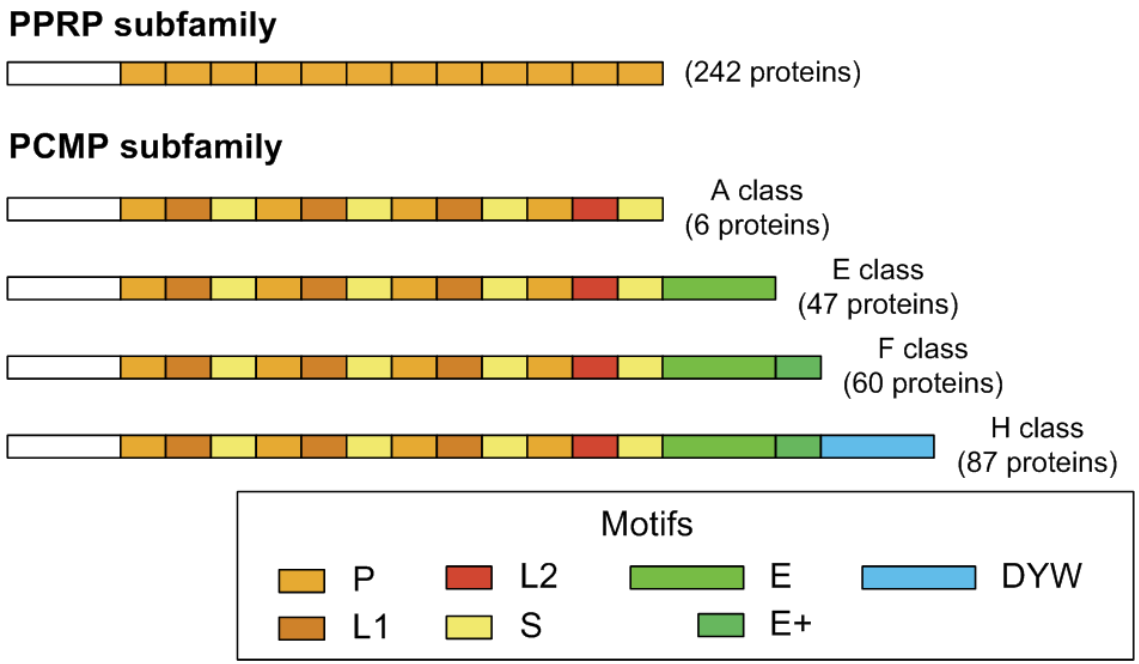
- capitalizes on already developed pHMMs (Pfam)
- deals with both :
  - variable number of repeated units
  - variable relative order of units
- takes in consideration :
  - sequence similarity
  - domain context
- built with or without prior knowledge
  $\longrightarrow$ can be adjusted & parameterised for a specific family

CpHMM : advantages (2)

- automatically performs both tagging and recognition
- yields a globally optimal "multiple tagging" of several domains
- gives a global E-value (*computed as in HMMER*)
- efficient in terms of computing time
  ($\approx$ 25 minutes for 45000 *poplar* proteins)

Summary

PPR protein family



Motif architecture of PPR proteins in Arabidopsis (Lurin et al., 2004)

PPR motifs : CpHMM versus expert manual annotation

Dataset

- PCMP subfamily in *Arabidopsis* : 197 protein sequences ;
- regular expression for PCMP subfamily :
  $(P - L - S^*)^* - (P - L_2 - S^*) - [E - [E^+ - [Dyw]]]$
- distribution of the number of PPR motifs per protein :
  avg = 15, variance = 15, min = 7, max = 28
  $\longrightarrow$ diversity of architectures ;

Results

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

- automatic annotation valid in 88% of the motifs ;

PPR analysis in plant genomes

plant genome	# proteins	# PPR proteins
Arabidopsis	40398	552
Poplar	45555	675
Rice	66710	592
Chlamydomonas (alga)	15143	22
Physcomitrella (moss)	572	8

Summary

Saposin protein superfamily

- small proteins, activate lipid-degrading enzymes in lysosomes
- several types of Saposin domains :

type A, type B region 1, type B region 2

$B_1 B_2 B_1 B_2 B_1 B_2 B_1 B_2 AAAA$

$B_1 B_1 B_2 B_1 B_2 B_1 B_2 B_1 B_2 B_1 B_2$

$AB_2 B_1 B_2 B_1 B_2 B_1 B_2 A$

Saposin domain architectures : examples (Pfam)

- different domain architectures, common structural features

Saposin domain : CpHMM versus HMMER annotation

Datasets

- set of proteins, representative for different domain architectures
  40 proteins *containing at least one B_1 domain*
- interesting test cases : difficult identification of relationships (Sisyphus)

Results

- detects all proteins with better E-values than *HMMER*
- retrieves expected domain architecture

- Support :
  ACI IMPBio REPEVOL
  http://www.lirmm.fr/~rivals/REPEVOL
- CpHMM for PPR annotation available at :
  http://atgc.lirmm.fr/PPR

Thanks for your attention !

Questions ?