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

Raluca Uricaru, Laurent Bréhélin and Eric Rivals

LIRMM, CNRS Université de Montpellier 2

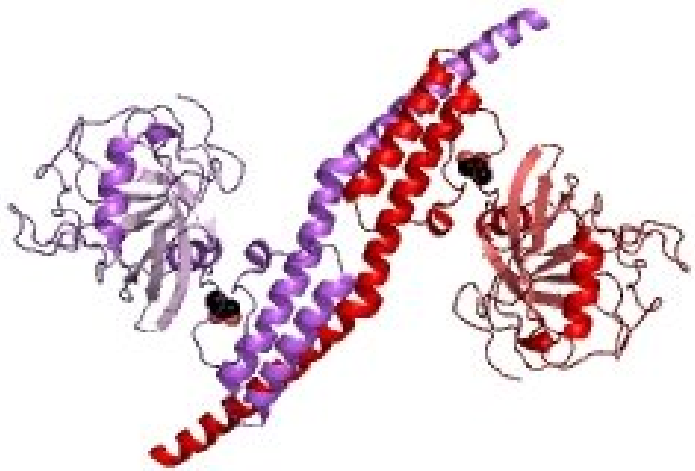
14 Juin 2007

Summary

General Context

Motifs, Domains, PFAM

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



many proteins are composed of several domains

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

pHMMs

Profile Hidden Markov Models (pHMMs)

- HMM = probabilistic automaton ;
- pHMM = HMM designed to represent domains/motifs ;
- test a protein's membership of a known family
  $\longrightarrow$ recognition
- tag the precise position of a domain in the sequence
  $\longrightarrow$ tagging

pHMMs (2)

Difficulties

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

pHMMs (2)

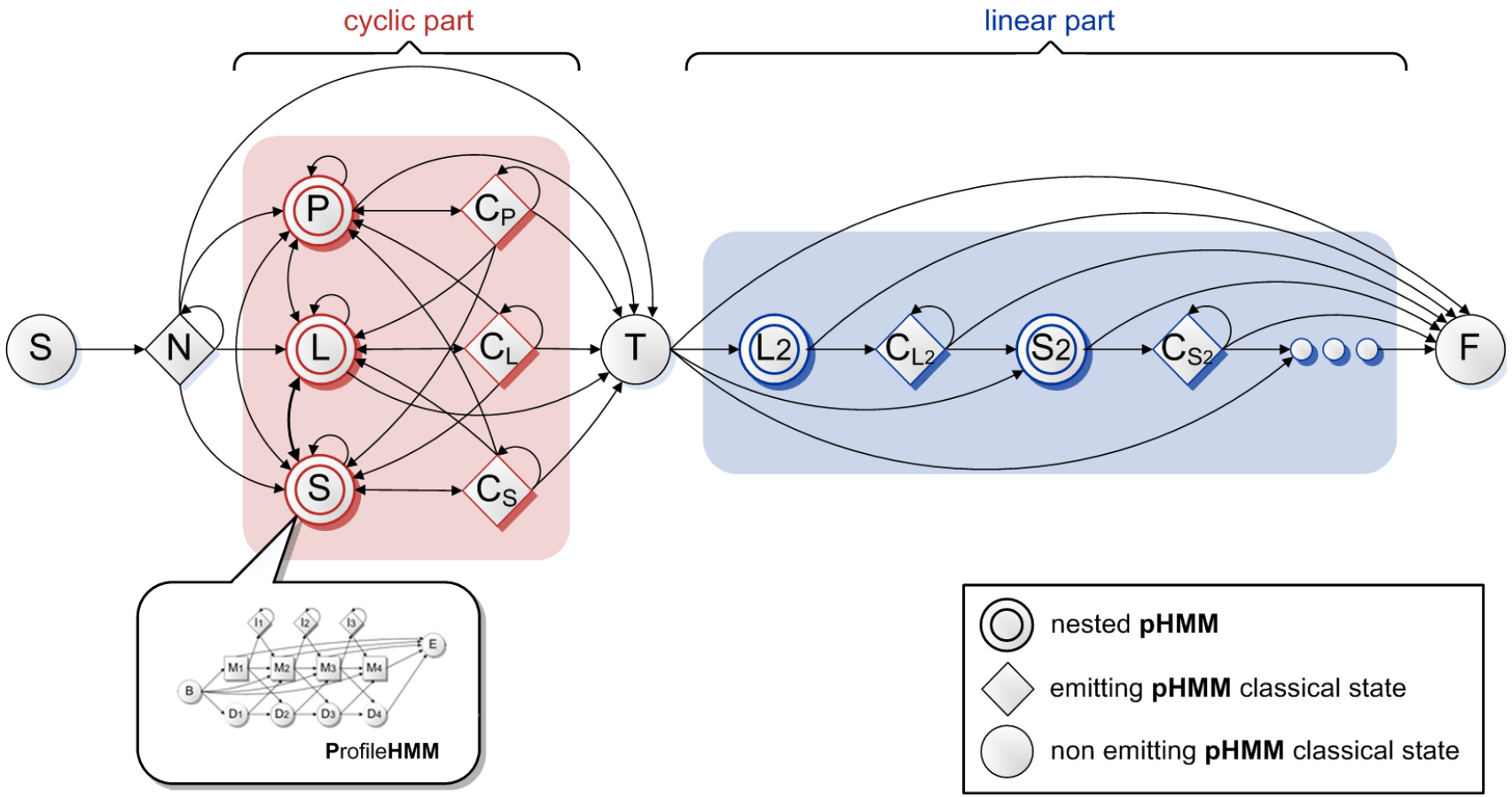
Difficulties

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

pHMMs : have a linear structure,
model a single domain
$\longrightarrow$ unable to deal with complex domain architectures

Proposed solution : generalized pHMM

Cyclic profile HMM (CpHMM) - *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 both :

  sequence similarity

  &

  domain context ;

CpHMM : advantages (2)

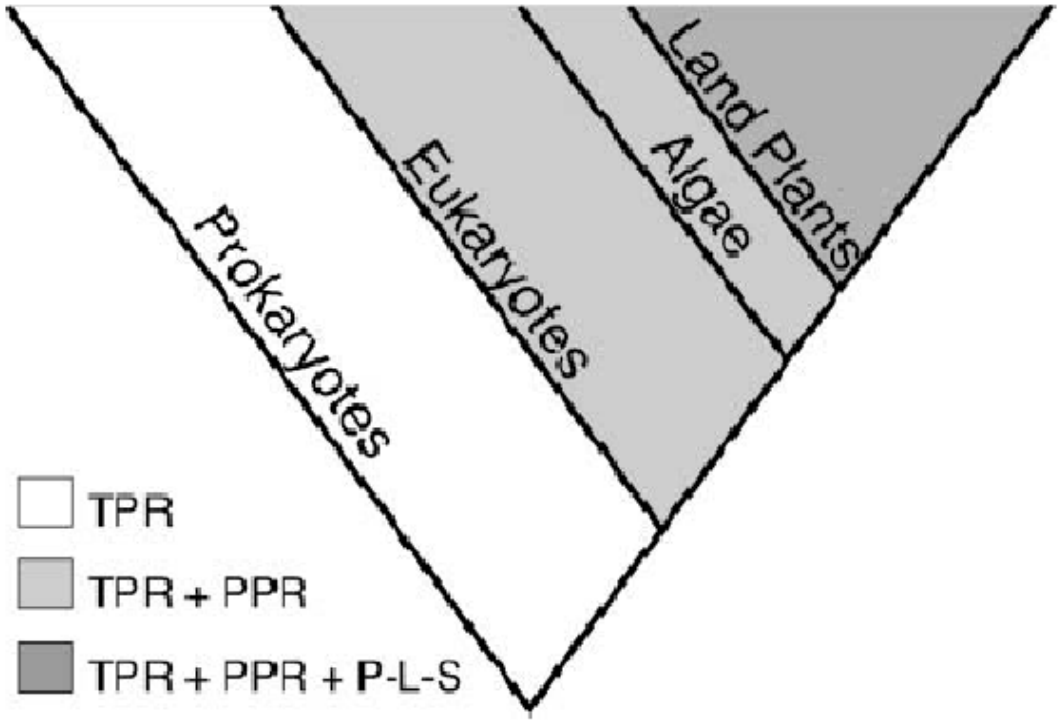
- built with or without prior knowledge
  $\longrightarrow$ can be adjusted & parameterised for a specific family ;
- automatically performs both tagging and recognition ;
- yields a globally optimal

  "multiple tagging"

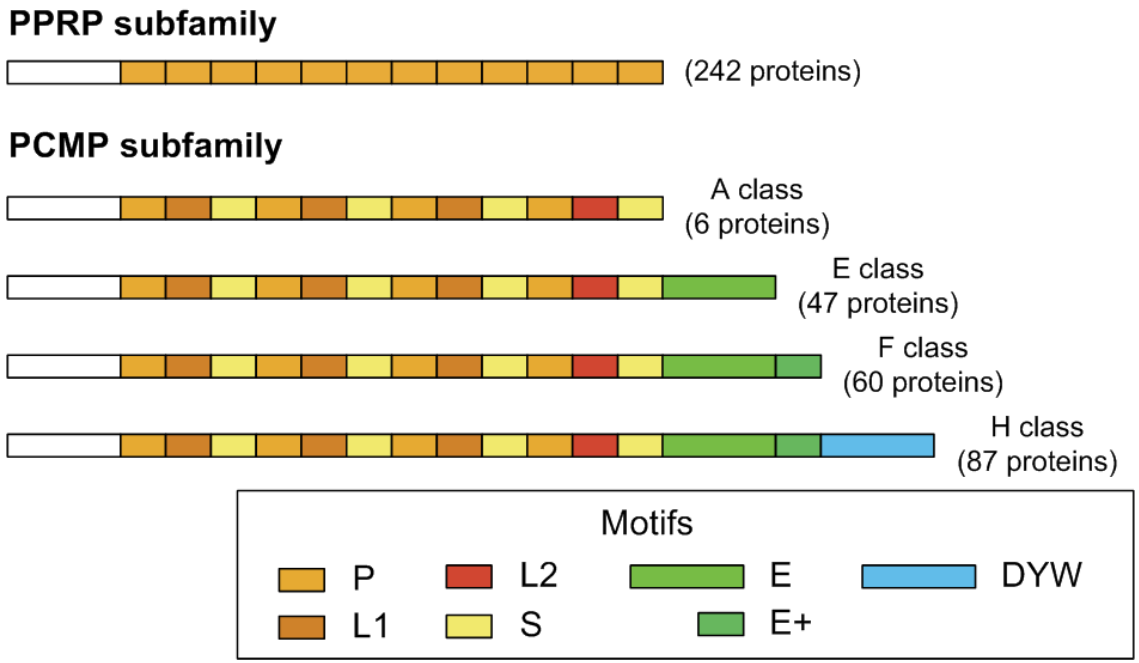
  of several domains ;
- gives a global E-value (*computed as in HMMER*) ;
- efficient in execution terms
  ($\approx$ 20 minutes for 40000 *arabidopsis* proteins)

PPR protein family

Protein family with complex multi-domain architecture
The PPR protein family is particularly large in plants



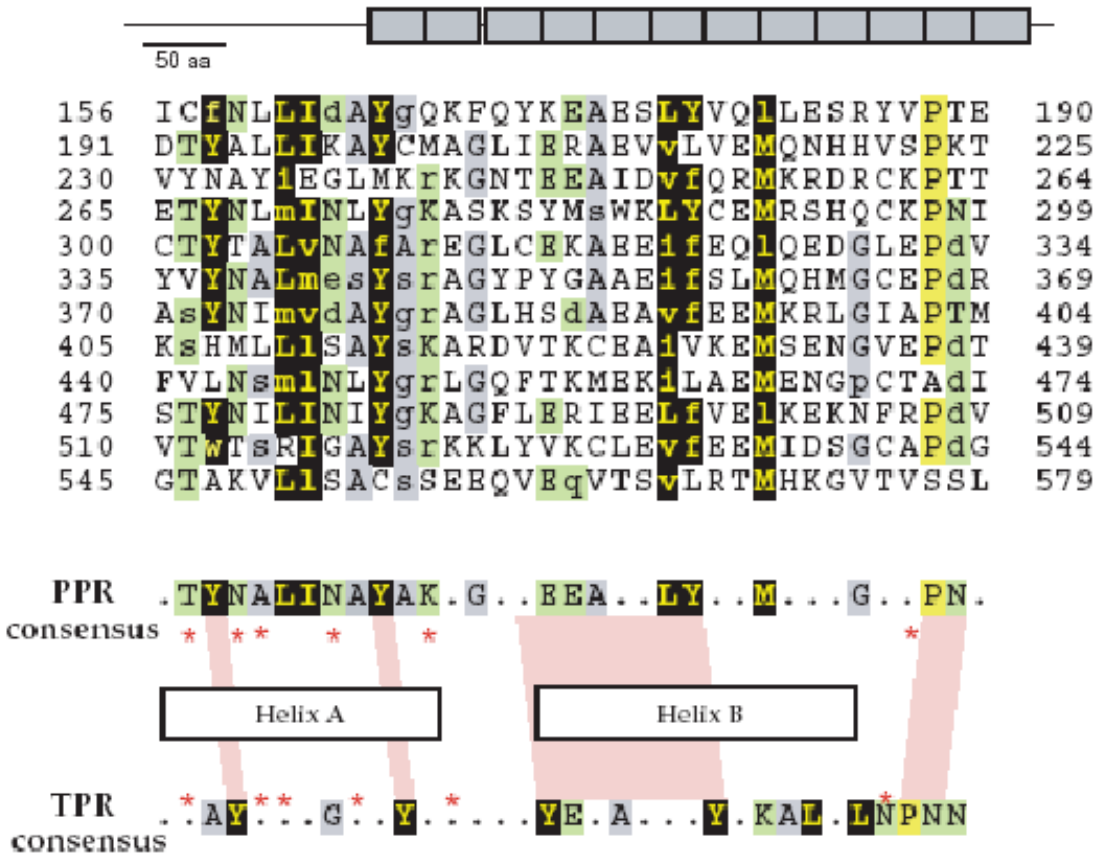
PPR protein family (2)



Motif structure of PPR proteins for *Arabidopsis*

PPR protein family (3)

PPR motif multiple alignment



Validation

- comparison between automatic annotation (*CpHMM*) & expert manual annotation
  on 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, min = 7, max = 28 $\longrightarrow$ diversity of architectures ;

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

- automatic annotation valid in 88% of the motifs ;