A new phylo-HMM paradigm to search for sequences

Jean-Baka DOMELEVO ENTFELLNER
&
Olivier GASCUEL

LIRMM (CNRS - UM2), Montpellier

June 10th, 2008

What is at stake?

Goal

Search a databank for sequences homologous to a query protein family.

Existing approaches

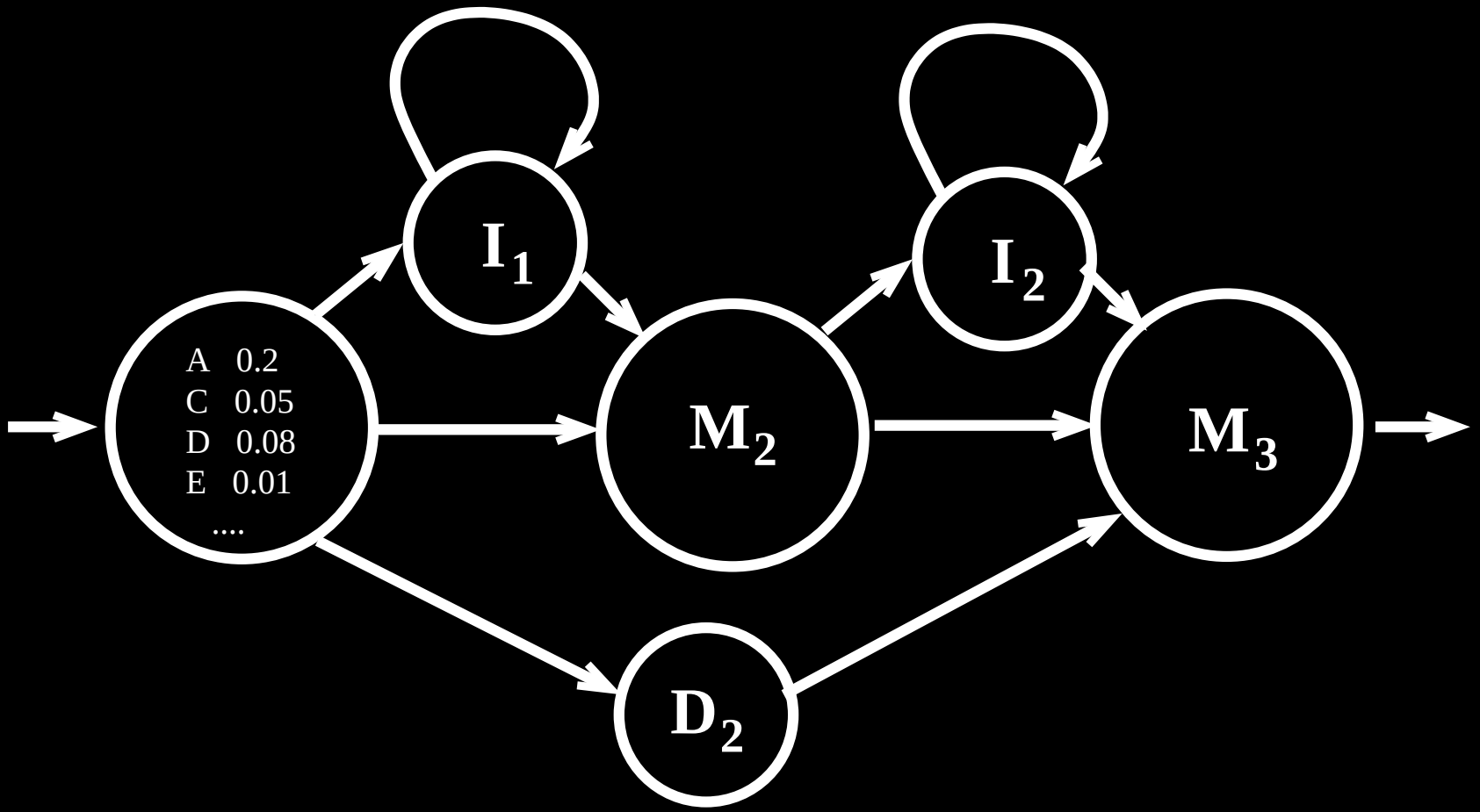
1. Blast: poor results when identity rate is too low ($\lesssim$30%)
2. Profile HMMs:
   - allow lower percentage of identity between query & target
   - but make no use of the phylogeny

Proposed solution

Design a model which takes advantage of:

1. the possible presence in the family of a sequence close to the target
2. the global information (e.g. hydrophilic/phobic columns) conveyed by the alignment

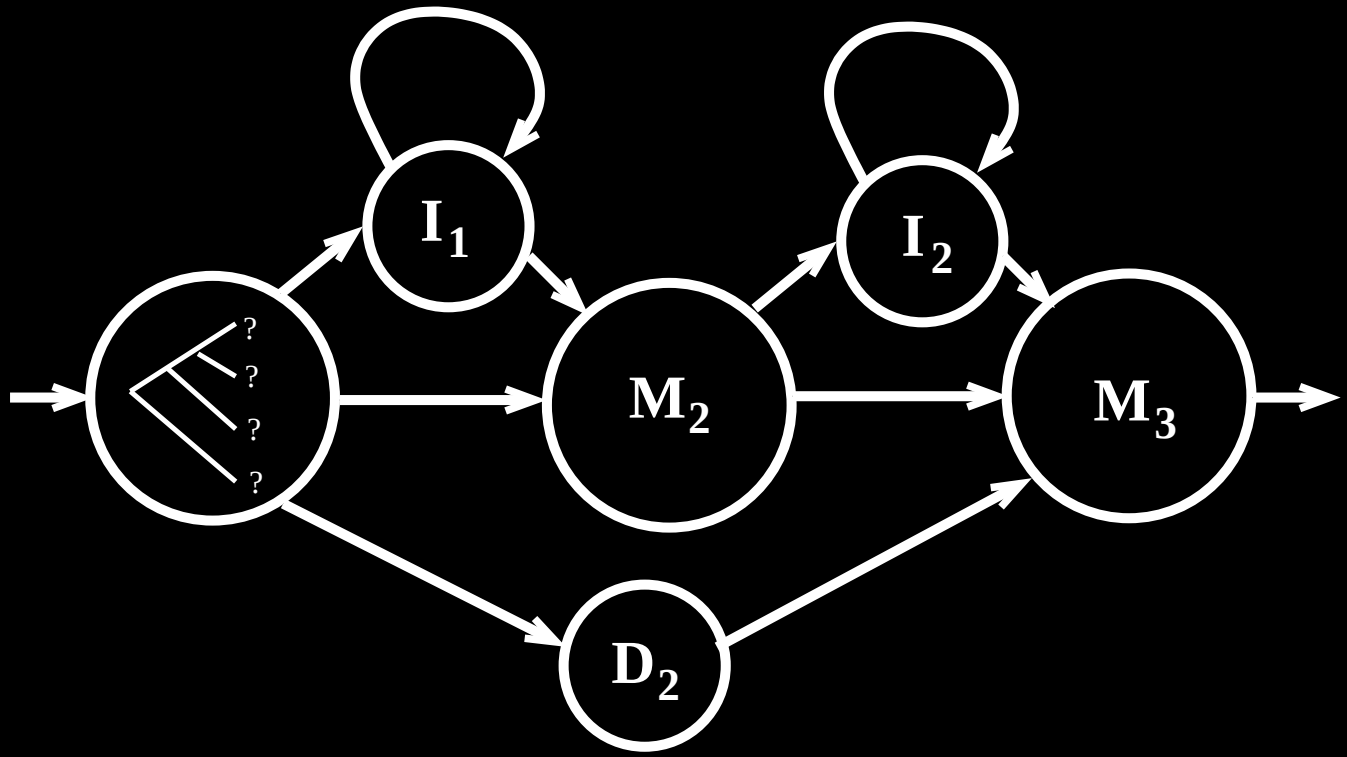
Profile HMMs



Each match and insertion state generates a single a.a.

phylo-HMMs

Seminal works: Goldman et al. 1996, Siepel & Haussler 2003

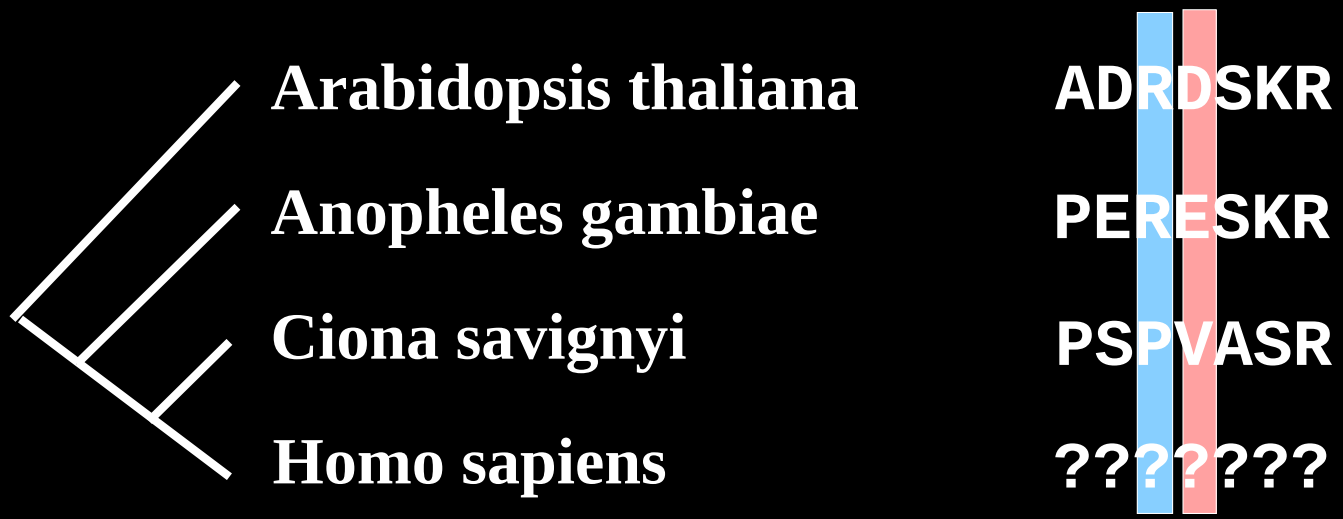


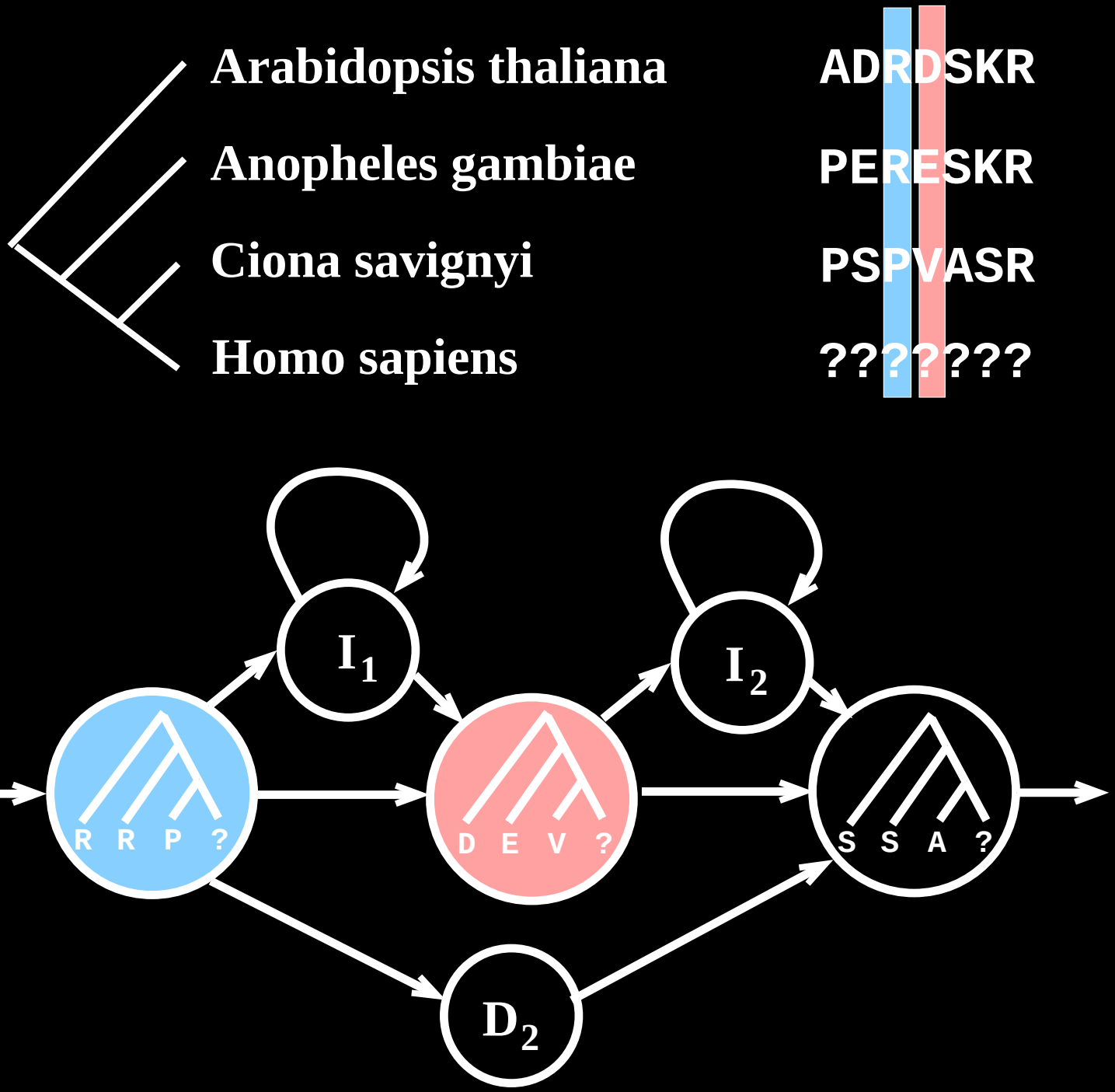
- each node is populated by a phylogeny which defines a probability distribution over a column of the alignment
- typical use: prediction of the conservation or secondary structure of the sites

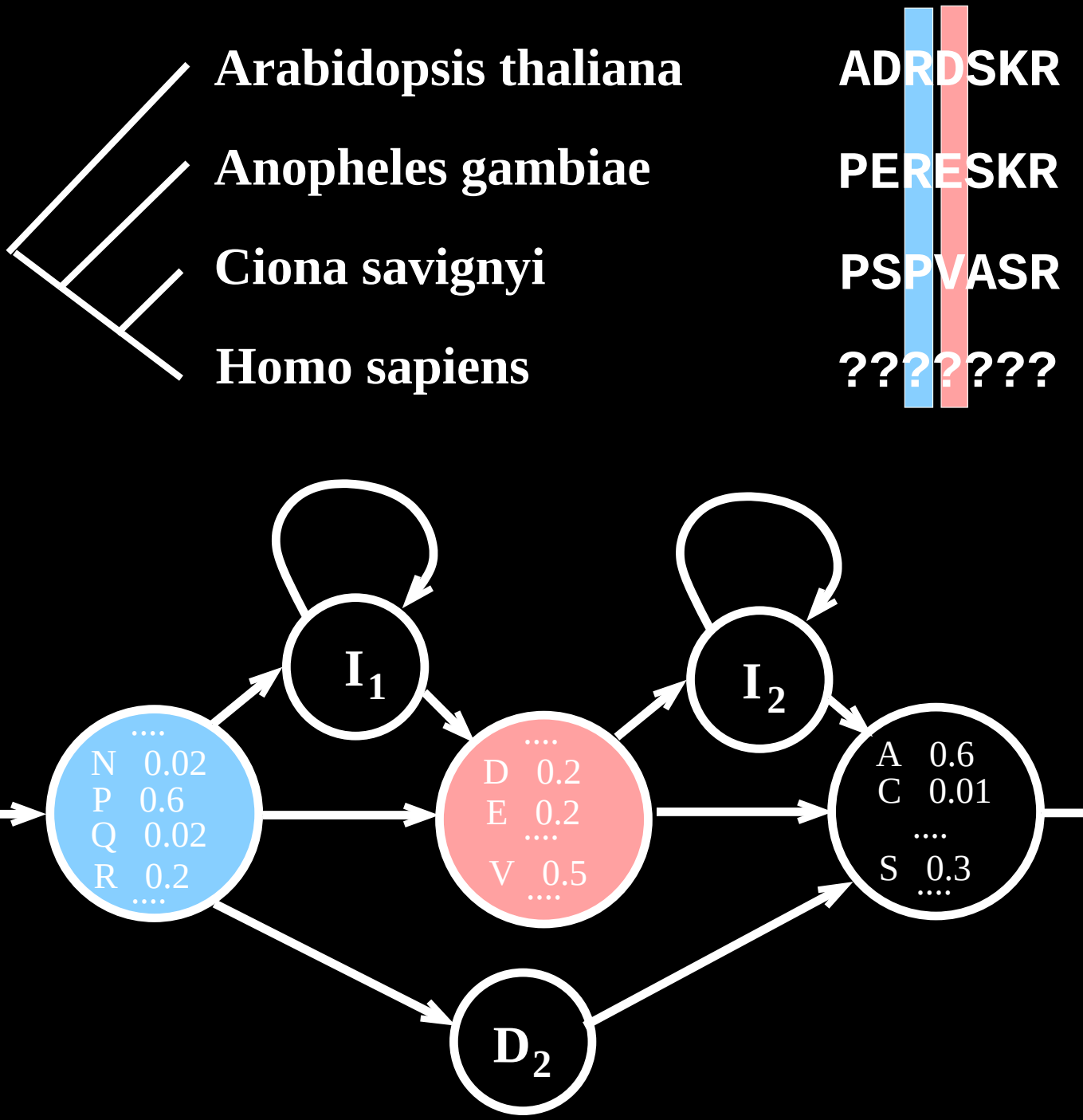
How we use phylo-HMMs

Knowing the phylogeny, we fill in each match state with the distribution of posterior probas of a.a. for the target, given the corresponding column of the alignment.

$\rightarrow$ Felsenstein's pruning algorithm




Arabidopsis thaliana
Anopheles gambiae
Ciona savignyi
Homo sapiens
ADRDSKR
PERESKR
PSPVASR
???????
I1
I2
D2
R R P ?
D E V ?
S S A ?

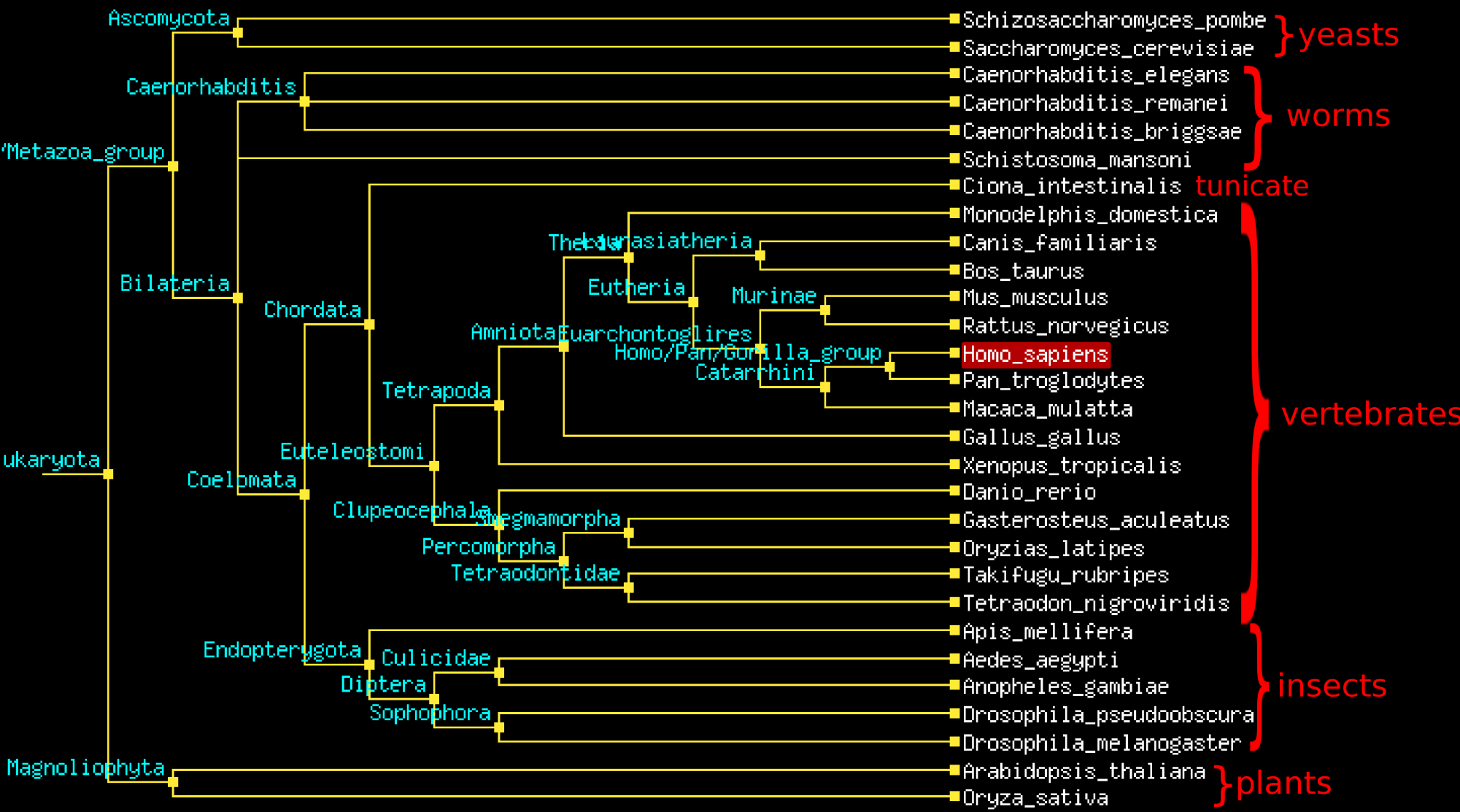

Arabidopsis thaliana
Anopheles gambiae
Ciona savignyi
Homo sapiens
ADRDSKR
PERESKR
PSPVASR
???????
I1
I2
D2
N 0.02
P 0.6
Q 0.02
R 0.2
D 0.2
E 0.2
V 0.5
A 0.6
C 0.01
S 0.3

Experimenting

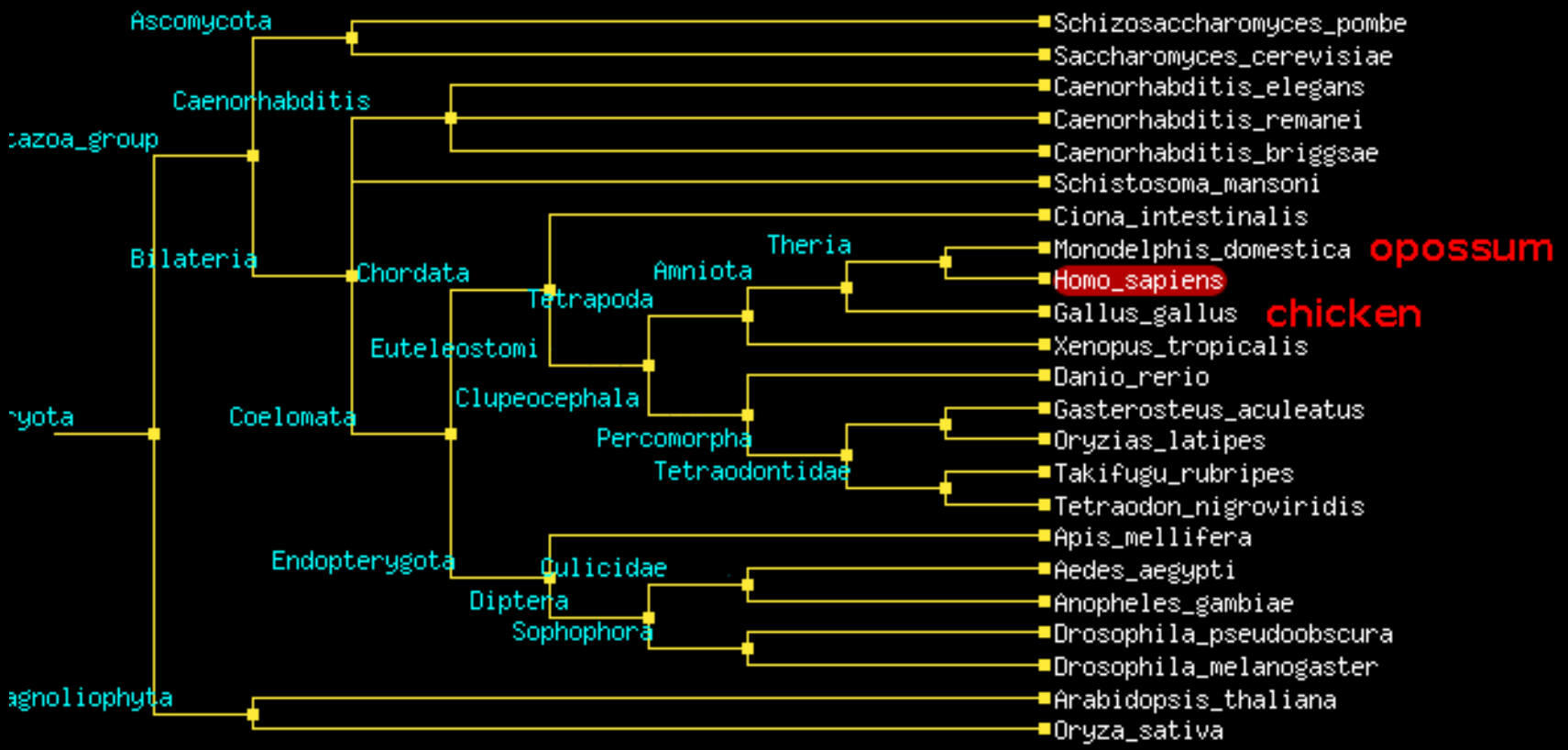
- test data: 690 protein families from the Treefam database (Vertebrates + Insects + 1 Tunicate, 4 worms, 2 yeasts and 2 plants).
- phylogeny is assumed (calculated with PhyML, matches NCBI consensus).

Experimental setup:

1. take those 690 complete families from Treefam
2. gradually prune to remove all Vertebrates, Insects, ...
3. realign the remaining sequences
4. build the profile HMM with `hmmbuild`
5. phylogenise it to scan for human proteins
6. scan the human proteome with resulting phylo-HMM to find the original protein

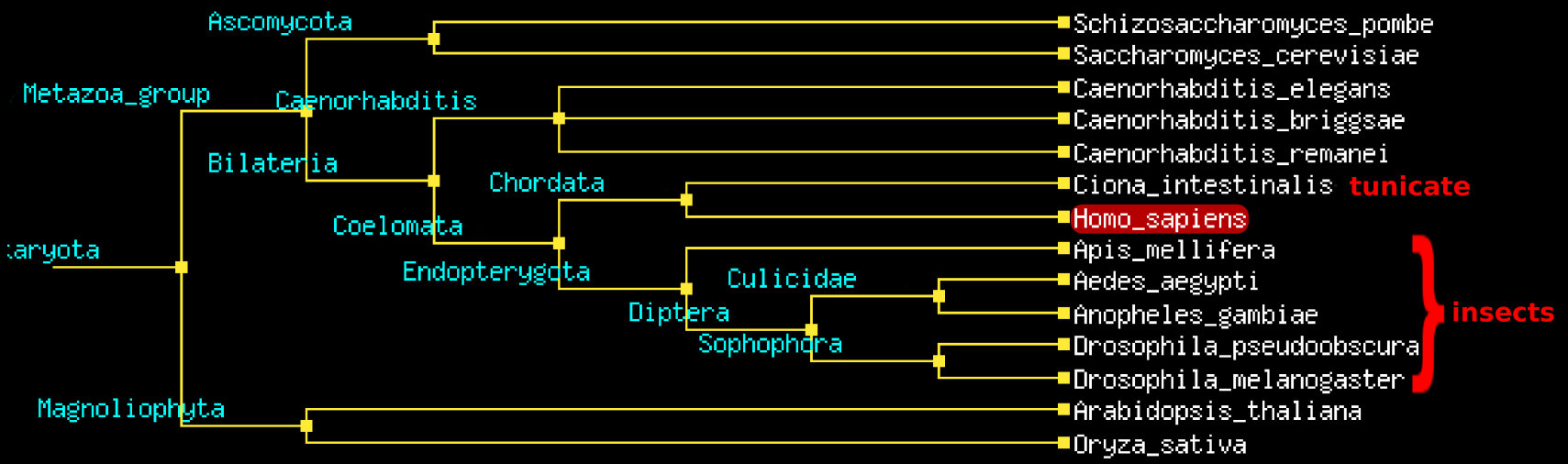

Ascomycota
Caenorhabditis
Metazoa_group
Bilateria
Chordata
Eukaryota
Coelomata
Euteleostomi
Tetrapoda
Amniota
Theria
Eutheria
Murinae
Euarchontoglires
Homo/Pan/Gorilla_group
Catarrhini
Clupeocephala
Neoteleostei
Percomorpha
Tetraodontidae
Endopterygota
Culicidae
Diptera
Sophophora
Magnoliophyta
Schizosaccharomyces_pombe
Saccharomyces_cerevisiae
Caenorhabditis_elegans
Caenorhabditis_remanei
Caenorhabditis_briggsae
Schistosoma_mansoni
Ciona_intestinalis
Monodelphis_domestica
Laurasiatheria
Canis_familiaris
Bos_taurus
Mus_musculus
Rattus_norvegicus
Homo_sapiens
Pan_troglodytes
Macaca_mulatta
Gallus_gallus
Xenopus_tropicalis
Danio_rerio
Gasterosteus_aculeatus
Oryzias_latipes
Takifugu_rubripes
Tetraodon_nigroviridis
Apis_mellifera
Aedes_aegypti
Anopheles_gambiae
Drosophila_pseudoobscura
Drosophila_melanogaster
Arabidopsis_thaliana
Oryza_sativa
yeasts
worms
tunicate
vertebrates
insects
plants

Pruned trees (1/3)



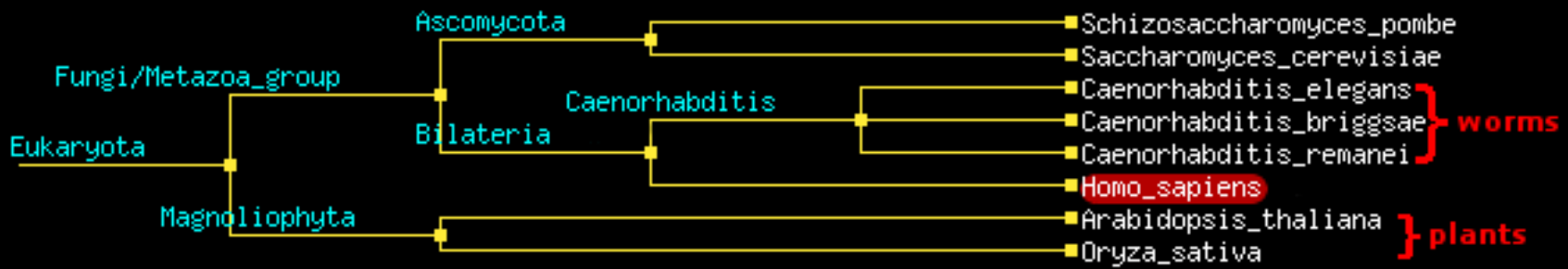
	# of true positives	sensitivity
standard profile HMM	1345	0.88
Blast	1434	0.94
phylo-HMM	1435	0.94
# expected detections	1526	

Pruned trees (2/3)



	# of true positives	sensitivity
standard profile HMM	1280	0.86
Blast	1293	0.87
phylo-HMM	1348	0.91
# expected detections	1489	

Pruned trees (3/3)



	# of true positives	sensitivity
Blast	25	0.38
standard profile HMM	38	0.58
phylo-HMM	52	0.80
# expected detections	65	

Conclusion

Our model uses phylogenetic information to *contextualize* a profile HMM.

- first results look promising
- good combination of Blast and profile HMMs paradigms, robust to remote phylogenetic relations