Identification and classification of protein subfamilies using top-down phylogenetic tree reconstruction

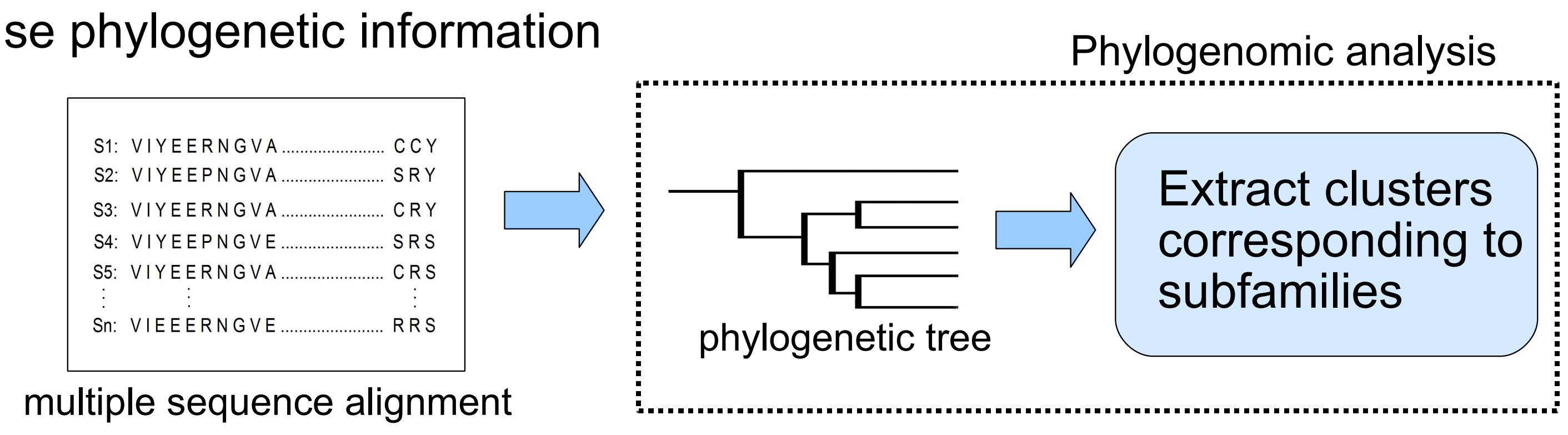
Eduardo P Costa – Celine Vens – Hendrik Blockeel
Dept. of Computer Science, Katholieke Universiteit Leuven

Proteins in a subfamily usually share a specific function that is not common to the entire family. We investigate the use of clustering trees to identify such subfamilies.

Protein function prediction

Several computational methods have been designed to assist scientists in the context of protein function prediction:

- Homology-based methods
  - Error prone: error propagation; proteins can change functions
- Supervised learning approach
  - Large amount of training data needed
- Phylogenomic methods
  - Use phylogenetic information

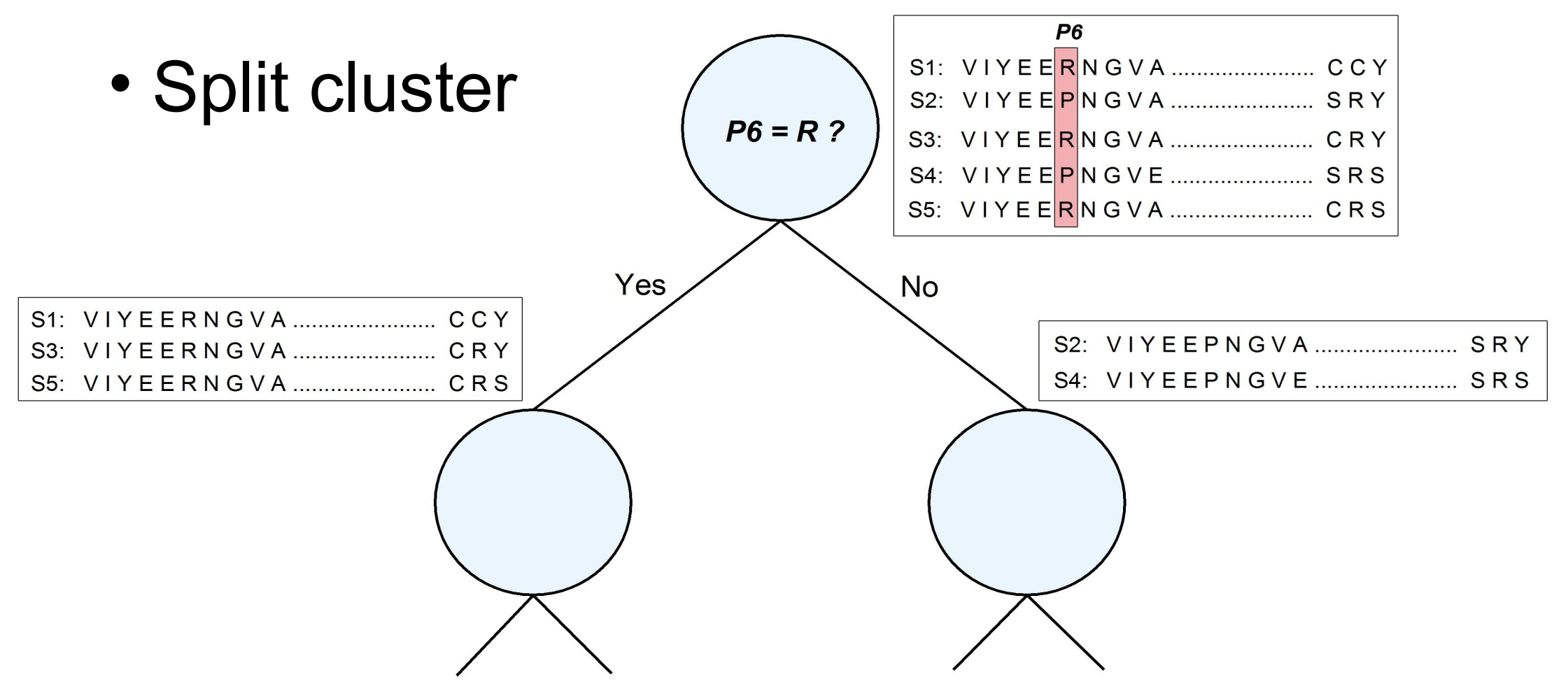

S1: VIYEERNGVA .... CCY
S2: VIYEEPNGVA .... SRY
S3: VIYEERNGVA .... CRY
S4: VIYEEPNGVE .... SRS
S5: VIYEERNGVA .... CRS
Sn: VIEEERNGVE .... RRS
multiple sequence alignment
Phylogenomic analysis
phylogenetic tree
Extract clusters corresponding to subfamilies

  - Example: SCI-PHY (Brown et al. 2007)

Top-down phylogenetic tree reconstruction

Divisive clustering algorithm: **Clus-φ**

- Start with one cluster containing all sequences
- Repeat
  - Split cluster


P6 = R ?
P6
S1: VIYEERNGVA .... CCY
S2: VIYEEPNGVA .... SRY
S3: VIYEERNGVA .... CRY
S4: VIYEEPNGVE .... SRS
S5: VIYEERNGVA .... CRS
Yes
No
S1: VIYEERNGVA .... CCY
S3: VIYEERNGVA .... CRY
S5: VIYEERNGVA .... CRS
S2: VIYEEPNGVA .... SRY
S4: VIYEEPNGVE .... SRS

Clustering method based on decision tree learning approach (CLUS).

Tests in the nodes check for **polymorphic positions**.

Clus-φ uses the **minimum evolution hypothesis**, namely, constructing a tree with minimal total branch length, to choose the best split.

- Until there is only one sequence per cluster

Applying Clus-φ to protein subfamily identification

Problem: how to extract clusters?

- Stop criterion (e.g. entropy reduction, f-test)
- Post-processing pruning (e.g. category utility)
- Use subfamily information (semi-supervised learning)
  - Functional information may be available for some of the sequences

Future work

ADVANTAGES over existing phylogenomic methods

- No need to build the complete tree if stop criterion is used
- Produces evolutionary trace
- Allows to identify functional sites
  - Amino acids that are discriminating among different subfamilies
- Allows to directly classify new sequences into subfamilies

Experiments

Scenario 1

Goal: check if trees that have splits based on polymorphic positions are useful for protein subfamily identification

Setting: we added the subfamily information to the data and induced a classification tree using CLUS (supervised classification task)

Table 1. Number of leaves in the classification tree

	# Subfamilies	Classification tree (# leaves)
Enolase	8	8
Crotonase	10	11
Secretin	15	15
Amine level 1	7	14
Amine level 2	31	34
NHR level 1	8	11
NHR level 2	27	30
NHR level 3	77	79

✓ Results: Subfamilies can be perfectly separated from one another using compact trees. The results show that the solution we are looking for is part of the search space.

Scenario 2

Goal: evaluate the quality of the trees being produced, regardless of the stop criterion

Setting: we grew the tree completely until each node was a singleton, and then cut the tree in a way that all clusters were pure and that the tree was as compact as possible. We did the same for Neighbor Joining (NJ) and SCI-PHY.

Table 2. Number of leaves in the post-processed phylogenetic tree

	Clus-φ	NJ	SCI-PHY
Enolase	12	38	28
Crotonase	33	35	70
Secretin	19	20	22
Amine level 1	30	27	36
Amine level 2	49	52	52
NHR level 1	22	12	30
NHR level 2	43	31	38
NHR level 3	90	97	105

✓ Results: Clus-φ produced more compact trees than NJ for 5 datasets, and more compact trees than SCI-PHY for 7 datasets. This shows that our method can yield good results if an adequate stop criterion is used.

Scenario 3

Goal: test the whole procedure

Setting: We defined as stop criterion the point where the entropy reduction given by best test, according to the total branch length heuristic, is less than 5%.

Evaluation measure: rand index (cfr. accuracy)

rand index = 1 - [probability that 2 random proteins are in the same predicted cluster but have different subfamilies, or the other way around].

Table 3. Rand index for the subfamily prediction task

	Clus-φ	SCI-PHY
Enolase	0.98	0.86
Crotonase	0.57	0.80
Secretin	0.61	0.96
Amine level 1	0.17	0.87
Amine level 2	0.06	0.96
NHR level 1	0.64	0.81
NHR level 2	0.65	0.99
NHR level 3	0.62	0.96

✗ Results: for most of the cases the Clus-φ tree stopped growing to soon, what explains the bad results.

We are now investigating new possibilities to define the stop-criterion.

KATHOLIEKE UNIVERSITEIT LEUVEN

DTAI

Contact:
<Eduardo.Costa@cs.kuleuven.be>
<Celine.Vens@cs.kuleuven.be>
<Hendrik.Blockeel@cs.kuleuven.be>