SEQUENCE ANALYSES.
PROTEIN FAMILIES

UCM-CURSO DE VERANO
JULY 2007

Ana M. Rojas
arojas@cnio.es

Outline

- **Background**
  - Domain Shuffling
  - Paralogy vs orthology
  - superfamilies, families & subfmailies
- **Why** we study protein families?
- Some real examples

Some Concepts

- **Homology**: implies an evolutionary relationship

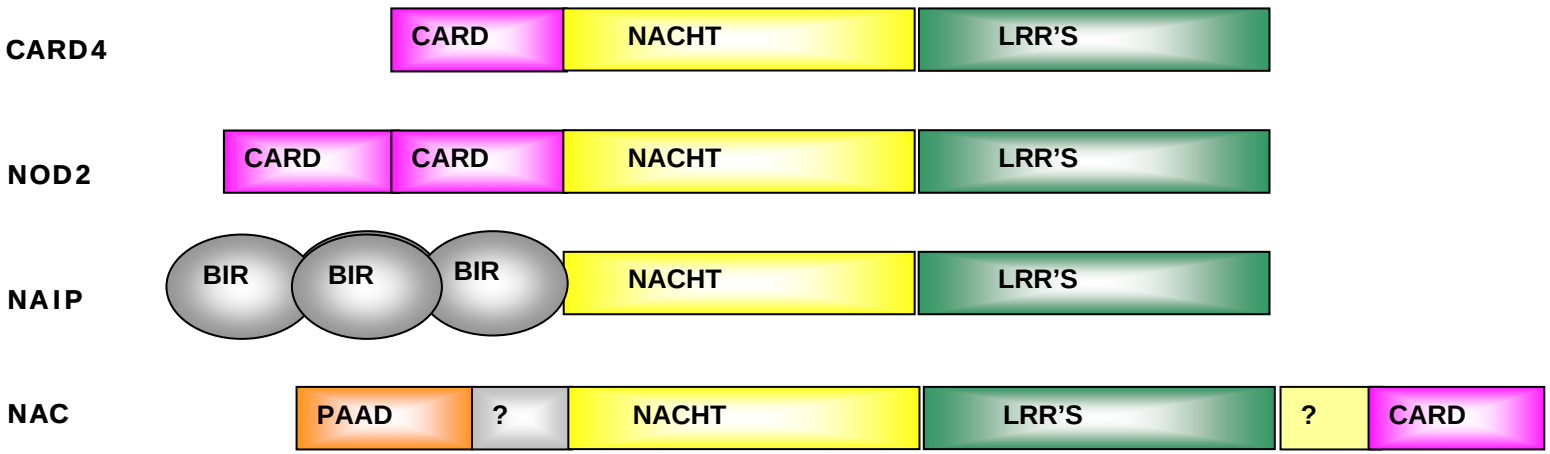
- How protein families appear?

  By domain shuffling
  By Gene duplication

- Technically: proteins sharing the same function are closely related.

Domain shuffling

- Homologues protein can have different domain architectures
- Protein function is a result of the domain individual functions.
- By domain function we can explain certain properties BUT NOT the protein function.

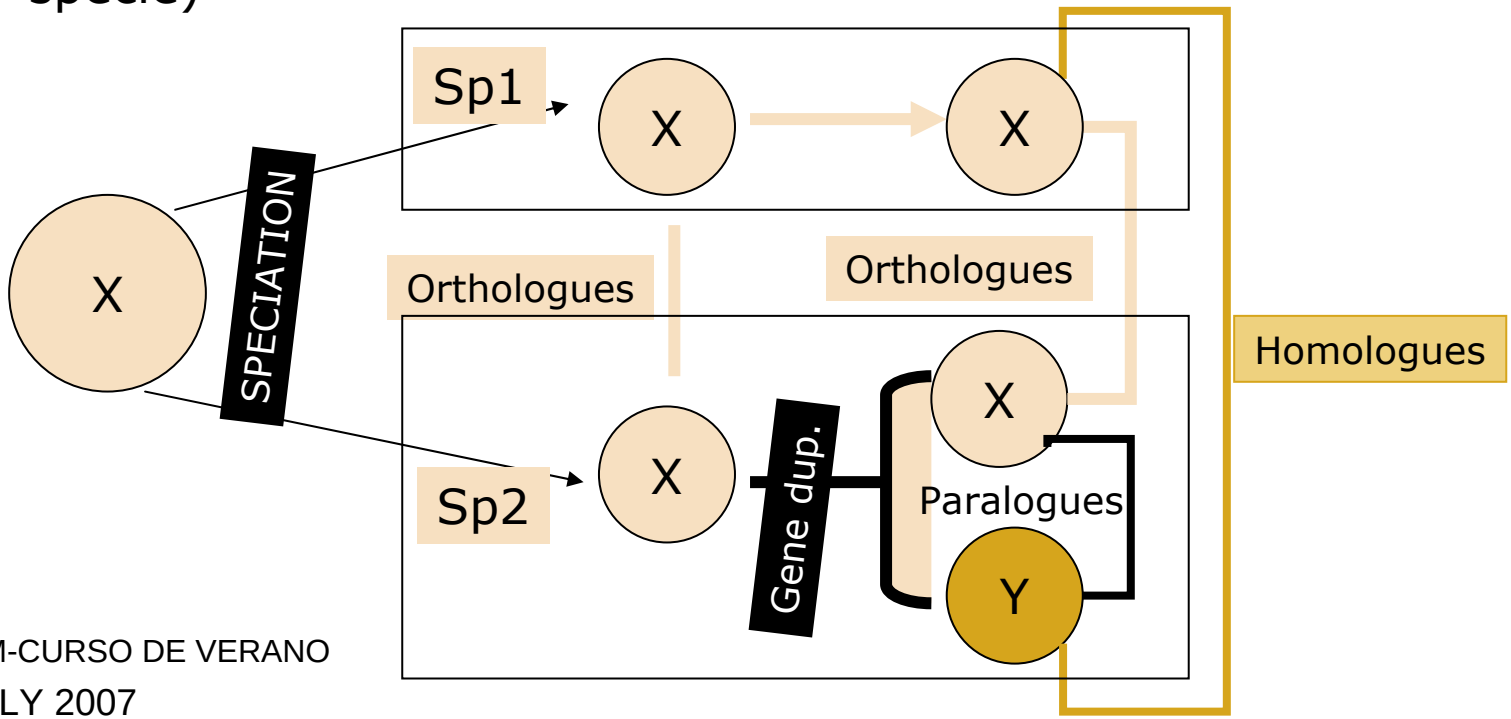


Gene Duplication

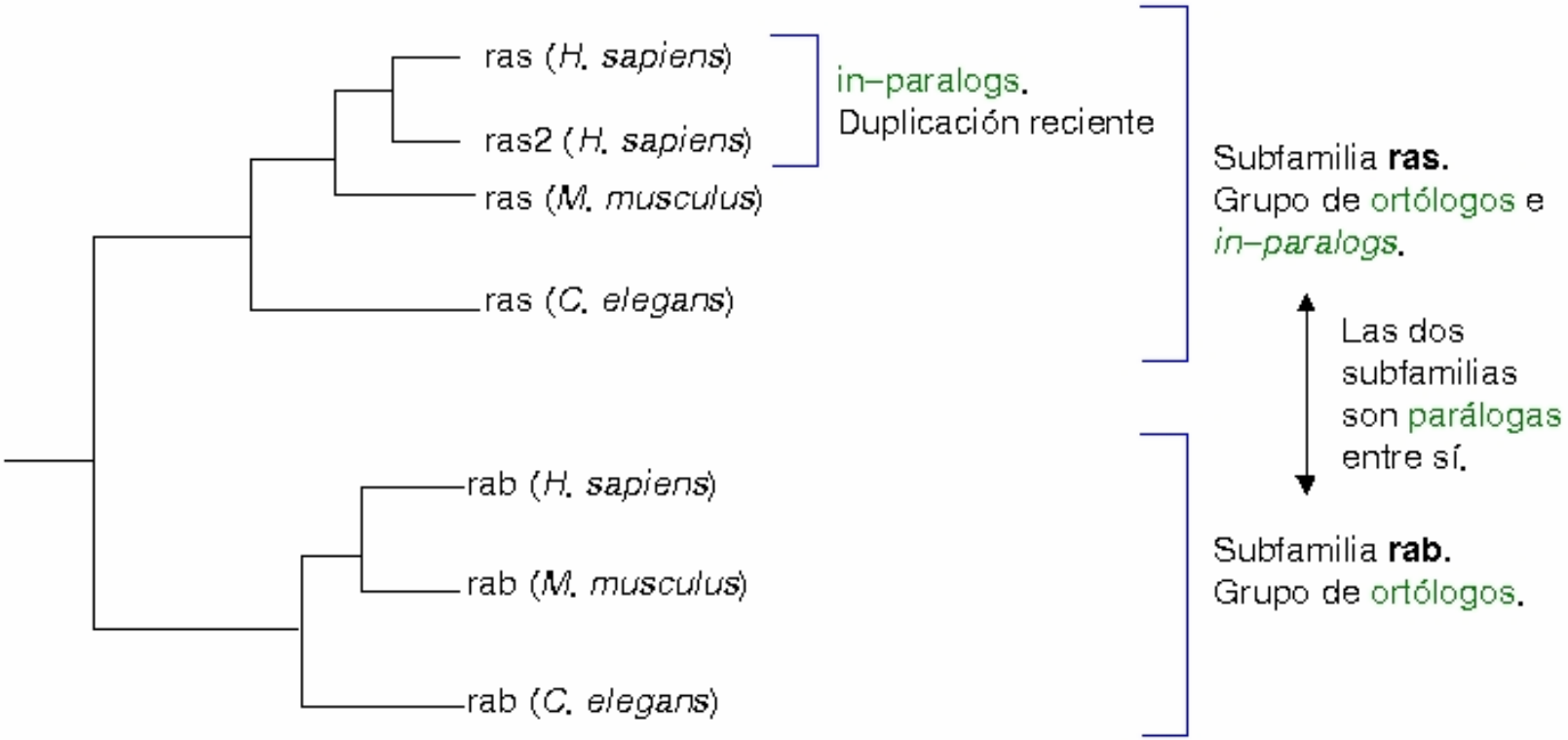
- Homologues protein are **Orthologues or / and paralogues**

- Orthologues: Gene duplication before speciation (same gene in different Species)

- Paralogues: gene duplication after speciation (several genes in the same specie)



Gene Duplication I

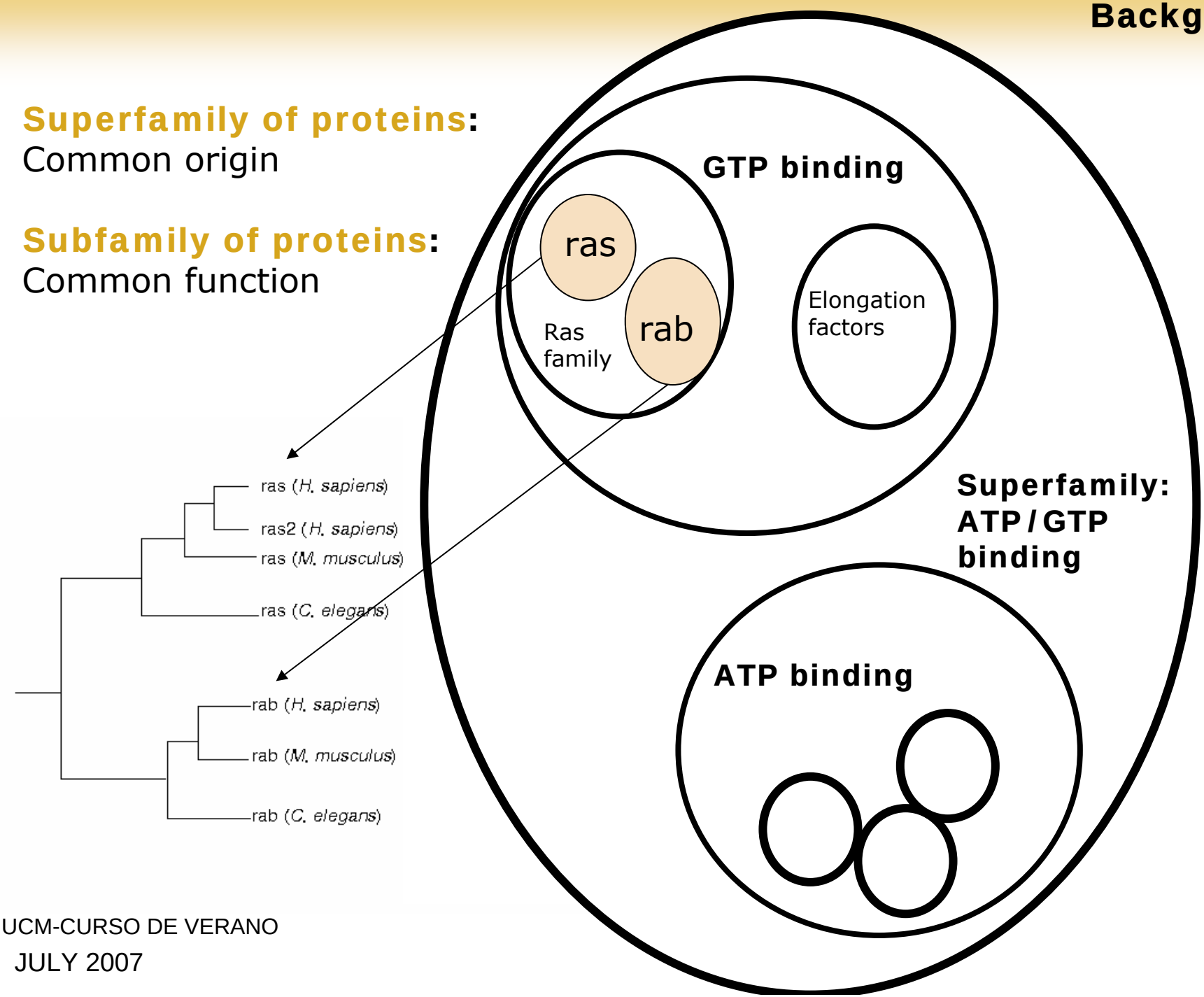


Example: Hp21-elongation factor EF-Tu of Ecoli
General function: signal transduction-protein synthesis
Functional feature: GTP binding

(Taken from F. Abascal)

Superfamily of proteins:
Common origin

Subfamily of proteins:
Common function



Why do we want to study protein families?

Function Prediction
Phylogenetic analyses
Functional specificity

Places where I can find information: Protein classification.

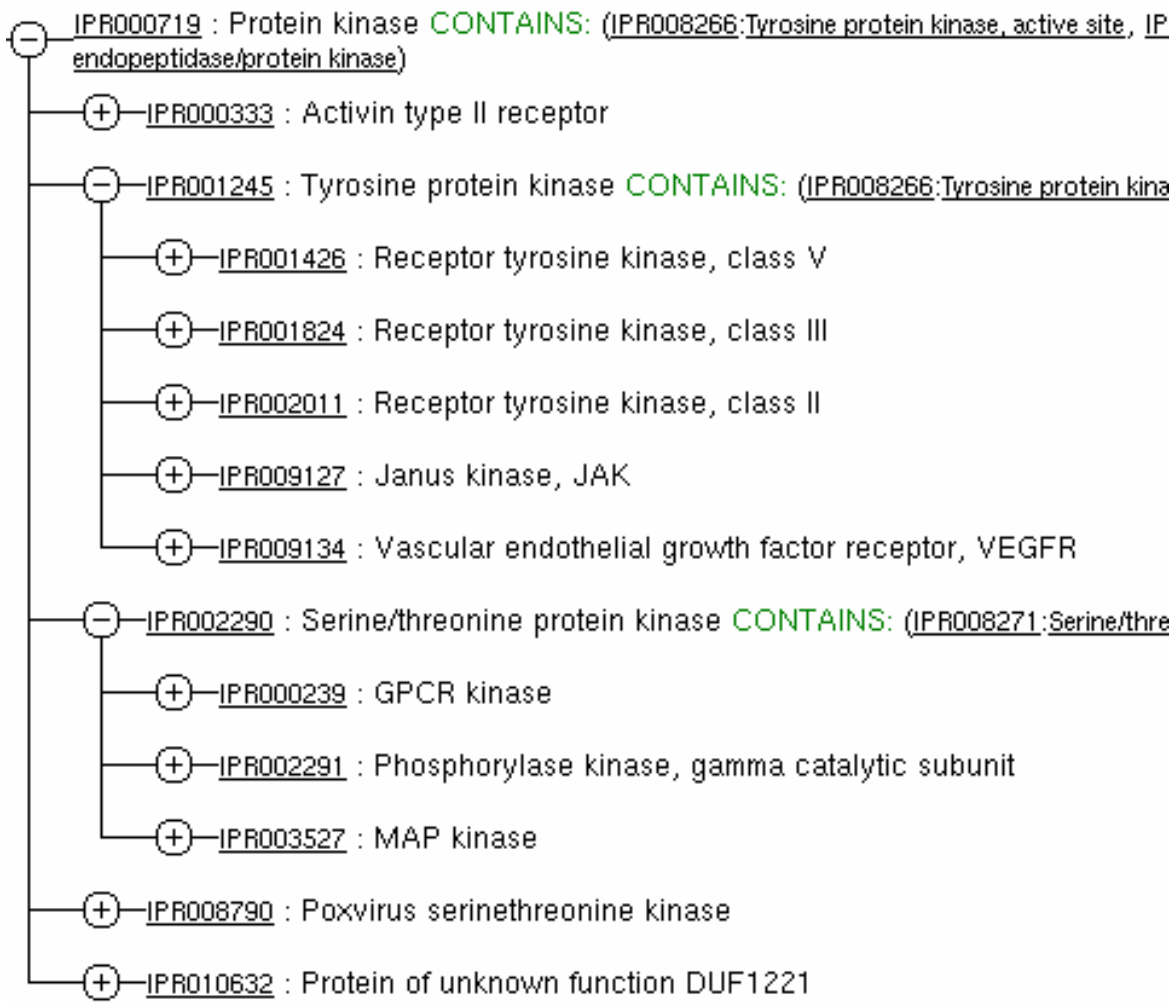
- PROSITE: http://us.expasy.org/prosite

  Motifs, regular expressions (low coverage ~1200 families)

- PFAM: domain database (HMM profiles, high coverage ~7300 fam)
- Interpro: huge information. High coverage, integrates all the DB's.

Database	Version	Entries
SWISS-PROT	42.5	138.992
PRINTS	37.0	1.850
TrEMBL	25.5	1.013.263
Pfam	**11.0**	**7.255**
PROSITE patterns	18.10	1.659
PROSITE preprofiles	N/A	131
ProDom	2002.1	1.021
InterPro	**7.1**	**10.403**
Smart	3.4	654
TIGRFAMs	3.0	1.977
PIR Superfamily	2.3	219
SUPERFAMILY	1.63	552

Interpro


IPR000719 : Protein kinase CONTAINS: (IPR008266:Tyrosine protein kinase, active site, IP
endopeptidase/protein kinase)
IPR000333 : Activin type II receptor
IPR001245 : Tyrosine protein kinase CONTAINS: (IPR008266:Tyrosine protein kina
IPR001426 : Receptor tyrosine kinase, class V
IPR001824 : Receptor tyrosine kinase, class III
IPR002011 : Receptor tyrosine kinase, class II
IPR009127 : Janus kinase, JAK
IPR009134 : Vascular endothelial growth factor receptor, VEGFR
IPR002290 : Serine/threonine protein kinase CONTAINS: (IPR008271:Serine/thre
IPR000239 : GPCR kinase
IPR002291 : Phosphorylase kinase, gamma catalytic subunit
IPR003527 : MAP kinase
IPR008790 : Poxvirus serinethreonine kinase
IPR010632 : Protein of unknown function DUF1221

Automatic methods to classify proteins: ProtoMap

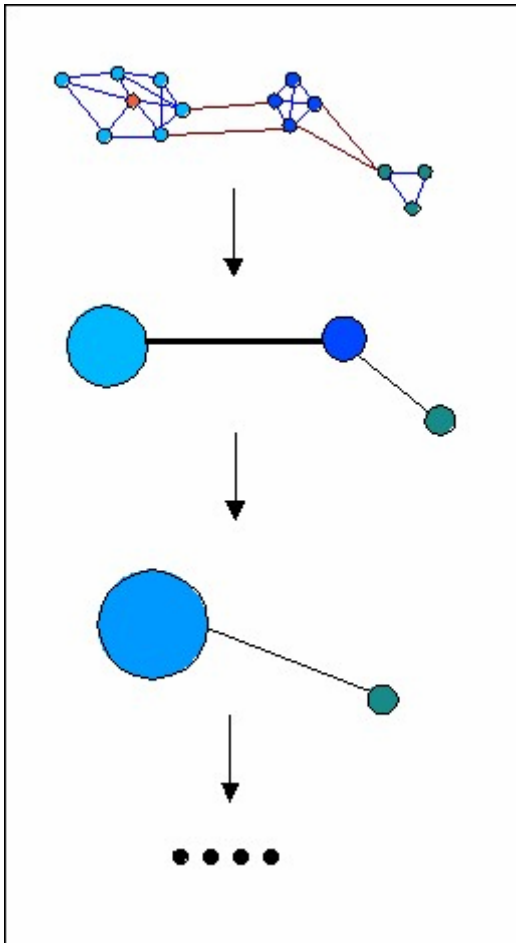
Where?

Based on **sequence distance**
Search: FASTA, BLAST…for each protein of SwissProt+ Trembl

Graph: nodes are prots, edges (weighted by e-value)

Clustering algorithm to find the groups.

Problem? **Domains!**



CLUSTERING ALGORITHM

0º.- Get sequence distances=> graph
1º.- Grouping close related sequences (e-value < 1e-100)
2º.- Initialise T = 1e-95.
3º.- Computing cluster distances:
geometrical mean of e-values between each cluster pair.
If no edges: assignment of e-value=1

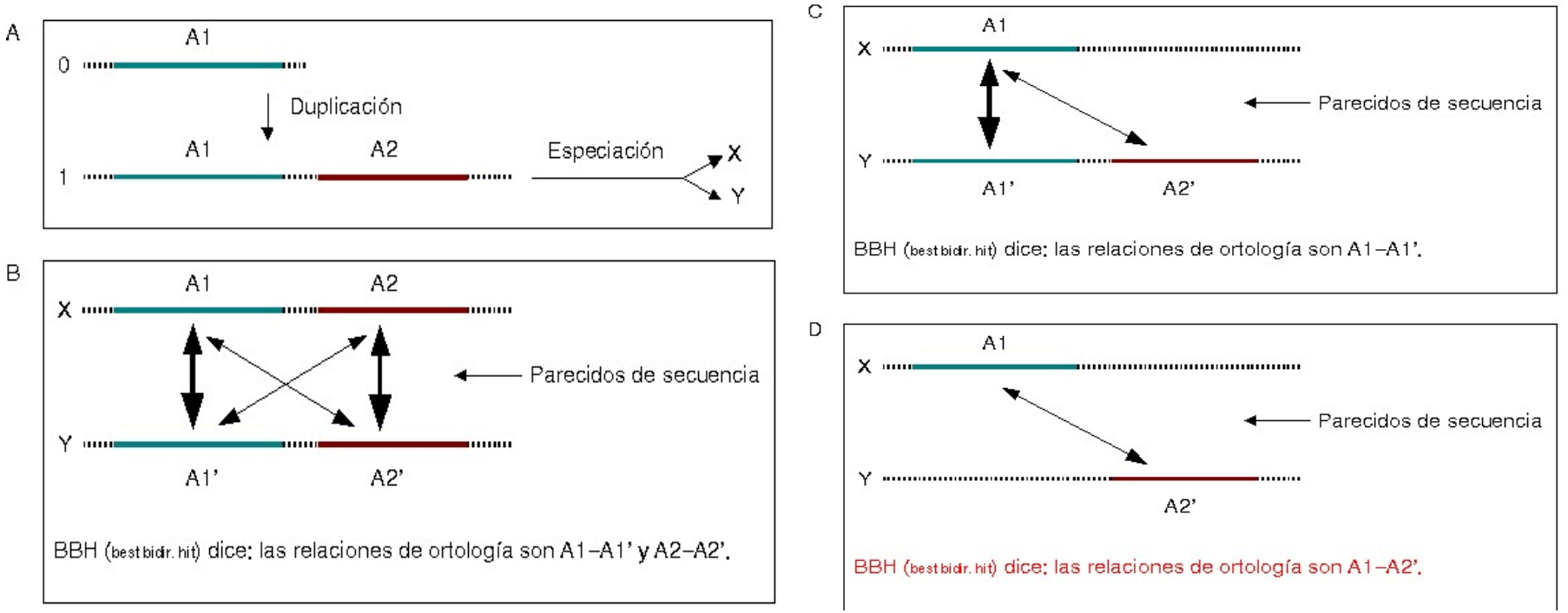
4º.- If the e-values mean is lower than rootsquare of T, clusters are joined.

5º.- Decrease the T value T: T = T*1e+05.
6º.- If T > 1 => stop. Else => go back to 3º.

Sequential implementation of T values (1e-95 -> 1e-90 -> 1e-85 ... 1e-00=1) allows a hierarchical classification of the proteins.

Automatic methods to classify proteins: GOGS Where?

Based on **B**e**T**s (best bidirectional hits-> best similarity in both directions)

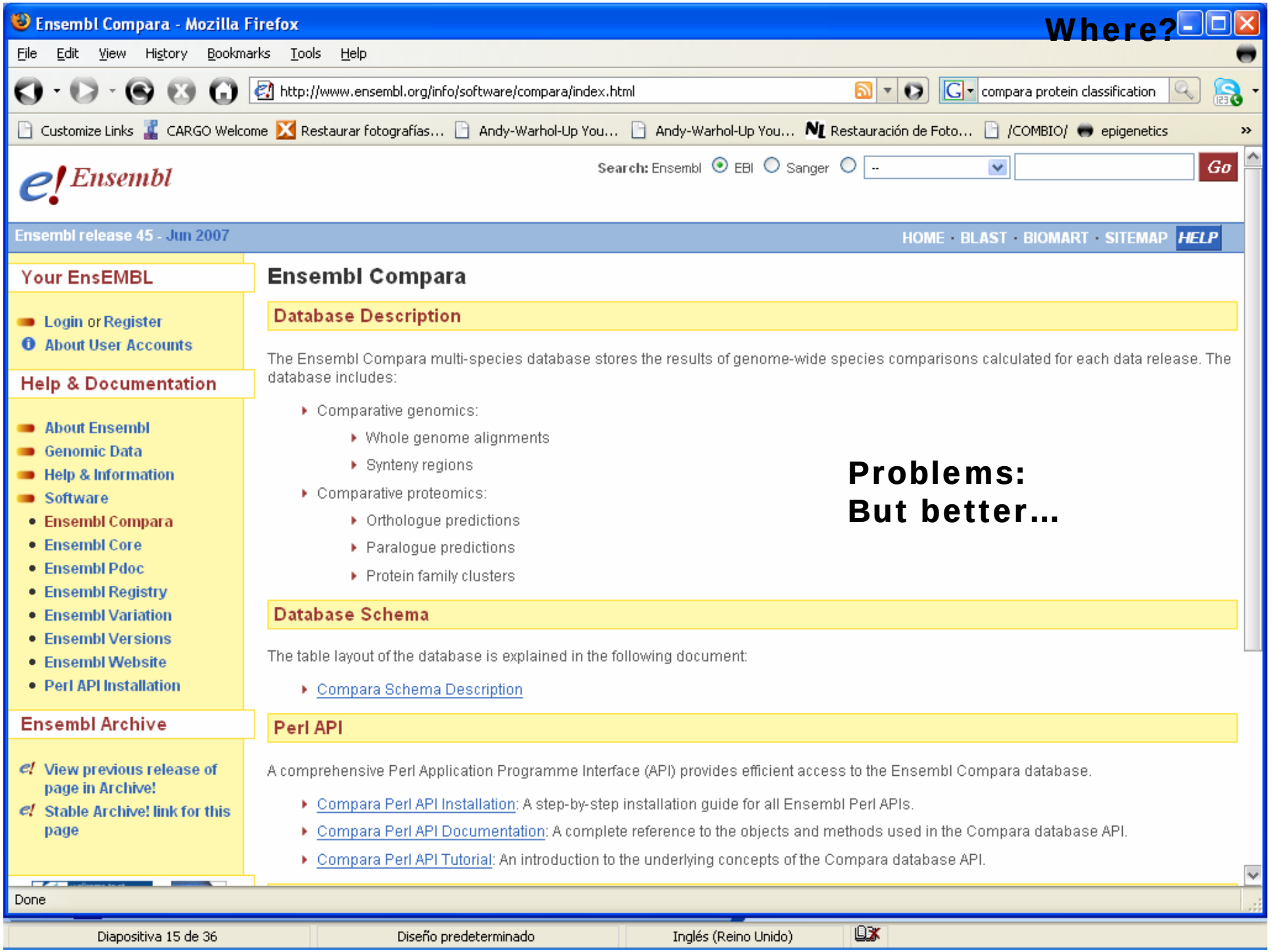


Paralogues fusion!

Problem? **Domains!**

Complete genomes


Inparanoid: Eukaryotic Ortholog Groups - Mozilla Firefox
Where?
File Edit View History Bookmarks Tools Help
http://inparanoid.sbc.su.se/
inparanois
Customize Links CARGO Welcome Restaurar fotografías... Andy-Warhol-Up You... Andy-Warhol-Up You... Restauración de Foto... /COMBIO/ epigenetics
Inparanoid
Home
All species vs all
Human vs all
Gene search
Text search
Blast search
Downloads
Old version
Summary
FAQ
Stockholm Bioinformatics Centre SBC
InParanoid: Eukaryotic Ortholog Groups
26 organisms: 509,483 sequences
Version 5.1, Updated January 2007
BROWSE the database - Select two species and view all their orthologs
BROWSE the human results - Select a species to compare against Human
SEARCH BY SEQUENCE IDs - View orthologs of a specific gene or protein
TEXT SEARCH - Query InParanoid by keywords
BLAST SEARCH - Find orthologs in InParanoid similar to your protein sequence
REQUEST A PAIRWISE GENOME COMPARISON
Problems:
Only pairs of seqs.
Close species
Done


Ensembl Compara - Mozilla Firefox
Where?
File Edit View History Bookmarks Tools Help
http://www.ensembl.org/info/software/compara/index.html
compara protein classification
Customize Links CARGO Welcome Restaurar fotografías... Andy-Warhol-Up You... Andy-Warhol-Up You... Restauración de Foto... /COMBIO/ epigenetics
Ensembl
Search: Ensembl EBI Sanger -- Go
Ensembl release 45 - Jun 2007
HOME · BLAST · BIOMART · SITEMAP HELP
Your EnsEMBL
Login or Register
About User Accounts
Help & Documentation
About Ensembl
Genomic Data
Help & Information
Software
Ensembl Compara
Ensembl Core
Ensembl Pdoc
Ensembl Registry
Ensembl Variation
Ensembl Versions
Ensembl Website
Perl API Installation
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Ensembl Compara
Database Description
The Ensembl Compara multi-species database stores the results of genome-wide species comparisons calculated for each data release. The database includes:
Comparative genomics:
Whole genome alignments
Synteny regions
Comparative proteomics:
Orthologue predictions
Paralogue predictions
Protein family clusters
Problems:
But better...
Database Schema
The table layout of the database is explained in the following document:
Compara Schema Description
Perl API
A comprehensive Perl Application Programme Interface (API) provides efficient access to the Ensembl Compara database.
Compara Perl API Installation: A step-by-step installation guide for all Ensembl Perl APIs.
Compara Perl API Documentation: A complete reference to the objects and methods used in the Compara database API.
Compara Perl API Tutorial: An introduction to the underlying concepts of the Compara database API.
Done
Diapositiva 15 de 36
Diseño predeterminado
Inglés (Reino Unido)

- How do I represent my protein family?

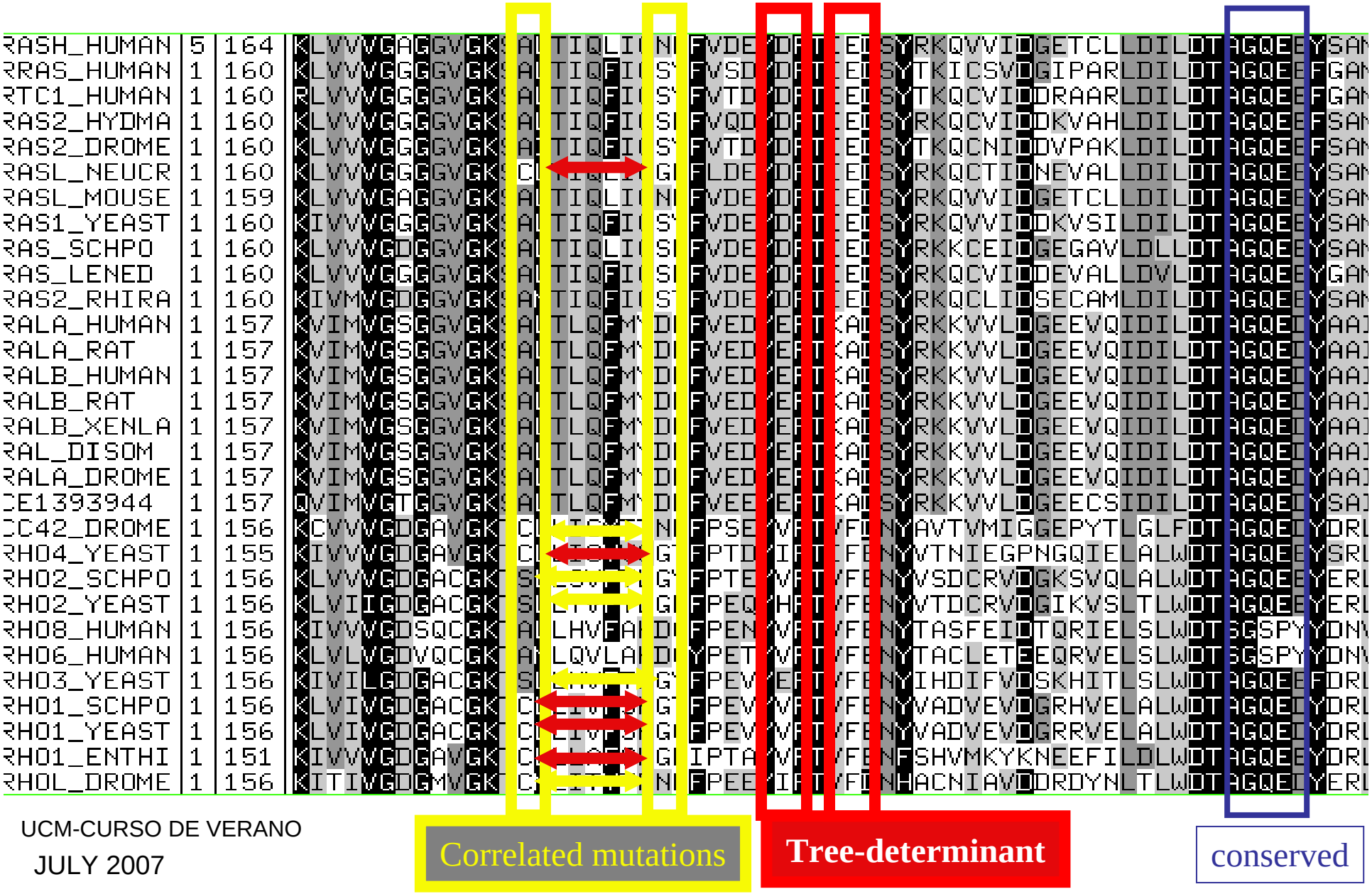
  By a multiple sequence alignment

- How do I align my sequences? Very important.

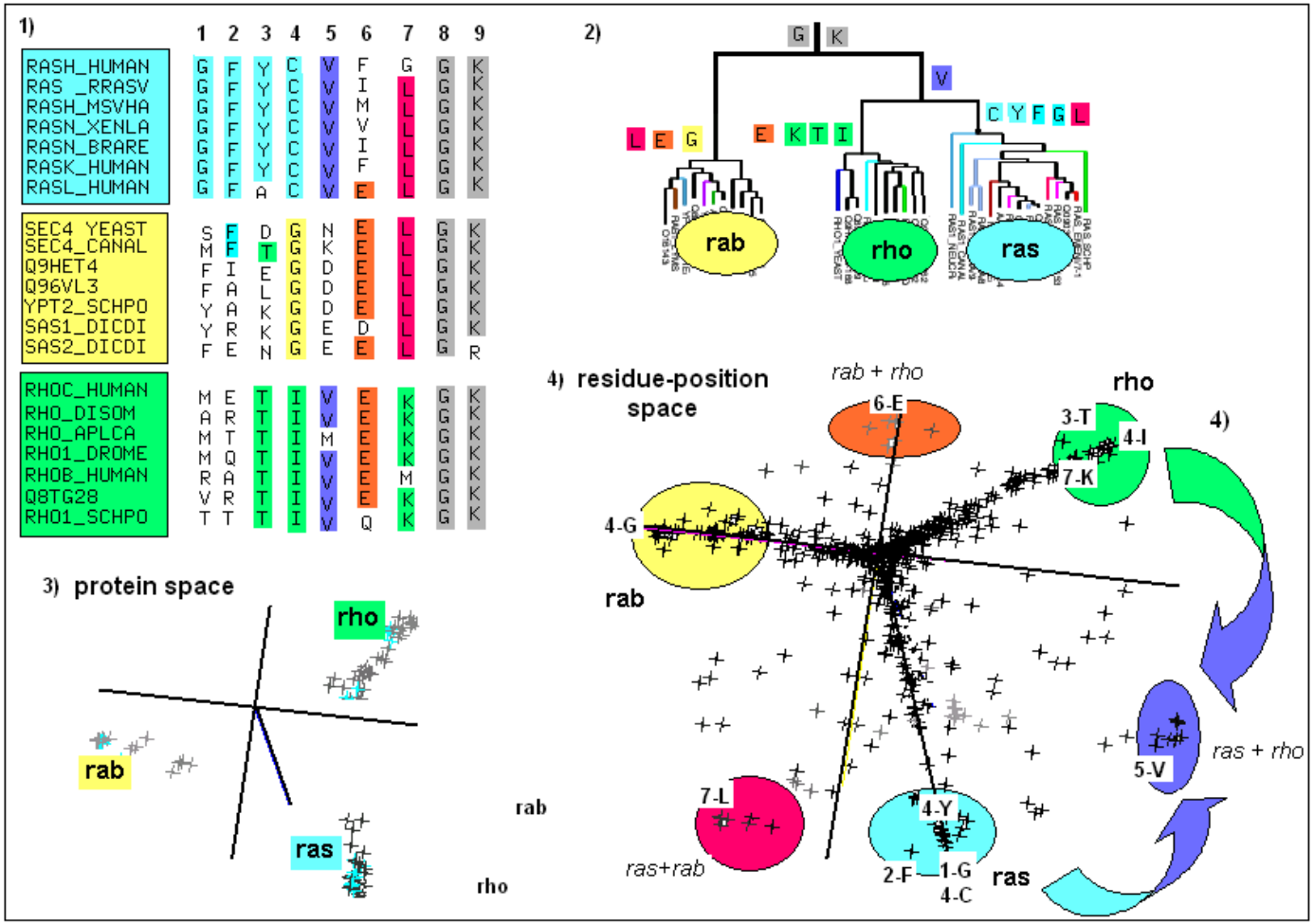
  Probcons/Muscle/T-Coffee/ClustalW …

- WHAT CAN I GET FROM MY ALIGNMENTS?

  A profile to do sensitive searches…
  A distance matrix to analyse trees
  Important conserved residues (structures)
  Important trends within subfamilies (specificity)
  Important residues indicating co-evolution


Correlated mutations
Tree-determinant
conserved

WHAT CAN I LEARN FROM MY ALIGNMENT?



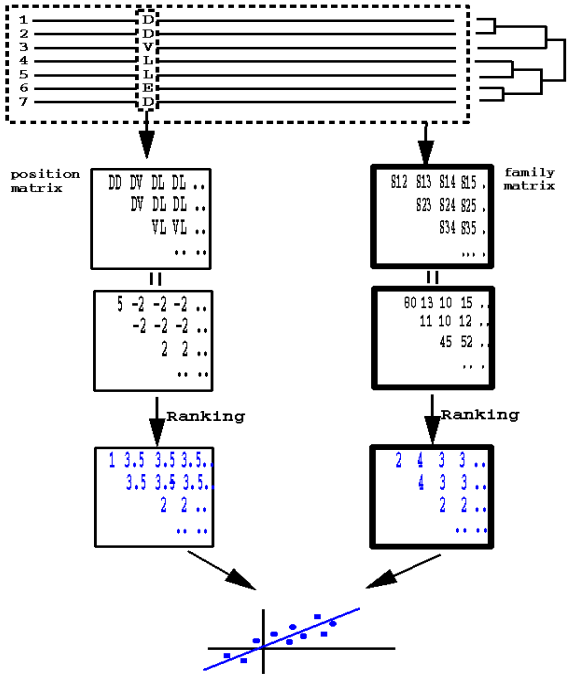
Casari, Sander, Valencia Nature Str. Biol. 95

Pazos, Valencia 2003

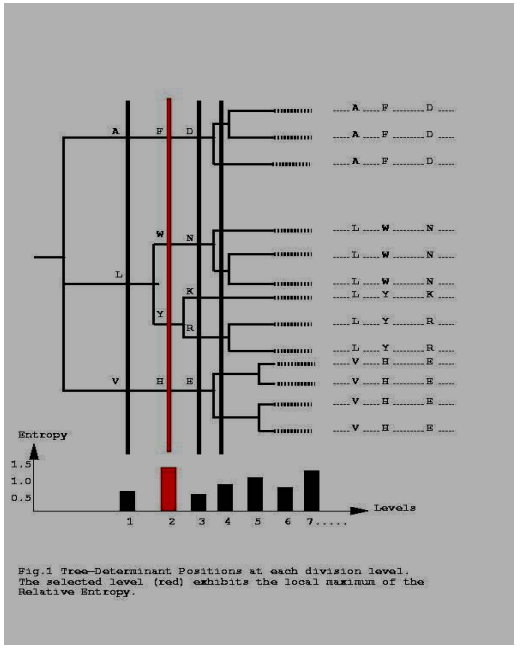
Romero, Valencia 04

WHAT CAN I LEARN FROM MY ALIGNMENT?

How to's?



Mutational behaviour
Pazos Valencia, 2001



$$H(n) = \sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log\frac{p(x_1,x_2,\ldots,x_n)}{\prod_{i=1}^{n} p(x_i)}$$

Relative entropy cut,
del Sol, Valencia 2002

Del Sol, Pazos, Valencia JMB 03

THE PROBLEM OF THE EUKARYA LINEAGE

WHAT TO DO THEN?

DOMAIN ANALYSES

CHECK CONSISTENCY BETWEEN DOMAIN DISTRIBUTION AND PHYLOGENETIC DISTRIBUTION

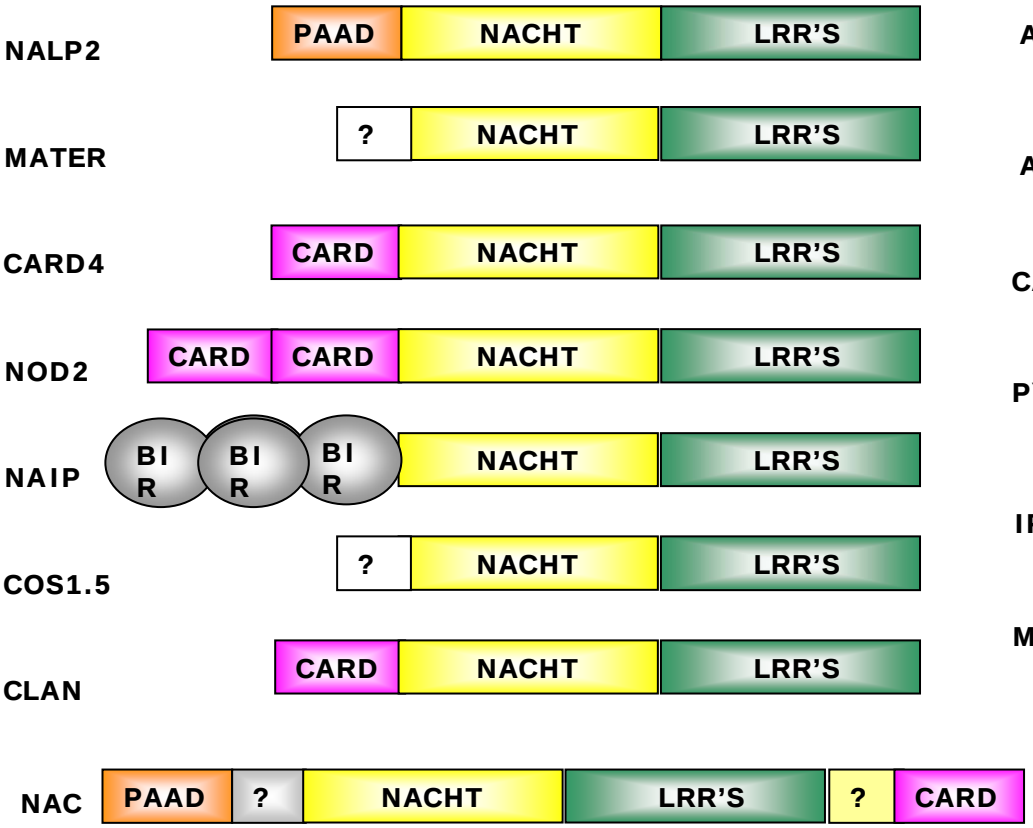
CHECK IF SHUFFLING IS RECENT OR OLD…

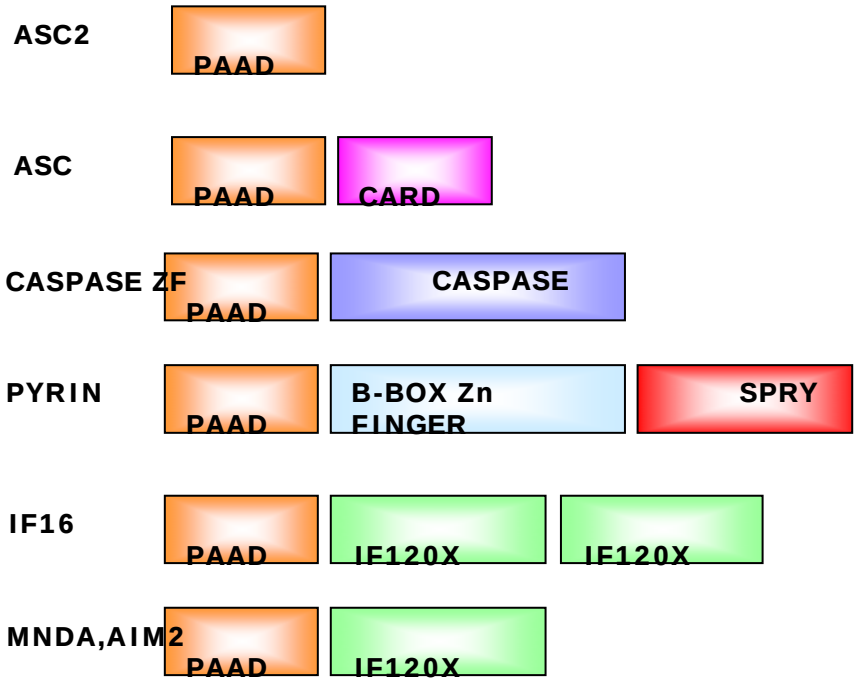
{Reed et al, Gen Res. 2003}

Case 1: domain shuffling

Get's real!

DOMAIN ARCHITECTURES





NACHT FAMILY

PAAD FAMILY

{Reed et al, Gen Res. 2003}

What are these PAAD and NACHT proteins?

They are involved in inflammation and apoptosis!!!!!

Nacht family: PAN/NALPs/DEFCAP/PYCARD, CATERPILLER
(Tschopp et al, Nature, 2003)

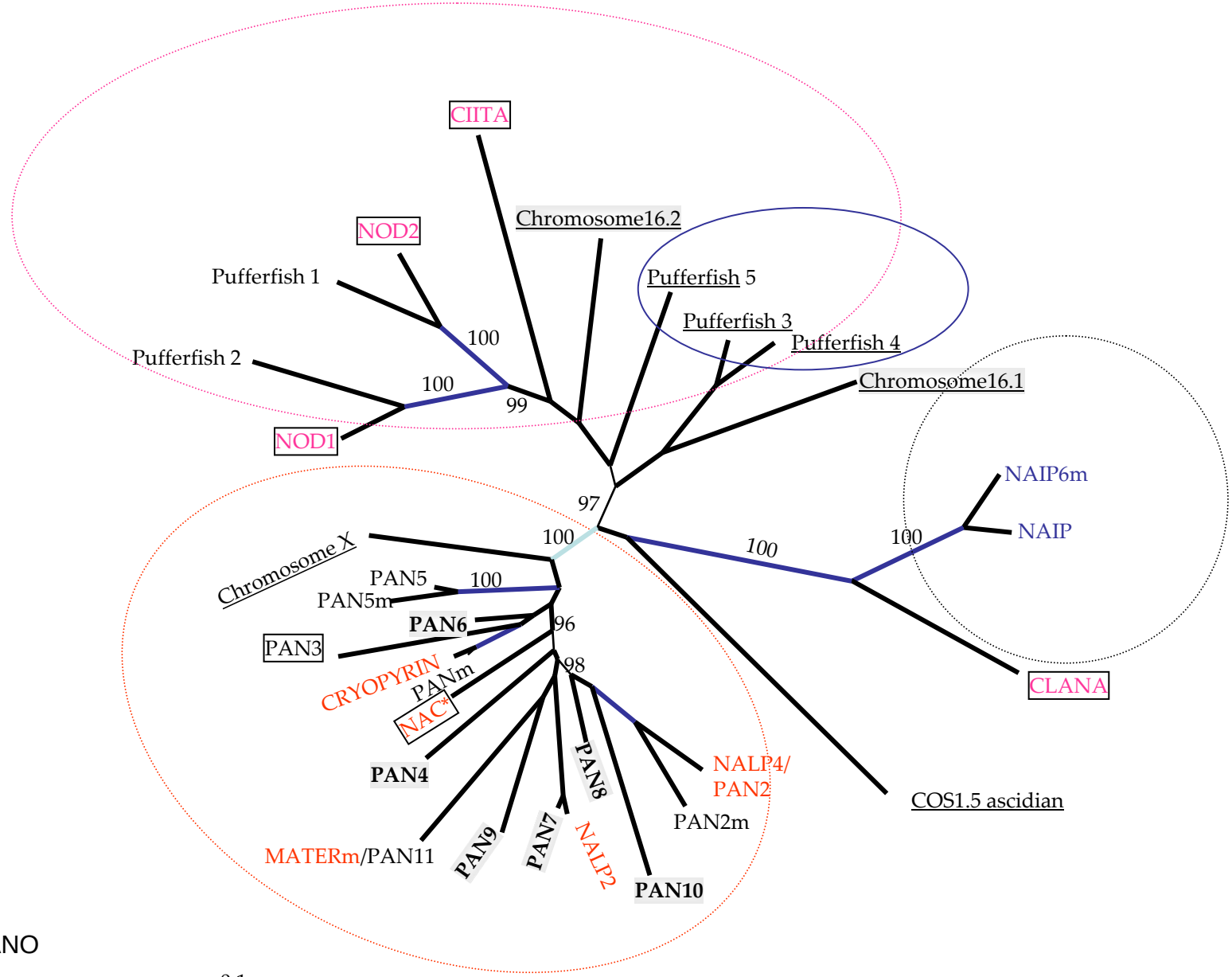
PAAD family: MEFV/PYRIN (Pawlowski, et.al., 2001 , others)

{Reed et al, Gen Res. 2003}

Case 1: domain shuffling

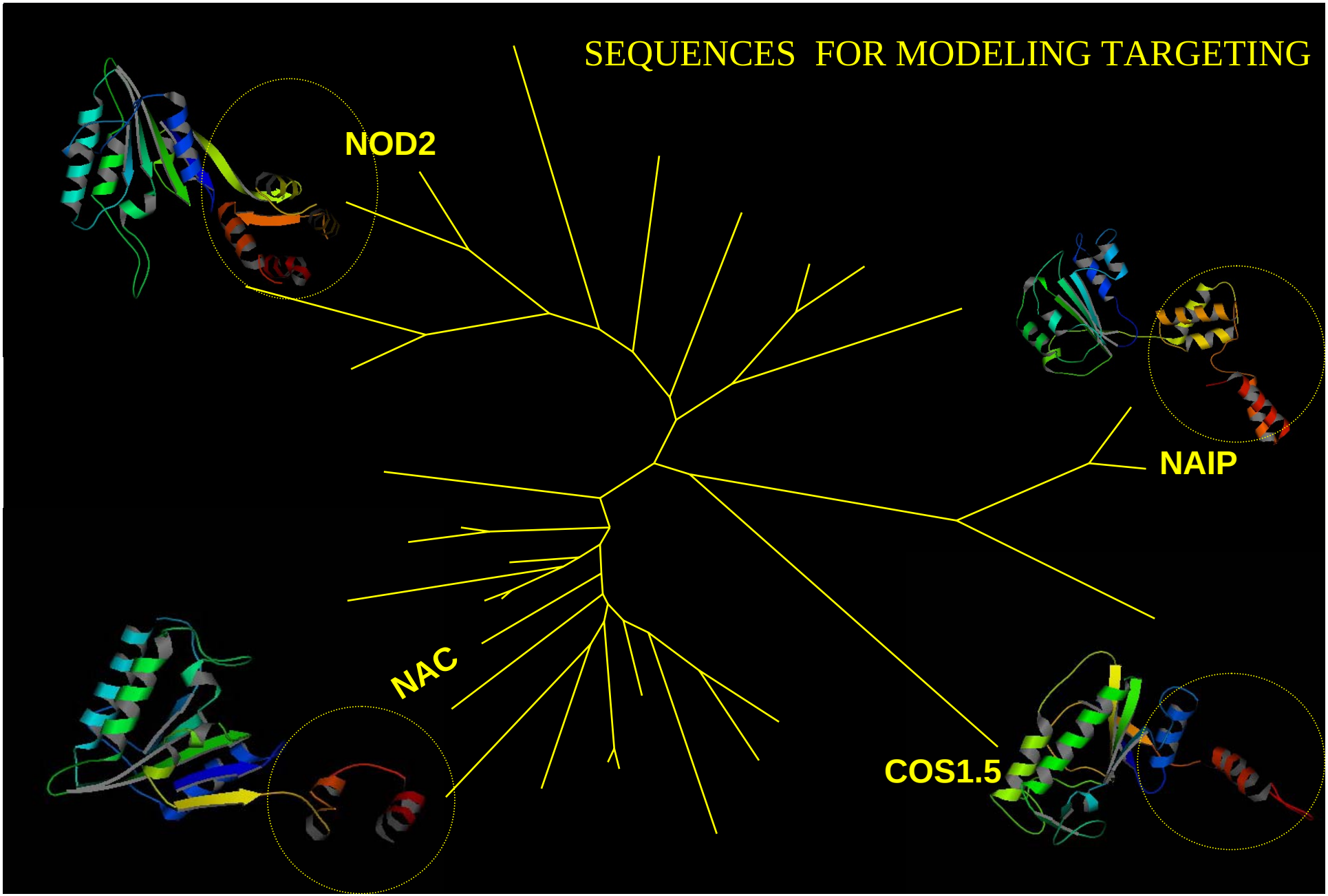
Get's real!

Phylogenetic tree of the NACHT family of proteins based on the NACHT domain.



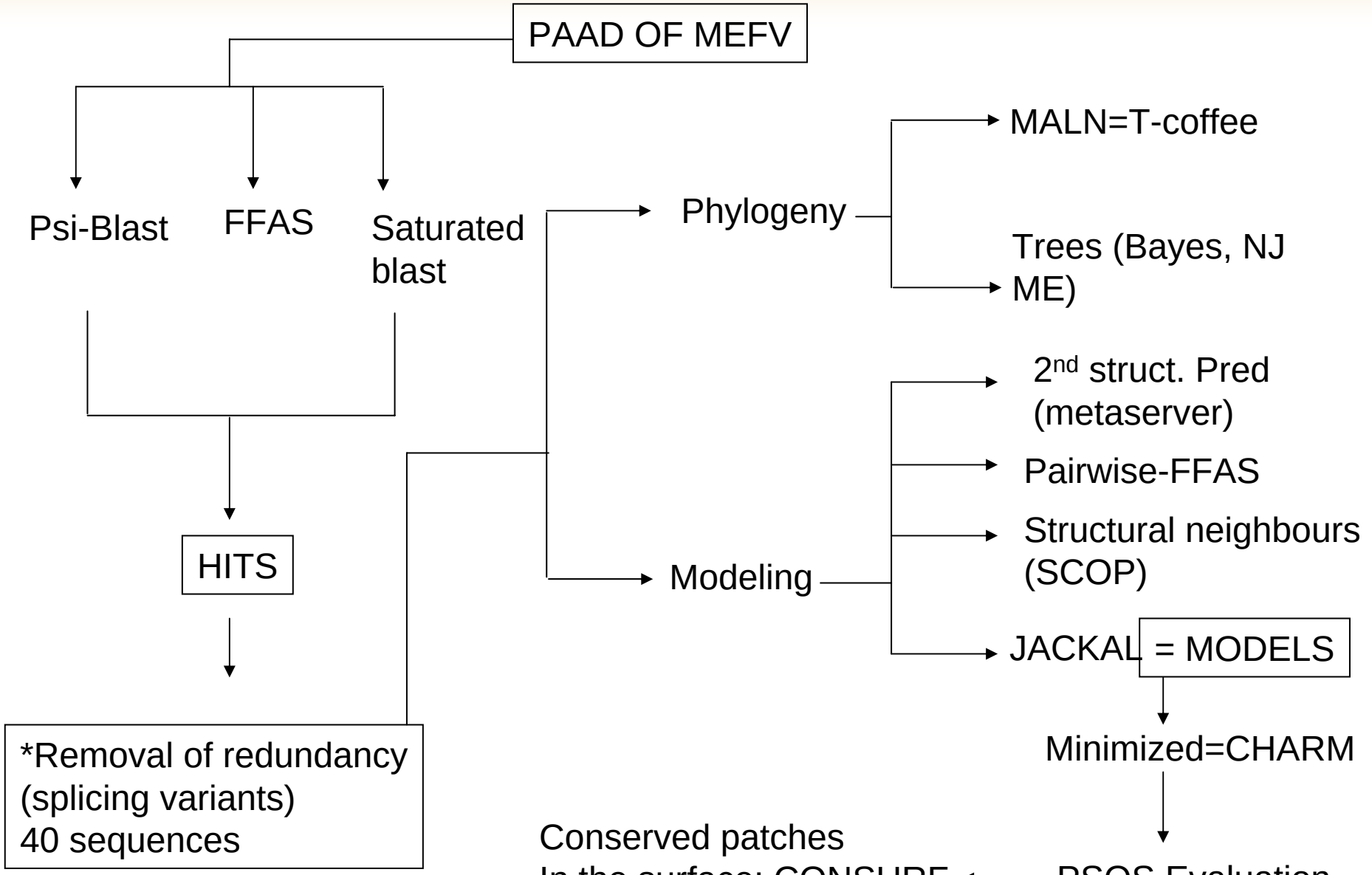
Case 1: domain shuffling

Get's real!



Case 1: domain shuffling

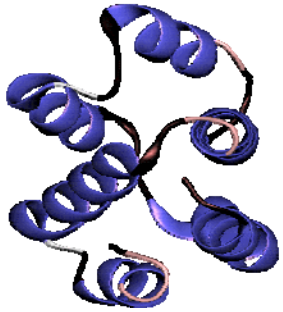
Get's real!



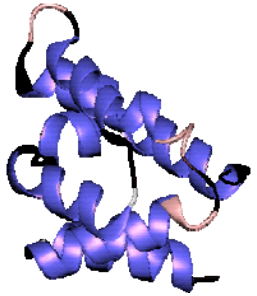
{Reed et al, Gen Res. 2003}

ANCESTORAL DOMAIN

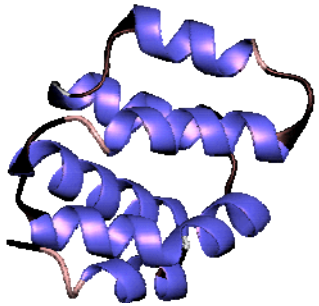
Get's real!



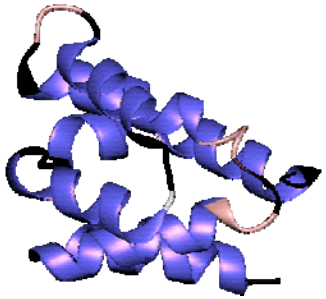
PAAD



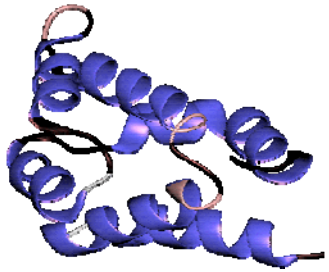
CARD



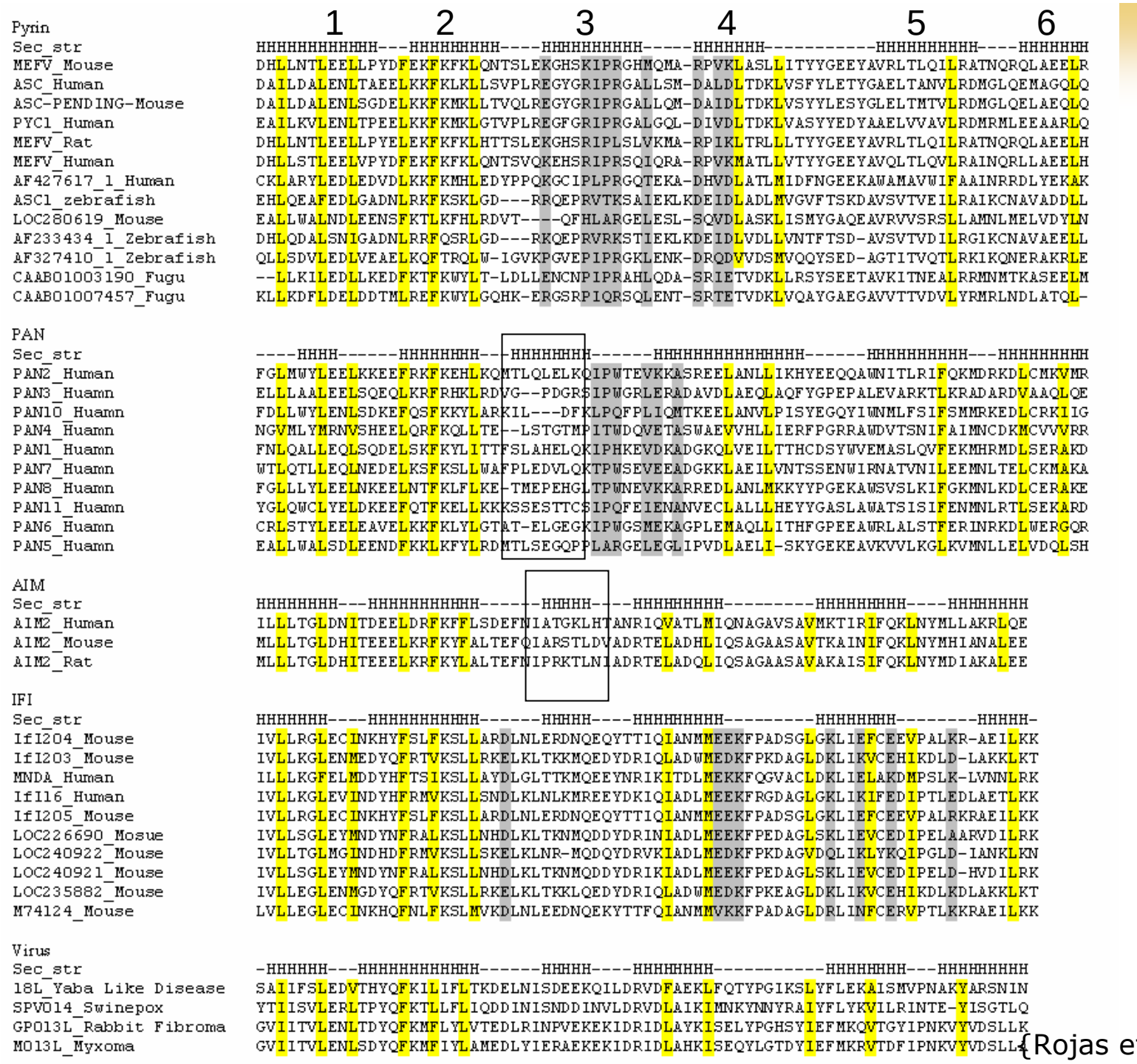
DD



DED



{Reed et al, Gen Res. 2003}

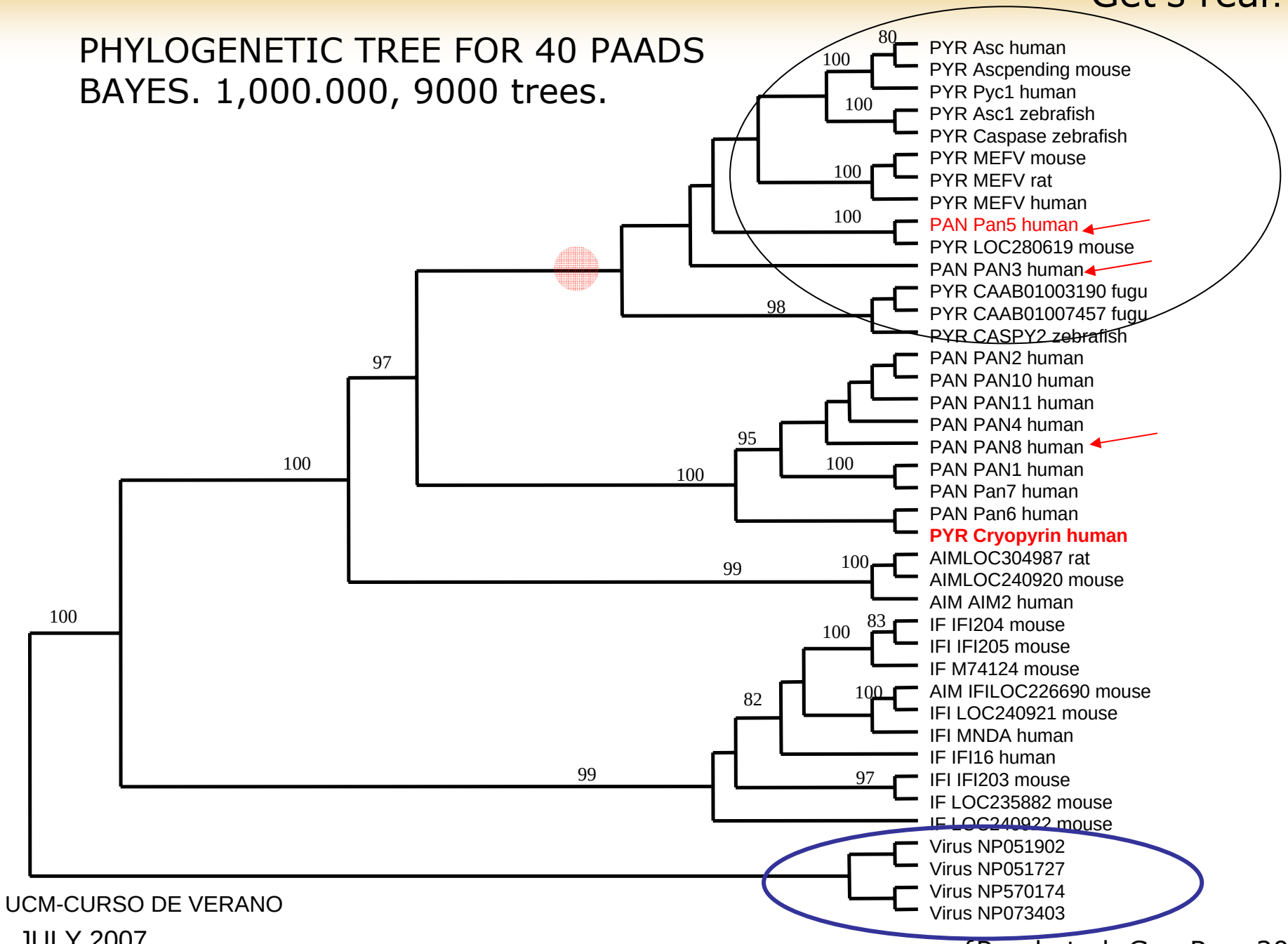


Get's real!

{Rojas et al, Prot. Sci. 2003}

Get's real!

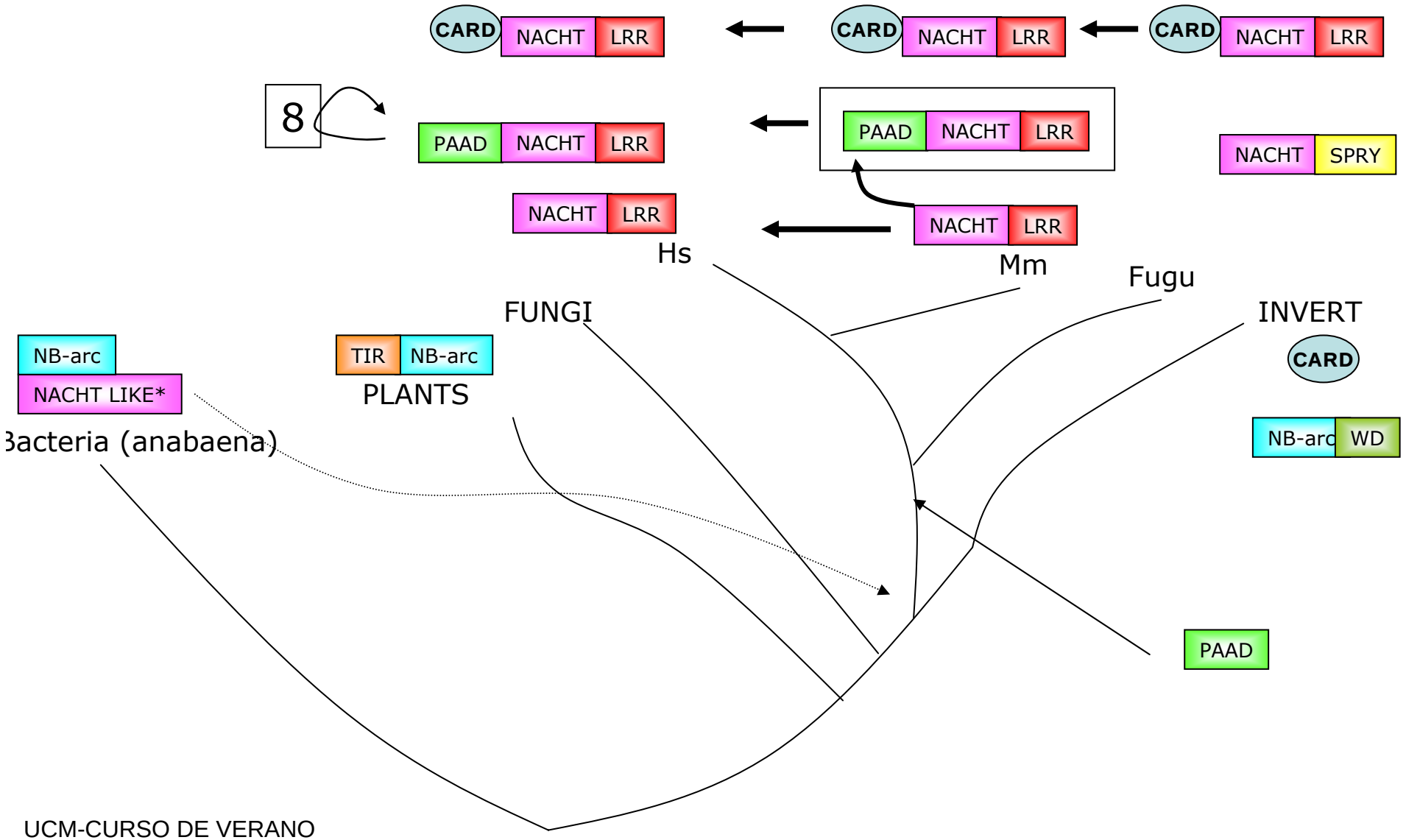
PHYLOGENETIC TREE FOR 40 PAADS
BAYES. 1,000.000, 9000 trees.



{Reed et al, Gen Res. 2003}

Case 1: domain shuffling

NACHT DISTRIBUTION: POSSIBLE SCENARIO



Case 2: novel species

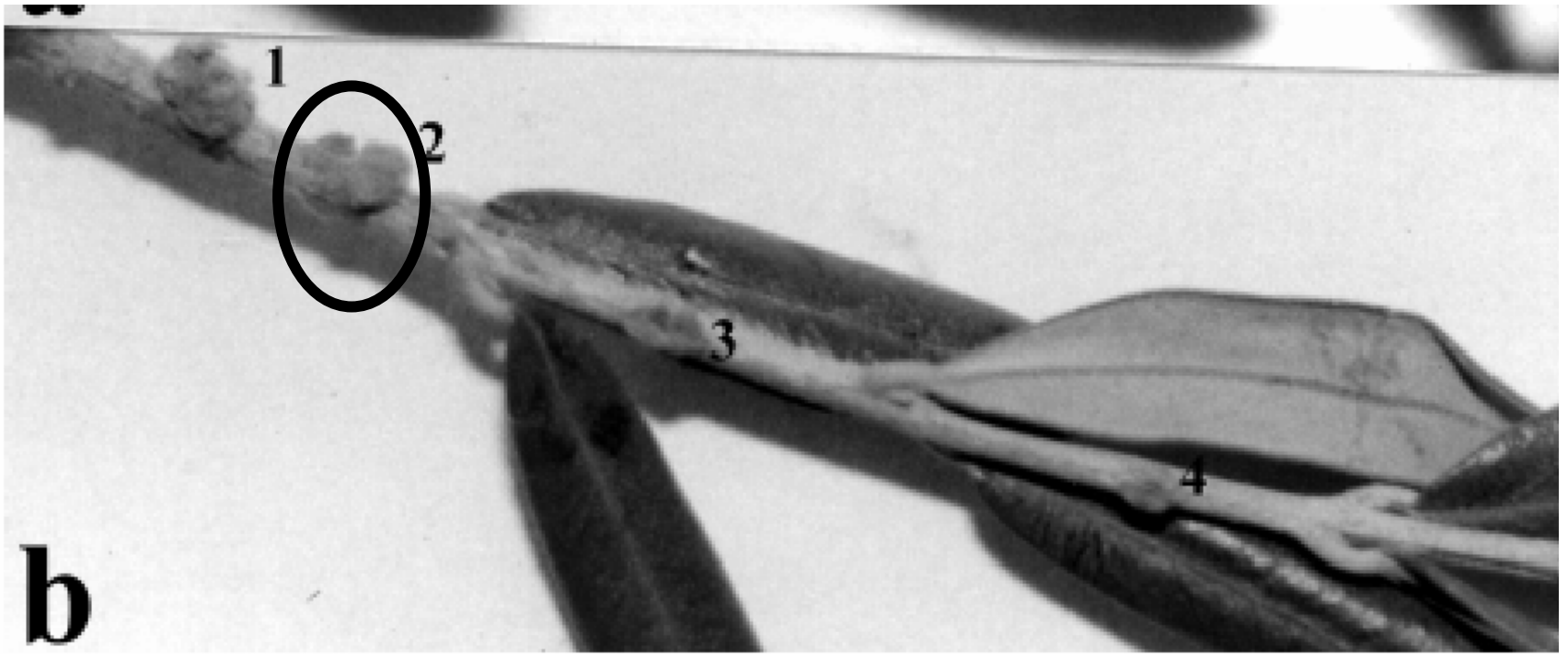
- DESCRIPTION OF NEW SPECIES
  Erwinia toletana sp. nov.

{Rojas et al, IJESM. 2002}

Case 2: novel species

Goal: to obtain a natural antagonist of *P. savastanoi*.

Data: Bacterial species isolated from wild trees' knots (Olives, oleander…)



total of 81 bacterial strains!

{Rojas et al, IJESM. 2002}

Case 2: novel species

The problem: Resemble phenotypically to several…

What to do?:

- Choose an universal conserved marker: i.e. 16SRNA

Extract similar sequences

Build phylogenetic trees

Gene sequencing:

16SRNA, 23SRNA, gnd, mdh

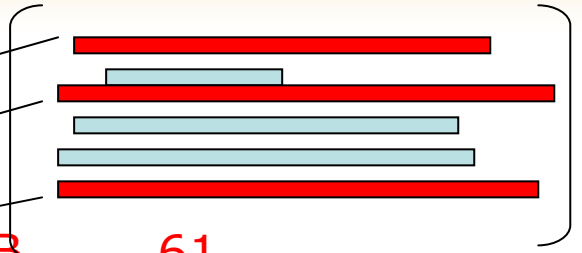
WHY THESE GENES? ?????????

{Rojas et al, IJESM. 2002}

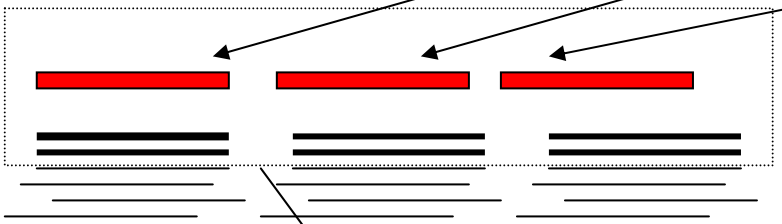
Case 2: novel species

METHOD FOR 16SRNA

From 81 sequences only the longest retained (61)



stand-alone blasted against a **filtered** EMBL DB



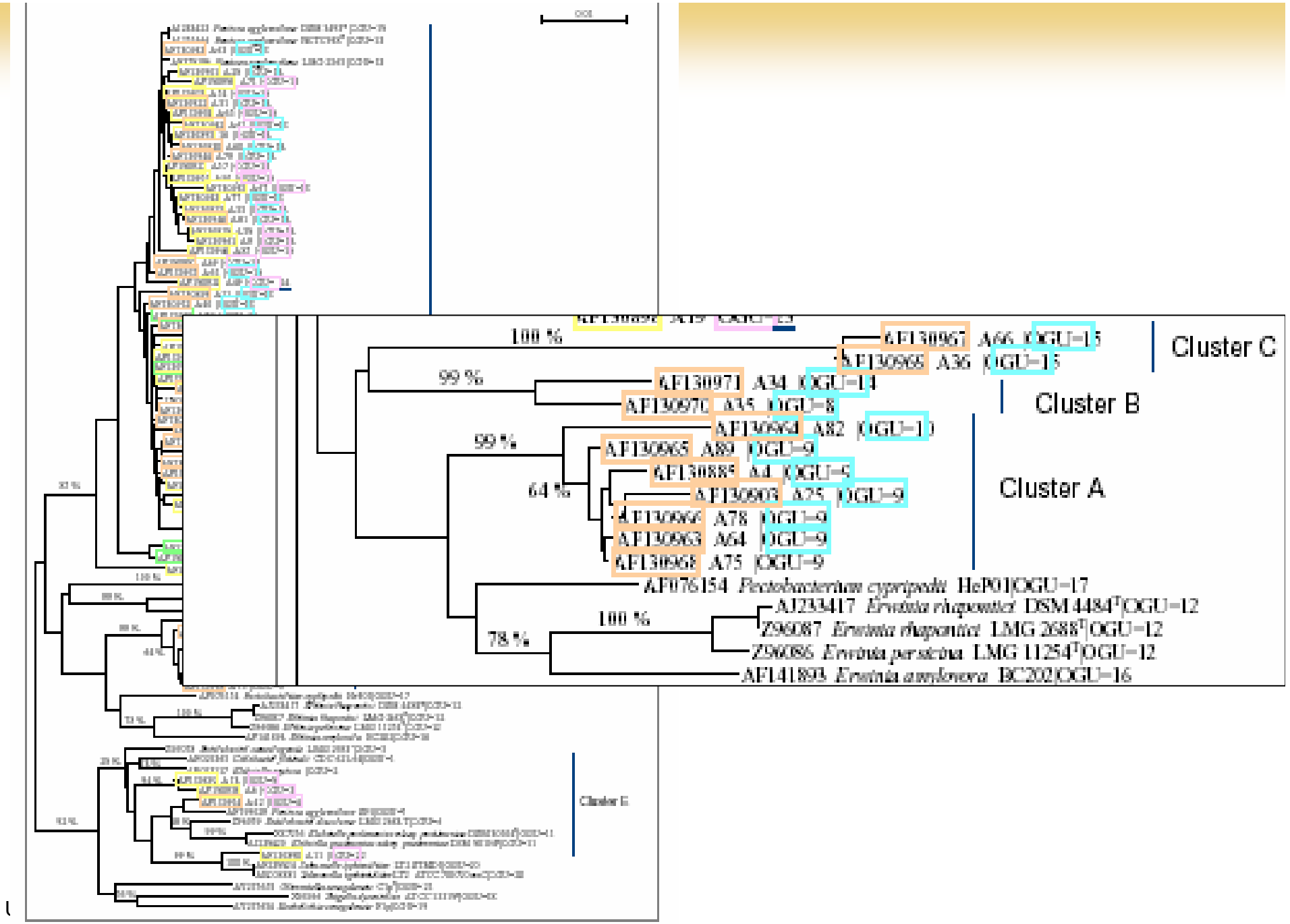
A total of 19,184 sequences retained (from 80,807 initial sequences). .

The 2 most similar are retained to phylogenetic tree reconstruction

Parsimony Maximum likelihood BioNJ 1000 boostrap

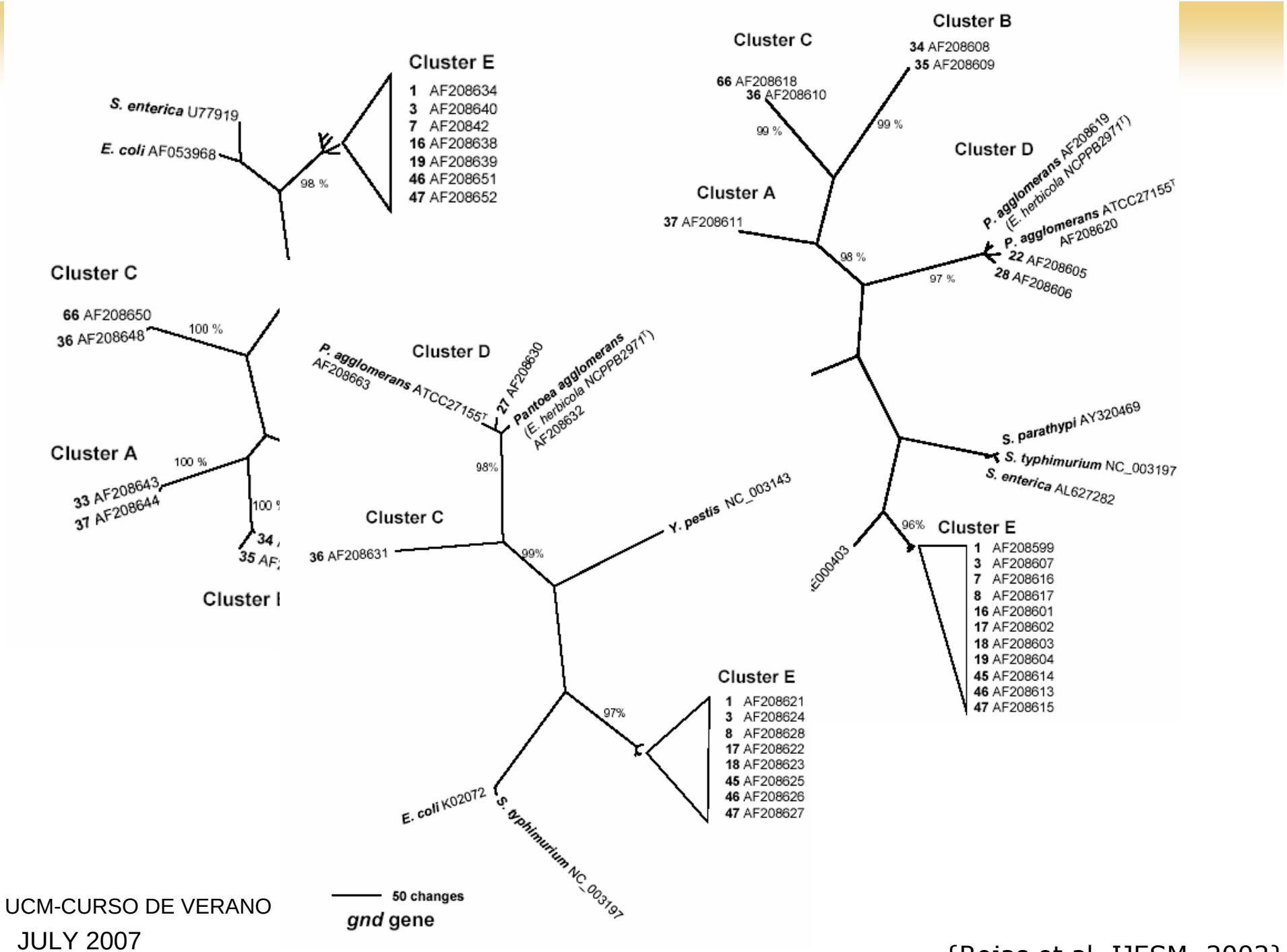
CONSENSUS!

{Rojas et al, IJESM. 2002}



JULY 2007

{Rojas et al, IJESM. 2002}



UCM-CURSO DE VERANO
JULY 2007

{Rojas et al, IJESM. 2002}

Case 2: Extending the family

- PLACEMENT OF NEW ISOLATED GENES
  Ocurrence of serin proteases in sponge and jellyfish

{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

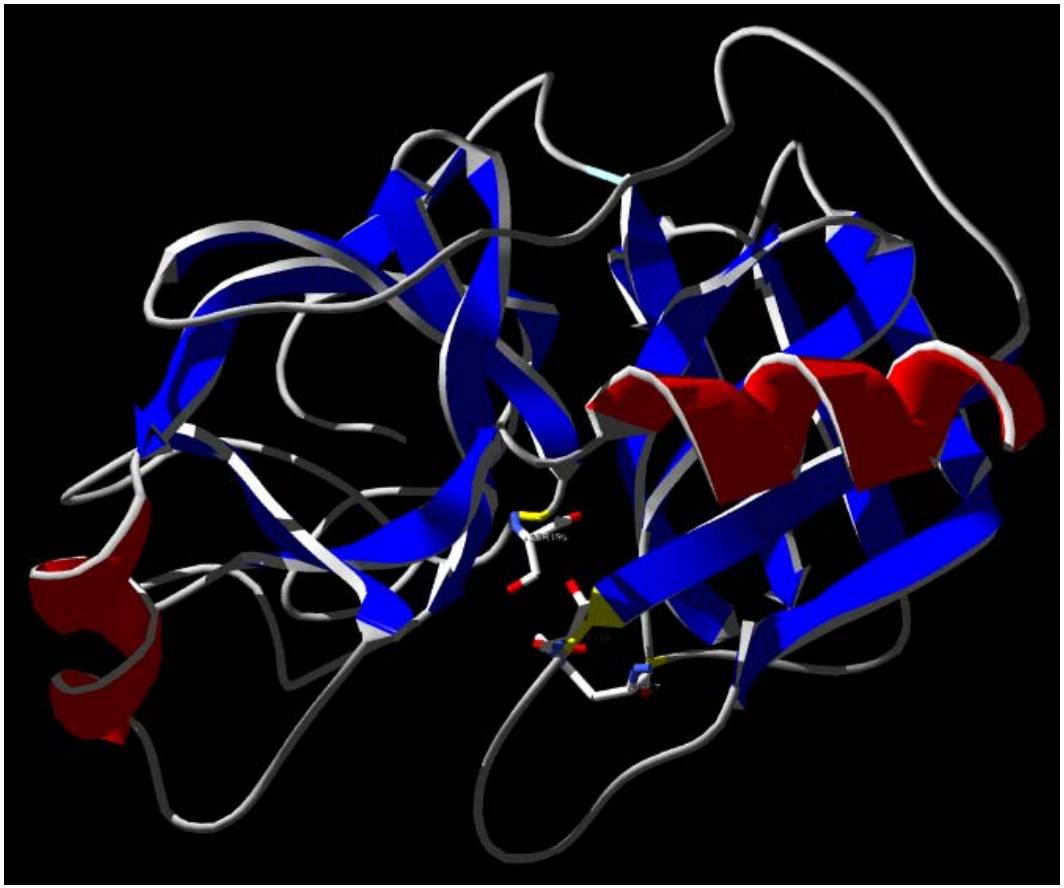
Goal: Confirm the existence of serine proteases in early-divergent phyla, *cnidaria* and *porifera.*
Where they come from?

Data: SP are absent in plants, and protists and in fungi are restricted to *Streptomyces*. However, there are hundreds in animals!

{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

What are serine proteases?



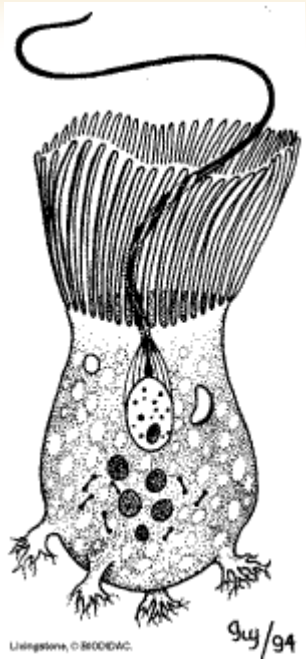
- Hundreds of entries
- Disulfide bonds
- Cleavage peptide
- Digestive:
trypsin, chymotrypsin
- no digestive:
blood clotting
elastases

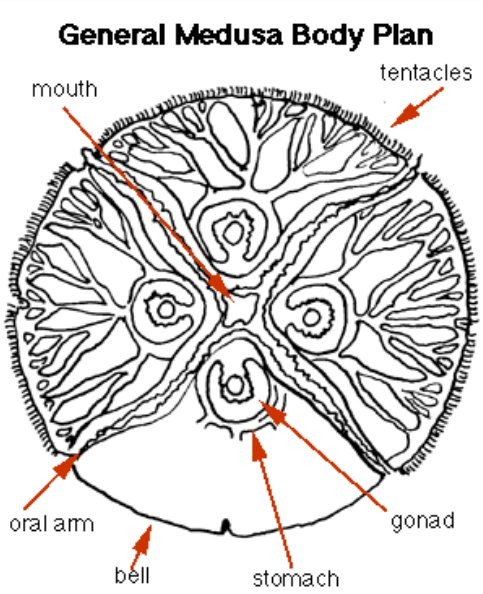
- catalytic triad
H-D-S
- Several structures.

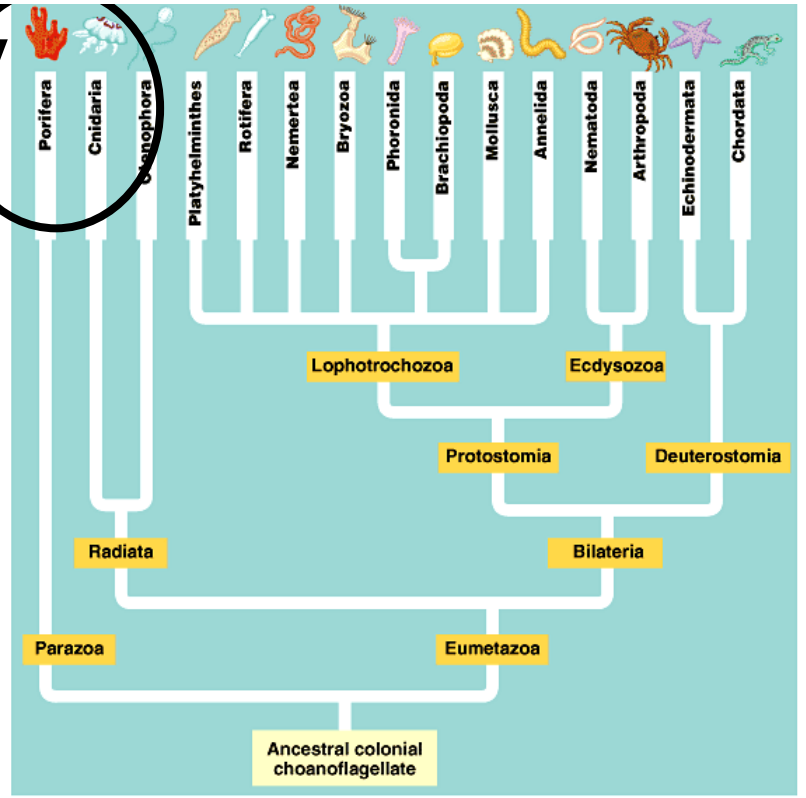
Why are they important? Fundamental question: how animals developed the ability to digest food?

{Rojas & Doolittle, 2002, JME}

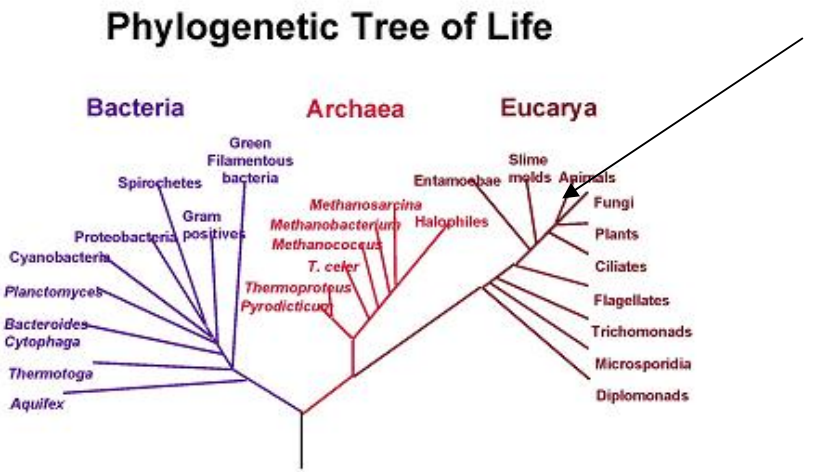
Case 2: Extending the family





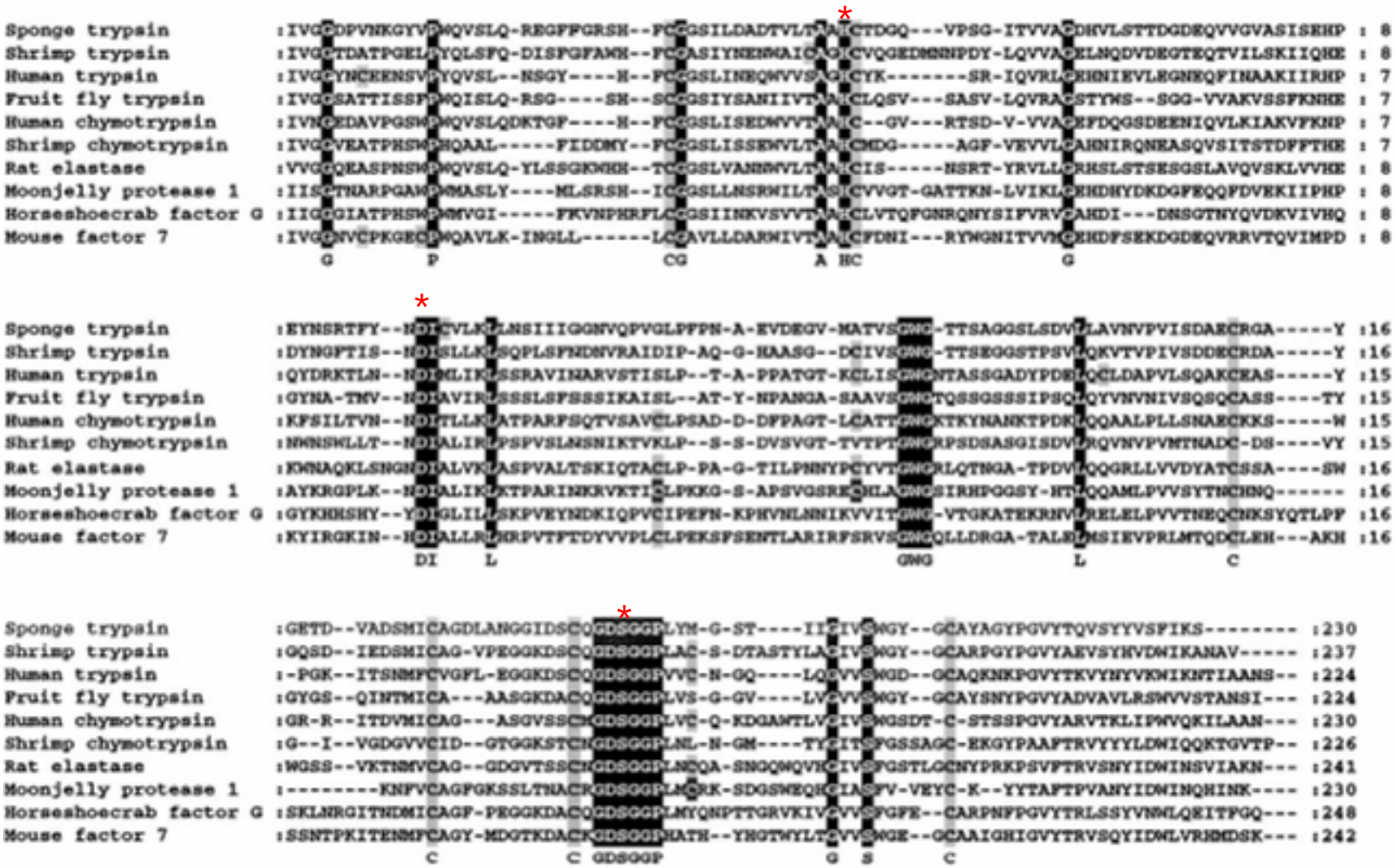


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.



{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

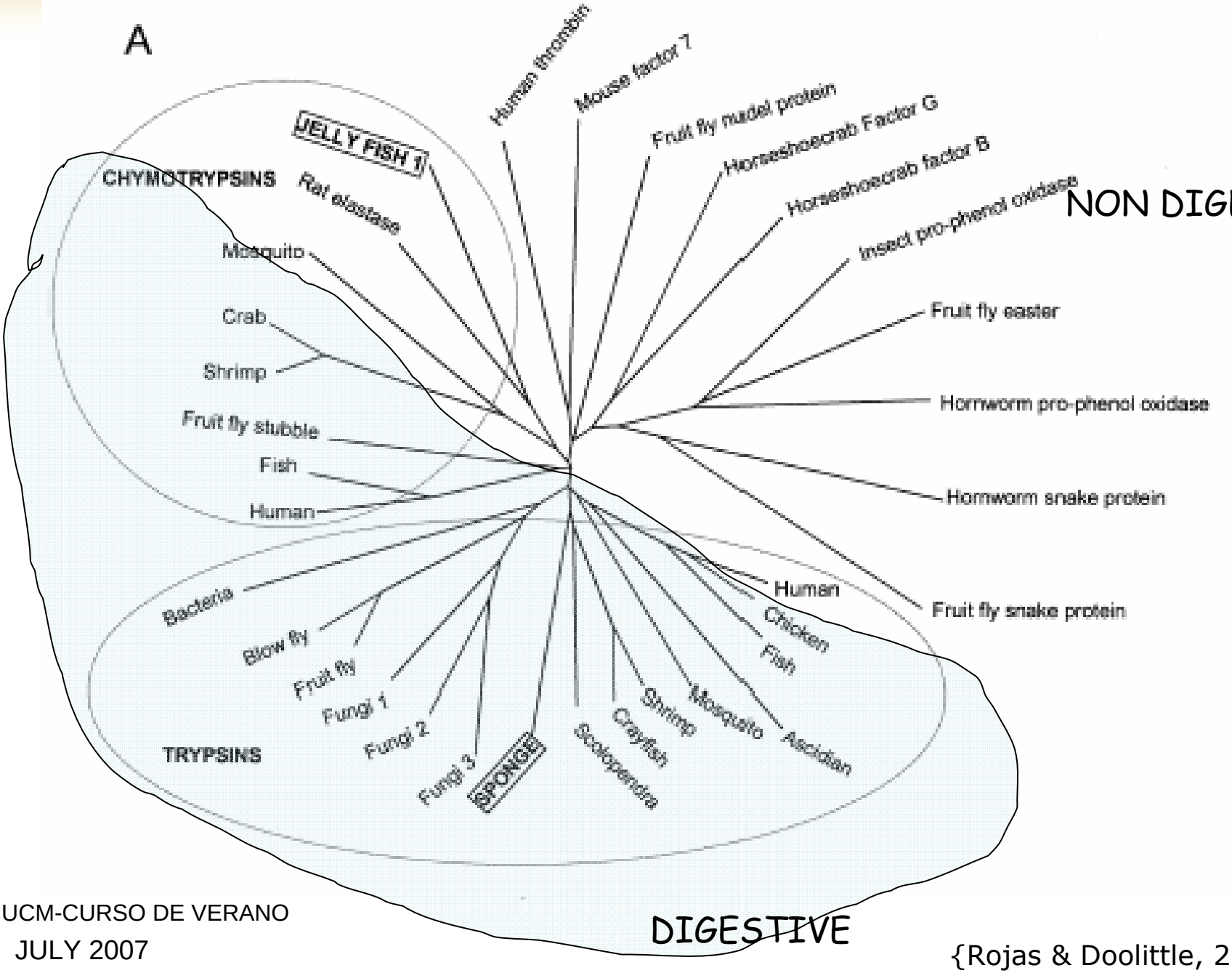


UCM-CURSO DE VERANO

JULY 2007

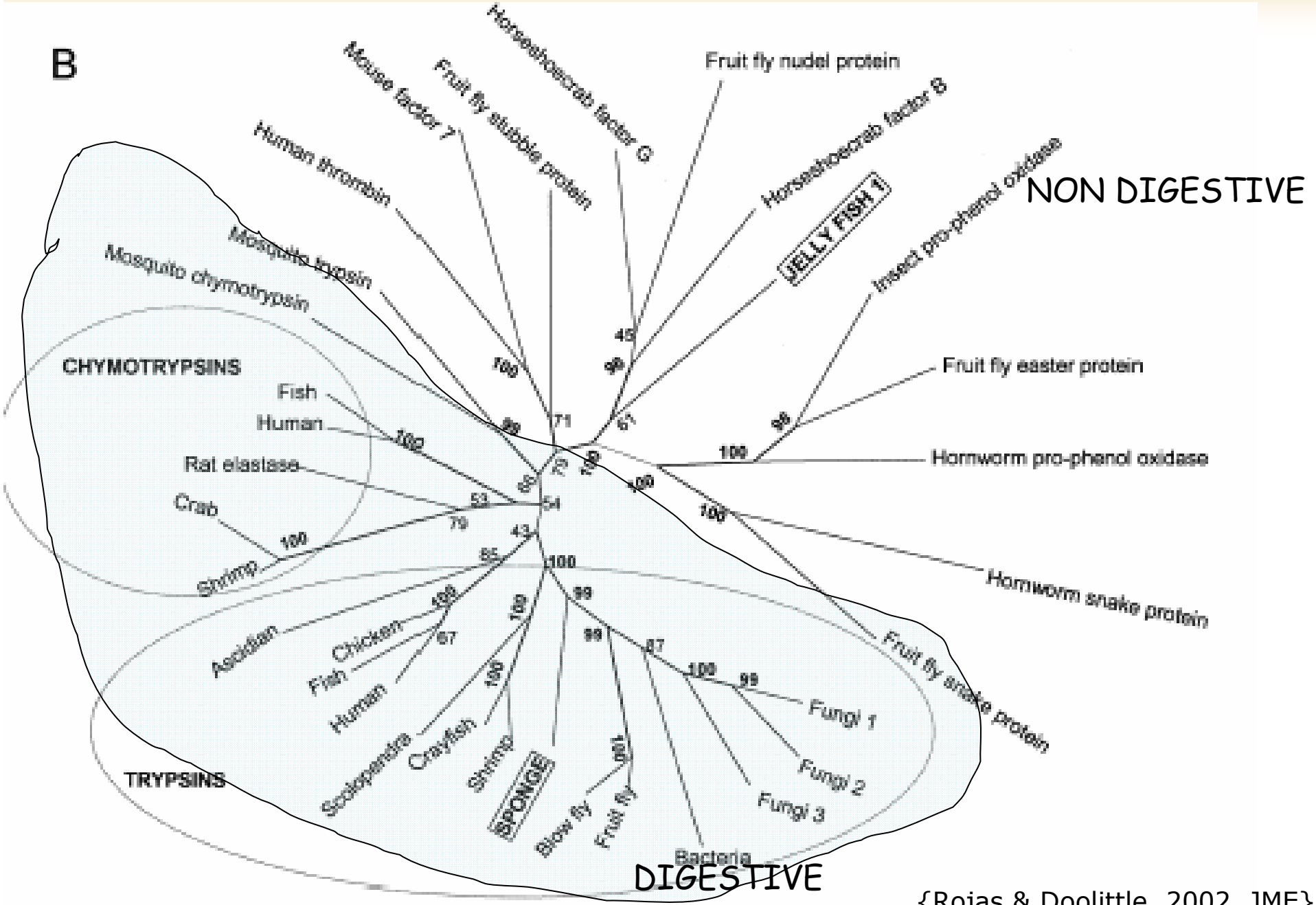
{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family



{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family



{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

WHAT IS THE ORIGIN OF THE CHYMOTRYPSIN FAMILY?

ADDITIONAL INFORMATION:

- Sponge has a D189 diagnostic for trypsin (Hannenshalli & Russell, 2000) Jelly has N189.

- Codon for Serine at the active site:
  sponge signature for trypsin: TCT
  jelly: AGT,AGC

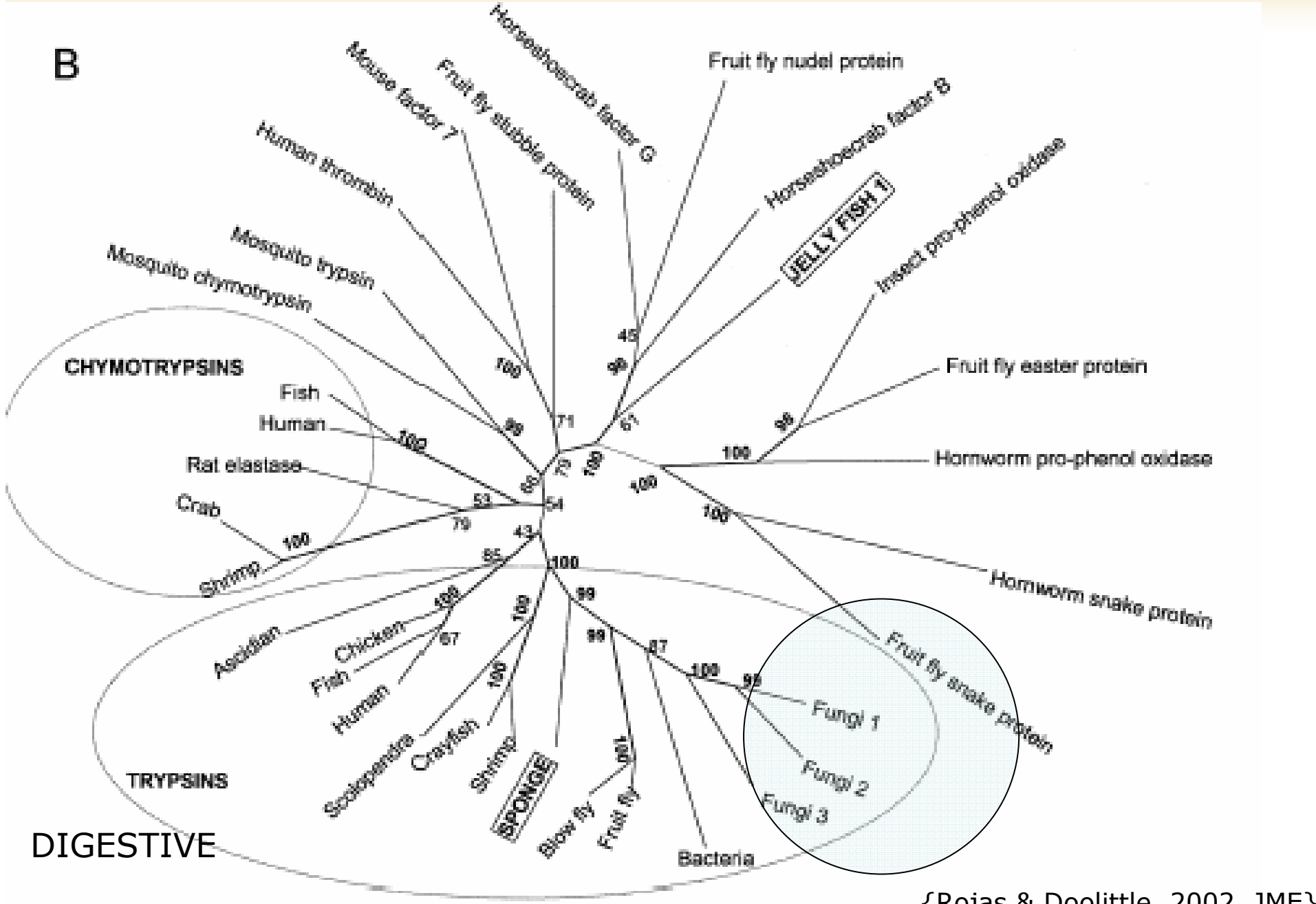
- When blasted against NR:
  sponge 48% with arthropod trypsin
  jelly 36% with RAT elastase

Disulfide bonds:
sponge 5 disulfide bonds and cys match with chymotrypsin-elastase (first tree)

Jelly has digestive system with organs, sponge are loose cells.

{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family



{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

WHY THE FUNGAL ONES CLADE WITH ANIMALS?

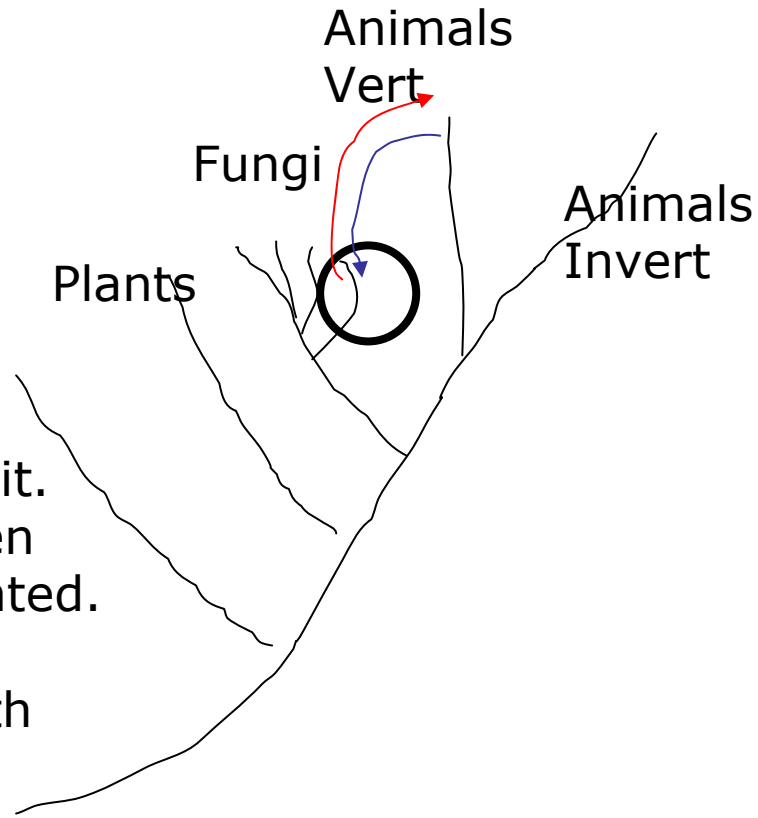
H.G.T!

SCENARIO 1

Plants and
all fungi-except
Streptomyces
lost it!
Fungi should be
more similar to
jelly and sponge

SCENARIO 2

then Plants and
all fungi never had it.
They appeared when
digestion was invented.
Fungi have them
because HGT in both
directions.



{Rojas & Doolittle, 2002, JME}

The DIO Family of Proteins

{Rojas et al, FEBS J. 2004}

Case 3: Revisiting the function

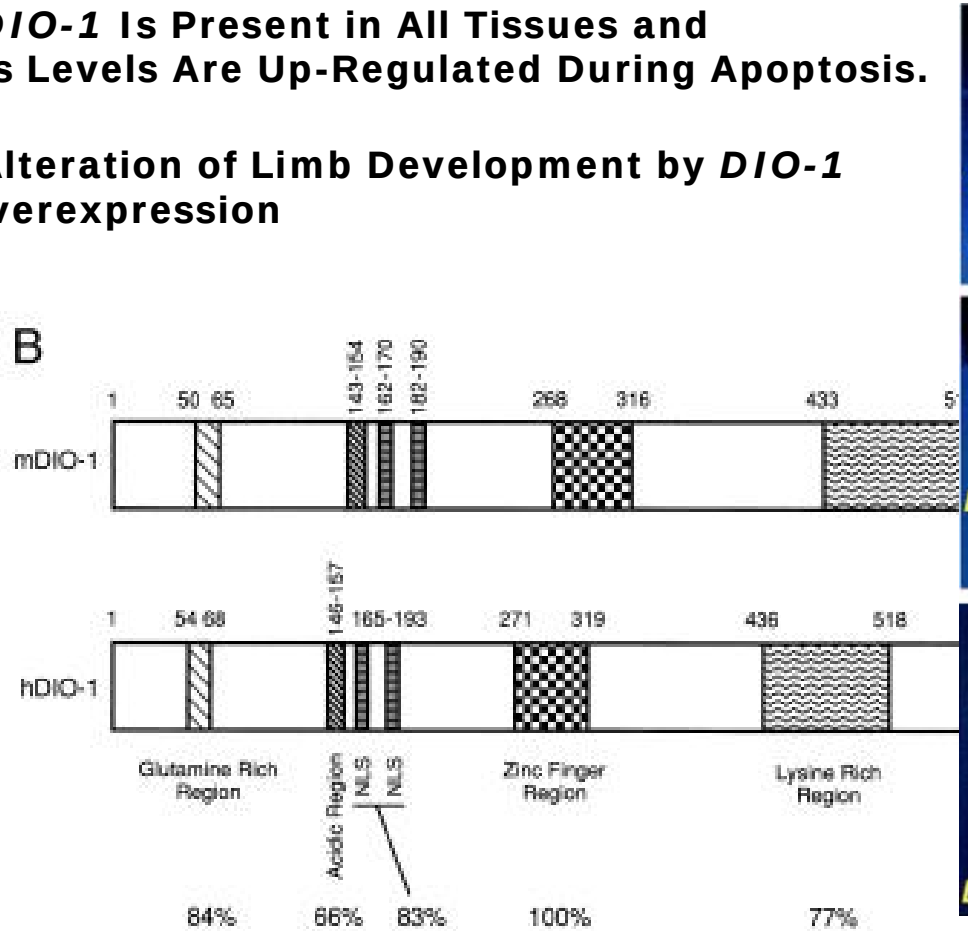
Get's real!

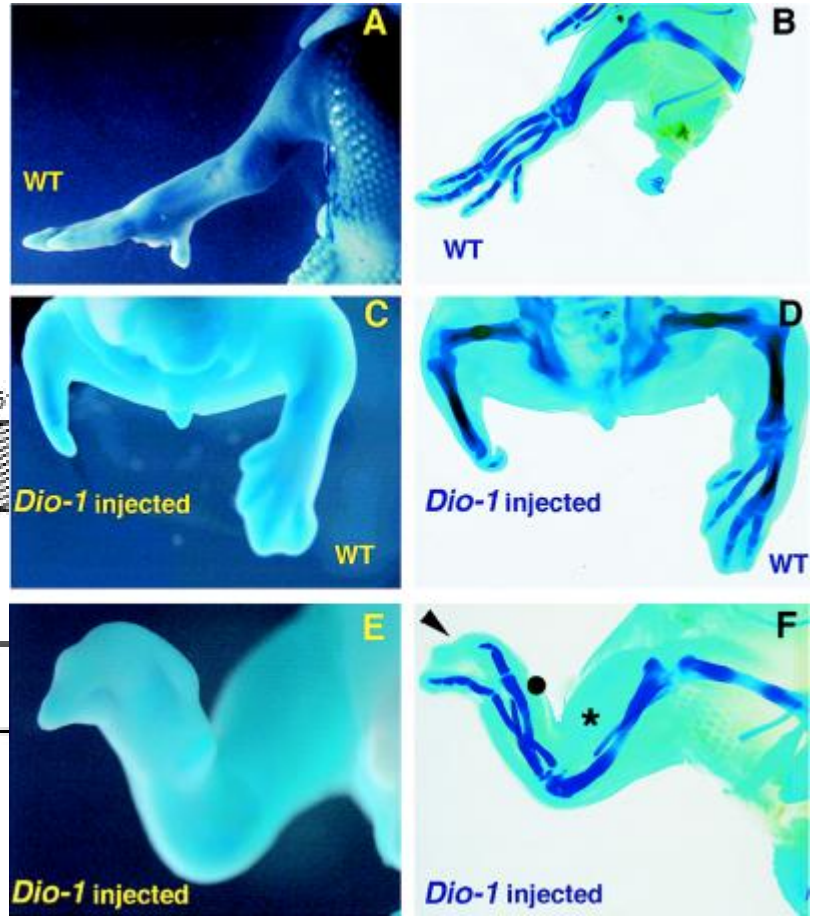
BACKGROUND DEATH INDUCER OBLITERATOR GENE (DIO)

DISRUPTS LIMB DEVELOPEMENT (Garcia-Domingo et al,. 1999)

- ***DIO-1* Is Present in All Tissues and Its Levels Are Up-Regulated During Apoptosis.**

- **Alteration of Limb Development by *DIO-1* Overexpression**





Suggests that the gene is a putative transcription factor…

Case 3: Revisiting the function

Get's real!

BACKGROUND DEATH INDUCER OBLITERATOR GENE (DIO)

INVOLVED IN APOPTOSIS (Garcia-Domingo et al,. 2003)

- DIO-1 nuclear translocation following apoptotic stimulation requires the NLS.
- DIO-1 forms oligomers.
- DIO-1 is present in multiple forms with distinct subcellular localizations.
- DIO-1 overexpression upregulates procaspase levels, leading to increased caspase activity.
- DIO-1ΔNLS is a dominant negative mutant that protects cells from apoptosis.

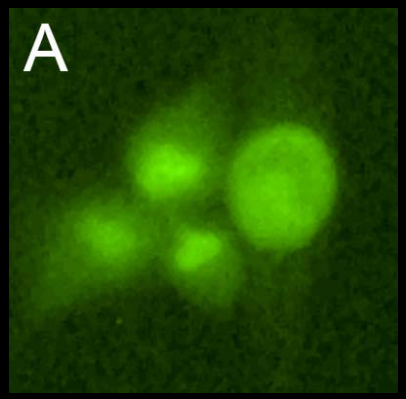
{Rojas et al, FEBS J. 2004}

Case 3: Revisiting the function

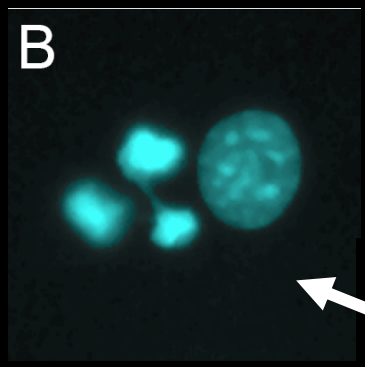
Get's real!

NEW DATA

DEATH INDUCER OBLITERATOR GENE (DIO)

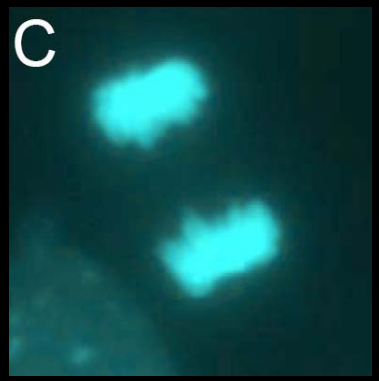


DIO-1 is present in mitotic chromosomes

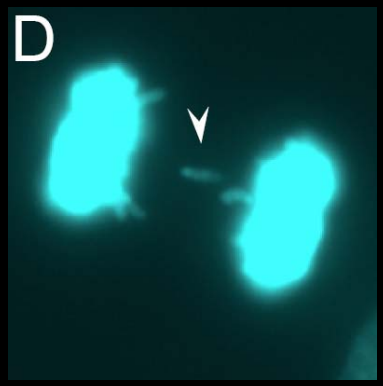


Mitosis on DIO overexpressed-cells

Asymmetric divisions!



Normal anaphase



DIO-targeted cells show abnormal anaphases: lagging chromosomes

TARGETED MICE SHOW SEVERE SUB-FERTILITY!!

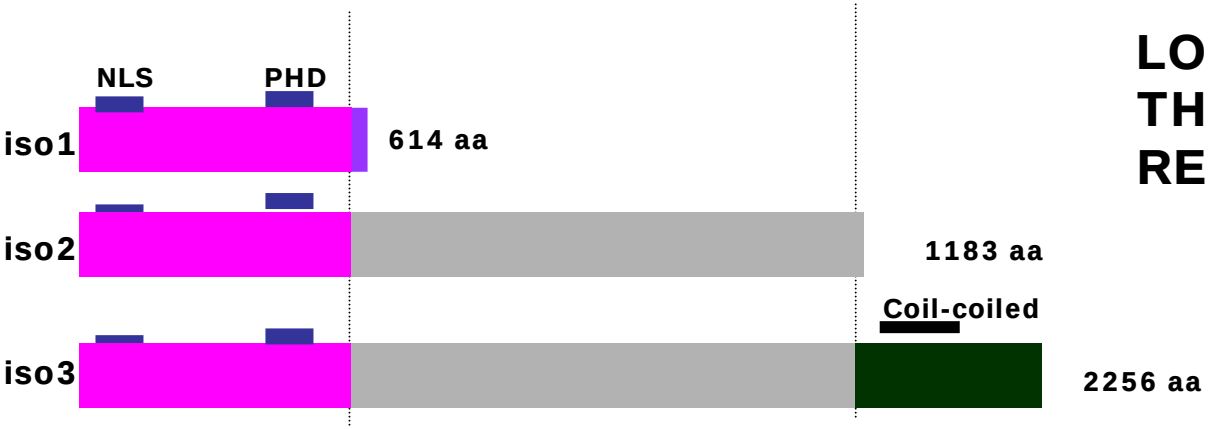
Case 3: Revisiting the function

Get's real!

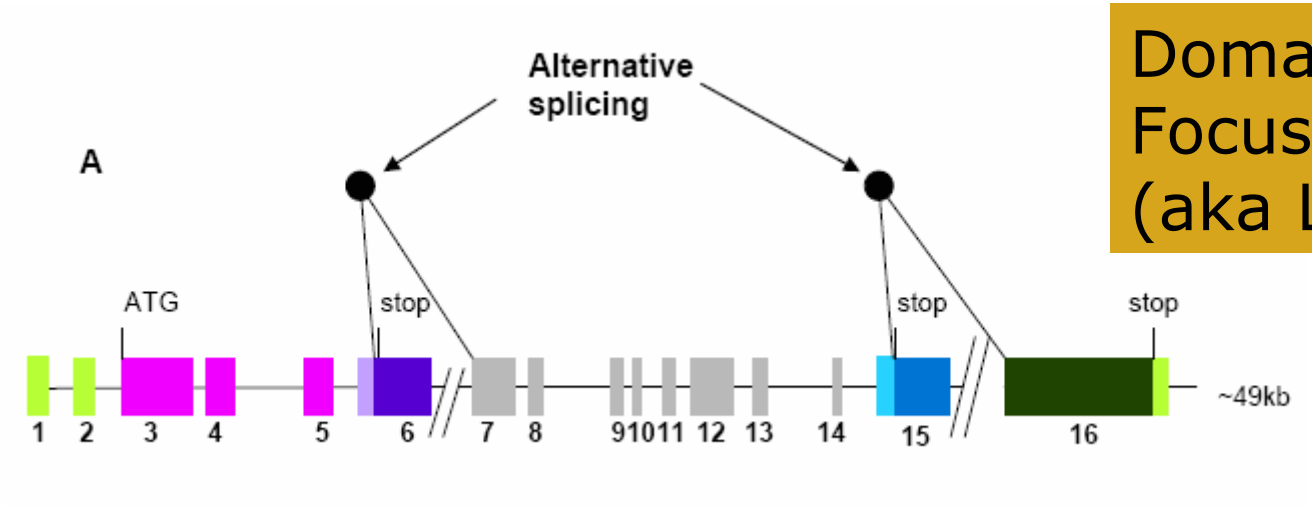
NEW DATA

DEATH INDUCER OBLITERATOR GENE (DIO)

DIO gene contains 3 splicing variants



LONG PARTS OF THE PROTEIN REMAIN UNCOVERED!



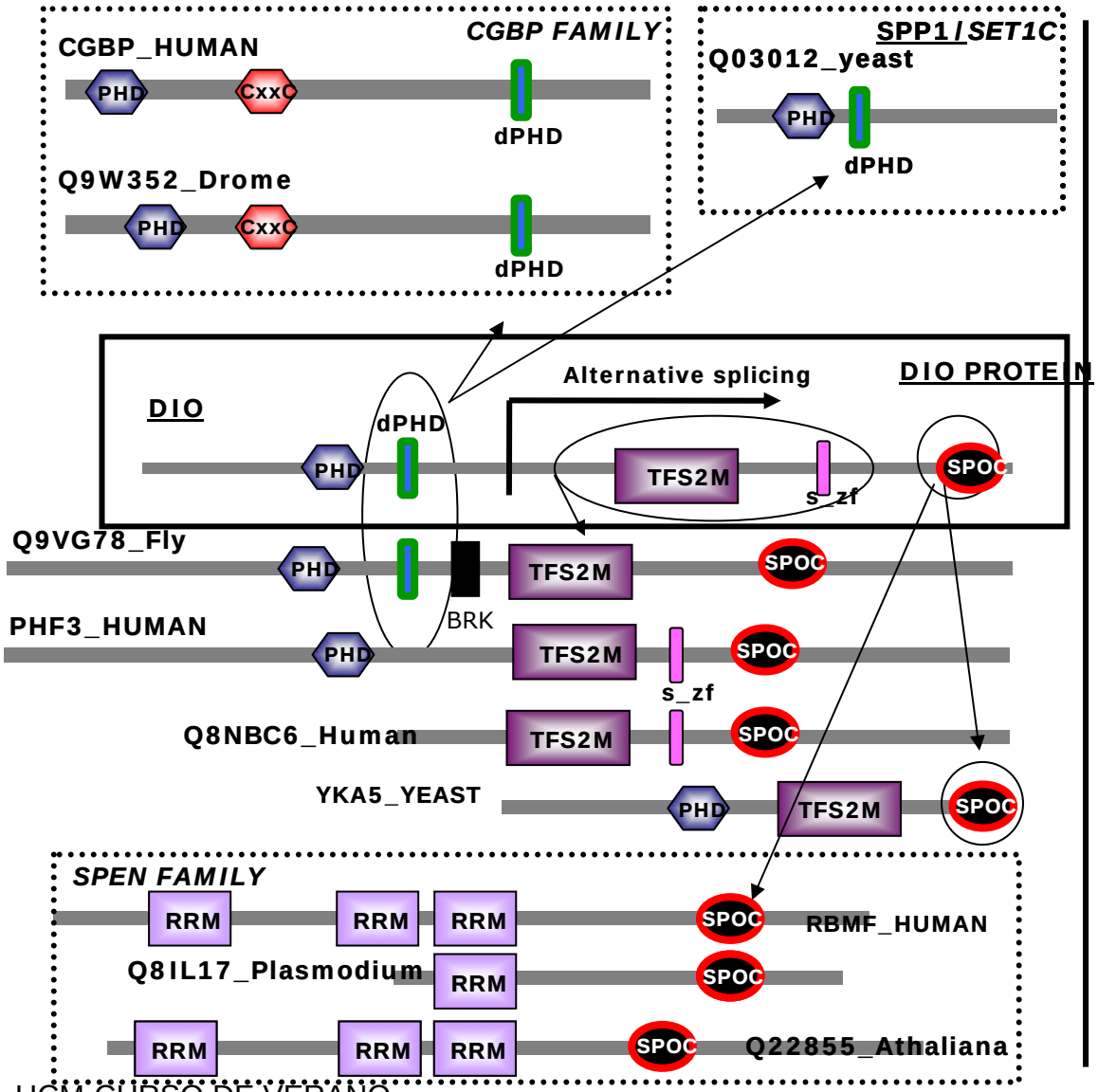
Domain Focus analysis (aka Luis ;-))

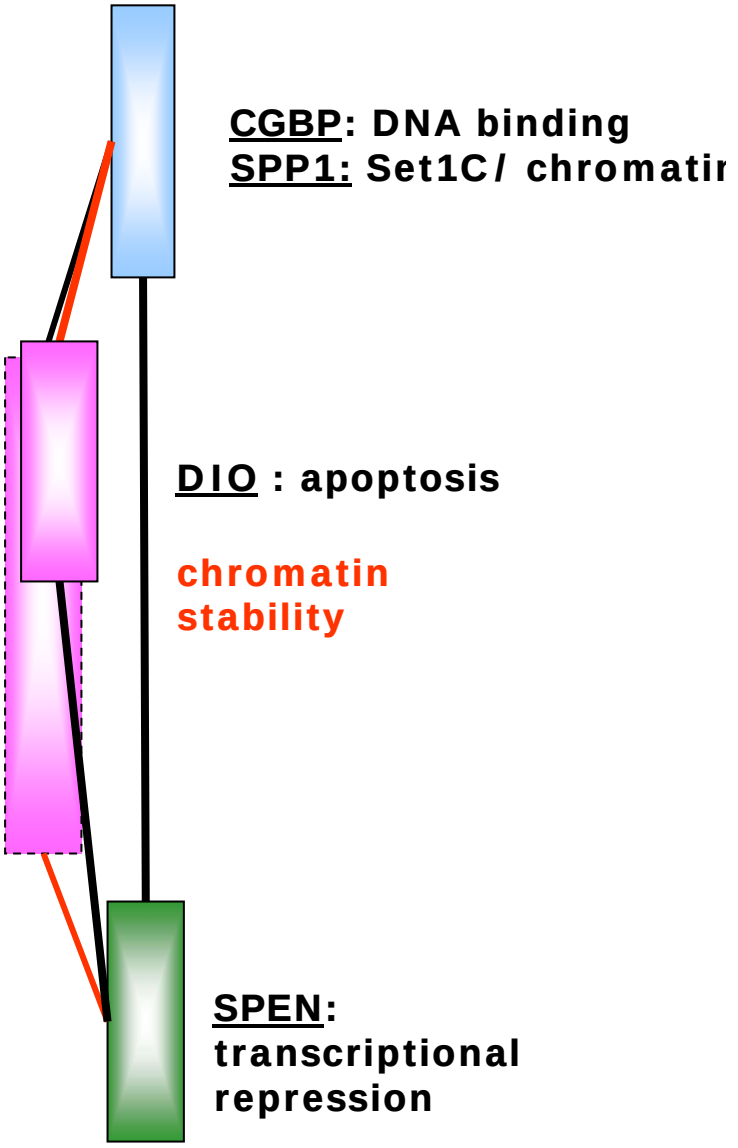
{Rojas et al, FEBS J. 2004}

Case 3: Revisiting the function

Get's real!

OVERVIEW





CGBP: DNA binding
SPP1: Set1C/ chromatir

DIO : apoptosis

chromatin stability

SPEN: transcriptional repression

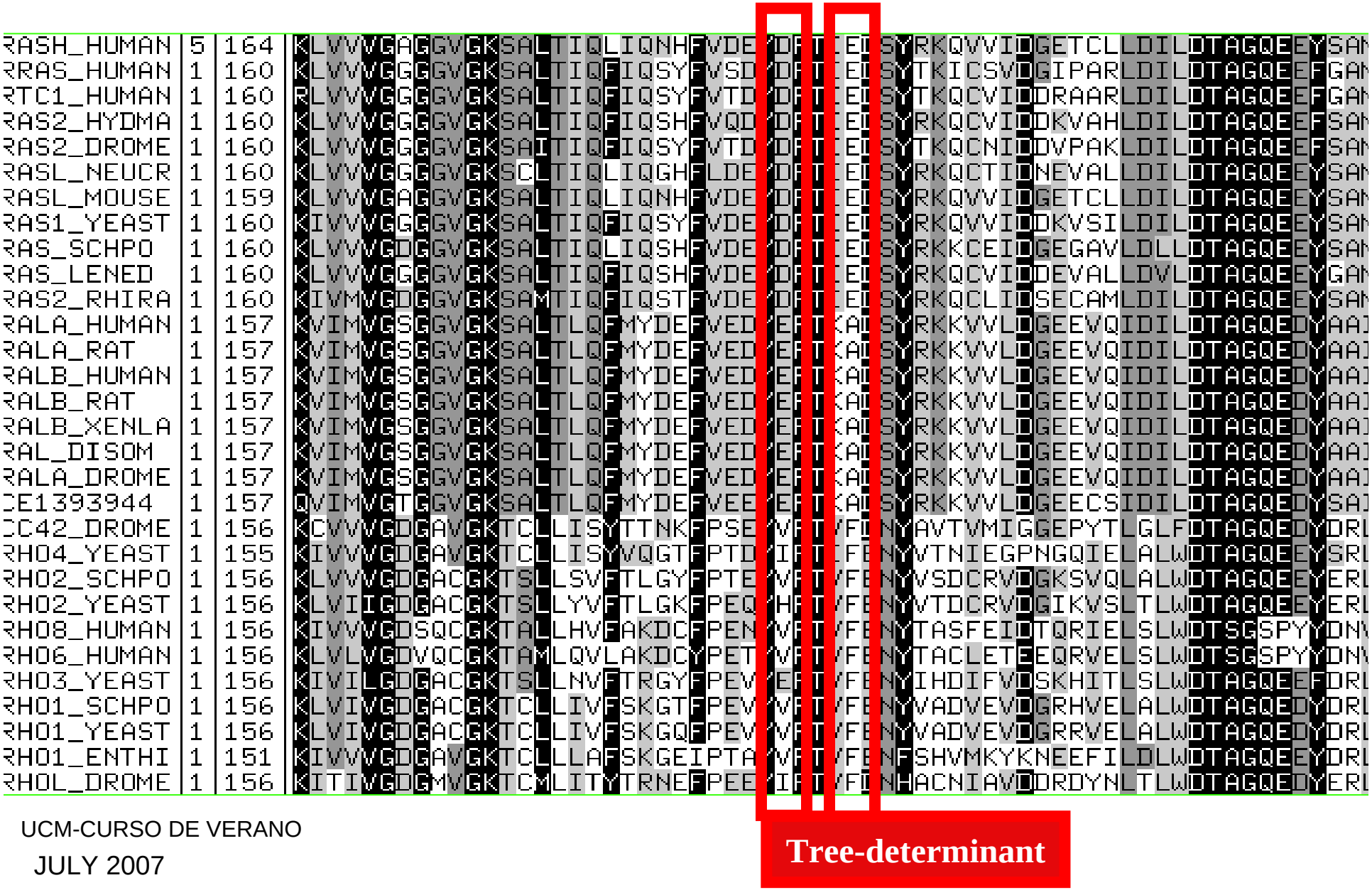
{Rojas et al, FEBS J. 2004}

Identifying Dimerization Residues in CCR chemokine receptors

{de Juan et al, Bioinformatics. 2005}

Case 4: Function Specificity

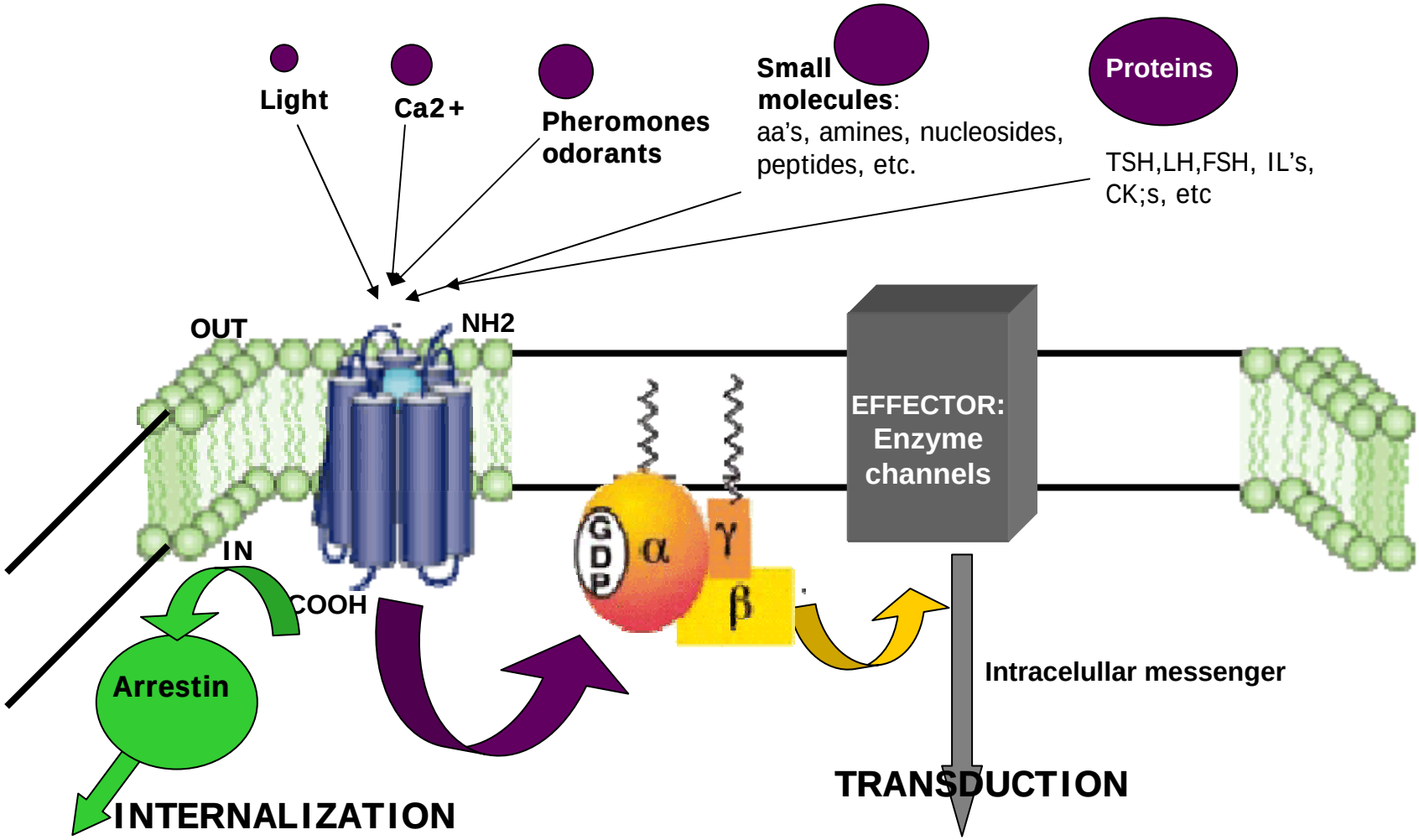
Get's real!



Case 4: Function Specificity

Get's real!

G - Coupled Receptor Proteins bind different ligands.

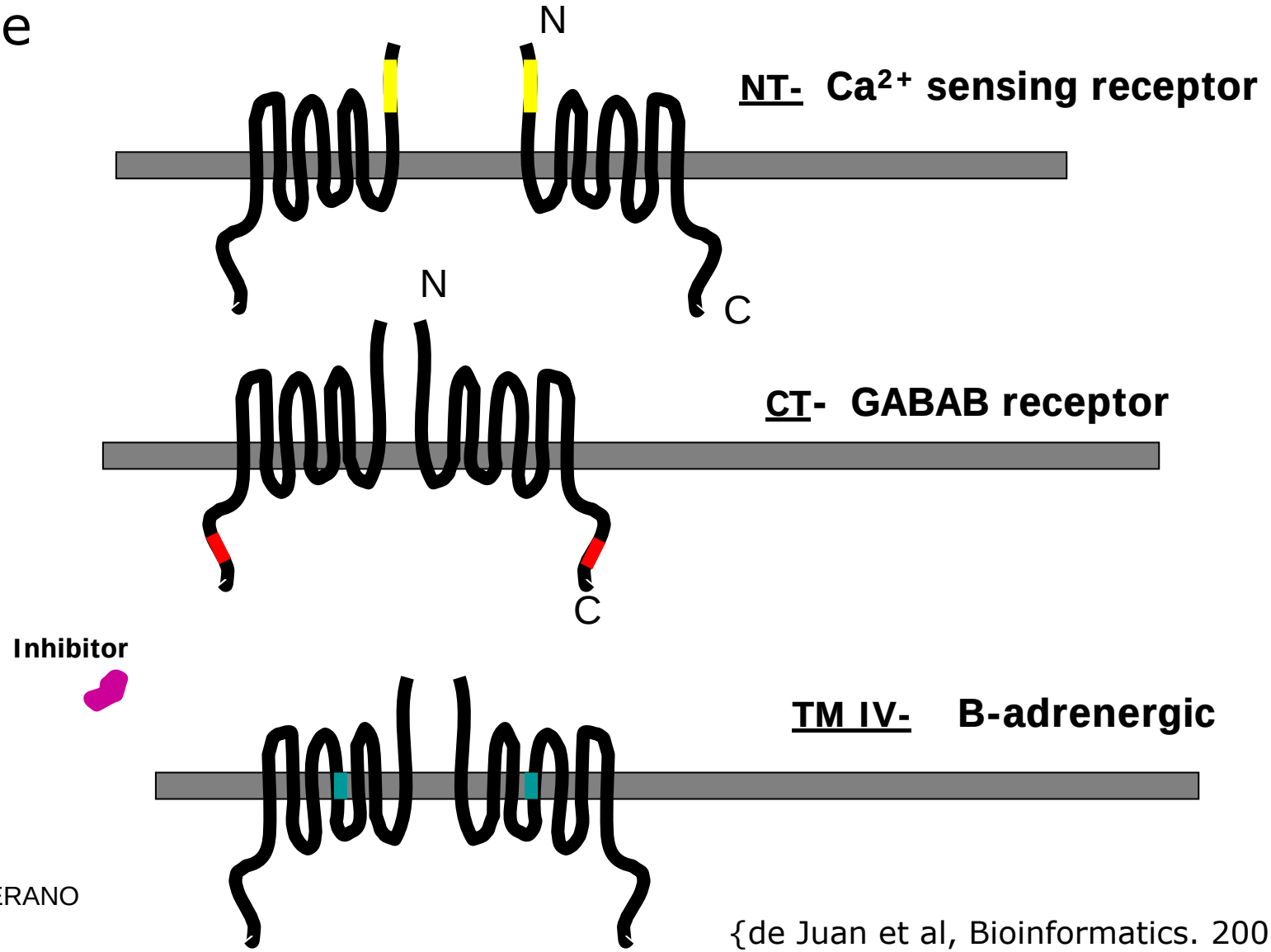


{de Juan et al, Bioinformatics. 2005}

Case 4: Function Specificity

Get's real!

The GCPR's dimerize



{de Juan et al, Bioinformatics. 2005}

The two main events here are:

- Binding specificity.
- Dimerization/Oligomerization.

Then, we have two aims:

- Can we predict the signals and **distinguish** them at the sequence level?
- Which residues are involved in **dimerization**?

Case 4: Function Specificity

Get's real!

- *Existing methods to detect important residues:*

THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 2004 by The American Society for Biochemistry and Molecular Biology, Inc.

Evolutionary Trace of G Protein-c
of Residues That Determine Globa

Published, JB

Srinivasan Madabushi‡§, Alecia K. Gross¶‖, Theodore G. Wensel‡¶, and Olivier Lichtarg

From the ‡Program in Structural and Computational Biology, and §§Program in Cellular and Molecular Bio Molecular and Human Genetics, Baylor College of Med and Molecular Pharmacology, University of California,

G protein-coupled receptor (GPCR) activation mediated by ligand-induced structural reorganization of its helices is poorly understood. To determine the universal elements of this conformational switch, we used evolutionary tracing (ET) to identify residue positions commonly important in diverse GPCRs. When mapped onto the rhodopsin structure, these trace residues cluster into a network of contacts from the retinal binding site to the G protein-coupling loops. Their roles in a generic transduction mechanism were verified by 211 of 239 published mutations that caused functional defects. When grouped according to the nature of the defects, these residues subdivided into three striking sub-clusters: a trigger region, where mutations mostly affect ligand binding, a coupling region near the cytoplasmic interface to the G protein,

14522 *Biochemistry* **2003,** *42,* 14522−14531

Dimerization in Aminergic G-Protein-Coupled Receptors: Application of a Hidden-Site Class Model of Evolution†

Orkun S. Soyer,‡ Matthew W. Dimmic,§ Richard R. Neubig,‖ and Richard A. Goldstein*,⊥

Department of Chemistry, Biophysics Research Division, and Department of Pharmacology, University of Michigan, Ann Arbor, Michigan 48109, and Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW71AA, U.K.

Received June 25, 2003; Revised Manuscript Received October 1, 2003

ABSTRACT: G-Protein-coupled receptors (GPCRs) are an important superfamily of transmembrane proteins involved in cellular communication. Recently, it has been shown that dimerization is a widely occurring phenomenon in the GPCR superfamily, with likely important physiological roles. Here we use a novel hidden-site class model of evolution as a sequence analysis tool to predict possible dimerization interfaces in GPCRs. This model aims to simulate the evolution of proteins at the amino acid level, allowing the analysis of their sequences in an explicitly evolutionary context. Applying this model to aminergic GPCR sequences, we first validate the general reasoning behind the model. We then use the model to perform a family specific analysis of GPCRs. Accounting for the family structure of these proteins, this approach detects different evolutionarily conserved and accessible patches on transmembrane (TM) helices 4−6 in different families. On the basis of these findings, we propose an experimentally testable dimerization mechanism, involving interactions among different combinations of these helices in different families of aminergic GPCRs.

fact, GPCRs are so ubiquitous that, although they are the targets of nearly 50% of current drugs (2), this is still a small fraction of their pharmacological potential (3).

Some of the major questions relevant to GPCR pharmacology include the following: What residues are critical for ligand binding and G protein activation? What do different receptor families have in common with regard to their activation mechanism? From a structural perspective, it is known that all GPCRs form a seven transmembrane (TM) α-helical bundle,

Hannenhalli & Russell. **JMB** (2000). 306:61-76.

Our strategy

<u>TEST CASE</u>: CHEMOKINES, known to dimerize.

Steps:

1.- Alignment selection.

2.- Tree determinants searching.

3.- Selecting regions.

4.- Mapping and rough model generation based on Rhodopsin (to visually represent the results).

Alignment selection

TEST CASE: CHEMOKINES

(http://www.gpcr.org/7M/)

- **Clustering**: to obtain a representative alignment containing groups: CCR1-9, CXCR3-5, and IL8A-B (**total 61**).

- **Different levels** of redundancy tested (75-100%). A redundancy level of 95% selected to compensate the number of sequences and alignment bias reduction

- **Realignment** using T-COFFEE with secondary structure predictions taking into account the rhodopsin model.

Case 4: Function Specificity

Get's real!

Finding residues

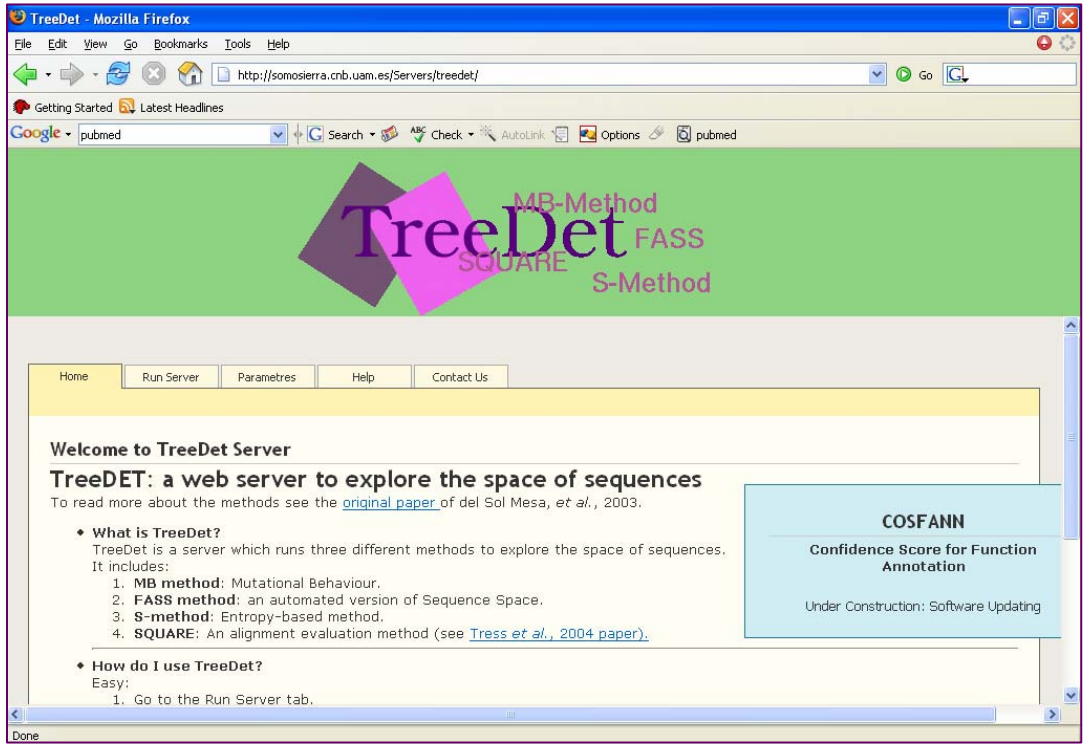
Basics: Homodimerization specificity is trying to avoid promiscuous dimerization between homologous sequences!

Dimerization-focused strategy: obtaining the best subfamily division (as many subfamily groups as possible).

TREE DETERMINANT SEARCHING

- Level entropy method
- Mutational behaviour method (MB)
- Sequence Space Automated Method (FASS)

POSTER AT ECCB2005



{Carro et al, NAR. 2006}
{de Juan et al, Bioinformatics. 2005}

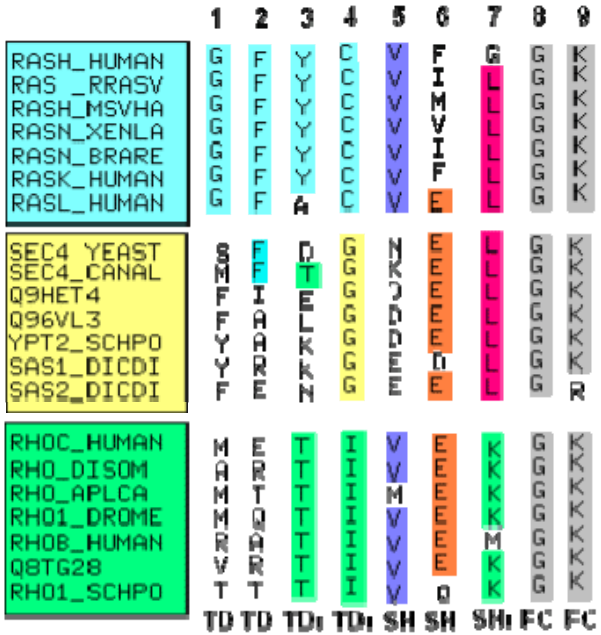
Case 4: Function Specificity

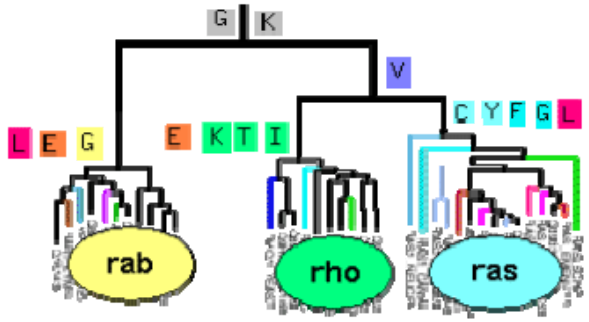
Get's real!

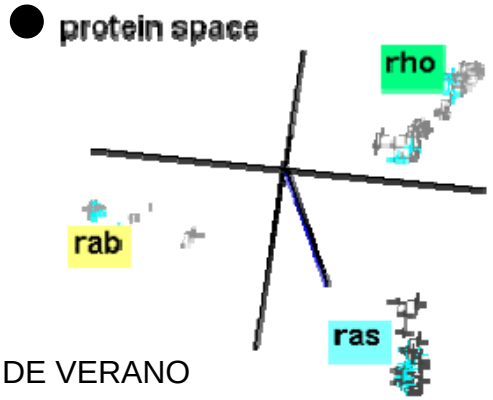
Sequence Space: overview

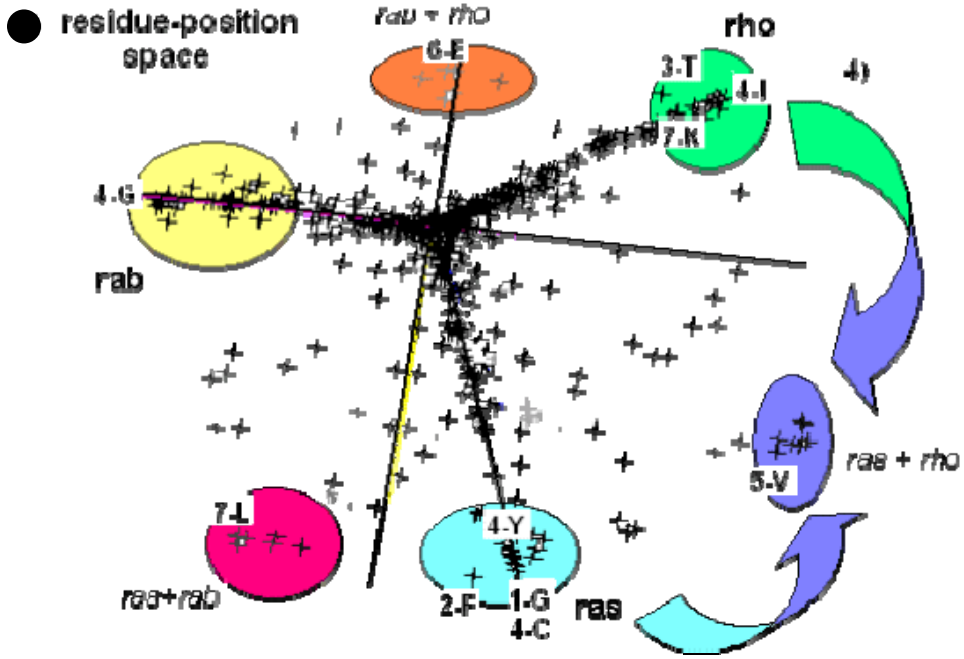
Casari, G. et al. **Nat. Struct. Biol** (1995). 2:171-178.

An example:





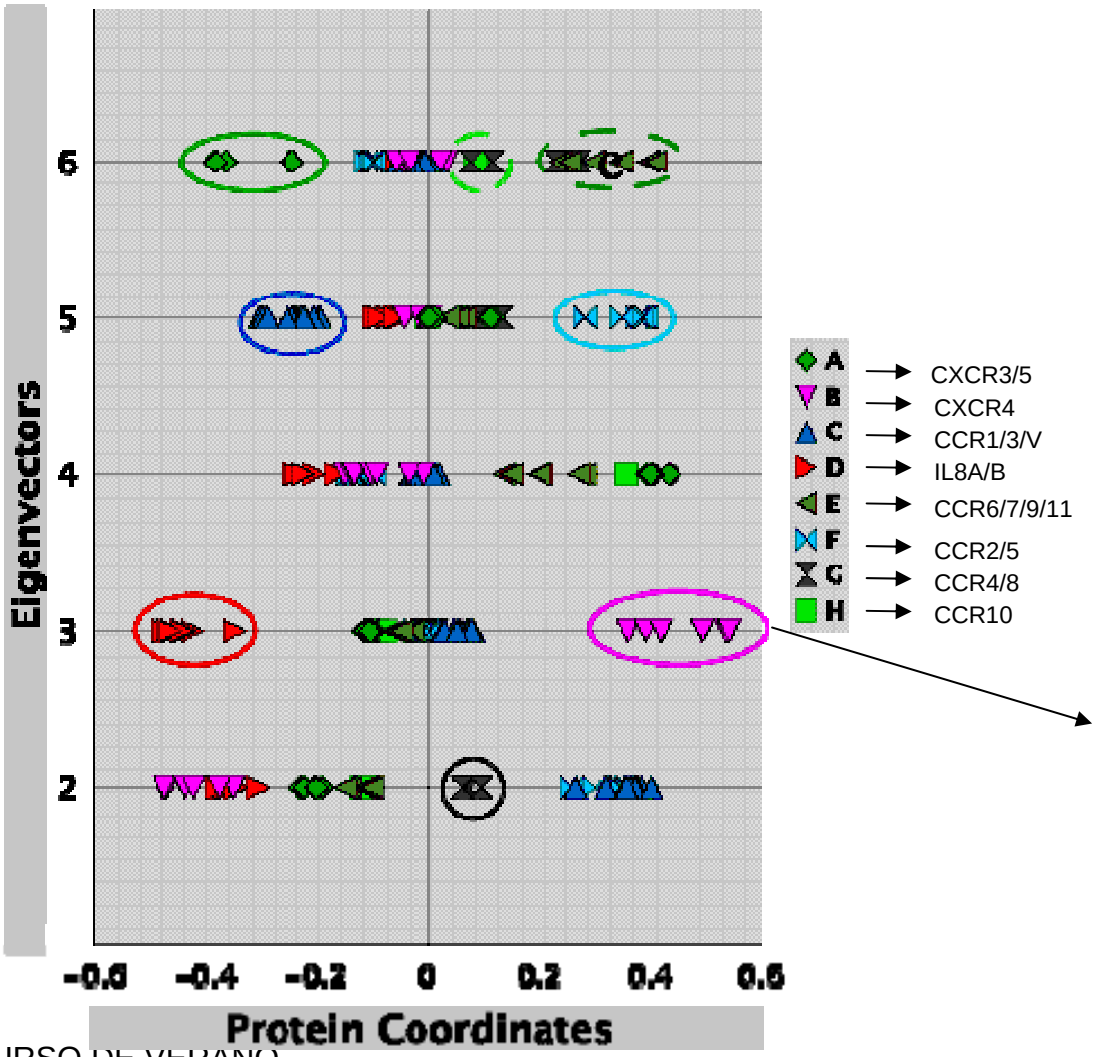


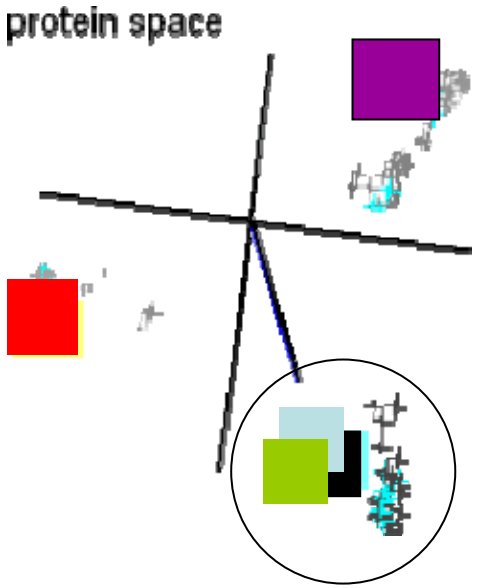


Case 4: Function Specificity

Get's real!

Sequence Space: Clustering results



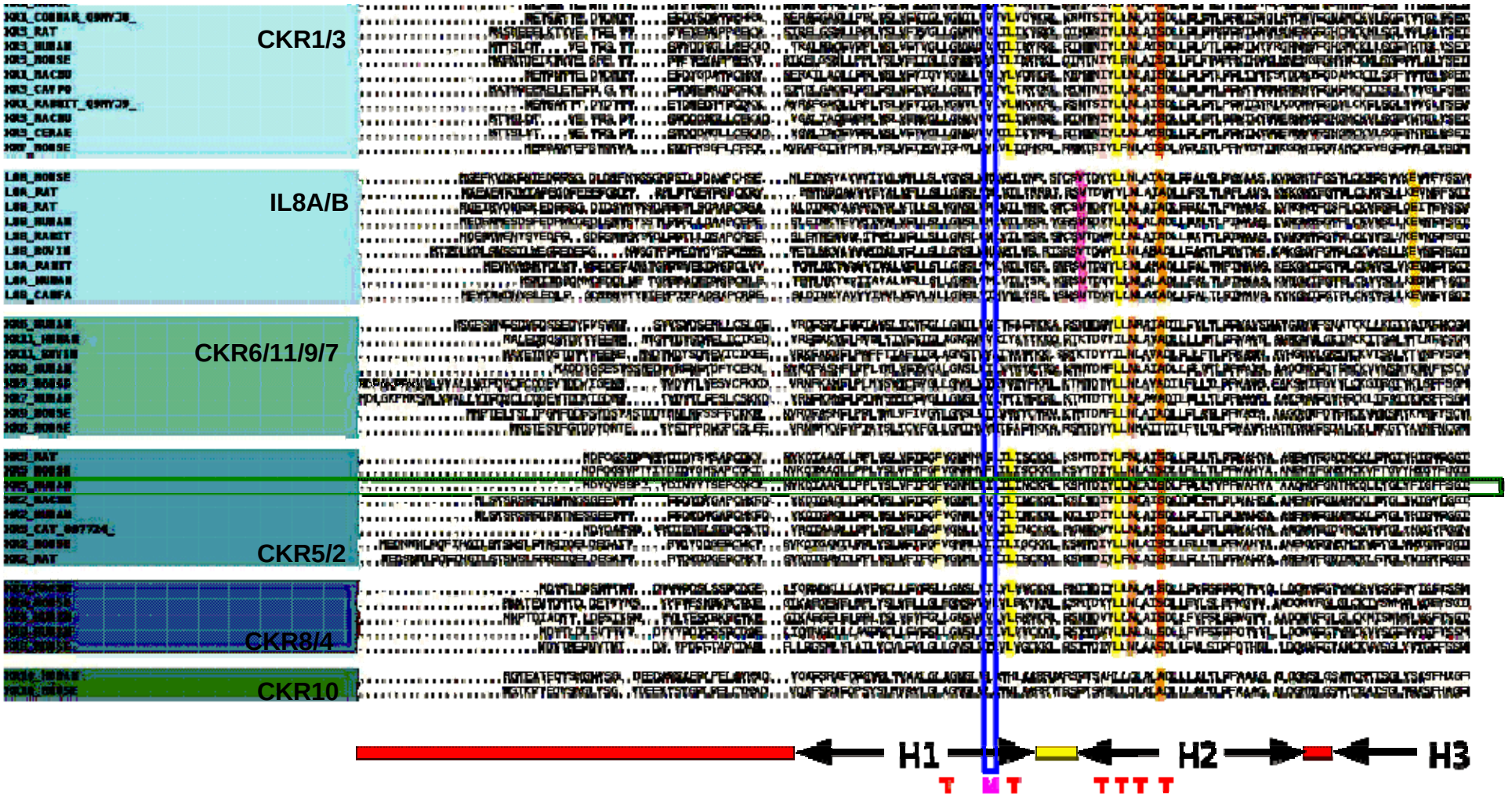


Case 4: Function Specificity

Get's real!

Sequence Space: Clustering results

Residues obtained by Sequence-Space family division.



Bioinformatics: Conclusions

- *The automated version is capable to detect the Functional signal*

- *The dimerization signal still needs extensive human supervision.*

- *Not all the obtained pairs were tested so, functional signals could very well be dimer/oligomerization ones.*

- *... But experimental validation of certain pairs confirmed the predicitive power of this approach.*

Acknowledgements

Luis Sanchez-Pulido — PDG
Mario Mellado — SCOMP
Karel van Wely

Adam Godzik
Y. Zhe
R.F. Doolittle
J.E. Garcia de los Rios
M. McClelland

Fede Abascal

You!