

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 & subfamilies

- **Why** we study protein families?

- Some real examples

Some Concepts

- **Homology**: implies an **evolutionary** relationship
- How protein families appear?

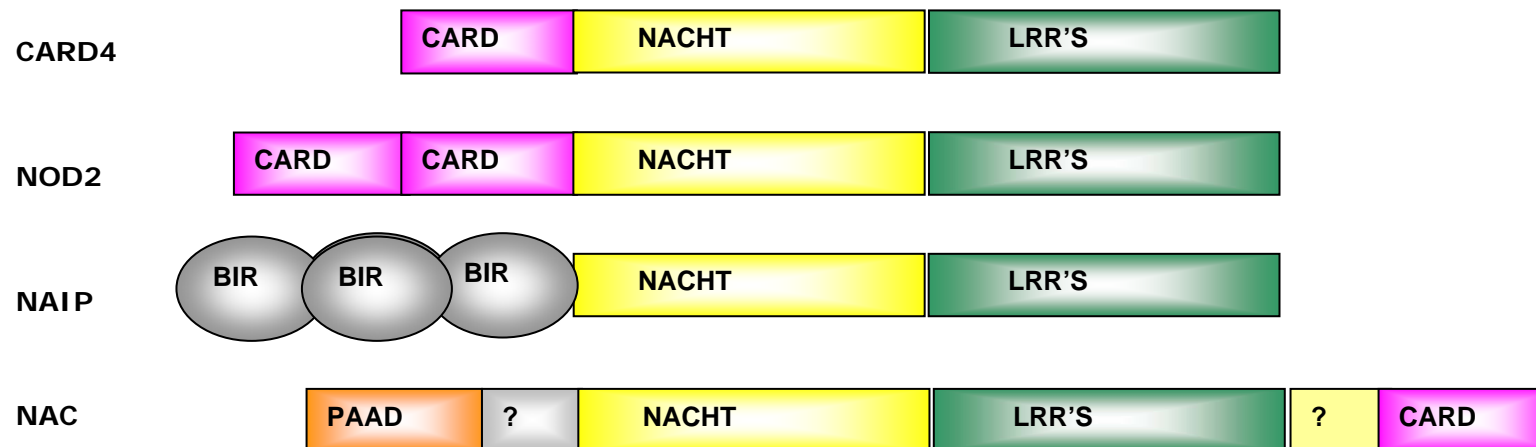
By domain shuffling

By Gene duplication

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

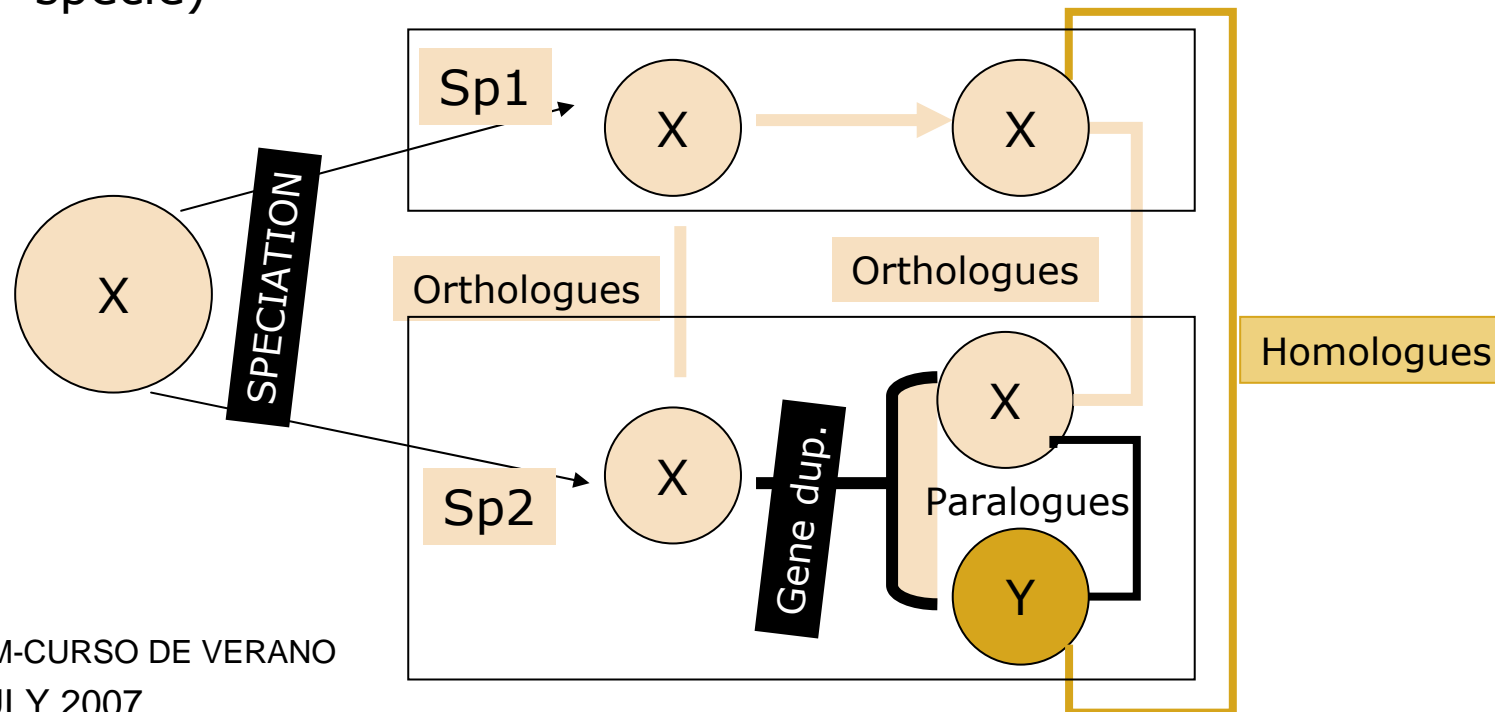
Domain shuffling

- 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.

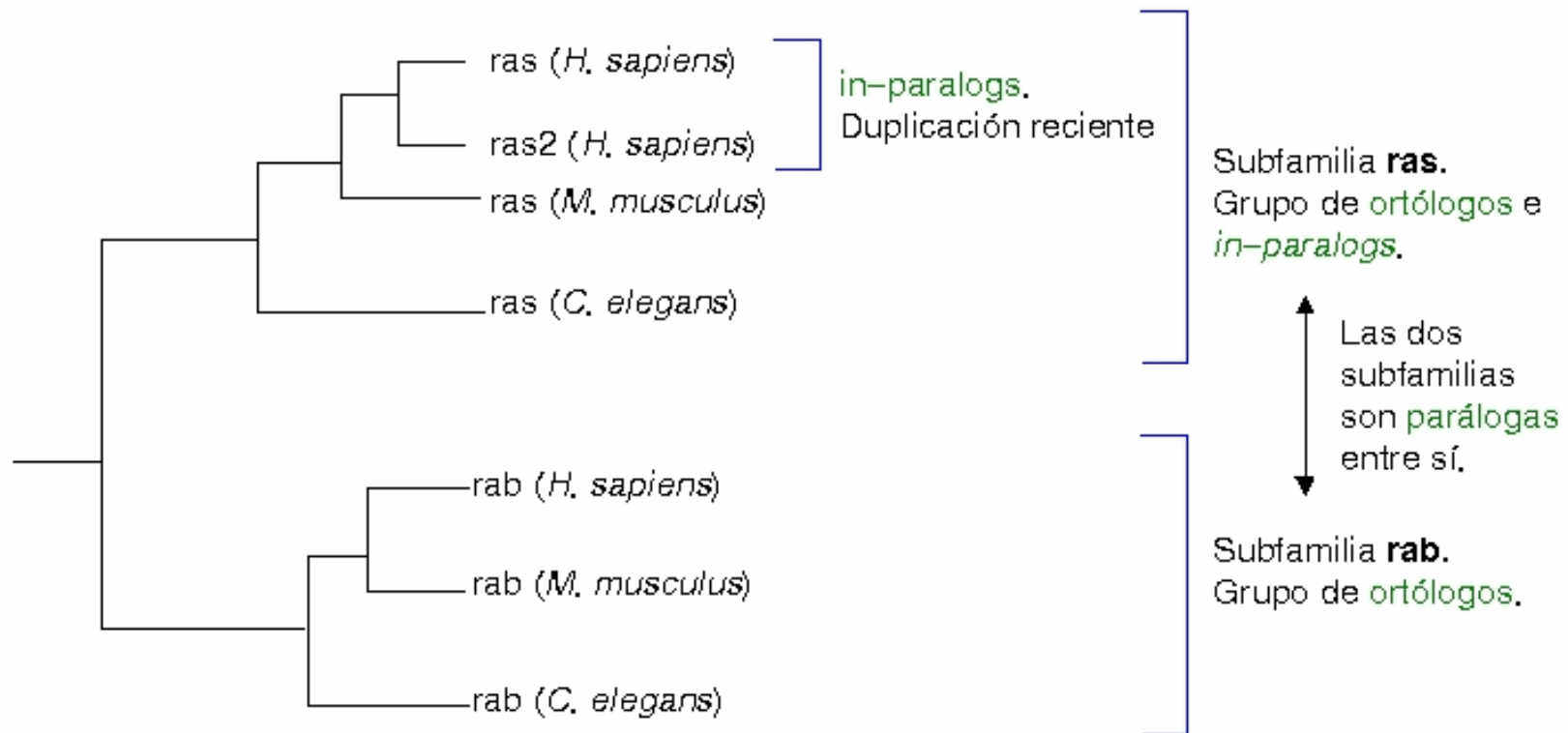


Gene Duplication

- 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:

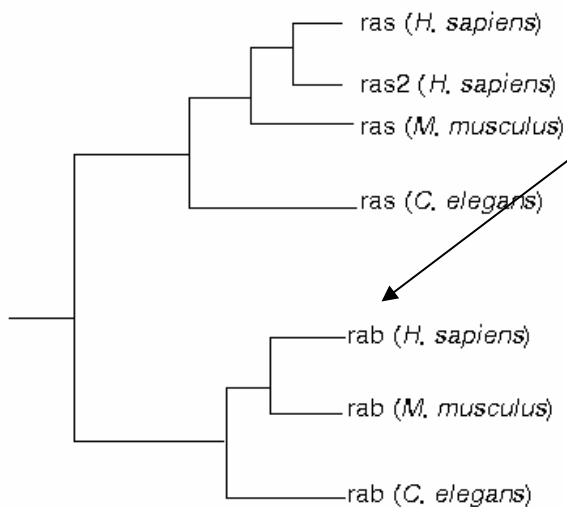
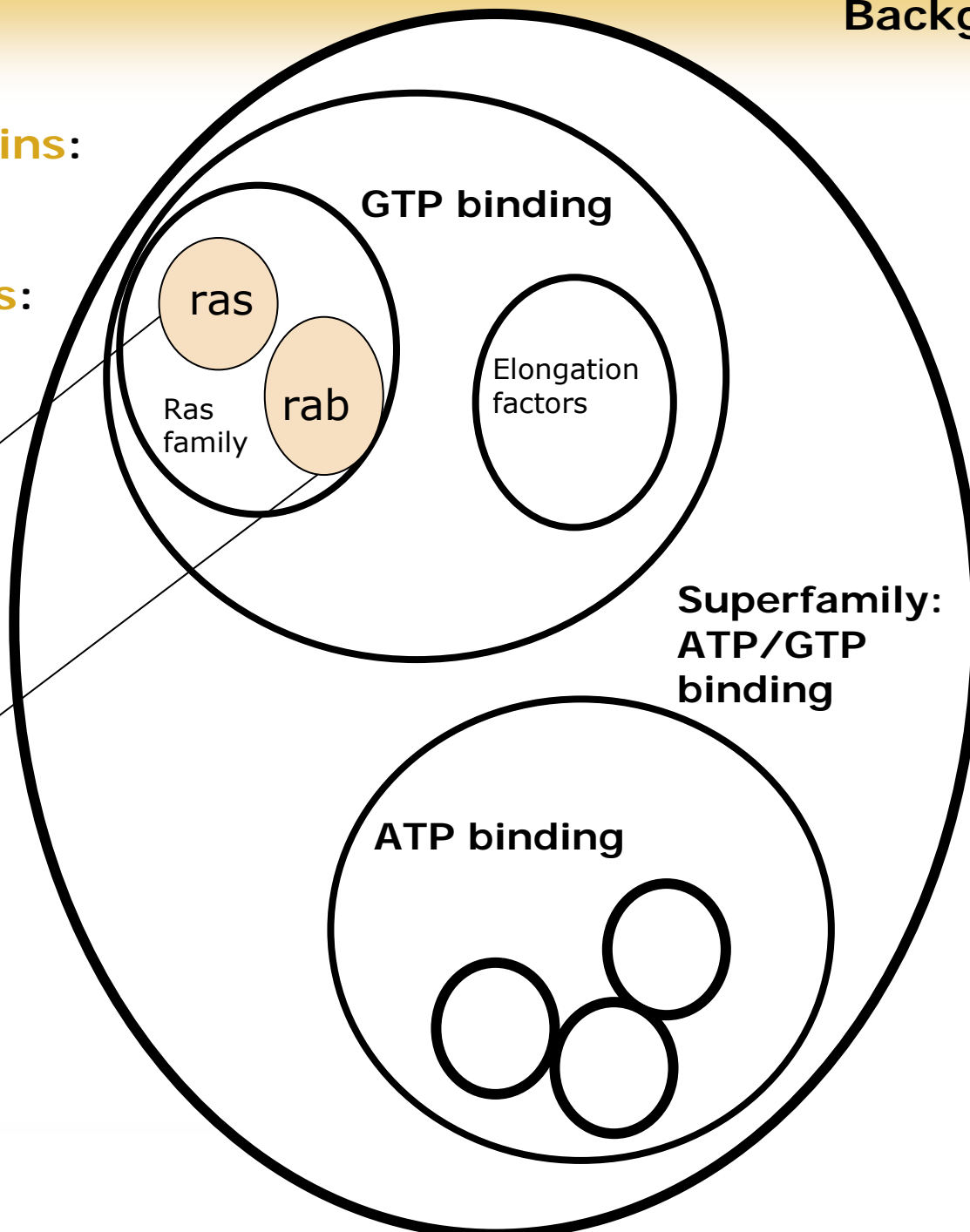
General function:

Functional feature:

Hp21-elongation factor EF-Tu of Ecoli
 signal transduction-protein synthesis
 GTP binding

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>

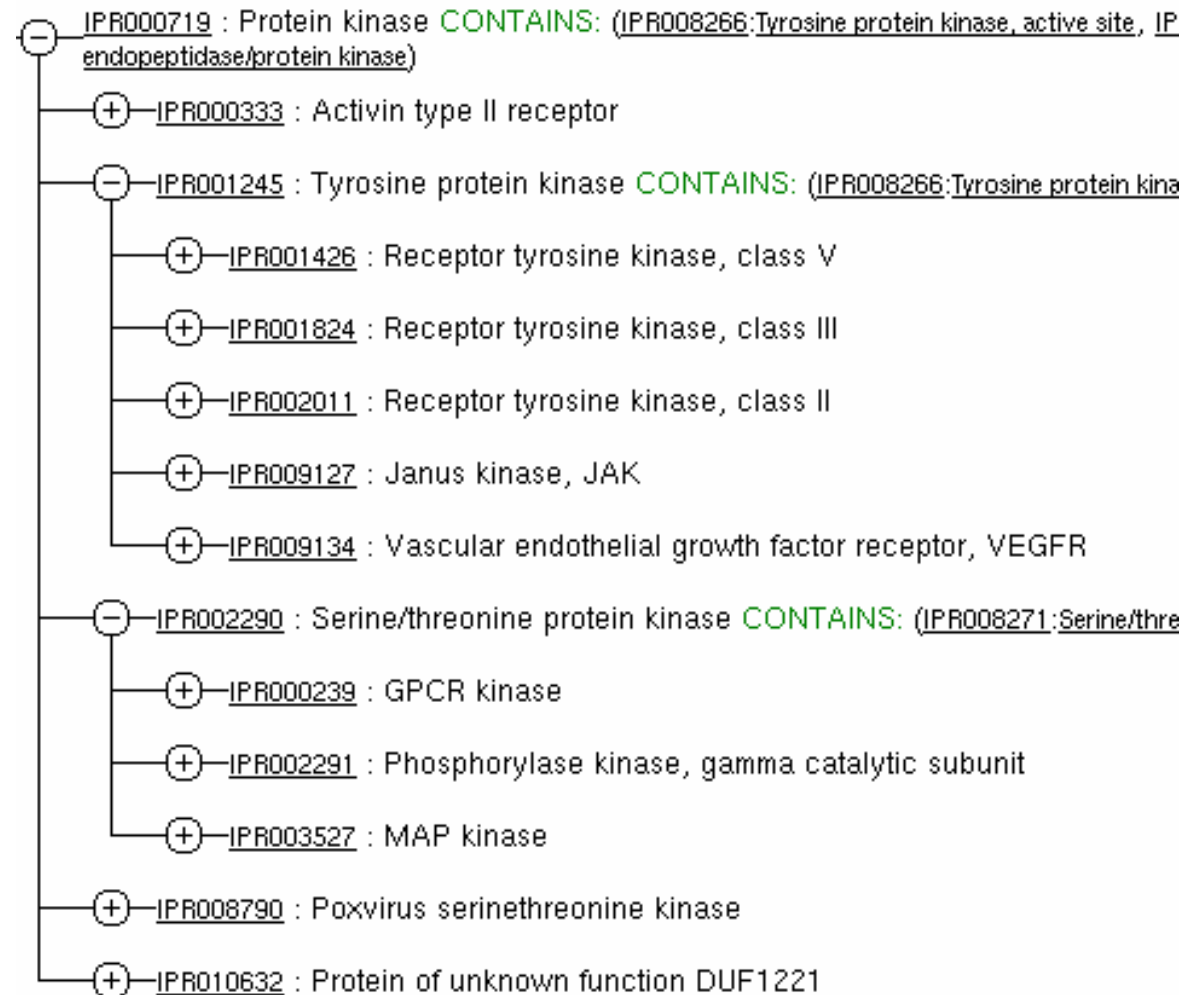
Motifs, regular expressions (low coverage ~1200 families)

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

Where?

| 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



Automatic methods to classify proteins: ProtoMap

Where?

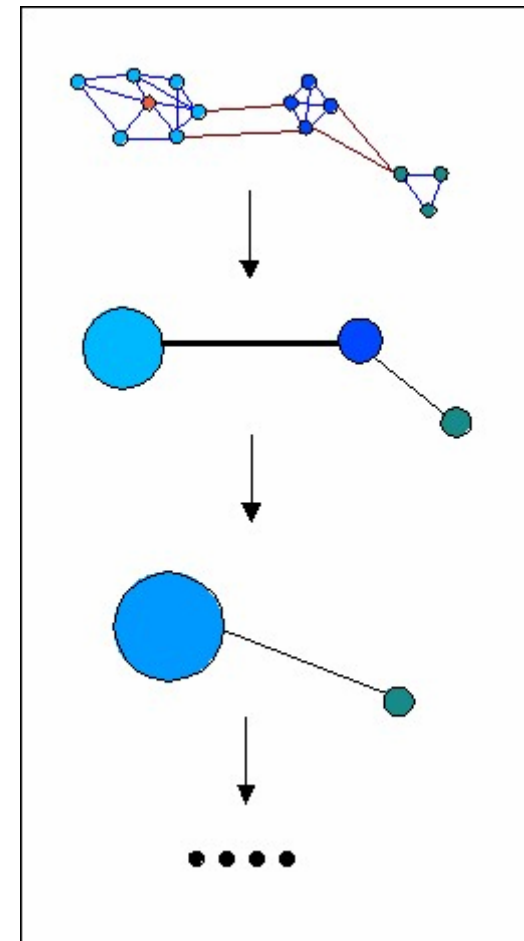
Based on **sequence distance**

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

- 0^o.- Get sequence distances=> graph
 - 1^o.- Grouping close related sequences (e-value < 1e-100)
 - 2^o.- Initialise $T = 1e-95$.
 - 3^o.- Computing cluster distances:
 - geometrical mean of e-values between each cluster pair.
 - If no edges: assignment of e-value=1
 - 4^o.- If the e-values mean is lower than rootsquare of T , clusters are joined.
 - 5^o.- Decrease the T value $T: T = T*1e+05$.
 - 6^o.- If $T > 1$ => stop. Else => go back to 3^o.
- Sequential implementation of T values (1e-95 -> 1e-90 -> 1e-85 ... 1e-00=1)
allows a hierarchical classification of the proteins.



- Home
- All species vs all
- Human vs all
- Gene search
- Text search
- Blast search
- Downloads
- Old version
- Summary
- FAQ



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



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
 - ▶ Parologue 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.



- 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

How to's?

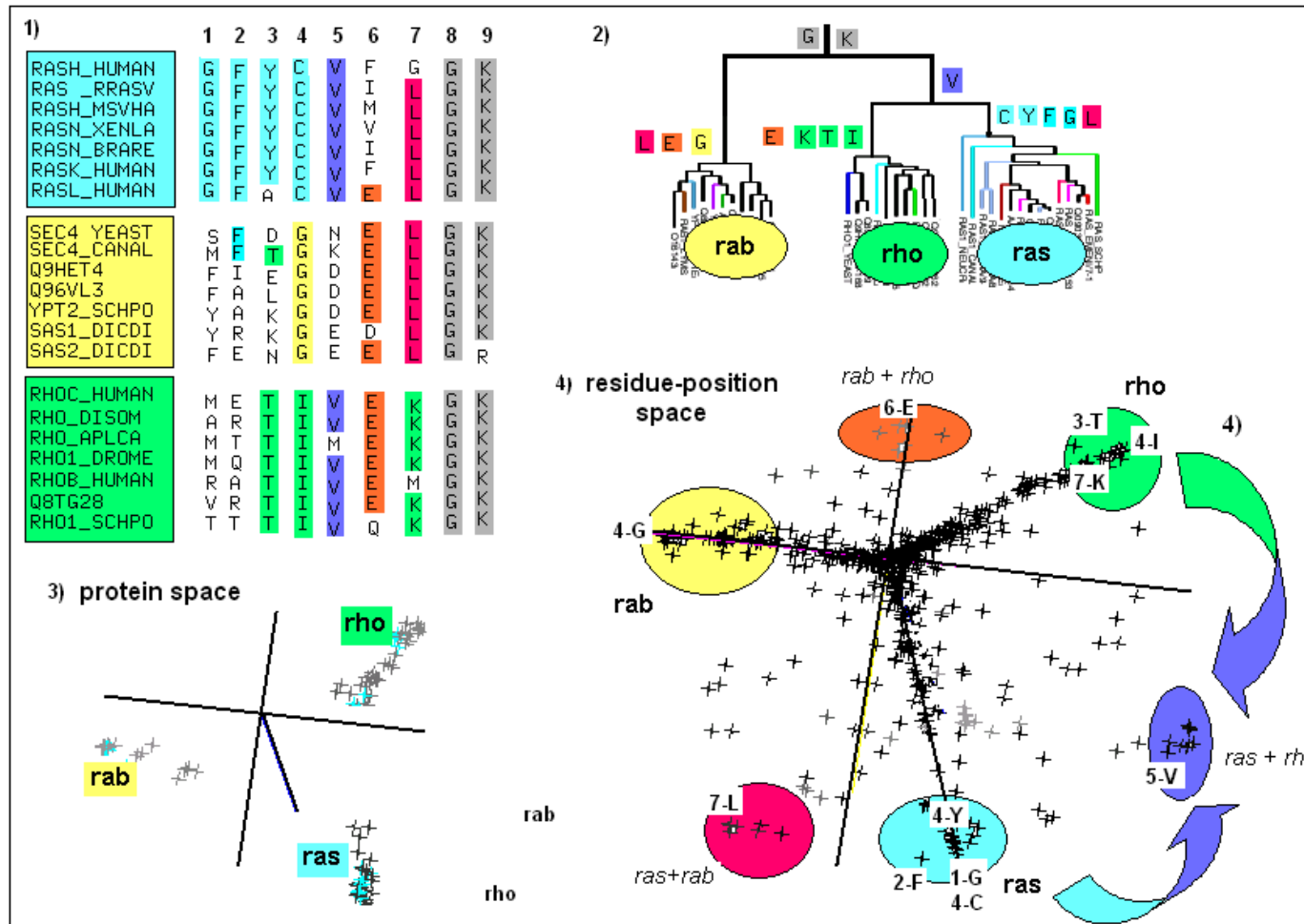
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|------------|---|-----|-------------|-----|----|-------|-------|------|------|---|---|---|----|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RASH_HUMAN | 5 | 164 | KLVVWGAGGGV | GK | A | TIQLI | N | FVDE | D | F | E | S | YR | KQ | V | I | D | G | E | T | C | L | L | D | I | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | | | | | |
| RRAS_HUMAN | 1 | 160 | KLVVWGAGGGV | GK | A | TIQFI | S | FVSD | D | F | E | S | Y | T | K | I | C | S | V | D | G | I | P | A | R | L | D | I | L | D | T | A | G | Q | E | E | F | G | A | N | | | | | | | | | | |
| RTC1_HUMAN | 1 | 160 | RLVVWGAGGGV | GK | A | TIQFI | S | FVTD | D | F | E | S | Y | T | K | Q | C | V | I | D | D | R | A | A | R | L | D | I | L | D | T | A | G | Q | E | E | F | G | A | N | | | | | | | | | | |
| RAS2_HYDMA | 1 | 160 | KLVVWGAGGGV | GK | A | TIQFI | S | FVQD | D | F | E | S | Y | R | K | Q | C | V | I | D | D | K | V | A | H | L | D | I | L | D | T | A | G | Q | E | E | F | S | A | N | | | | | | | | | | |
| RAS2_DROME | 1 | 160 | KLVVWGAGGGV | GK | A | TIQFI | S | FVTD | D | F | E | S | Y | T | K | Q | C | N | I | D | D | V | P | A | K | L | D | I | L | D | T | A | G | Q | E | E | F | S | A | N | | | | | | | | | | |
| RASL_NEUCR | 1 | 160 | KLVVWGAGGGV | GK | C | ← | → | G | FLDE | D | F | E | S | Y | R | K | Q | C | T | I | D | N | E | V | A | L | L | D | I | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | |
| RASL_MOUSE | 1 | 159 | KLVVWGAGGGV | GK | A | TIQLI | N | FVDE | D | F | E | S | Y | R | K | Q | V | I | D | G | E | T | C | L | L | D | I | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | | | |
| RAS1_YEAST | 1 | 160 | KIVVWGAGGGV | GK | A | TIQFI | S | FVDE | D | F | E | S | Y | R | K | Q | V | I | D | D | K | V | S | I | L | D | I | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | | | |
| RAS_SCHPO | 1 | 160 | KLVVWGDGGV | GK | A | TIQLI | S | FVDE | D | F | E | S | Y | R | K | K | C | E | I | D | G | E | G | A | V | L | D | L | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | | |
| RAS_LENED | 1 | 160 | KLVVWGAGGGV | GK | A | TIQFI | S | FVDE | D | F | E | S | Y | R | K | Q | C | V | I | D | D | E | V | A | L | L | D | V | L | D | T | A | G | Q | E | E | Y | G | A | N | | | | | | | | | | |
| RAS2_RHIRA | 1 | 160 | KIVVWGDGGV | GK | A | TIQFI | S | FVDE | D | F | E | S | Y | R | K | Q | C | L | I | D | S | E | C | A | M | L | D | I | L | D | T | A | G | Q | E | E | Y | S | A | N | | | | | | | | | | |
| RALA_HUMAN | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RALA_RAT | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RALB_HUMAN | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RALB_RAT | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RALB_XENLA | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RAL_DISOM | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| RALA_DROME | 1 | 157 | KVIMVGS | GGV | GK | A | TLQFM | D | FVED | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | V | Q | I | D | I | L | D | T | A | G | Q | E | Y | A | A | I | | | | | | | | | |
| CE1393944 | 1 | 157 | QVIMVGT | GGV | GK | A | TLQFM | D | FVEE | Y | E | T | K | A | S | Y | R | K | K | V | L | D | G | E | E | C | S | I | D | I | L | D | T | A | G | Q | E | Y | S | A | I | | | | | | | | | |
| CC42_DROME | 1 | 156 | KCVVWGD | GAV | GK | C | ← | → | N | F | P | S | E | M | F | T | N | Y | A | V | T | V | M | I | G | G | E | P | Y | T | L | G | L | F | D | T | A | G | Q | E | Y | D | R | L | | | | | | |
| RH04_YEAST | 1 | 155 | KIVVWGD | GAV | GK | C | ← | → | G | F | P | T | D | I | F | F | N | Y | V | T | N | I | E | G | P | N | G | Q | I | E | L | A | L | W | D | T | A | G | Q | E | Y | S | R | L | | | | | | |
| RH02_SCHPO | 1 | 156 | KLVIIGD | GAC | GK | S | ← | → | G | F | P | T | E | M | F | T | V | F | N | Y | V | S | D | C | R | V | D | G | K | S | V | Q | L | A | L | W | D | T | A | G | Q | E | Y | E | R | L | | | | |
| RH02_YEAST | 1 | 156 | KLVIIGD | GAC | GK | S | ← | → | G | F | P | E | Q | H | F | T | V | F | N | Y | V | T | D | C | R | V | D | G | I | K | V | S | L | T | L | W | D | T | A | G | Q | E | Y | E | R | L | | | | |
| RH08_HUMAN | 1 | 156 | KIVVWGD | SQC | GK | A | L | H | V | F | A | D | I | F | P | E | N | M | F | T | V | F | N | Y | T | A | S | F | E | I | D | T | Q | R | I | E | L | S | L | W | D | T | A | G | S | P | Y | D | N | V |
| RH06_HUMAN | 1 | 156 | KLVLVGD | VQC | GK | A | L | Q | V | L | A | D | I | F | P | E | T | M | F | T | V | F | N | Y | T | A | C | L | E | T | E | E | Q | R | V | E | L | S | L | W | D | T | A | G | S | P | Y | D | N | V |
| RH03_YEAST | 1 | 156 | KIVILGD | GAC | GK | S | ← | → | G | F | P | E | V | Y | E | T | V | F | N | Y | I | H | D | I | F | V | D | S | K | H | I | T | L | S | L | W | D | T | A | G | Q | E | F | D | R | L | | | | |
| RH01_SCHPO | 1 | 156 | KLVIIGD | GAC | GK | C | ← | → | G | F | P | E | V | M | F | T | V | F | N | Y | V | A | D | V | E | V | D | G | R | H | V | E | L | A | L | W | D | T | A | G | Q | E | Y | D | R | L | | | | |
| RH01_YEAST | 1 | 156 | KLVIIGD | GAC | GK | C | ← | → | G | F | P | E | V | M | F | T | V | F | N | Y | V | A | D | V | E | V | D | G | R | R | V | E | L | A | L | W | D | T | A | G | Q | E | Y | D | R | L | | | | |
| RH01_ENTHI | 1 | 151 | KIVVWGD | GAV | GK | C | ← | → | G | I | P | T | A | M | F | T | V | F | N | F | S | H | V | M | K | Y | K | N | E | E | F | I | L | D | L | W | D | T | A | G | Q | E | Y | D | R | L | | | | |
| RH0L_DROME | 1 | 156 | KITIVGD | G | M | V | G | K | C | ← | → | N | F | P | E | E | I | F | T | V | F | N | H | A | C | N | I | A | V | D | D | R | D | Y | N | L | T | L | W | D | T | A | G | Q | E | Y | E | R | L | |

Correlated mutations

Tree-determinant

conserved

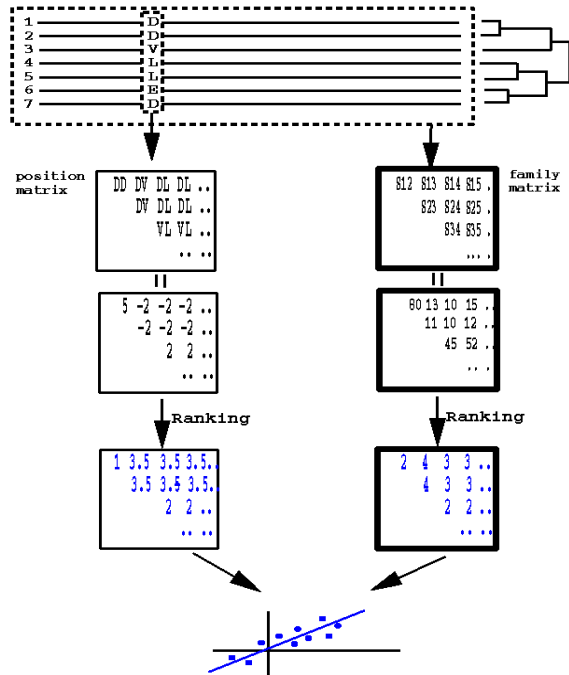
WHAT CAN I LEARN FROM MY ALIGNMENT?



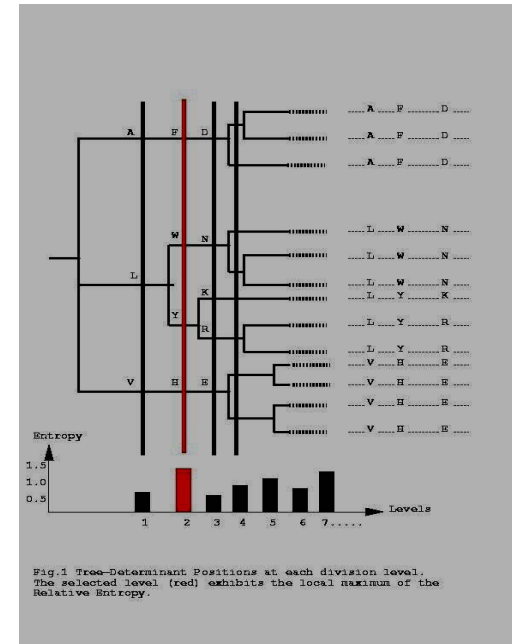
Casari, Sander, Valencia Nature Str. Biol. 95
 Pazos, Valencia 2003

WHAT CAN I LEARN FROM MY ALIGNMENT?

How to's?



Mutational behaviour
Pazos Valencia, 2001



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

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

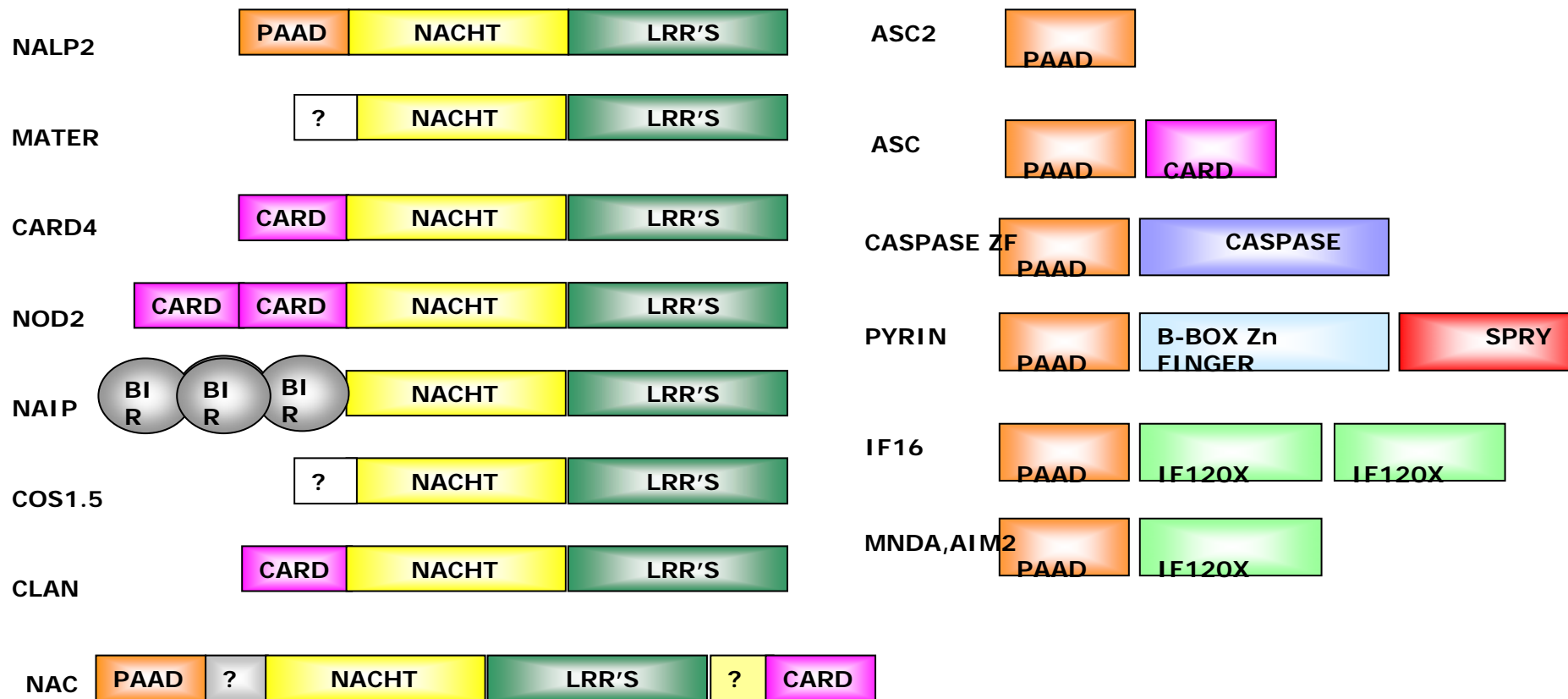
CHECK CONSISTENCY BETWEEN DOMAIN DISTRIBUTION
AND PHYLOGENETIC DISTRIBUTION

CHECK IF SHUFFLING IS RECENT OR OLD...

Case 1: domain shuffling

Get's real!

DOMAIN ARCHITECTURES



NACHT FAMILY

PAAD FAMILY

What are these PAAD and NACHT proteins?

They are involved in inflammation and apoptosis!!!!

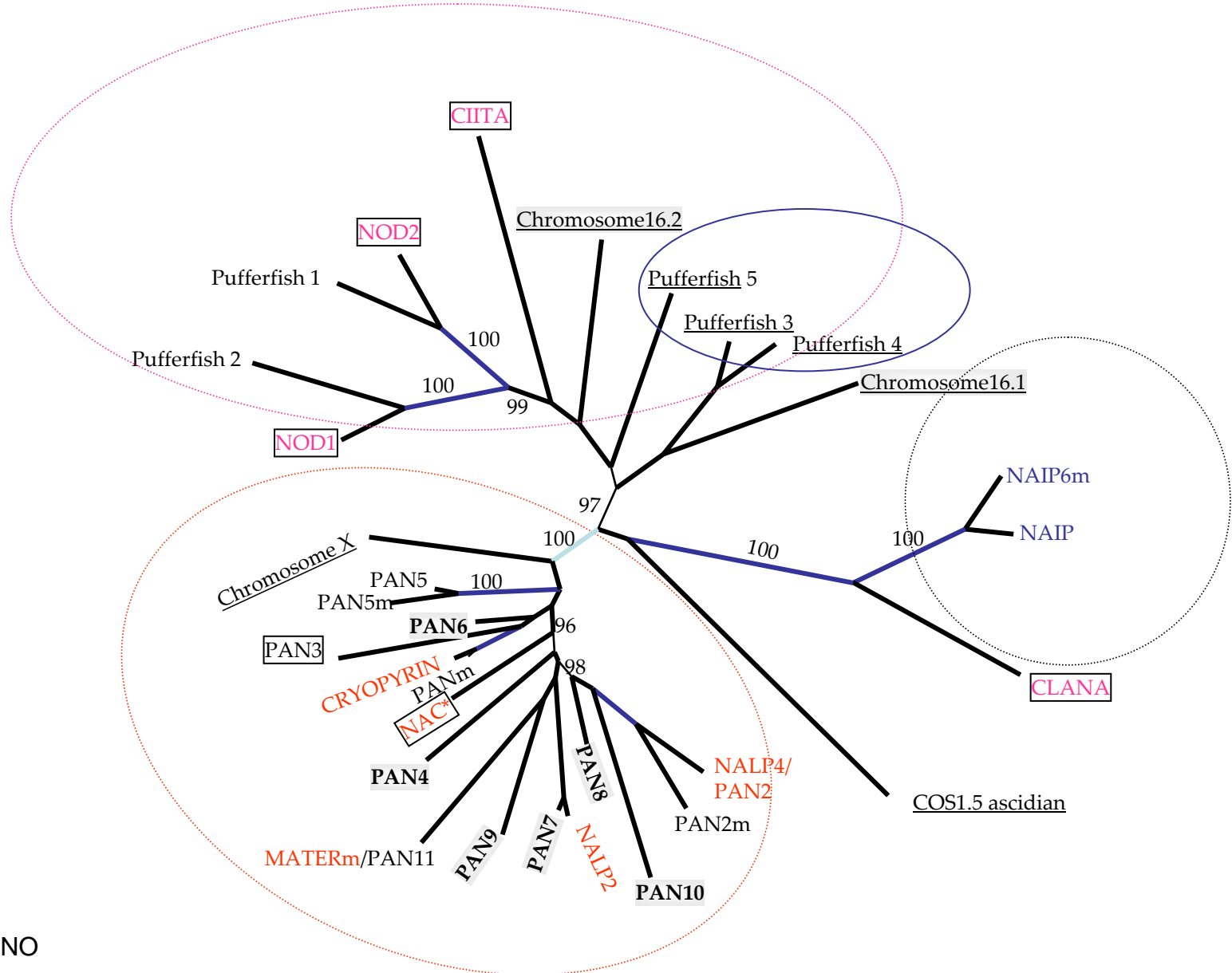
Nacht family: PAN/NALPs/DEFCAP/PYCARD,
CATERPILLER
(Tschopp et al, Nature, 2003)

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

Case 1: domain shuffling

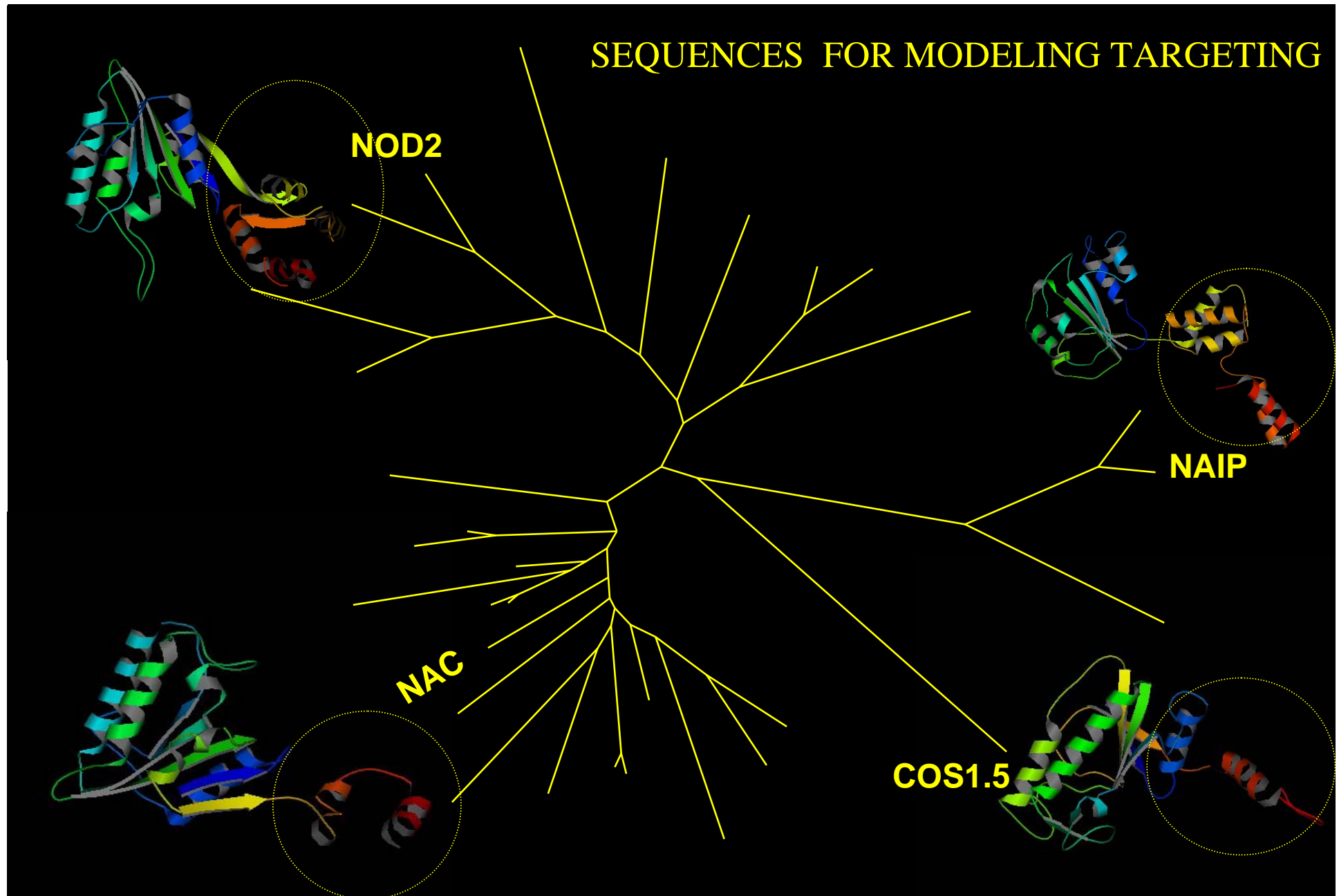
Get's real!

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



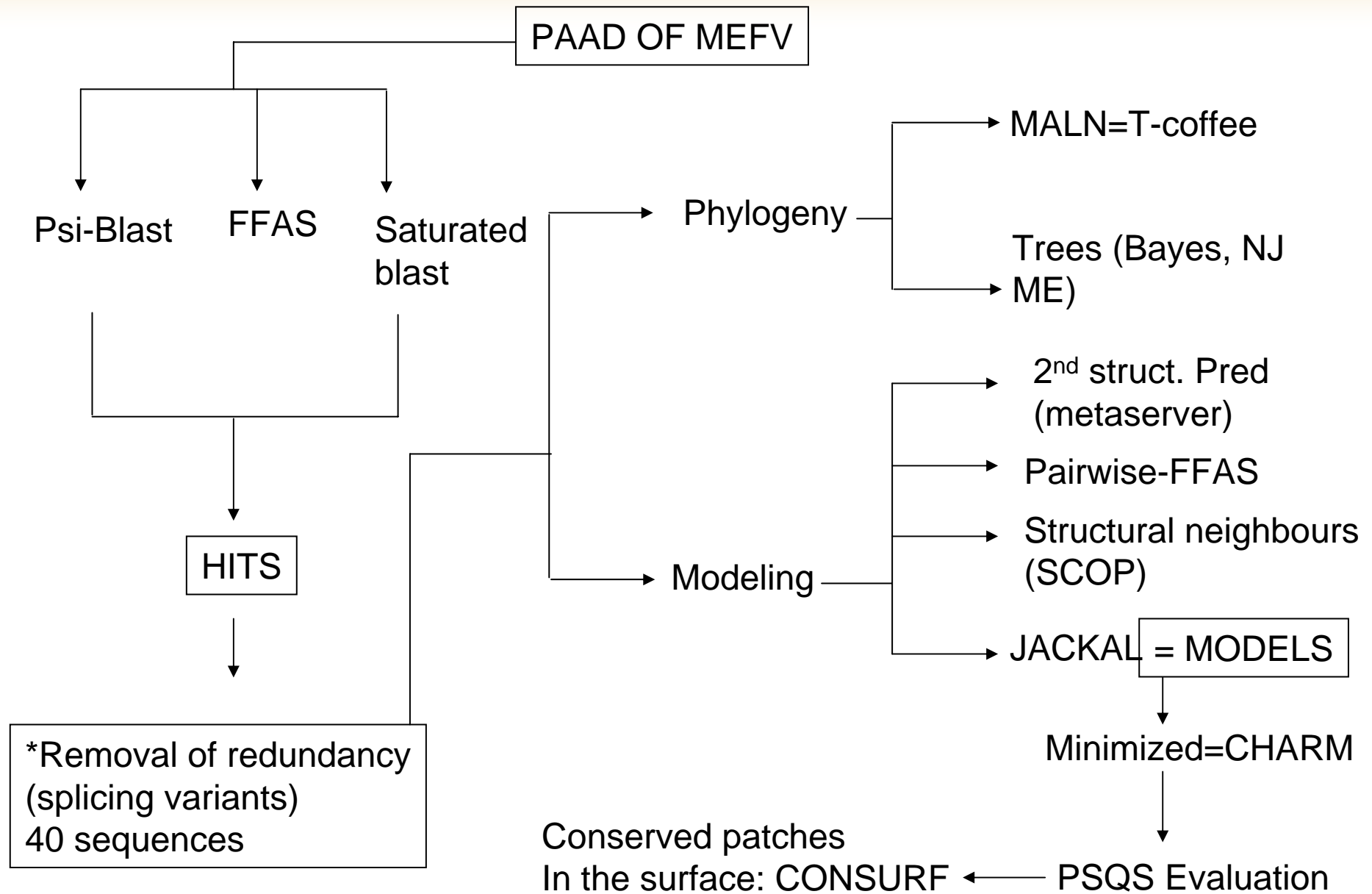
Case 1: domain shuffling

Get's real!



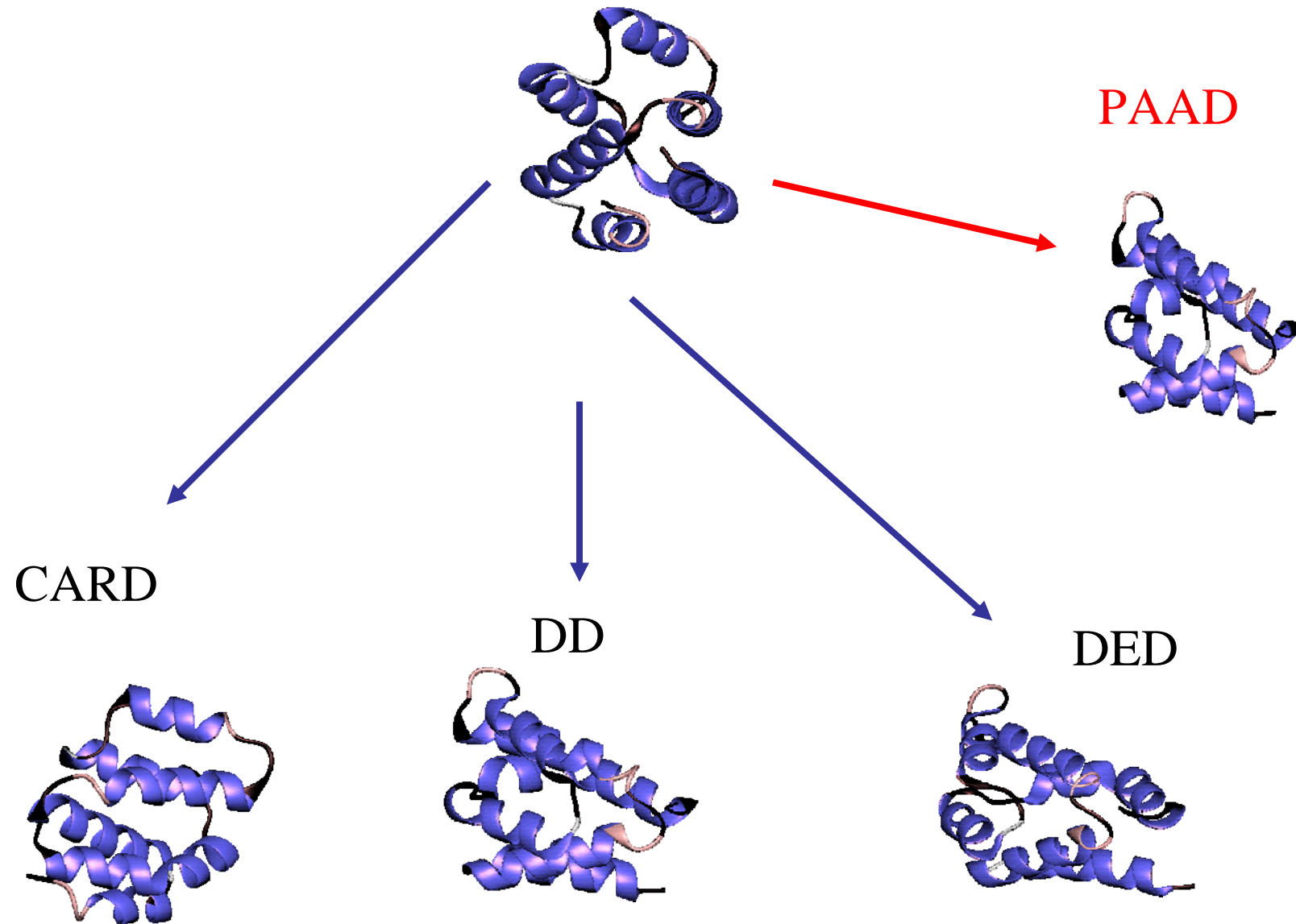
Case 1: domain shuffling

Get's real!



ANCESTORAL DOMAIN

Get's real!



Get's real!

| | | | | | | |
|----------------------|--------------|-----------|-------------|-------------|--------------|-------------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Pyrin | ----- | ----- | ----- | ----- | ----- | ----- |
| Sec_str | HHHHHHHHHH | HHHHHHHH | HHHHHHHH | HHHHHH | HHHHHHHH | HHHHHH |
| MEFV_Huamn | DHLLNLTLEELL | PYDFEKF | KFKLQNTSLEK | GHGSKIPRCHM | QMA-RPVK | LASLLITYYGE |
| ASC_Huamn | DAILDALENL | TAEELK | FKLKLKLLSV | PLREGYGRIP | RGALLSM-D | ALDLDTKL |
| ASC-PENDING-Mouse | DAILDALENL | SGDE | LKFKKMLL | TVQLREGYGR | IPRGALLOM-D | AIDLTDKL |
| PYCl_Huamn | EAILKVL | ENLTPEEL | LKFKKMLL | GTVPLE | RGFGRI | PRGALGOL-D |
| MEFV_Rat | DHLLNLTLEELL | PYDFEKF | KFKLHTTSLEK | GHGSRIP | PLSLVKMA-RP | IKLTRL |
| MEFV_Huamn | DHLLSTLEEL | VVPYDFEKF | KFKLQNTS | VQKEHSRIP | RSQIQRA-RPV | KMATLLV |
| AF427617_1_Huamn | CKLARYLE | DELDV | LKFKKMLLED | YPPQKCI | PLPRGQTEKA-D | HVDL |
| ASC1_zebrafish | EHLQEA | FEDL | GADNLR | KFKSKLGD--- | RRQEP | RVTKSAIEK |
| LOC280619_Mouse | EALLWAL | NDLEENS | FKTLKFHL | RDVT--- | QFHL | ARGELES |
| AF233434_1_Zebrafish | DHLQDAL | SNIGADN | LRRFQ | SRLGD--- | RRQEP | RVTKSTIEK |
| AF327410_1_Zebrafish | QLLSDV | LEDLVE | AELKQF | TROQLW- | IGVKP | GVPIPRGK |
| CAAB01003190_Fugu | --LLKIL | EDLLK | EDFKTF | KWYLT-LD | LLENC | NP |
| CAAB01007457_Fugu | KLLKDF | LDEL | DDTML | REFK | WYLGQHK- | ERGS |

| | | | | | | | | | | | | |
|-------------|---------|---------|--------|--------|----------|-------|---------|--------|---------|--------|-------|--------|
| PAN | ---- | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| Sec_str | ----- | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| PAN2_Huamn | FGLMWY | EELKKEE | FRKFK | HEHLKQ | MTLQEL | KQIPW | TEVKKAS | REELAN | LLIKHYE | EQQAMN | ITLRI | FQKMD |
| PAN3_Huamn | ELLLA | AELSE | QELK | FRHKL | RDV--- | PDGRS | IPWGR | LERAD | AVDLA | EOLAQ | FYGP | PEPA |
| PAN10_Huamn | FDLLWY | ENLSD | KFPQS | FKKYL | ARKIL--- | DFKLP | QFPLI | QMTKEE | LANVPL | ISYEG | QYI | WNML |
| PAN4_Huamn | NGVM | LYHRN | VSHEE | LQRF | KQLL | TE--- | LSTG | TMPI | TWDQ | VETAS | WAEV | VHLL |
| PAN1_Huamn | FNLQAL | LEQLS | QDEL | SKFK | YLIT | TFSLA | HELQK | IPHEV | DKAD | GKQ | LVEIL | TTCH |
| PAN7_Huamn | WTLQ | TLEQL | NEDEL | KSF | KSLW | AFPLE | DVLO | KTPW | SEVEE | ADG | KLAE | ILV |
| PAN8_Huamn | FGLLLY | EELNKEE | LNTFK | LFKE- | TMEPE | HGLTP | WNEV | KKAR | REDLAN | LMMK | KYYP | GEKA |
| PAN11_Huamn | YGLQ | WCLYEL | DKEEF | QTFK | ELLK | KSSES | ETCS | IPQF | EIENAN | VECL | ALLH | EYGA |
| PAN6_Huamn | CRLS | TYLE | ELEAVE | LKFK | LYL | GTAT- | ELGEG | KIPW | GSMEK | AGPLE | MAQLL | ITHF |
| PAN5_Huamn | EALLWAL | SDLEEN | DFK | LKFL | YLRD | MTLSE | QOPPL | ARGE | LEGL | IPVD | LAE | LI-SKY |

| | | | | | | | | | | | |
|------------|----------|-------|----------|-------|------|------|----------|-------|----------|-------|----------|
| AIM | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| Sec_str | HHHHHHHH | ---- | HHHHHHHH | ----- | HHHH | ---- | HHHHHHHH | ----- | HHHHHHHH | ---- | HHHHHHHH |
| AIM2_Huamn | ILLLT | TGLDN | ITDEE | LDRF | KFLS | DEFN | IATG | KLHT | ANR | IQVAT | LMION |
| AIM2_Mouse | MLLLT | TGLDH | ITEE | LKR | FKYF | ALTE | FQIAR | STLD | VADR | TE | LADHL |
| AIM2_Rat | MLLLT | TGLDH | ITEE | LKR | FKYF | ALTE | FQIAR | STLD | VADR | TE | LADHL |

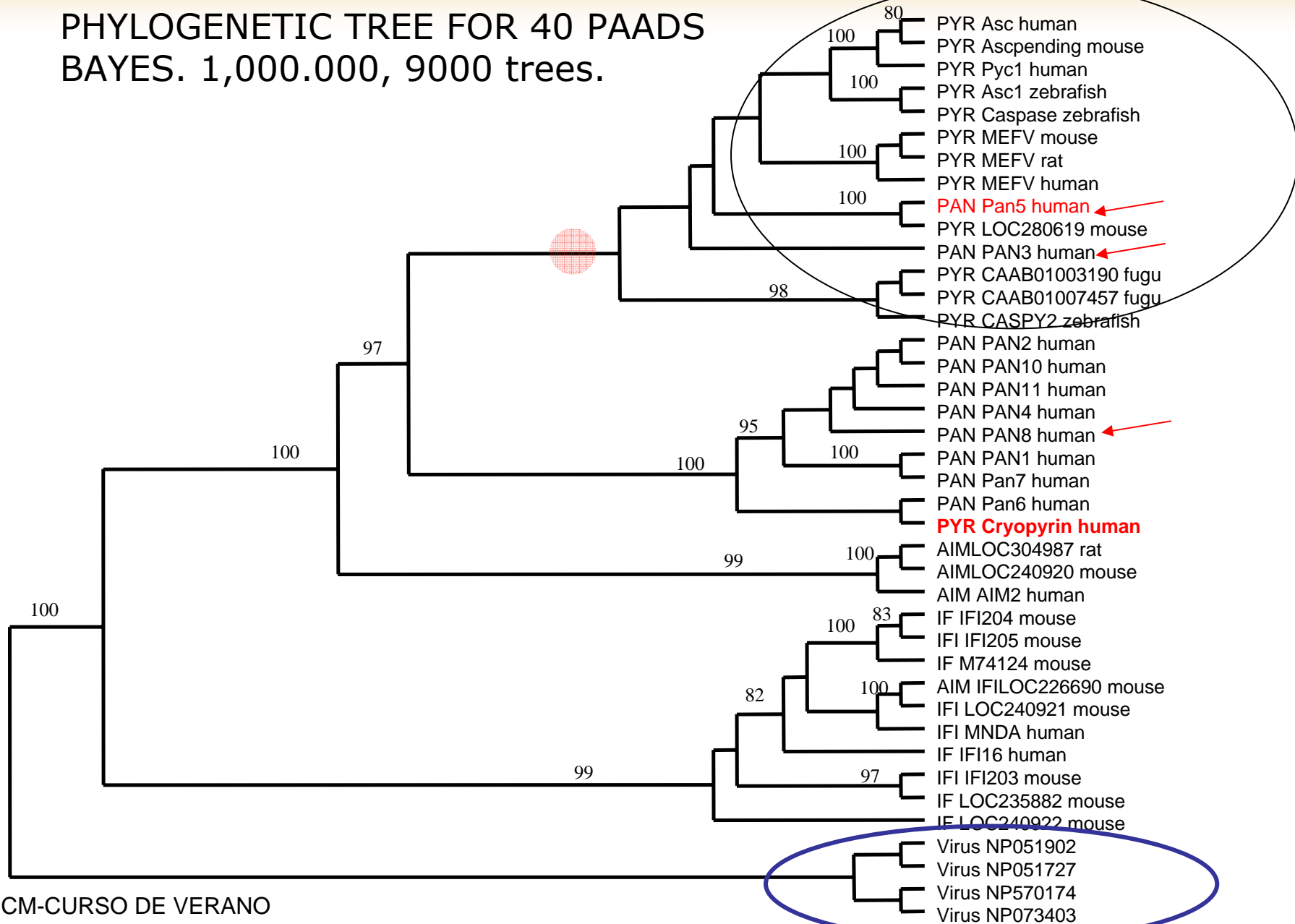
| | | | | | | | | | | | |
|-----------------|--------|------|----------|-------|--------|-------|----------|-------|--------|------|------|
| IFI | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| Sec_str | HHHHHH | ---- | HHHHHHHH | ----- | HHHH | ---- | HHHHHHHH | ----- | HHHHHH | ---- | HHHH |
| If1204_Mouse | IVLLR | GLEC | INKHY | SLFK | SLLARD | LNLER | DNQ | EYTTI | QIAN | MEEK | FAD |
| If1203_Mouse | IVLLK | GLEN | MEDY | QFR | TVK | SLLR | KELK | LTKK | MOED | YDR | IQLA |
| MNDA_Huamn | ILLLK | GFE | LMD | YH | TSIK | SL | LAYD | LGLT | TKM | QEE | YNR |
| If116_Huamn | IVLLK | GLEV | INDY | HFR | MV | KSL | SN | DLK | LN | LKM | REY |
| If1205_Mouse | IVLLR | GLEC | INKHY | SLFK | SLLARD | LNLER | DNQ | EYTTI | QIAN | MEEK | FAD |
| LOC226690_Mouse | IVLLS | GLEY | MNDY | NFR | AL | KSL | LN | HDLK | L | TKN | MQ |
| LOC240922_Mouse | IVLLT | G | MG | IND | HFR | MV | KSL | SK | EL | KLNR | -MQD |
| LOC240921_Mouse | IVLLS | GLEY | MNDY | NFR | AL | KSL | LN | HDLK | L | TKN | MQ |
| LOC235882_Mouse | IVLLE | GLEN | MGDY | QFR | TVK | SLLR | KELK | LTKK | LOED | YDR | IQLA |
| M74124_Mouse | IVLLE | GLEC | INKH | QNF | L | FK | S | L | NV | KDLN | LE |

| | | | | | | | | | | | | |
|-----------------------|-------|------|------|------|-------|------|-------|------|------|------|------|------|
| Virus | ----- | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| Sec_str | ----- | HHHH | ---- | HHHH | ----- | HHHH | ---- | HHHH | ---- | HHHH | ---- | HHHH |
| 18L_Yaba Like Disease | SALIF | SLED | VTHY | QKIL | IFLTK | DELN | ISDEE | KQL | DRVD | FAE | KL | |
| SPV014_Swinepox | YTIIS | VLER | LTPY | QFK | TLL | FLIQ | DDIN | ISN | DD | IN | V | |
| GP013L_Rabbit Fibroma | GVII | ITV | ENL | TDY | QK | MFLY | LV | TED | LR | IN | P | |
| M013L_Myxoma | GVII | ITV | ENL | SDY | QK | MFLY | LV | TED | LR | IN | P | |

{Rojas et al, Prot. Sci. 2003}

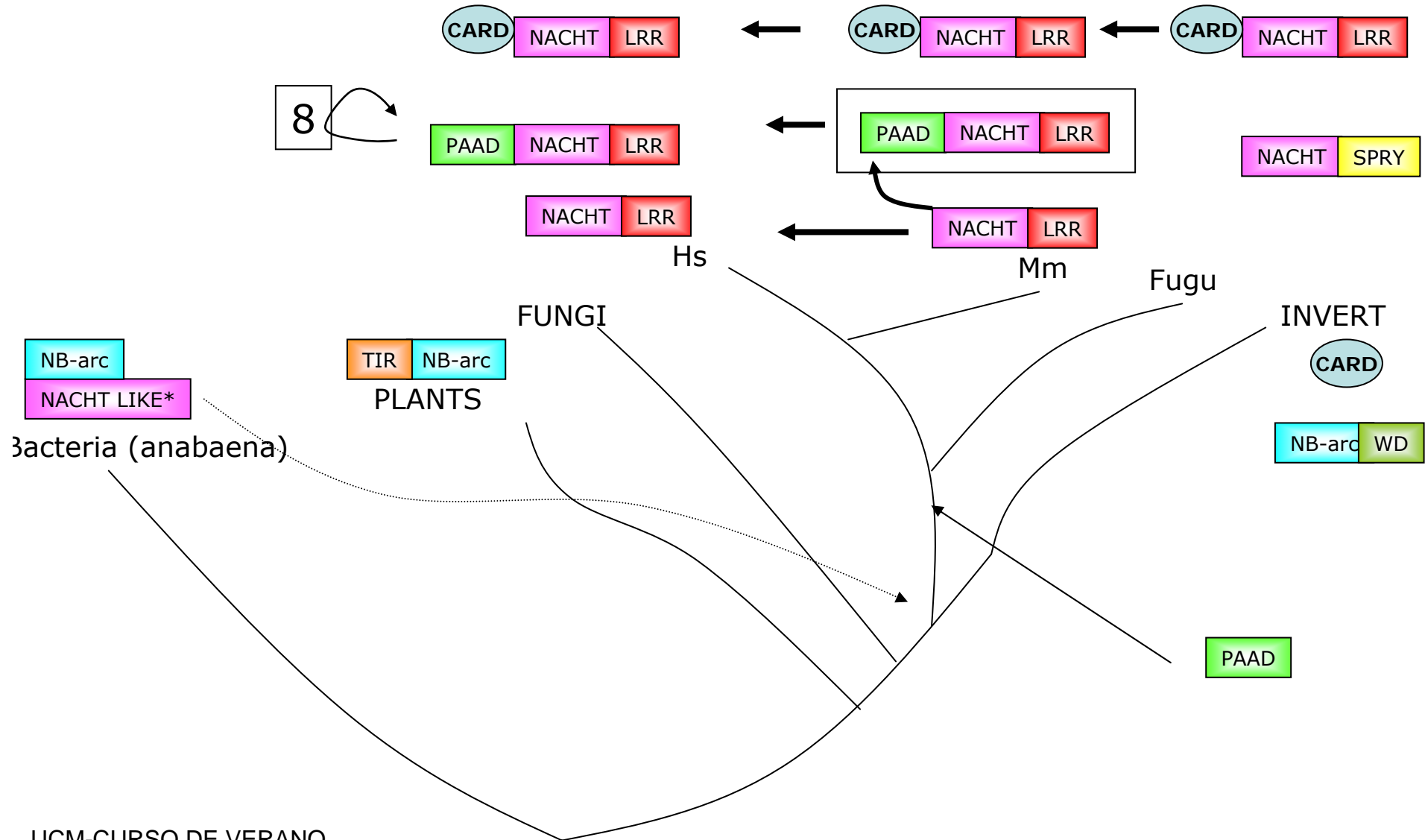
Get's real!

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



Case 1: domain shuffling

NACHT DISTRIBUTION: POSSIBLE SCENARIO



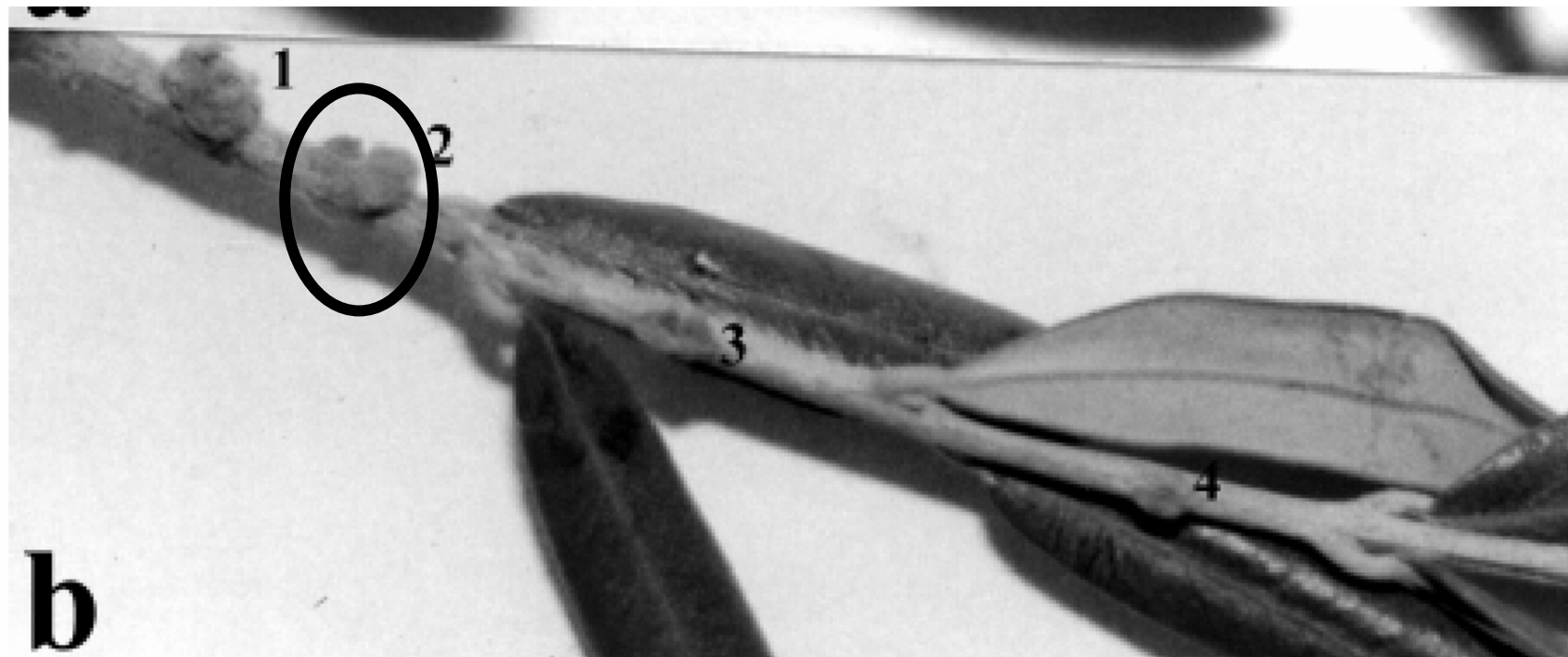
Case 2: novel species

- DESCRIPTION OF NEW SPECIES
Erwinia toletana sp. nov.

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!

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:

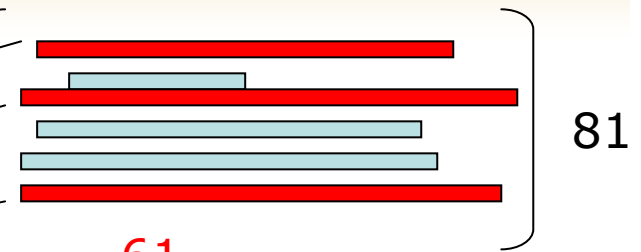
16SRNA, 23SRNA, gnd, mdh

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

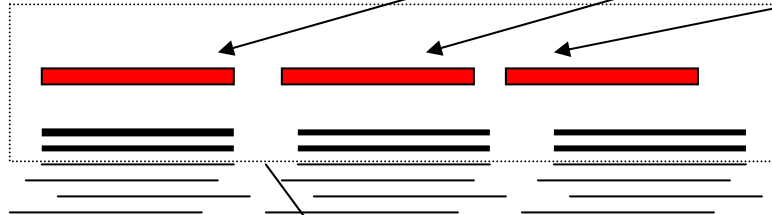
Case 2: novel species

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

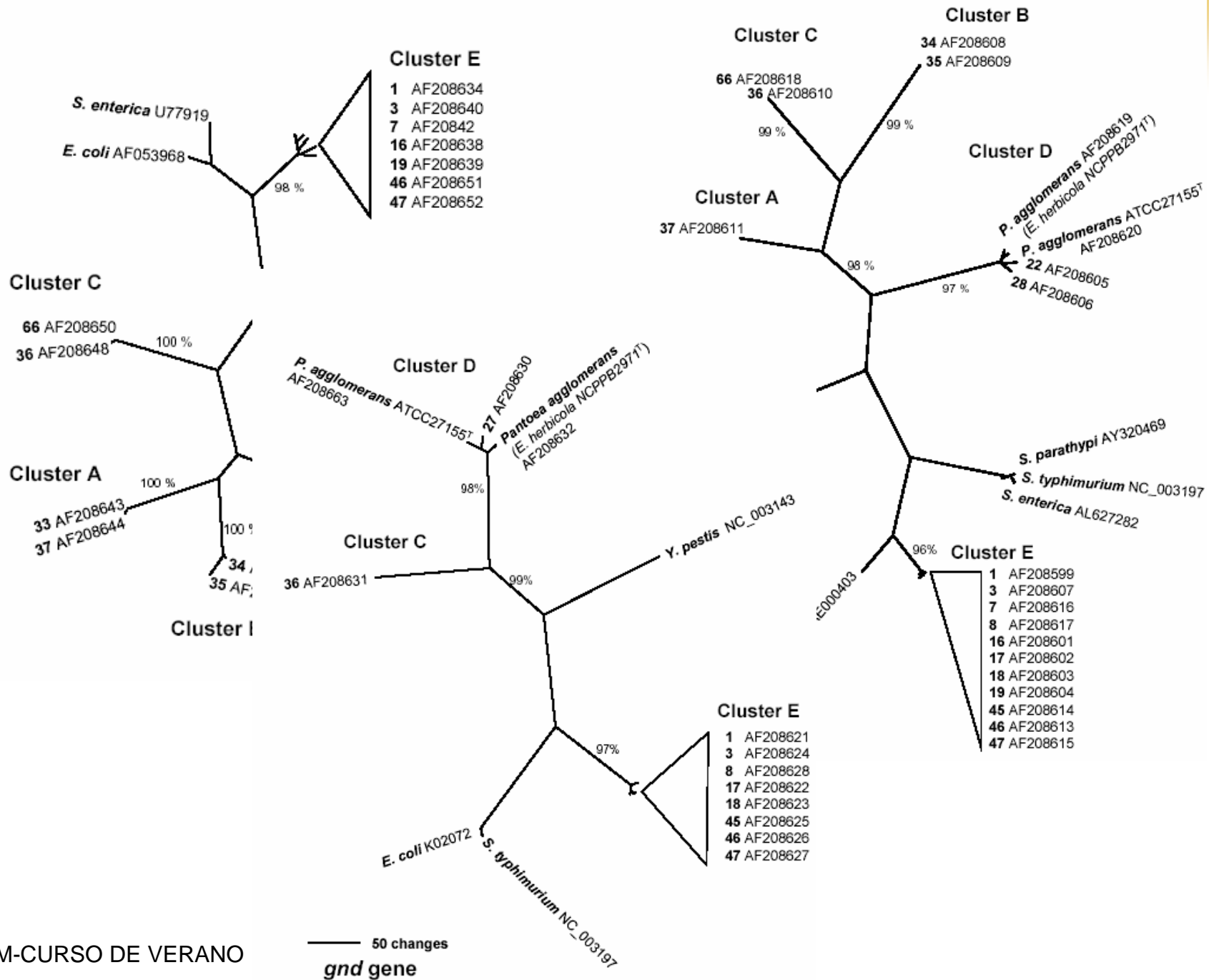
Parsimony

Maximum likelihood

BioNJ

1000 bootstrap

CONSENSUS!



Case 2: Extending the family

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

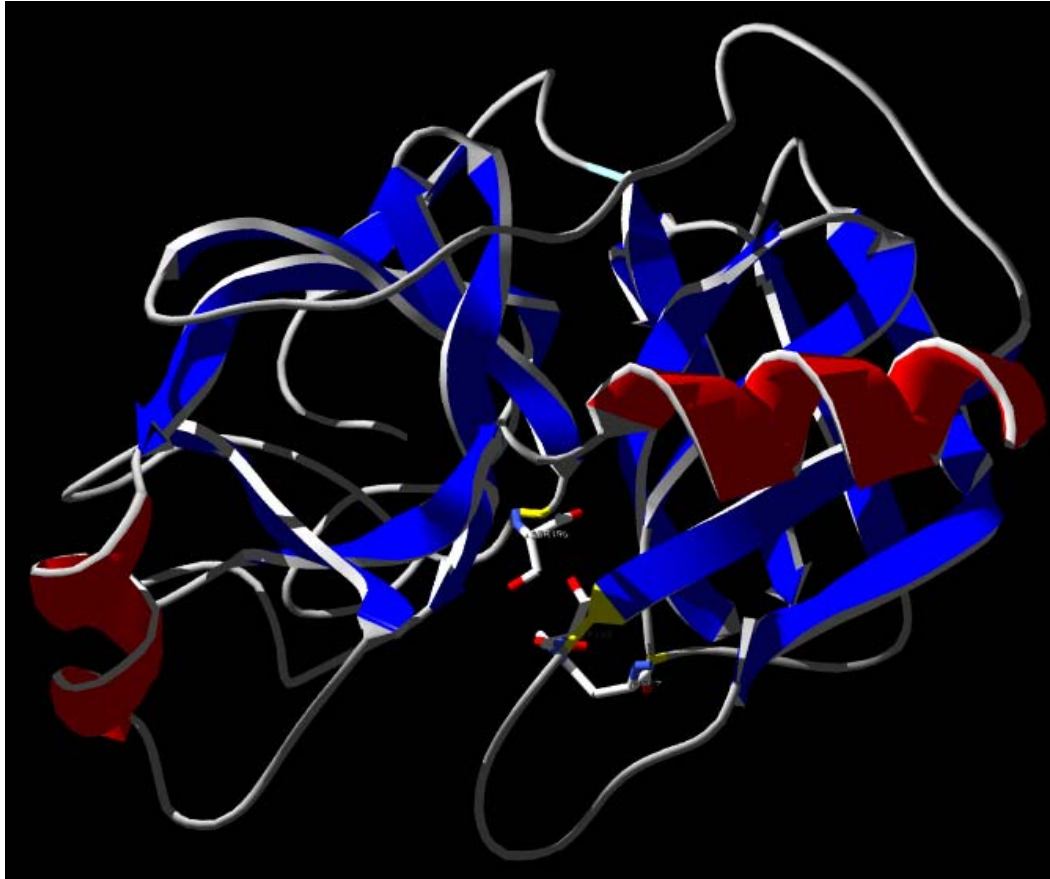
Case 2: Extending the family

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!

Case 2: Extending the family

What are serine proteases?

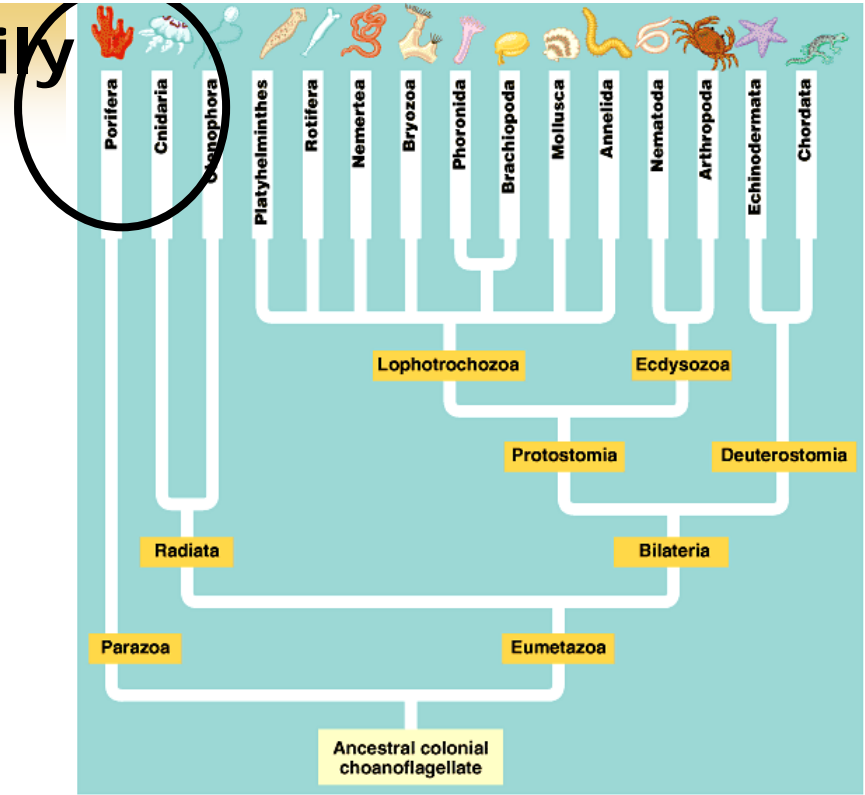
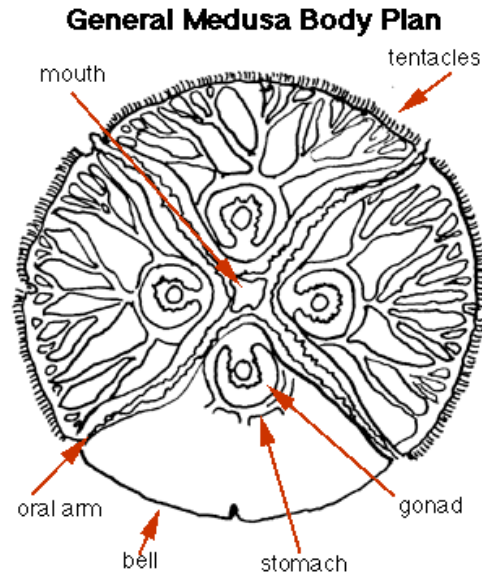
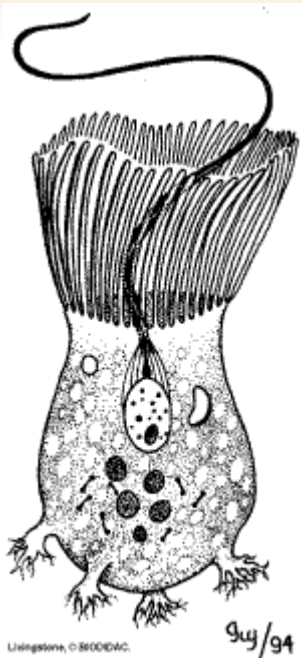


- Hundreds of entries
- Disulfide bonds
- Cleavage peptide
- Digestive:
trypsin, chymotrypsin
- no digestive:
blood clotting
elastases

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

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

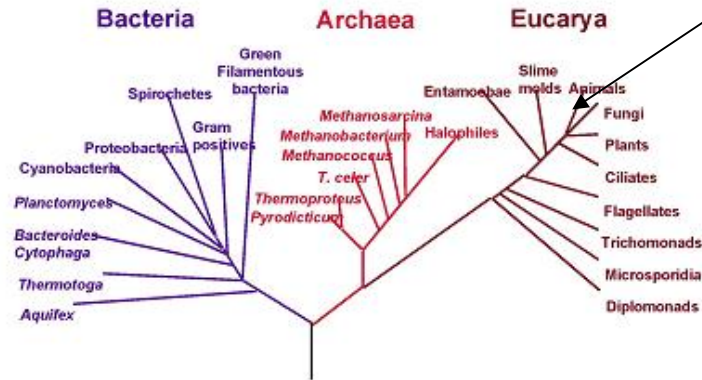
Case 2: Extending the family



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



Phylogenetic Tree of Life



{Rojas & Doolittle, 2002, JME}

Case 2: Extending the family

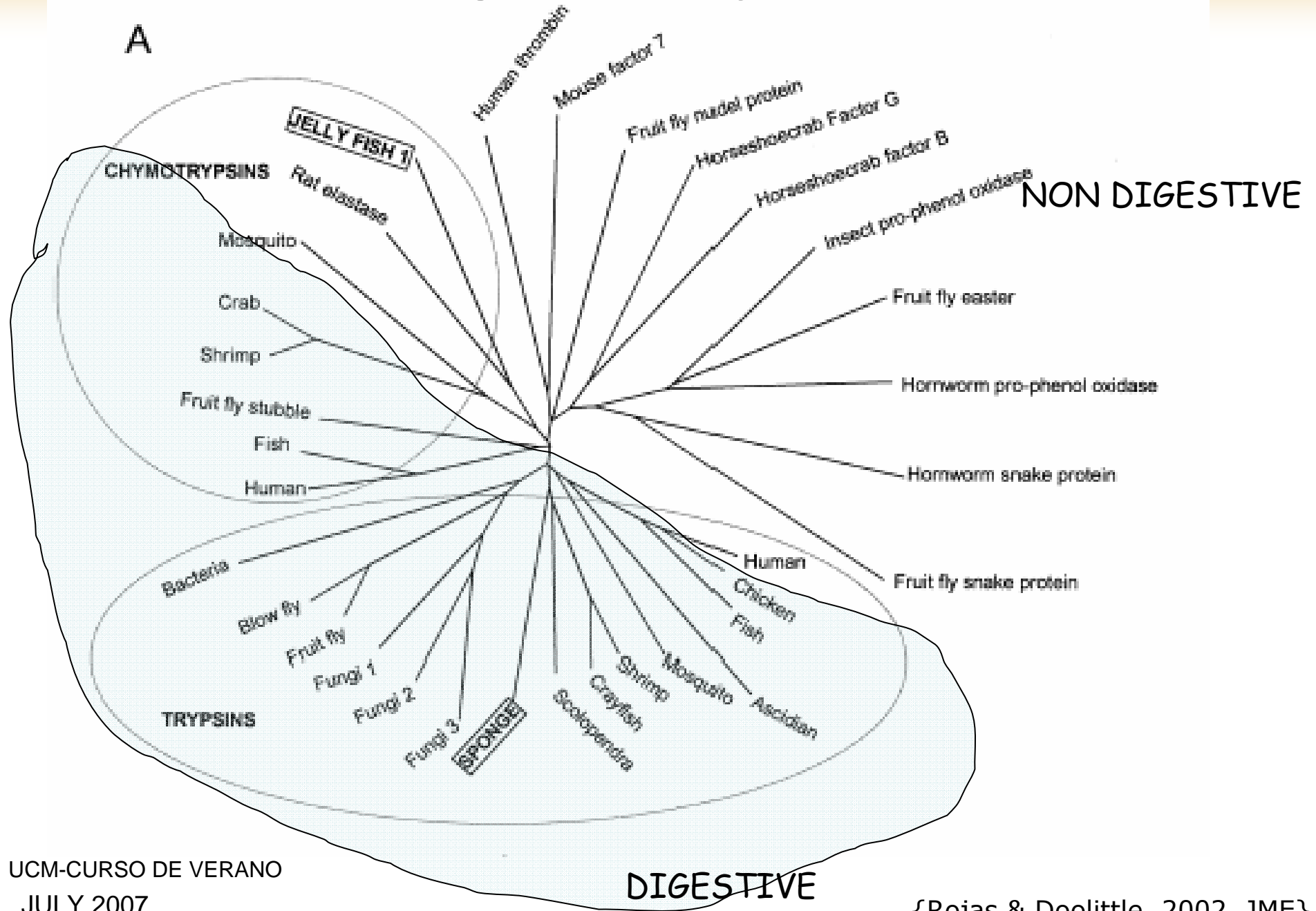
```

Sponge trypsin      :IVGQDPVNEKGYVWQVSLQ-REGFFGRSH--FCGGSLDADTVLTAACCTDQQ---VPSG-ITVVAQDHVLSSTTDGDRQVVGVASISRHP : 8
Shrimp trypsin     :IVGQTDATPGELAYQLSFQ-DISPGFANK--FCGASIIYNEHMAIGQICVQGEDSRPFDY-LQVVAQHLNODVDEGTEQTVILSKIIQHE : 8
Human trypsin      :IVGQYNCEENSVPYQVSL--NSGY----H--FCGGSLINEQWVVAACICYK-----SR-IQVRLQHHNIEVLEGNQFQFINAAKII RHP : 7
Fruit fly trypsin  :IVGQSATTISSFPWQISLQ-RSG----SH--SCGGSIYSANIIIVTAACCLQSV---SASV-LQVRAQSTYWS--SGG-VVAKVSSPKMHE : 7
Human chymotrypsin :IVGQEDAVPGSMPWQVSLQDKTGF----H--FCGGSLISEDWVVAACIC--GV---RTSD-V-VVAQDFDQGSDEENIQVLKIAKVPKMP : 7
Shrimp chymotrypsin :IVGQVEATPNSMPWQAAL----FIDDMY--FCGGSLISSEWVLTAAICMDG----AGF-VEVVLCAHNIHQNEASQVSIITSTDFPTHE : 7
Rat elastase       :VVGQGEASPNSEWQVSLQ-YLSSGKQKH--TCGGSLVANRWVLTAAICIS-----NSRT-YRVLLQKSLSTSESGSLAVQVSKLVVHE : 8
Moonjelly protease 1 :IIISQTHARPGAMPWASLY----MLSRSH--ICGGSLLSNRWILTAAICVVGQ--GATTKN-LVIKLCQHDHYDKDGFQQQFDVEKII PHP : 8
Horseshoecrab factor G :IIIGGIATPNSMPWVGI----FKVMPHRFLCGGSIINKVSVVTAACCLVTQFQNRQNYISIFVRVQAADI---DNSGQNYQVVKVIVHQ : 8
Mouse factor 7     :IVGQAVCPKEGEPWQAVLK-INGLL-----LQAVLLDARWIVTAACCFDNI---RYWGNITVVVQCHDFSEKDGDEQVRRVTVQVIMPD : 8
      G          P          CG          A HC          G

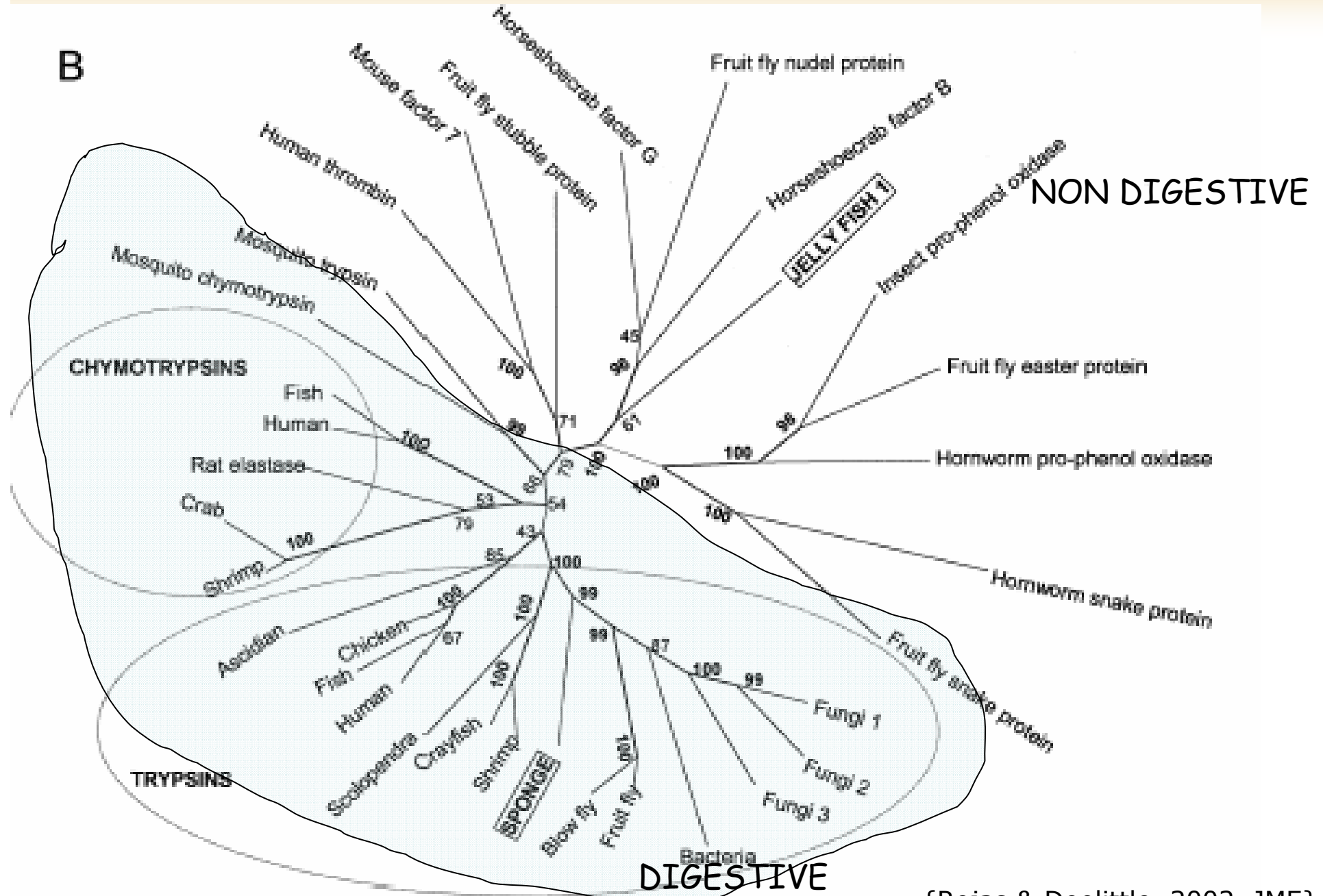
Sponge trypsin      :EYNSRTFY--NDAICVLEKLNLSIIIGGNVQPVGLPPFN-A-EVDEGV-MATVSNQ--TTSAGGSLSDVLLAVNVPVISDARCRGA-----Y :16
Shrimp trypsin     :DYNGPTIS--NDAISLLKQSPLSPNDNVRAIDIP-AQ-G-HAASG--DCIVSNQ--TTSEGGSTPSVQKQVTVPIVSDDECRDA-----Y :16
Human trypsin      :QYDRKTLN--NDAIILIKSSRAVINARVSTISLP--T-A-PPATQT-KCLISQNGTASSGADYDDEIQCCLDAPVLSQAKCEAS-----Y :15
Fruit fly trypsin  :GYNA-TMV--NDAVINLSSLSLSPSSIKAISL--AT-Y-NPANGA-SAAVSGNGTQSSGSSIPSCQYVNVNIVSQSQCASS-----TY :15
Human chymotrypsin :KFSILTVN--NDAITLLKATPARFSQTVSAVCLPSAD-D-DFFAQT-LCATTQNGKTKYNANKTPDKLQQAALPLLSNAECKKS-----W :15
Shrimp chymotrypsin :NWNHSLLT--NDAIILRQPSFVSLMSNIKTVKLP--S-S-DVSVGT-TVPTQNGRPSDSASGISDVRQVNVVPMNADC-DS-----VY :15
Rat elastase       :KQNAQKLSNGHDAIVKQASPVALTSKIQTACLPA-G-TILPNHYPCYVTVNGRQLQTNQA-TPDVQGGRLLVVDYATCSSA-----SW :16
Moonjelly protease 1 :AYKRGPLK--NDAIILIKLTPARINKRVKTI-SLPKQ-S-APSVGSRK-QLAGNSIRHPGGSY-HTLQOAMLPVVSYTHCENQ----- :16
Horseshoecrab factor G :GYKHSY--YDGLILLSKPVEYNDKIQPVCIPEFN-KPHVNLAKIKVVITNG--VTGKATEKRVVIRELELPVVTNEQCNKSYQTLPP :16
Mouse factor 7     :KYIRGKIN--NDAIALLRQRPVPTDYVVVPLCLPEKSFSENTLARIKRSRVSQNGQLDRGA-TALELMSIEVPRIMTQCCLN-----AKH :16
      DI          L          GNG          L          C

Sponge trypsin      :GETD--VADSMICAGDLANGGIDSCGGDGGPLM-G-ST----IIQIVSNQY--GCAYAGYGVYTVQVSYVYVSPFKS----- :230
Shrimp trypsin     :QQSD--IEDSMICAG-VPEGGKDCGGDGGPLAC-S-DTASTYLACIVSNQY--GCARFGYGVYAEVSYHVDWIKANAV----- :237
Human trypsin      :I-PGK--ITSNMPVGFLEGGKDCGGDGGPVLV-N-GQ----LQVVSNGD--GCAQKNKPGVYTKVYNYVVKMINKMTIAANS-- :224
Fruit fly trypsin  :GYGS--QINTMICA---AASGKDACGGDGGPLVS-G-GV----LVGVVSNQY--GCAYSNYPGVYADVAVLRSWVSTANSI--- :224
Human chymotrypsin :GR-R--ITDVMICAG---ASGVSSCGDGGPVLV-Q-KDQANTLVGIVSNQSDT-C-STSSPGVYARVTKLIPVQKILAAN--- :230
Shrimp chymotrypsin :G--I--VGDGVVCI--GTGGKSTCGDGGPLAL-N-GM----TYGITSPGSSAGC-EKGYPAAPTRVYVYLDWIQKQKTVTP-- :226
Rat elastase       :WGSS--YKTNMVCAG--GDGVTSSCGDGGPVLVQA-SNGQNVQVGVIVSGSTLQCNYPKPSVFPTRVSNYIDWINSVIAKN--- :241
Moonjelly protease 1 :-----KNFVCAGFGKSSLTNAFCGGDGGPLVCK-SDGSMEQKGLASV-VVEY-K--YTTAFTPVANYIDWINQHINK--- :230
Horseshoecrab factor G :SKLNRGITHDMICAGF-PEGGKDACGGDGGPLMYPQPTTGRVKIVGVVSGFPE--CARPNFPGVYTRLSYVYVNLQKITPQQ--- :248
Mouse factor 7     :SSNTPKITEINMFCAGY-MDGTKDACGGDGGPLATH--YHGTWYLTGVVSNQY--CCAAIGHIGVYTRVSYIDMLVRHEDSK--- :242
      C          C GDSGGP          G S          C
    
```

Case 2: Extending the family



Case 2: Extending the family



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.

Case 2: Extending the family

WHY THE FUNGAL ONES CLADE WITH ANIMALS?

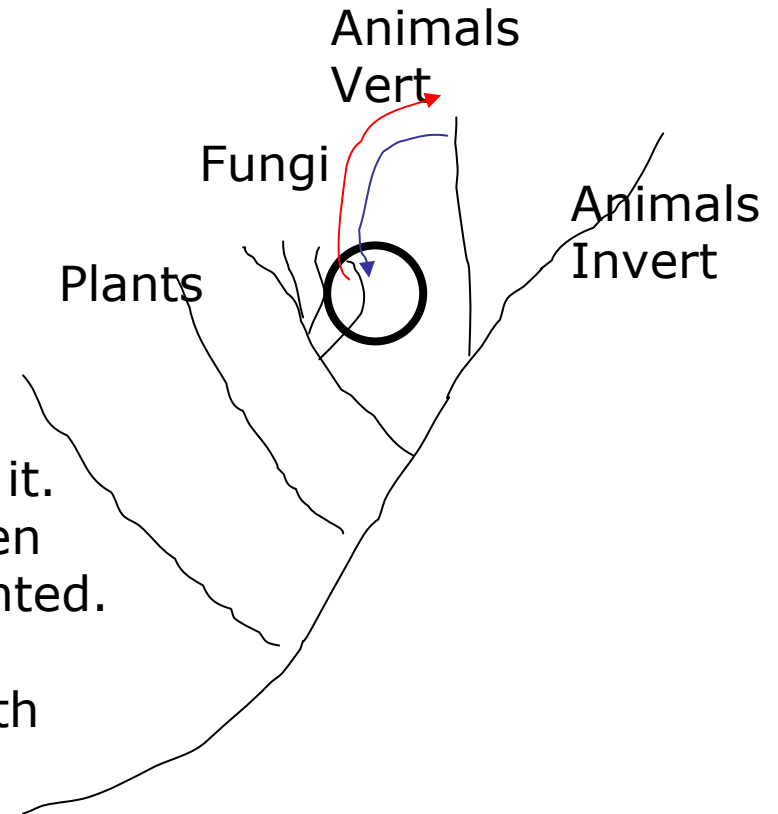
H.G.T!

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.



The DIO Family of Proteins

Case 3: Revisiting the function

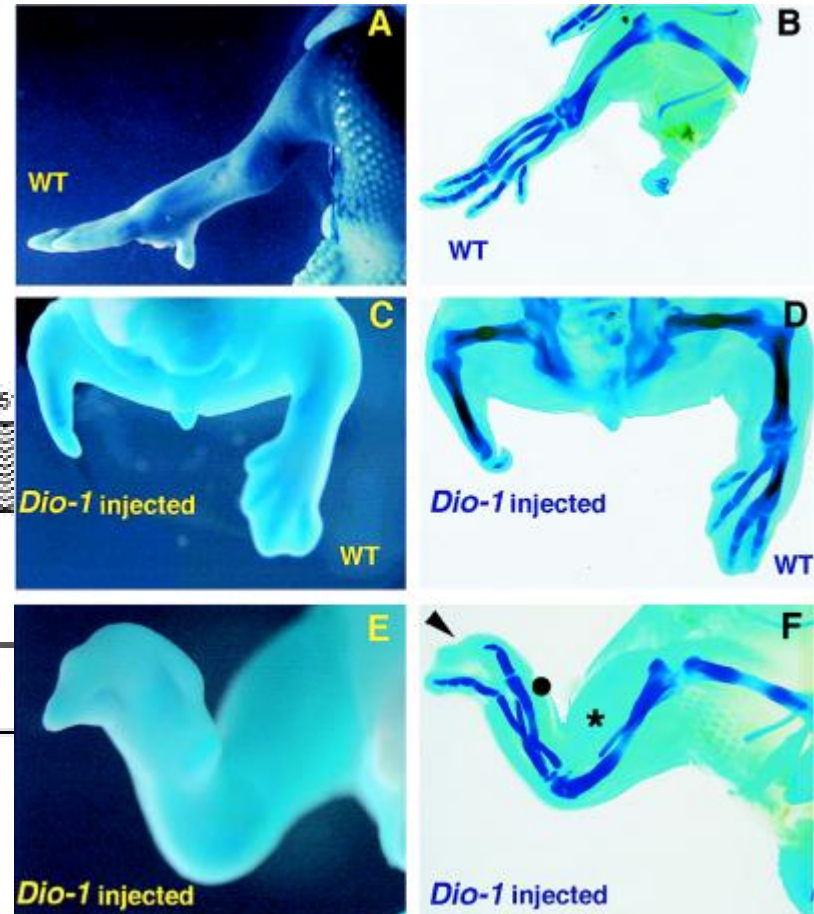
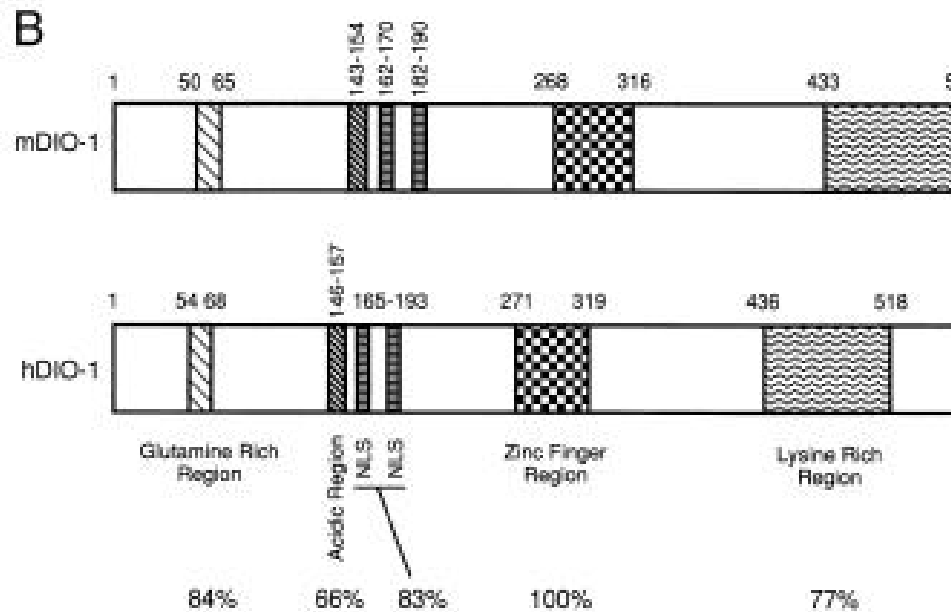
Get's real!

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



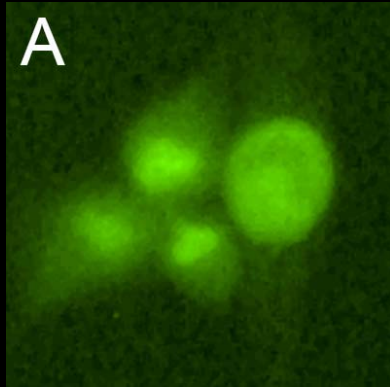
BACKGROUND DEATH INDUCER OBLITERATOR GENE (DIO)

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

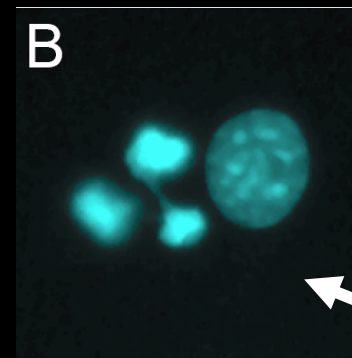
- DIO-1 nuclear translocation following apoptotic stimulation requires the NLS.
- DIO-1 forms oligomers.
- DIO-1 is present in multiple forms with distinct subcellular localizations.
- DIO-1 overexpression upregulates procaspase levels, leading to increased caspase activity.
- DIO-1 Δ NLS is a dominant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

NEW DATA

DEATH INDUCER OBLITERATOR GENE (DIO)

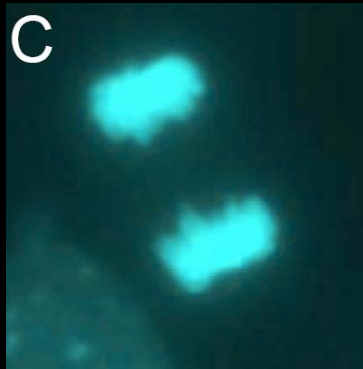


DIO-1 is present in mitotic chromosomes

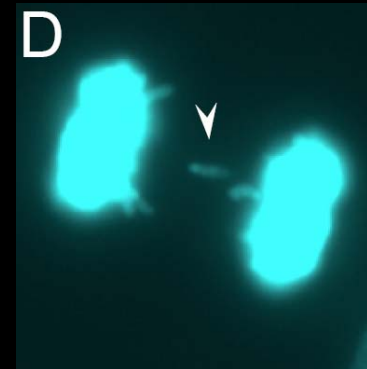


Mitosis on DIO overexpressed-cells

Asymmetric divisions!



Normal anaphase



DIO-targeted cells show abnormal anaphases: lagging chromosomes

TARGETED MICE SHOW SEVERE SUB-FERTILITY!!

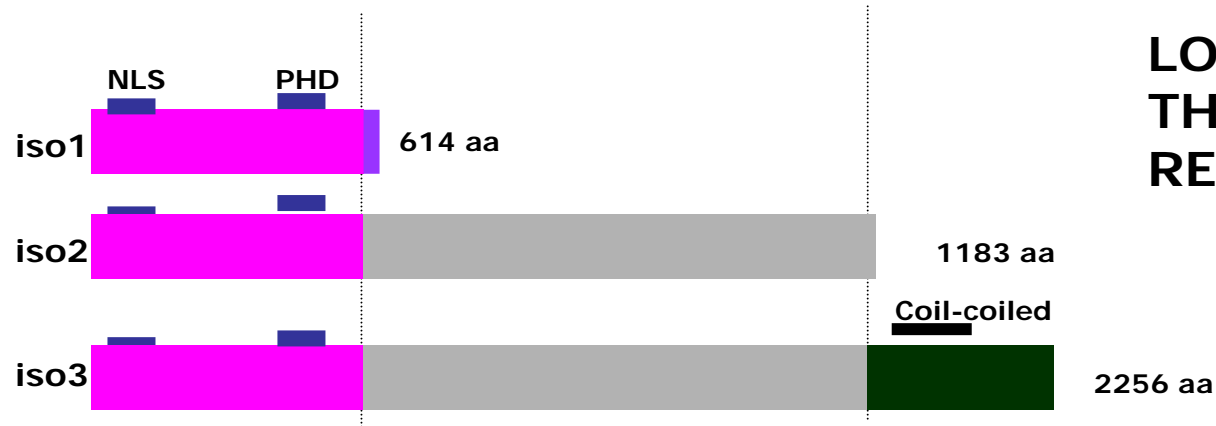
Case 3: Revisiting the function

Get's real!

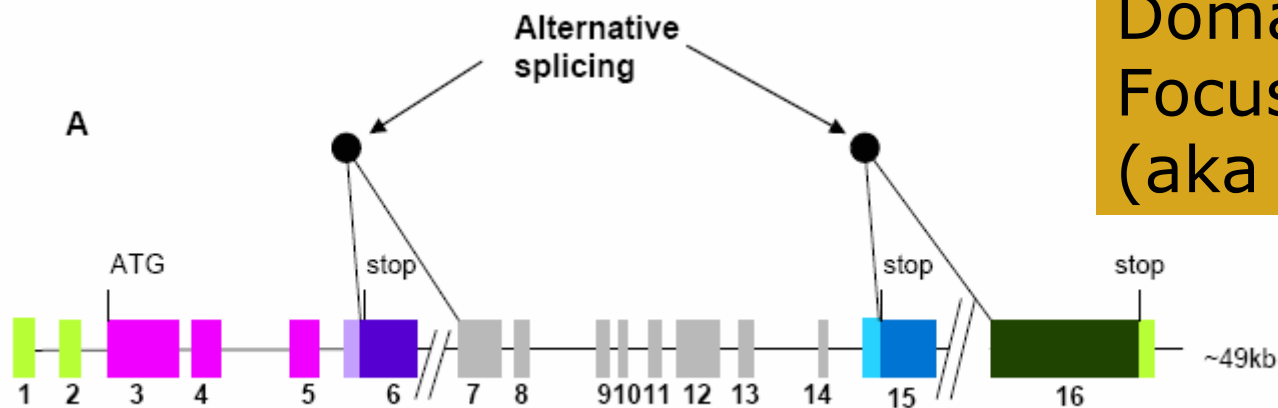
NEW DATA

DEATH INDUCER OBLITERATOR GENE (DIO)

DIO gene contains 3 splicing variants



LONG PARTS OF THE PROTEIN REMAIN UNCOVERED!

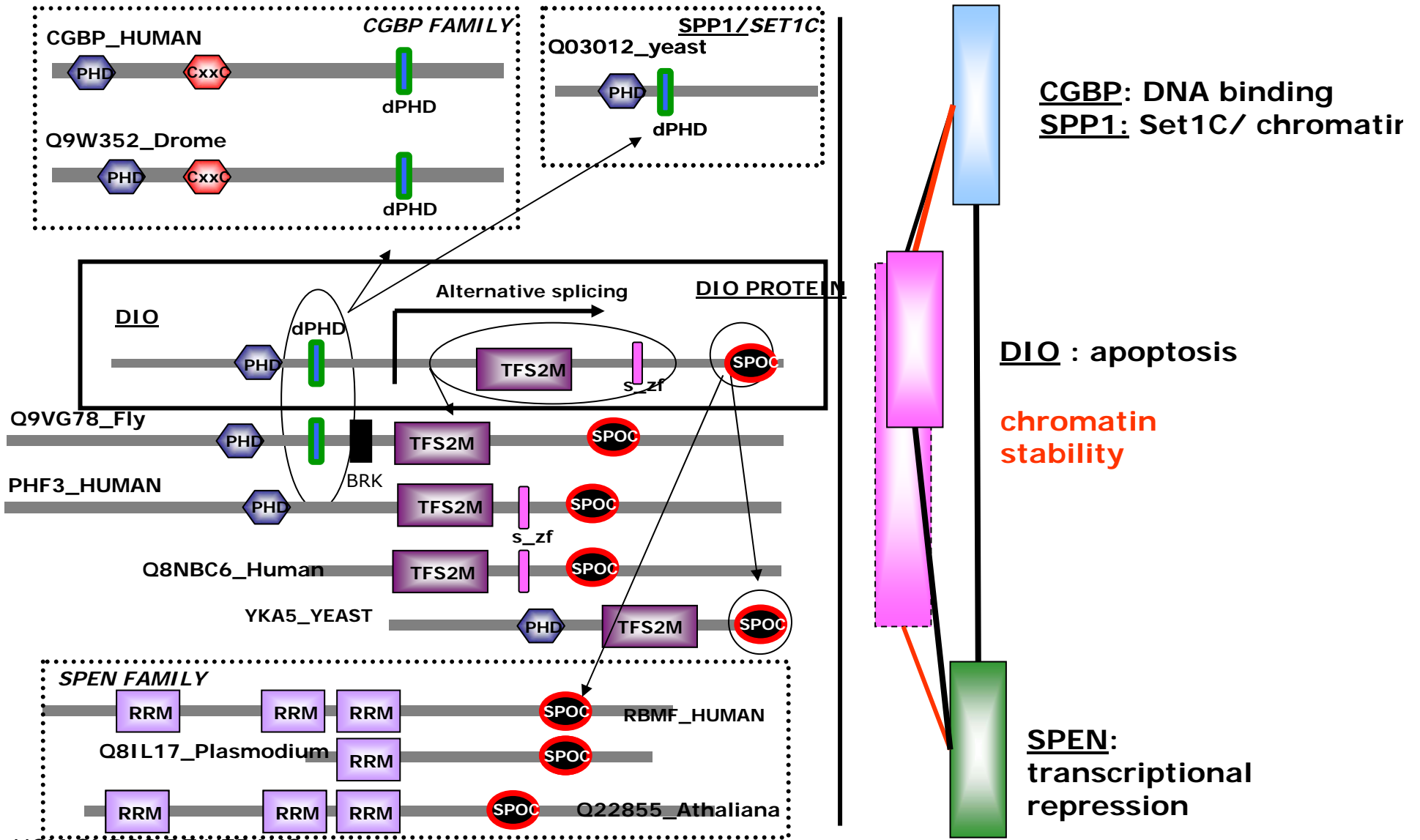


Domain Focus analysis (aka Luis ;-)

Case 3: Revisiting the function

Get's real!

OVERVIEW

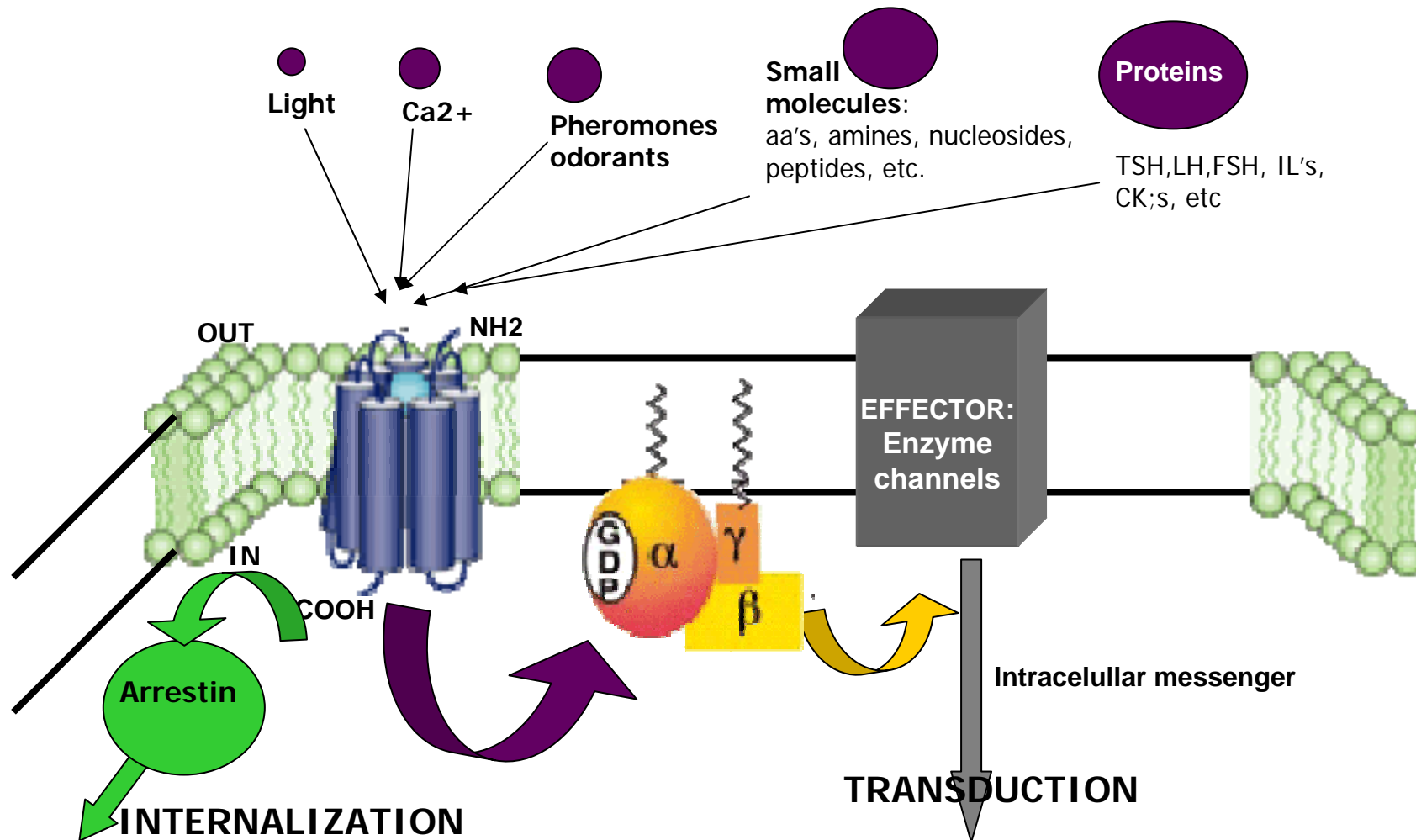


Identifying Dimerization Residues in CCR chemokine receptors

Case 4: Function Specificity

Get's real!

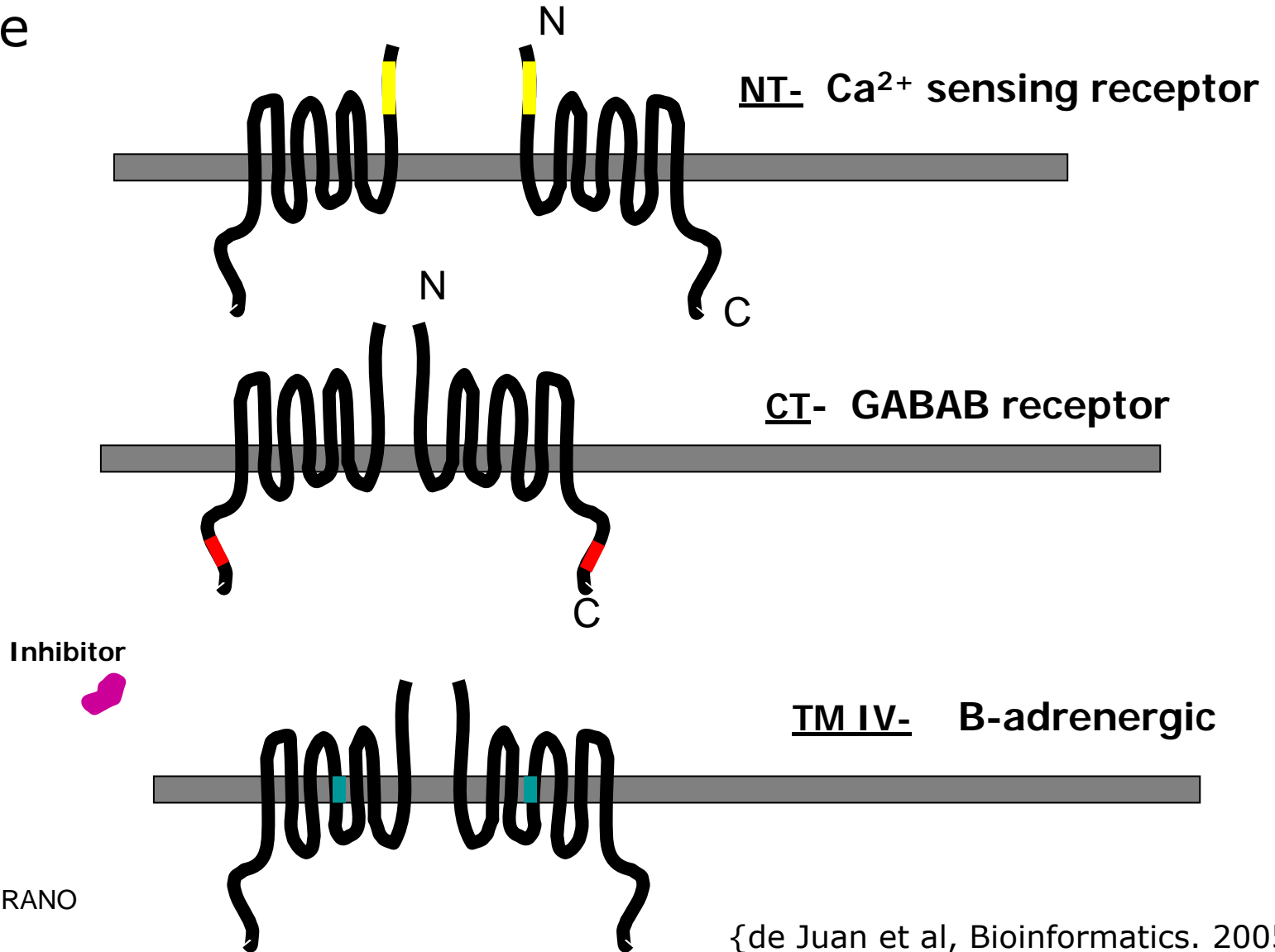
G - Coupled Receptor Proteins bind different ligands.



Case 4: Function Specificity

Get's real!

The GCPR's dimerize



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?

- 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-coupled Receptors of Residues That Determine Global Properties

Published, JBC

Srinivasan Madabushi^{1,2,3}, Alecia K. Gross^{1,4}, Theodore G. Wensel^{1,5}, and Olivier Lichtarge^{1,6}

¹From the Program in Structural and Computational Biology, and ²Program in Cellular and Molecular Biology, ³Molecular and Human Genetics, Baylor College of Medicine, and ⁴Department of Molecular Pharmacology, University of California, San Diego, La Jolla, California 92037, ⁵Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, and ⁶Department of Mathematics, University of California, San Diego, La Jolla, California 92037

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^{*-1}

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.

responds to most water-soluble hormones and neurotransmitters. In 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,

Our strategy

TEST CASE: 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.

Finding residues

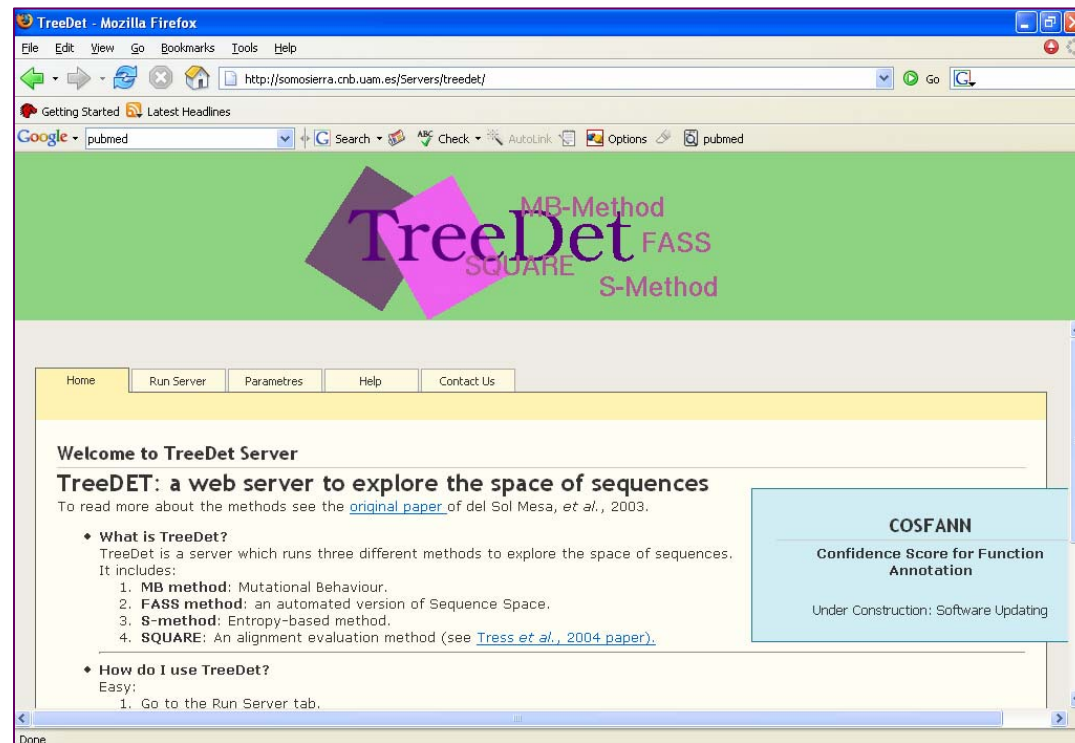
Basics: Homodimerization specificity is trying to avoid promiscuous dimerization between homologous sequences!

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



Case 4: Function Specificity

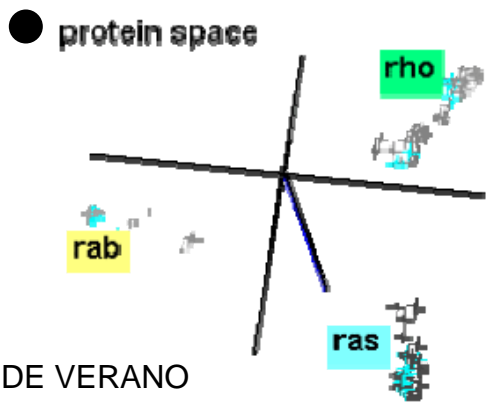
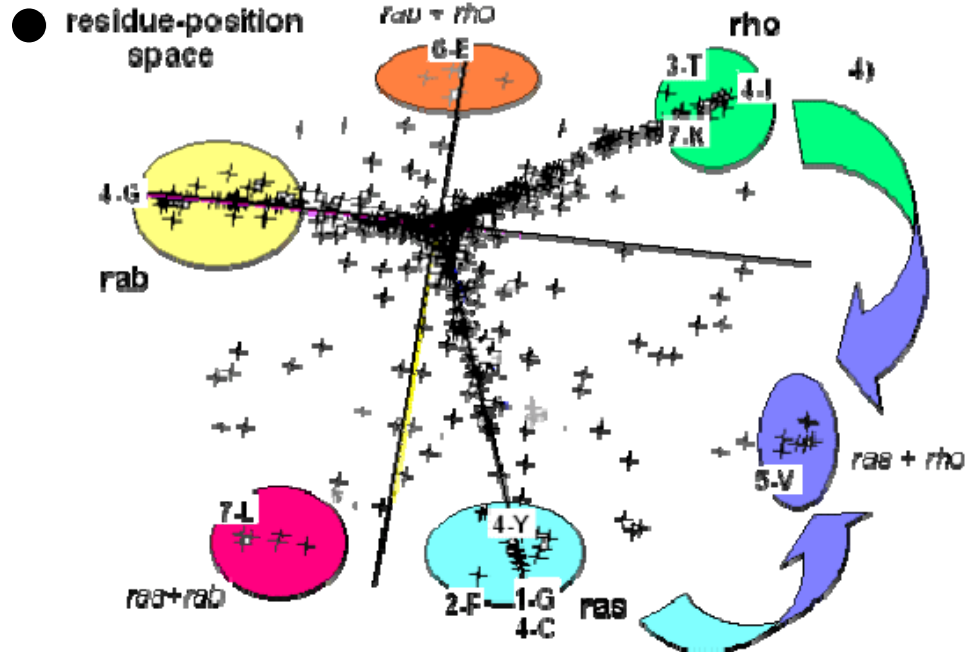
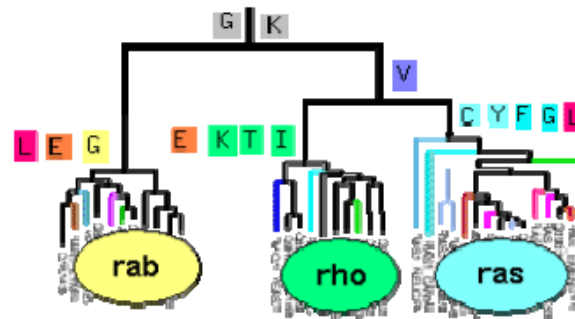
Get's real!

Sequence Space: overview

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

An example:

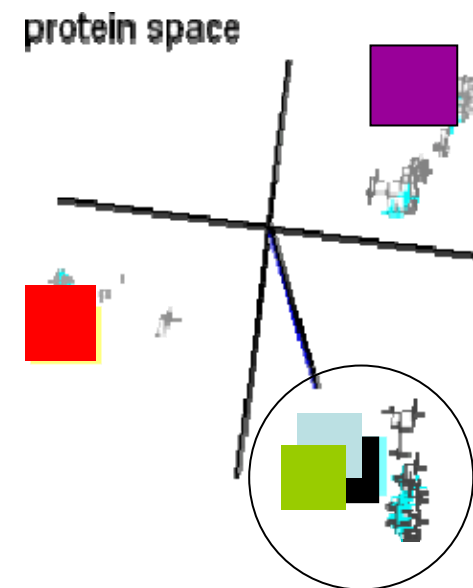
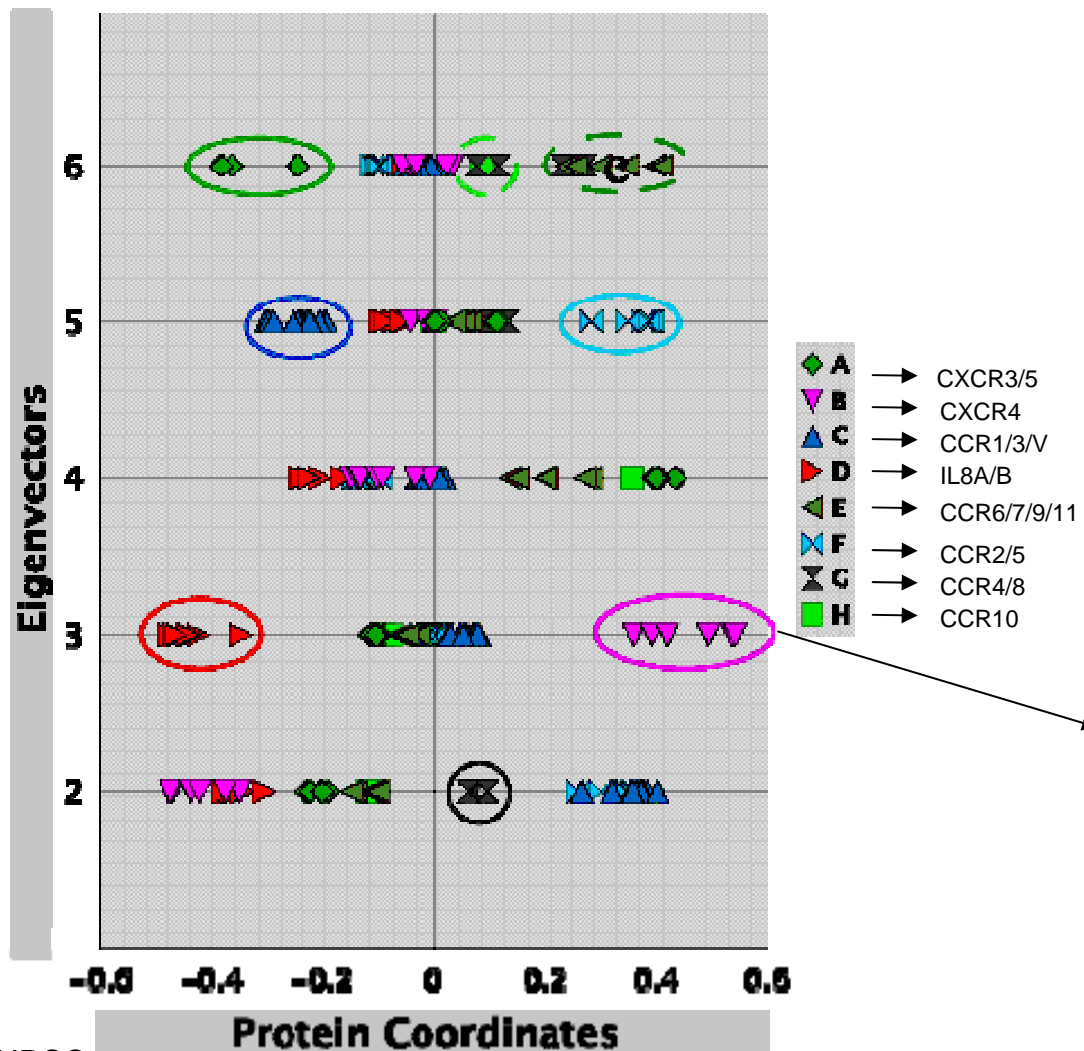
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------|---|---|---|---|---|---|---|---|---|
| RASH_HUMAN | G | F | Y | C | V | F | G | G | K |
| RAS_RRASV | G | F | Y | C | V | F | G | G | K |
| RASH_MSVHA | G | F | Y | C | V | F | G | G | K |
| RASN_XENLA | G | F | Y | C | V | F | G | G | K |
| RASN_BRARE | G | F | Y | C | V | F | G | G | K |
| RASK_HUMAN | G | F | Y | C | V | F | G | G | K |
| RASL_HUMAN | G | F | Y | C | V | F | G | G | K |
| SEC4_YEAST | S | E | D | G | K | F | F | G | K |
| SEC4_CANAL | S | E | D | G | K | F | F | G | K |
| Q9HET4 | S | E | D | G | K | F | F | G | K |
| Q96VL3 | S | E | D | G | K | F | F | G | K |
| YPT2_SCHPO | T | R | R | G | K | F | F | G | K |
| SAS1_DICDI | T | R | R | G | K | F | F | G | K |
| SAS2_DICDI | T | R | R | G | K | F | F | G | K |
| RHOC_HUMAN | M | E | T | I | V | E | K | G | K |
| RHO_DISOM | M | E | T | I | V | E | K | G | K |
| RHO_APLCA | M | E | T | I | V | E | K | G | K |
| RHO1_DROME | M | E | T | I | V | E | K | G | K |
| RHO2_HUMAN | M | E | T | I | V | E | K | G | K |
| Q8TG28 | V | R | R | T | V | E | K | G | K |
| RHO1_SCHPO | T | T | T | T | V | E | K | G | K |



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
Mario Mellado
Karel van Wely

PDG
SCOMP

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

Fede Abascal

You!