SEQUENCE ANALYSES.
PROTEIN FAMILIES
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.
Background

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)
Background

Example:
Hp21-elongation factor EF-Tu of Ecoli

General function:
signal transduction-protein synthesis

Functional feature:
GTP binding

Gene Duplication I

(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
Where?

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.
<table>
<thead>
<tr>
<th>Database</th>
<th>Version</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWISS-PROT</td>
<td>42.5</td>
<td>138.992</td>
</tr>
<tr>
<td>PRINTS</td>
<td>37.0</td>
<td>1.850</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>25.5</td>
<td>1.013.263</td>
</tr>
<tr>
<td><strong>Pfam</strong></td>
<td><strong>11.0</strong></td>
<td><strong>7.255</strong></td>
</tr>
<tr>
<td>PROSITE patterns</td>
<td>18.10</td>
<td>1.659</td>
</tr>
<tr>
<td>PROSITE preprofiles</td>
<td>N/A</td>
<td>131</td>
</tr>
<tr>
<td>ProDom</td>
<td>2002.1</td>
<td>1.021</td>
</tr>
<tr>
<td><strong>InterPro</strong></td>
<td><strong>7.1</strong></td>
<td><strong>10.403</strong></td>
</tr>
<tr>
<td>Smart</td>
<td>3.4</td>
<td>654</td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>3.0</td>
<td>1.977</td>
</tr>
<tr>
<td>PIR Superfamily</td>
<td>2.3</td>
<td>219</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>1.63</td>
<td>552</td>
</tr>
</tbody>
</table>
Automatic methods to classify proteins: ProtoMap

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

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

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

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

Based on **BeTs** (best bidirectional hits -> best similarity in both directions)

Paralogues fusion!

Problem? **Domains!**

Complete genomes

UCM-CURSO DE VERANO
JULY 2007
Problems:
Only pairs of seqs.
Close species
## 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.
• 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
WHAT CAN I LEARN FROM MY ALIGNMENT?

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

UCM-CURSO DE VERANO
JULY 2007

Romero, Valencia 04
WHAT CAN I LEARN FROM MY ALIGNMENT?

**How to's?**

**Mutational behaviour**

Pazos Valencia, 2001

\[
H(n) = \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log \frac{p(x_1, x_2, \ldots, x_n)}{\prod p(x_i)}
\]

**Relative entropy cut,**

del Sol, Valencia 2002

Del Sol, Pazos, Valencia JMB 03
Case 1: domain shuffling

Get’s real!

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!

{Reed et al, Gen Res. 2003}
Case 1: domain shuffling

**DOMAİN ARCHITECTURES**

- **NALP2**
  - PAAD
  - NACHT
  - LRR’S

- **MATER**
  - ?
  - NACHT
  - LRR’S

- **CARD4**
  - CARD
  - NACHT
  - LRR’S

- **NOD2**
  - CARD
  - CARD
  - NACHT
  - LRR’S

- **NAIP**
  - BI
  - R
  - BI
  - R
  - NACHT
  - LRR’S

- **COS1.5**
  - ?
  - NACHT
  - LRR’S

- **CLAN**
  - CARD
  - NACHT
  - LRR’S

**NACHT FAMILY**

- **ASC2**
  - PAAD

- **ASC**
  - PAAD
  - CARD

- **CASPASE ZF**
  - PAAD
  - CASPASE

- **PYRIN**
  - PAAD
  - B-BOX Zn FINGER
  - SPRY

- **IF16**
  - PAAD
  - IF120X
  - IF120X

- **MNDA,AIM2**
  - IF120X

**PAAD FAMILY**

UCM-CURSO DE VERANO
JULY 2007

{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  

**PAAD family:** MEFV/PYRIN (Pawlowski, et.al., 2001, others)
Case 1: domain shuffling

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

Get’s real!
Case 1: domain shuffling

Get’s real!

SEQUENCES FOR MODELING TARGETING

NOD2

NAIP

NAC

COS1.5
Case 1: domain shuffling

Psi-Blast → FFAS → Saturated blast → HITS → PAAD OF MEFV

*Removal of redundancy (splicing variants) 40 sequences

MALN=T-coffee → Trees (Bayes, NJME)

 Phylogeny → 2nd struct. Pred (metaserver)

Pairwise-FFAS → Structural neighbours (SCOP)

Modeling → JACKAL = MODELS

Minimized=CHARM → Conserved patches

In the surface: CONSURF ← PSQS Evaluation

{Reed et al, Gen Res. 2003}
ANCESTRAL DOMAIN

Get’s real!

CARD

DD

PAAD

DED

UCM-CURSO DE VERANO
JULY 2007

{Reed et al, Gen Res. 2003}
<table>
<thead>
<tr>
<th>Pos</th>
<th>Seq_str</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
<tr>
<td>2</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
<tr>
<td>3</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
<tr>
<td>4</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
<tr>
<td>5</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
<tr>
<td>6</td>
<td>HHHH...</td>
<td>HHHH...</td>
</tr>
</tbody>
</table>

**References:**

PHYLOGENETIC TREE FOR 40 PAADS
BAYES. 1,000,000, 9000 trees.

Get’s real!

UCM-CURSO DE VERANO
JULY 2007

{Reed et al, Gen Res. 2003}
Case 1: domain shuffling

NACHT DISTRIBUTION: POSSIBLE SCENARIO

PLANTS

Fungi

INVERT

Bacteria (anabaena)
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 bootstrap)

CONSENSUS!

---

UCM-CURSO DE VERANO
JULY 2007

{Rojas et al, IJESM. 2002}
{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!
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?

UCM-CURSO DE VERANO
JULY 2007

{Rojas & Doolittle, 2002, JME}
Case 2: Extending the family

General Medusa Body Plan

- mouth
- tentacles
- oral arm
- bell
- stomach
- gonad

Phylogenetic Tree of Life

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

{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}
WHY THE FUNGAL ONES CLADE WITH ANIMALS?

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}
Case 3: Revisiting the function \textit{Get’s real!}

The DIO Family of Proteins

\{Rojas et al, FEBS J. 2004\}
Case 3: Revisiting the function

BACKGROUND

DEATH INDUCER OBLITERATOR GENE (DIO)

DISRUPTS LIMB DEVELOPMENT (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

DEATH INDUCER OBLITERATOR GENE (DIO)

BACKGROUND

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.
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!
Case 3: Revisiting the function

Get’s real!

NEW DATA

DEATH INDUCER OBLITERATOR GENE (DIO)

DIO gene contains 3 splicing variants

iso1

iso2

iso3

LONG PARTS OF THE PROTEIN REMAIN UNCOVERED!

Domain Focus analysis (aka Luis ;-))

Rojas et al, FEBS J. 2004
Case 3: Revisiting the function

OVERVIEW

Get’s real!

CGBP: DNA binding

SPP1: Set1C/ chromatin

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!

<table>
<thead>
<tr>
<th>Protein</th>
<th>Accession</th>
<th>Start</th>
<th>End</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASH_HUMAN</td>
<td>5</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RRAS_HUMAN</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RTC1_HUMAN</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS2_HYDRA</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS2_DROME</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RASL_NEURC</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RASL_MOUSE</td>
<td>1</td>
<td>159</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS1_YEAST</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS_SCHPO</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS_LENED</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAS2_RHRA</td>
<td>1</td>
<td>160</td>
<td>166</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALA_HUMAN</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALA_RAT</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALB_HUMAN</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALB_RAT</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALB_XENLA</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RAL_DISOM</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RALA_DROME</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>CE1393944</td>
<td>1</td>
<td>157</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>CC42_DROME</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH04_YEAST</td>
<td>1</td>
<td>155</td>
<td>162</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH02_SCHPO</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH02_YEAST</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH08_HUMAN</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH06_HUMAN</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH03_YEAST</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH01_SCHPO</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH01_YEAST</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RH01_ENTHI</td>
<td>1</td>
<td>151</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
<tr>
<td>RHOL_DROME</td>
<td>1</td>
<td>156</td>
<td>164</td>
<td>KLVWGGGVGKSLTIIQIQNHFVDEDFTEESYRRKQVVIDGETCLLDLDIOTAGGEEFYSAN</td>
</tr>
</tbody>
</table>
**Case 4: Function Specificity**

**G - Coupled Receptor Proteins bind different ligands.**

Small molecules: aa’s, amines, nucleosides, peptides, etc.

- Light
- Ca²⁺
- Pheromones odorants

Proteins: TSH, LH, FSH, IL’s, CK; s, etc.

**Intracellular messenger**

**EFFECTOR:** Enzyme channels

**Arrestin**

**INTERNALIZATION**

**TRANSDUCTION**

(de Juan et al, Bioinformatics. 2005)
Case 4: Function Specificity

The GCPR’s dimerize

NT- Ca$^{2+}$ sensing receptor

CT- GABAB receptor

Inhibitor

TM IV- B-adrenergic

{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:

**Evolutionary Trace of G Protein-coupled Receptors: Identification of Residues That Determine Global Conformation**

G protein-coupled receptor (GPCR) activation mediated by ligand-induced conformational changes is a key step in the transduction of signals from the extracellular to the intracellular domain. To identify residues that are important in diverse GPCRs, we used evolutionary tracing (ET) to identify residues commonly important in different GPCRs when mapped onto the rhodopsin structure. These 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 270 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 extracellular interface to the G protein, and a region involved in activation. This approach finds differences in amino acid composition and accessibility of some of these residues. In fact, GPCRs are so ubiquitous that, although they are the targets of nearly 50% of current drugs, this is still a small fraction of their pharmacological potential.

**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 NW7 1AA, 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 a likely important physiological role. 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 interfaces in an explicitly-evolutionary context. Applying this model to aminergic GPCR sequences, we first validated the general reasoning behind the model. Then we 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.

**References**

Case 4: Function Specificity

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.
Case 4: Function Specificity

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

Sequence Space: overview

An example:

Case 4: Function Specificity

Sequence Space: Clustering results

UCM-CURSO DE VERANO
JULY 2007
Case 4: Function Specificity

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 predictive power of this approach.*
Acknowledgements

Luis Sanchez-Pulido
Mario Mellado
Karel van Wely

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

PDG
SCOMP

Fede Abascal
You!