Protein function prediction

(focusing on methods not based on the detection of sequence similarity)

Manuel J. Gómez
Bioinformatics Lab
Centro de Astrobiología, CSIC-INTA

ONTOLOGIES IN MOLECULAR BIOLOGY

Controlled and structured vocabularies, constructed with TWO purposes: proposing standard collections of terms, and, organizing the knowledge of a given field around its language: the relations between the terms are supposed to reflect the biological reality.

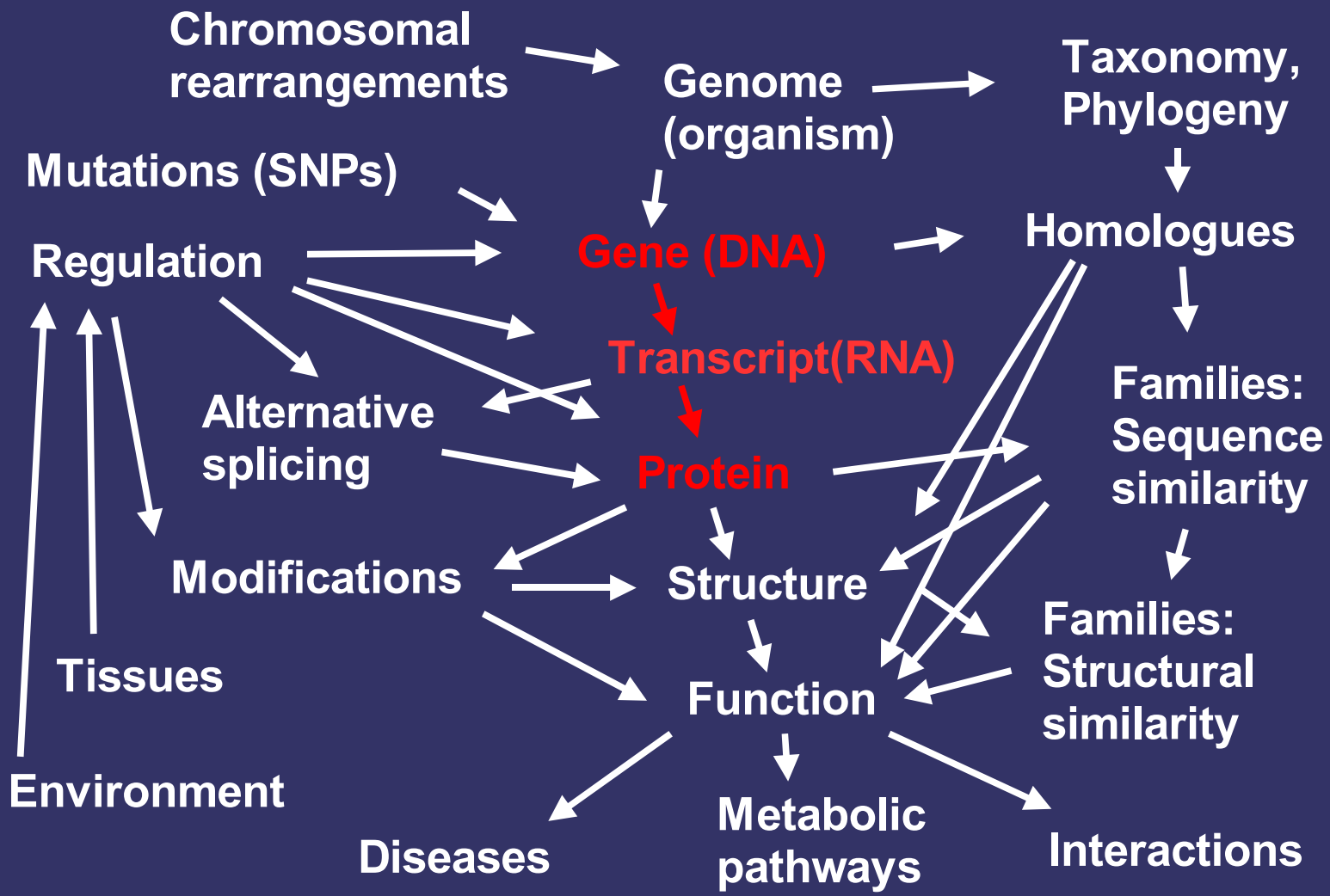
- Enzyme Commission Nomenclature.

  EC 1. -. -.- Oxidoreductases.
  EC 1. 1. -.- Acting on the CH-OH group of donors.
  EC 1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
  EC 1. 1. 2.- With a cytochrome as acceptor.

- MeSH (Medical Subject Headings) terms: NLM controlled vocabulary.

- Gene Ontology: developed and maintained by a consortium (GO Consortium) of laboratories and institutions involved in molecular biology database management.

ONTOLOGIES IN MOLECULAR BIOLOGY



GENE ONTOLOGY

Biological terms have been groupd in three ontologies:

- Molecular functions
- Cellular processes
- Cellular components

Within each ontology, the terms are related hierarchically.

The relation between parent and child terms can be of two types:

- Part of
- Instance of

Most molecular biology databases have joined this initiative, and have included annotations following this standard.

Gene Ontology BROWSER

GO:0003673 : Gene_Ontology (46199)
GO:0008150 : biological_process (30188)
GO:0007610 : behavior (291)
GO:0000004 : biological_process unknown (3665)
GO:0007154 : cell communication (6212)
GO:0008151 : cell growth and/or maintenance (20547)
GO:0016265 : death (525)
GO:0007275 : development (3620)
GO:0008371 : obsolete (1640)
GO:0007582 : physiological processes (854)
GO:0016032 : viral life cycle (27)
GO:0005575 : cellular_component (22371)
GO:0003674 : molecular_function (37018)

Gene Ontology

Mycoplasma pneumoniae proteome.
Distribution of GO anotations.

GO Classification for *M. pneumoniae*			
Term		Proteins	
GO:0003674	**molecular_function**	**431**	**62.7%**
GO:0003676	nucleic acid binding	119	17.3%
GO:0030528	transcription regulator activity	5	0.7%
GO:0003754	chaperone activity	10	1.4%
GO:0003824	catalytic activity	245	35.6%
GO:0015070	toxin activity	1	0.1%
GO:0005194	cell adhesion molecule activity	1	0.1%
GO:0005198	structural molecule activity	56	8.1%
GO:0005215	transporter activity	64	9.3%
GO:0005488	binding	211	30.7%
GO:0005554	molecular_function unknown	65	9.4%
GO:0008150	**biological_process**	**343**	**49.9%**
GO:0008152	metabolism	280	40.7%
GO:0006810	transport	65	9.4%
GO:0006950	response to stress	16	2.3%
GO:0007049	cell cycle	22	3.2%
GO:0007154	cell communication	16	2.3%
GO:0007275	development	5	0.7%
GO:0007582	physiological process	335	48.7%
GO:0005575	**cellular_component**	**277**	**40.3%**
GO:0005576	extracellular	1	0.1%
GO:0005623	cell	274	39.8%
GO:0005941	unlocalized	5	0.7%

FUNCTION PREDICTION PROTOCOL: SEQ SIM

Based on the transfer of function from proteins that are similar at sequence level

Protein primary sequence

Primary Database similarity search
- SwissProt / UniProt
- nr / SP+SPTrEMBL
- COG / KOG
- PDB

Orthologs / paralogs
MSA
Family discovery or assignment
Functional residues
Phylogenetic profile
Gene neighbourhood
Function prediction?

Some of these approaches have been used for the development of automatic annotation servers

Secondary Database similarity search
- Prosite
- Pfam
- SMART
- PRINTS
- BLOCKS
- InterPro

Protein motifs
Domain organization
Family assignment
Function prediction?

AUTOMATIC ANNOTATION OF GENOMES

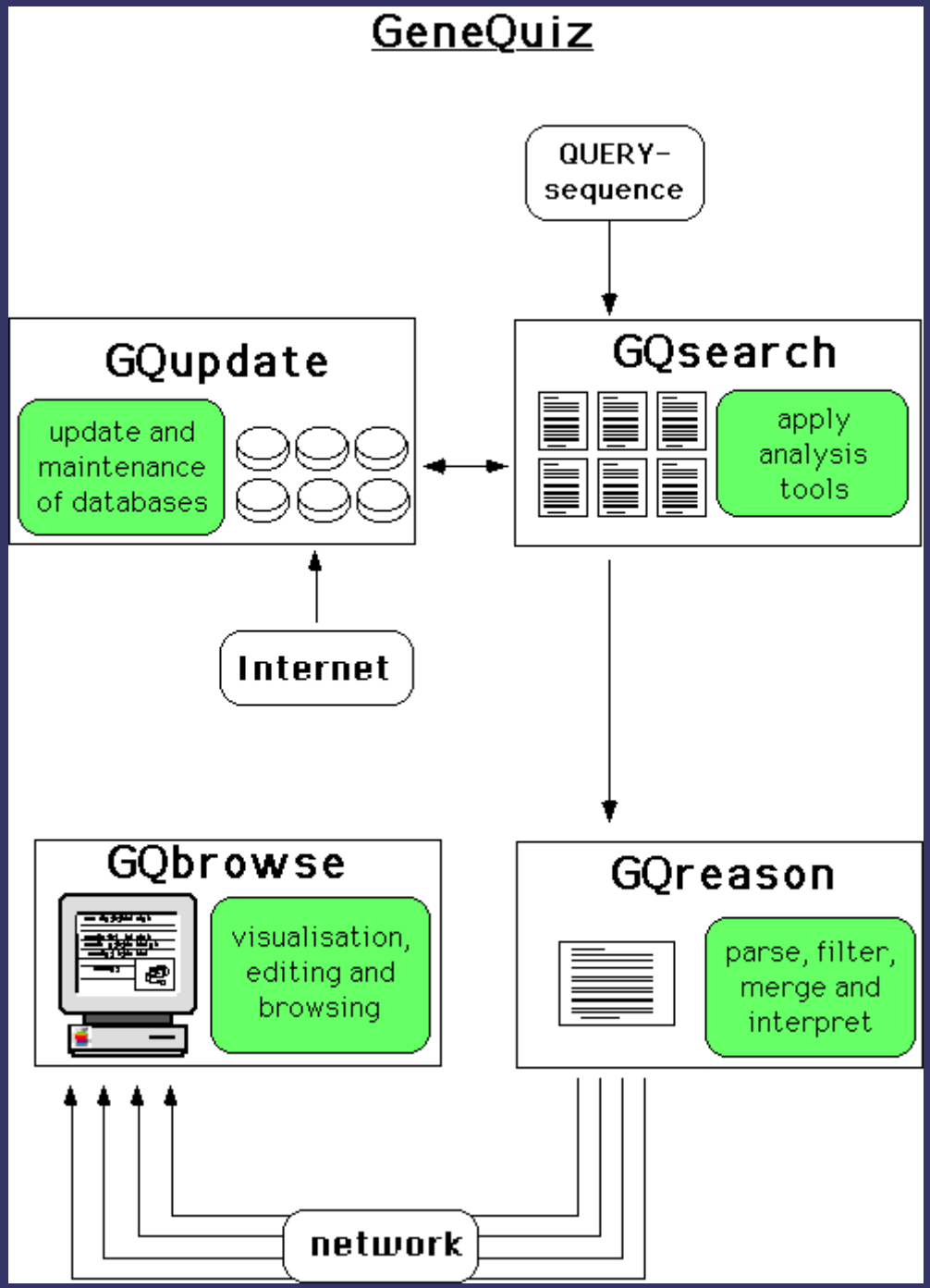
GeneQuiz (EBI)
(Andrade et al 1999)

PEDANT
(Andrade et al 1999)

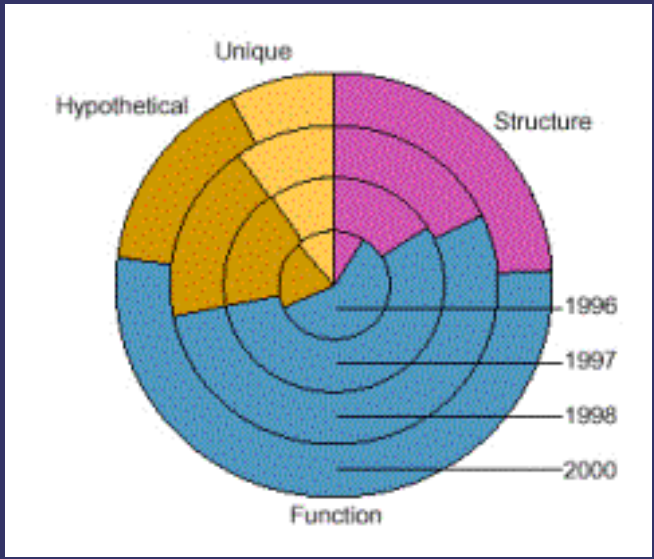
MAGPIE
(Gaasterland et al, 1996)

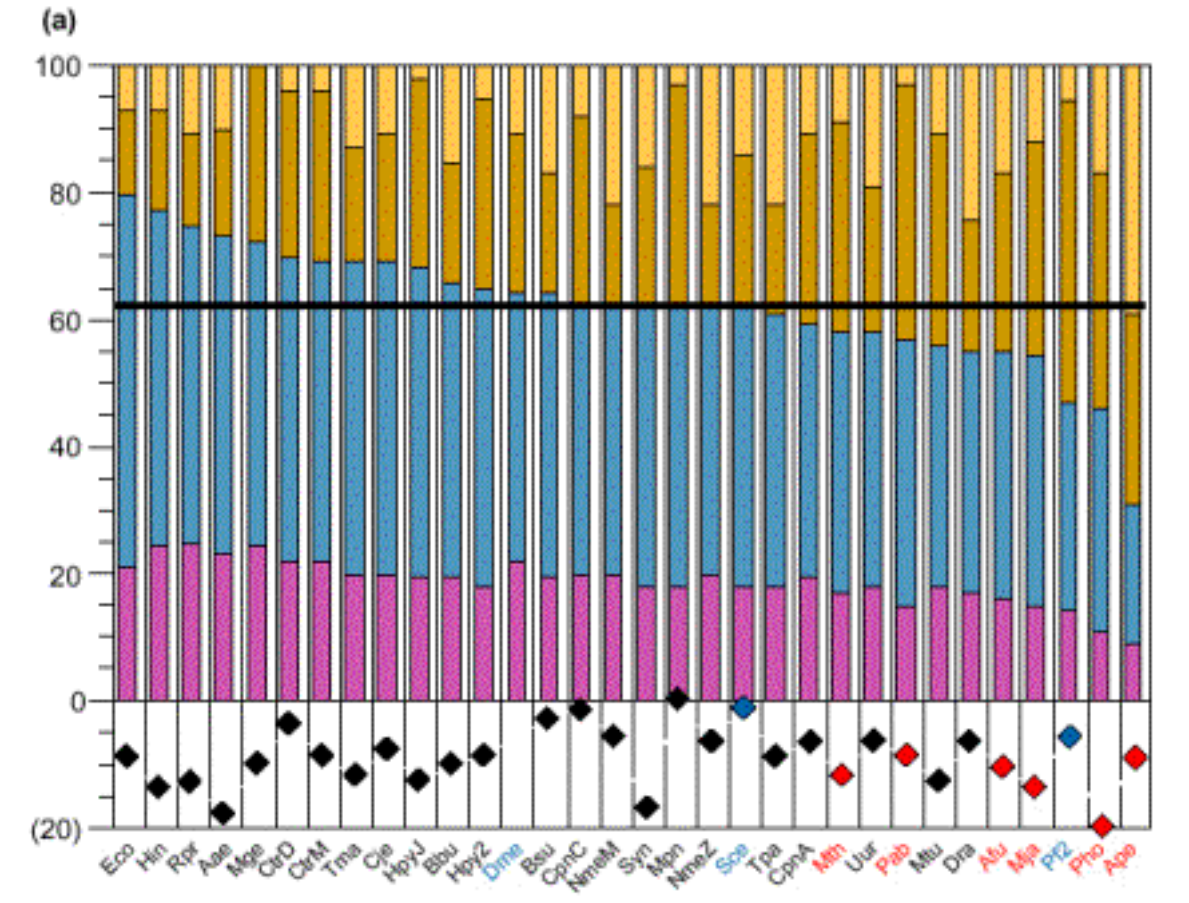
ENSEMBL
(Clamp et al, 2003)

RiceGAAS
(Sakata el al, 2002)



DISTRIBUTION OF DIFFERENT LEVELS OF ANNOTATION IN MICROBIAL GENOMES





FUNCTION PREDICTION PROTOCOL: SEQ SIM + STRUCT SIM

Based on sequence similarity and structural analyses, mainly to identify remote homologs

Protein primary sequence

Primary Database similarity search
- SwissProt / UniProt
- nr / SP+SPTrEMBL
- COG / KOG
- PDB

Orthologs / paralogs
MSA
Family discovery or assignment
Functional residues
Phylogenetic profile
Gene neighbourhood
Function prediction?

Secondary Database similarity search
- Prosite
- Pfam
- SMART
- PRINTS
- BLOCKS
- InterPro

Protein motifs
Domain organization
Family assignment
Function prediction?

Protein structure analyses
- SCOP / CATH classification
- Functional sites mapped on structure

Function prediction (molecular level)?

Protein structure prediction
- 1D features
- 3D structure / fold prediction

Known / Predicted structure

DETECTION OF SIMILARITY OF PROTEIN 1D FEATURES

- Several protein features that can be calculated or predicted from the primary sequence are used frequently to get hints about protein function.
- These may include:
  - Amino acid composition and isoelectric point.
  - Presence of signal peptides and trans-membrane segments.
  - Secondary structure elements.
  - Post translational modification sites.
- The challenge is to devise methods to find correlations between these properties and protein functions.

Around 5500 human protein sequences, classified according to:

- EC number
- Euclid
- Gene Ontology

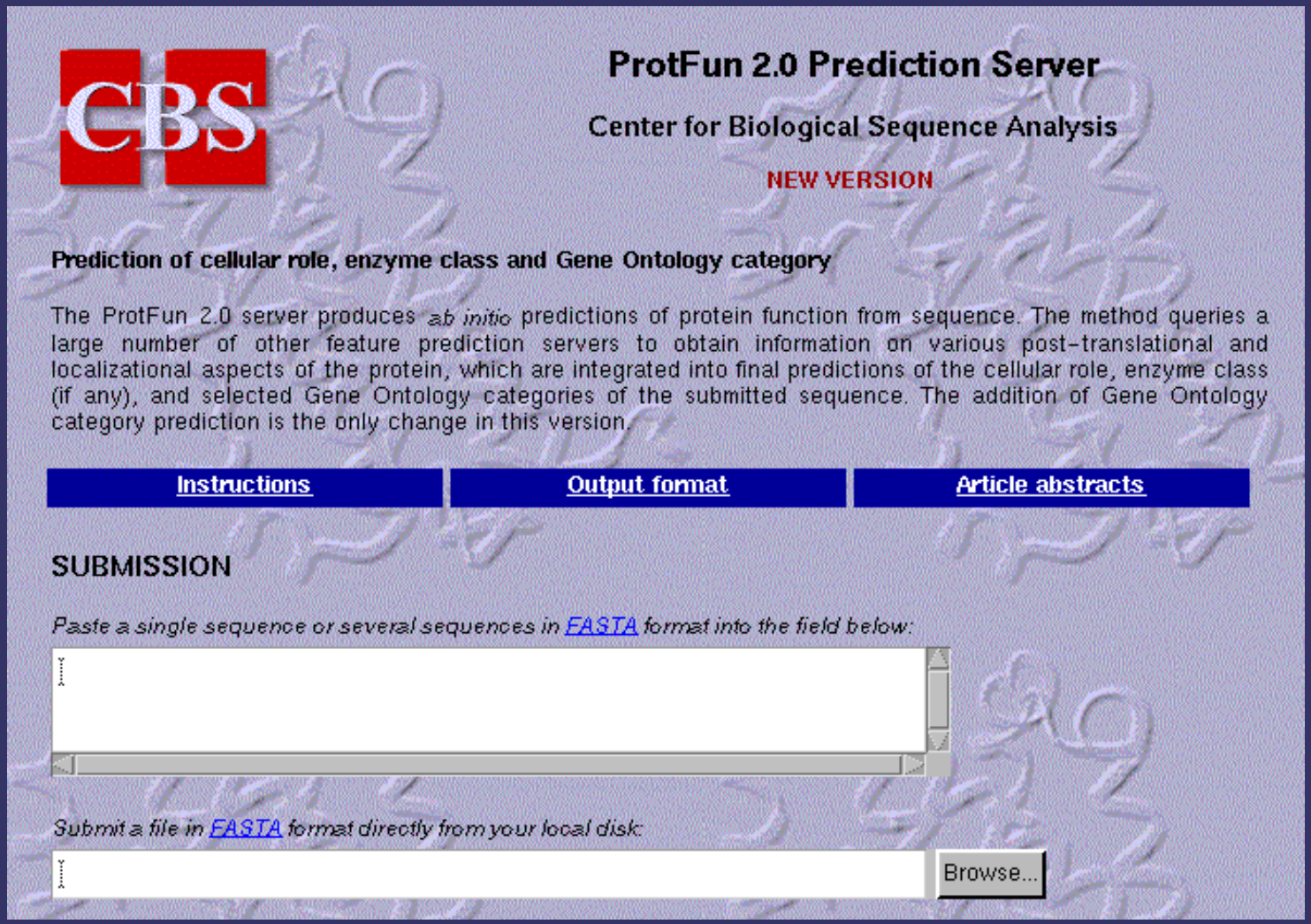
→ Prediction or calculation of features that can be derived directly from the primary sequence

→ Training of a Neural Network in the association between features and function

Jensen et al. JMB (2002)

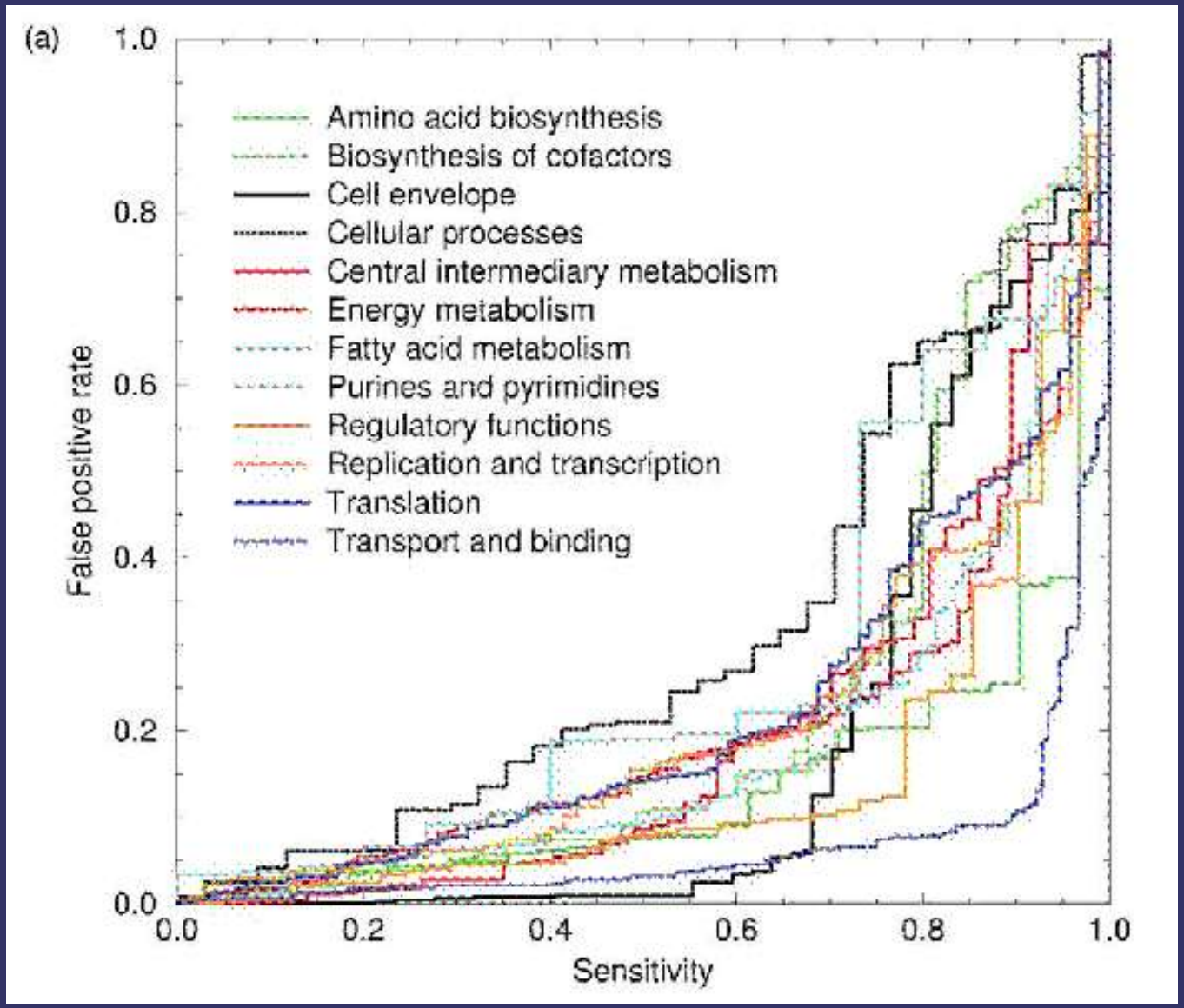
Abbriviation	Encoding	Description
ec	single value	Extinction coefficient predicted by ExPASy ProtParam
gravy	single value	Hydrophobicity predicted by ExPASy ProtParam
mneg	single value	Number of negatively charged residues counted by ExPASy ProtParam
npos	single value	Number of positively charged residues counted by ExPASy ProtParam
nglyc	potential in 5 bins	N-glycosylation sites predicted by NetNGlyc
oglyc	potential-threshold in 10 bins	GalNAc O-glycosylations predicted by NetOGlyc
pest	fraction in 10 bins	PEST rich regions identified by PESTfind
phosST	potential in 10 bins	Serine and threonine phosporylations predicted by NetPhos
phosY	potential in 10 bins	Tyrosine phosporylations predicted by NetPhos
psipred	helix, sheet, coil in 5 bins	Predicted secondary structure from PSI-Pred
psort	20 probabilities	Subcellular location predtions by PSORT
signalp	meanS, maxY, log(cleavage pos)	Signal peptide predictions made by SignalP
tmhmm	inside, outside, membrane in 5 bins	Transmembrane helix predictions made by TMHMM

DETECTION OF SIMILARITY OF PROTEIN 1D FEATURES



ProtFun Server

DETECTION OF SIMILARITY OF PROTEIN 1D FEATURES



Sensitivity vs specificity in the prediction of functional roles, according to Euclid

METHODS BASED ON ANALYSES OF PROTEIN STRUCTURE

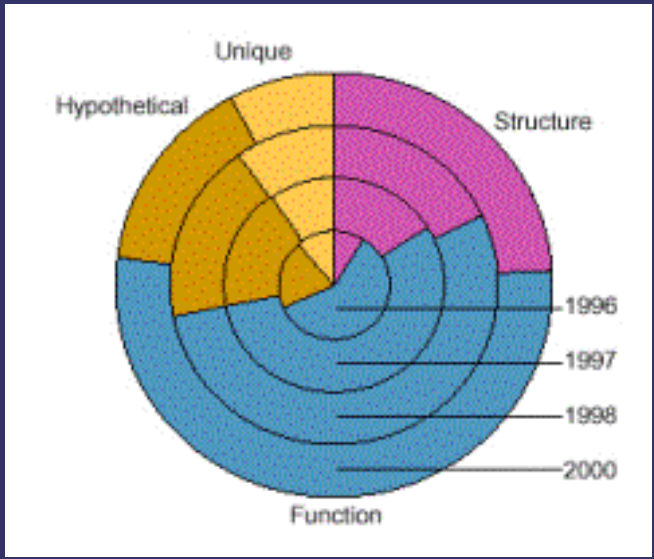
Several methods have been proposed or used:

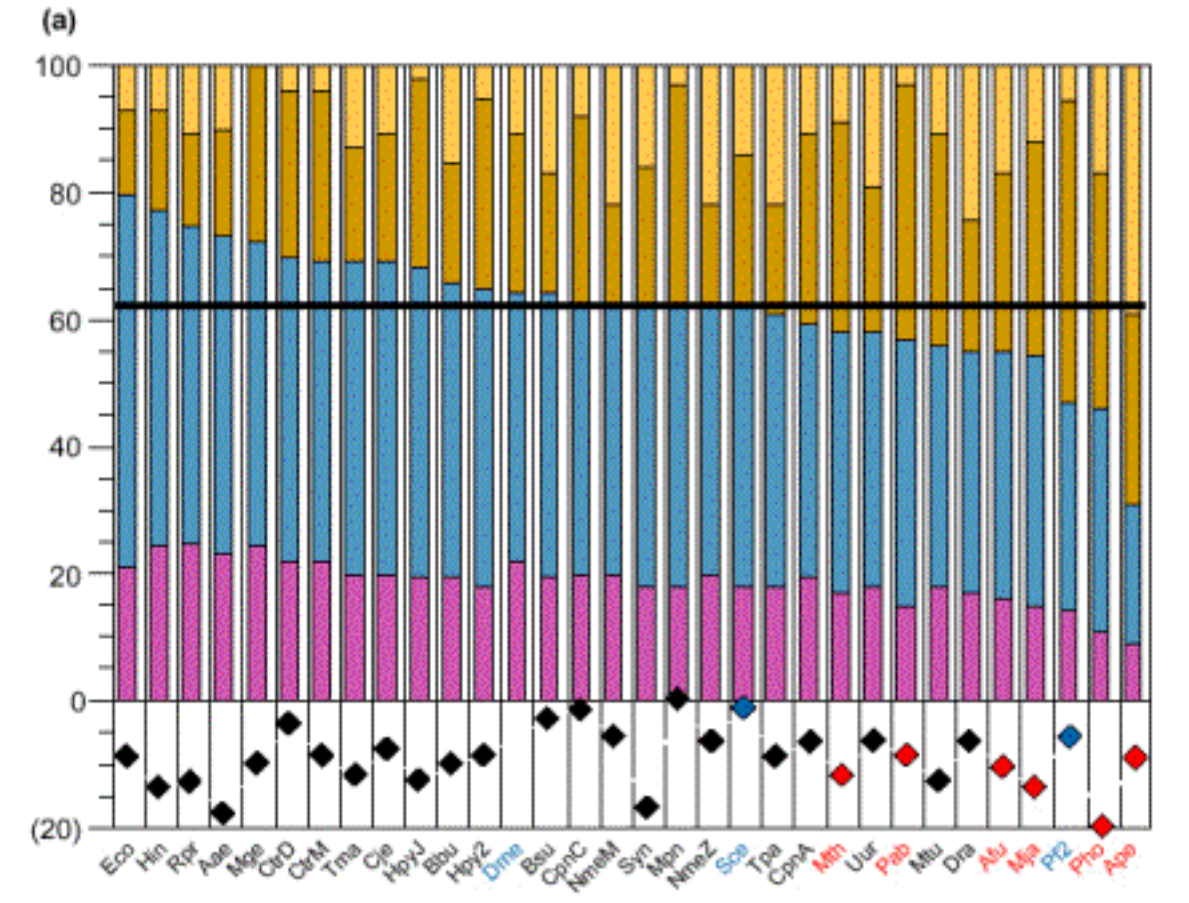
- Similarity, at a general structural level, to proteins of known function (detection of remote homologs by fold recognition).
- Similarity at local structural level (common structure of the active site, in enzymes that are not related)
- Prediction of binding sites (cavities) followed by prediction of ligands with docking algorithms.
- Prediction of interactions with other proteins by analysis of protein surfaces to identify potential interaction sites, followed prediction of interaction partners by docking methods.

The utility of the last three types of methods to predict function is, for the moment, anecdotical.

Eisenstein Curr Op Biotec (2000)

DISTRIBUTION OF DIFFERENT LEVELS OF ANNOTATION IN MICROBIAL GENOMES





FUNCTION PREDICTION PROTOCOL: SEQ & STRUCT SIM + INT

Based on sequence similarity, structural analyses and information about interacting partners.

Protein primary sequence

Primary Database similarity search
- SwissProt / UniProt
- nr / SP+SPTrEMBL
- COG / KOG
- PDB

Orthologs / paralogs
MSA
Family discovery or assignment
Functional residues
Phylogenetic profile
Gene neighbourhood
Function prediction?

Protein interactions characterization

Function prediction (cellular level)?

Secondary Database similarity search
- Prosite
- Pfam
- SMART
- PRINTS
- BLOCKS
- InterPro

Protein motifs
Domain organization
Family assignment
Function prediction?

Protein structure analyses
- SCOP / CATH classification
- Functional sites mapped on structure

Function prediction (molecular level)?

Protein structure prediction
- 1D features
- 3D structure / fold prediction

Known / Predicted structure

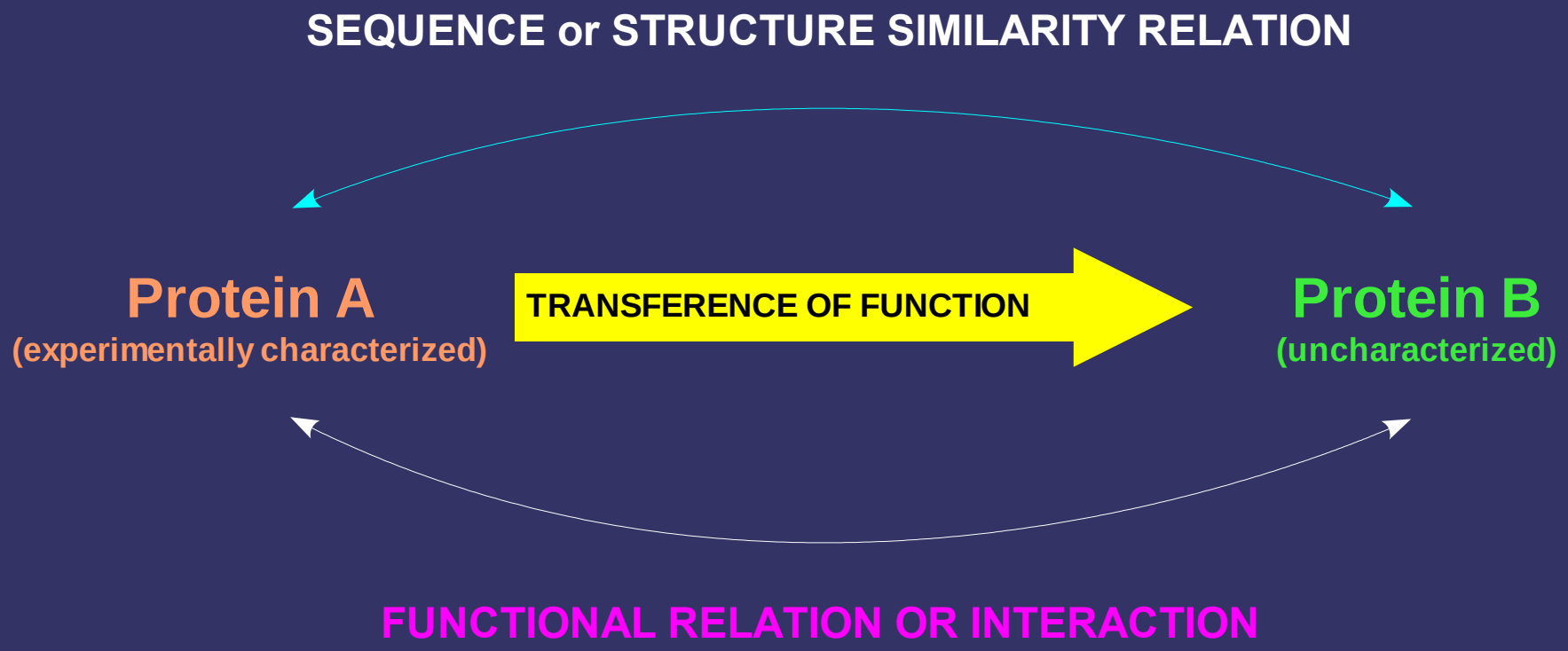
METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

- In general the prediction of function based on the detection of functional interactions follows the concept presented often as "guilty by association".

- In the absence of any detectable indication of homology:
  - If *protein A* is a regulator of, or it is regulated by, *protein B*;
  - or p*rotein A* interacts physically with *protein B*;
  - or *protein A* and *protein B* are co-expressed;

  then, we can assume that the function of both proteins is also related.

FUNCTION PREDICTION METHODS



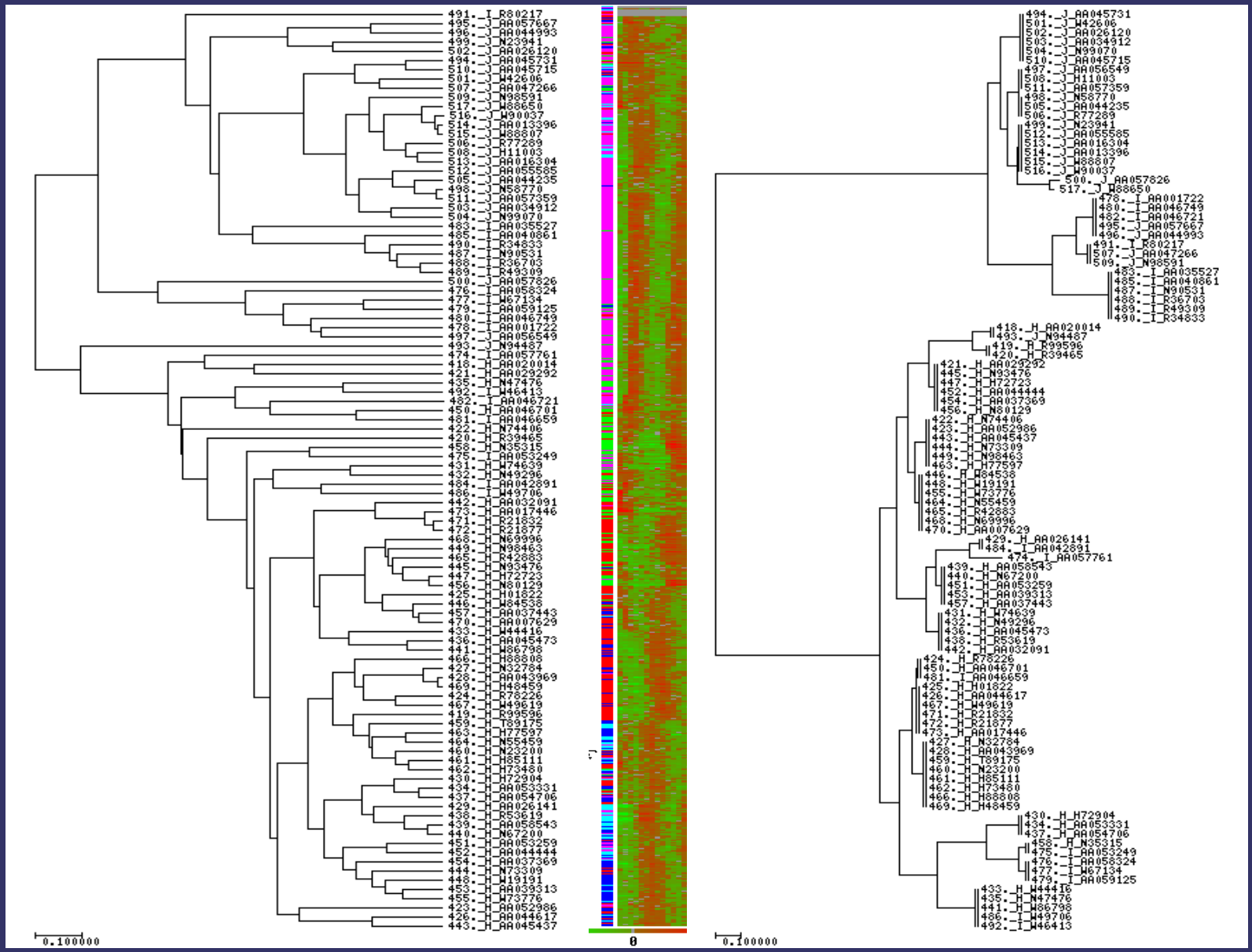
METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

We can consider four sources of information about functional interactions:

- Functional Genomics
- Proteomics
- Text mining (information extraction)
- Comparative sequence genomics

METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

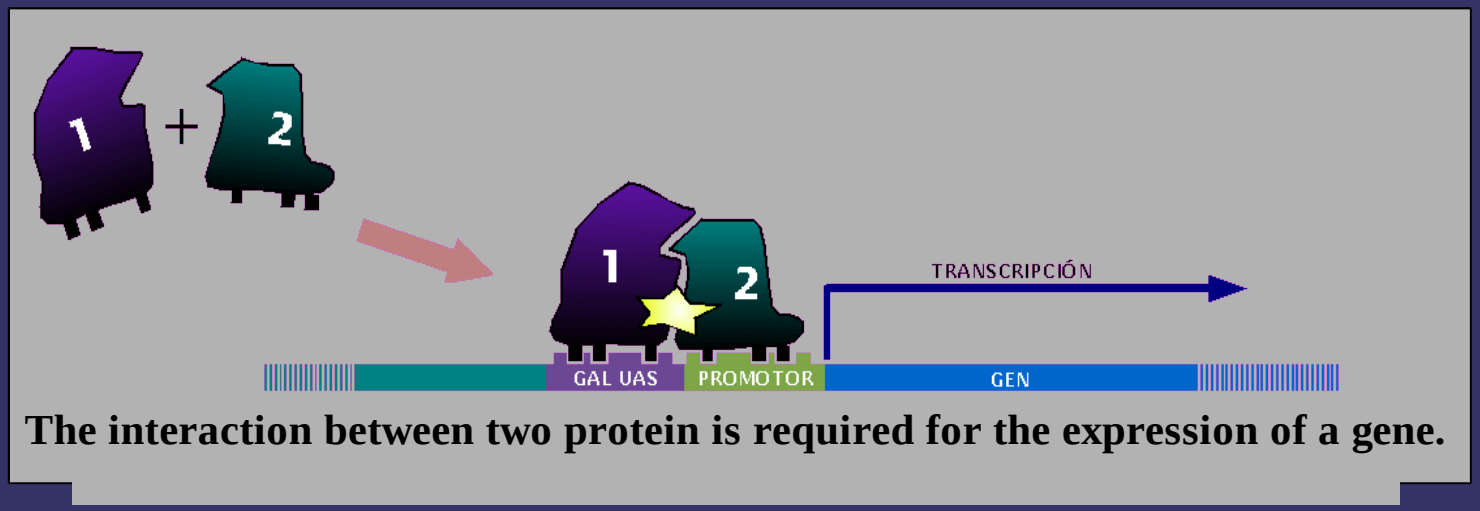
Functional Genomics



Clustering of genes according to their pattern of expression: UPGMA versus SOTA

J. Dopazo

METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

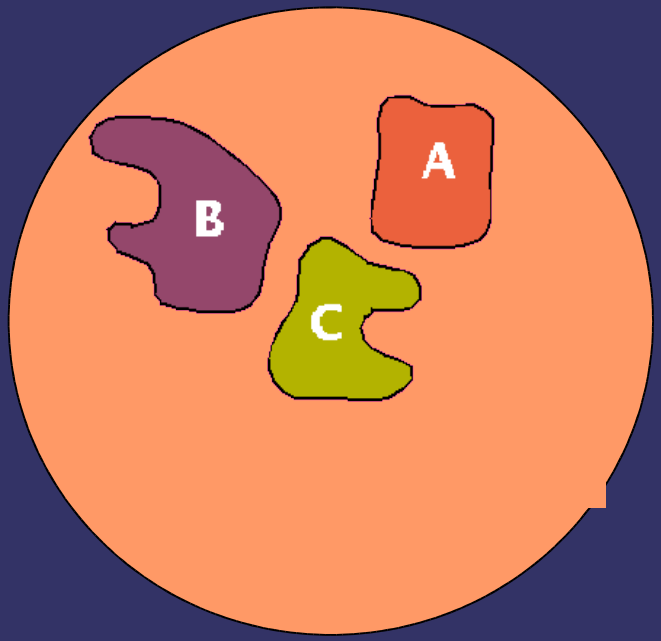


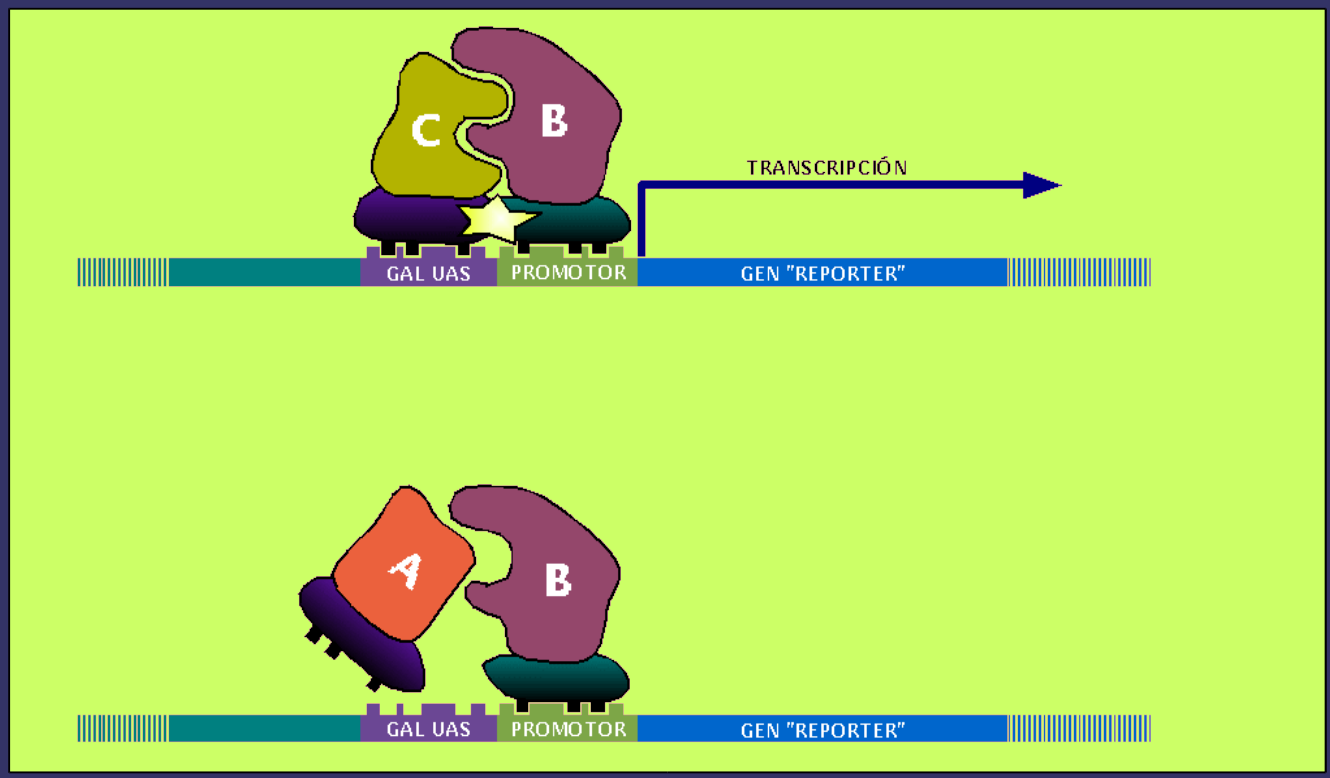
The interaction between two protein is required for the expression of a gene.

Proteomics:

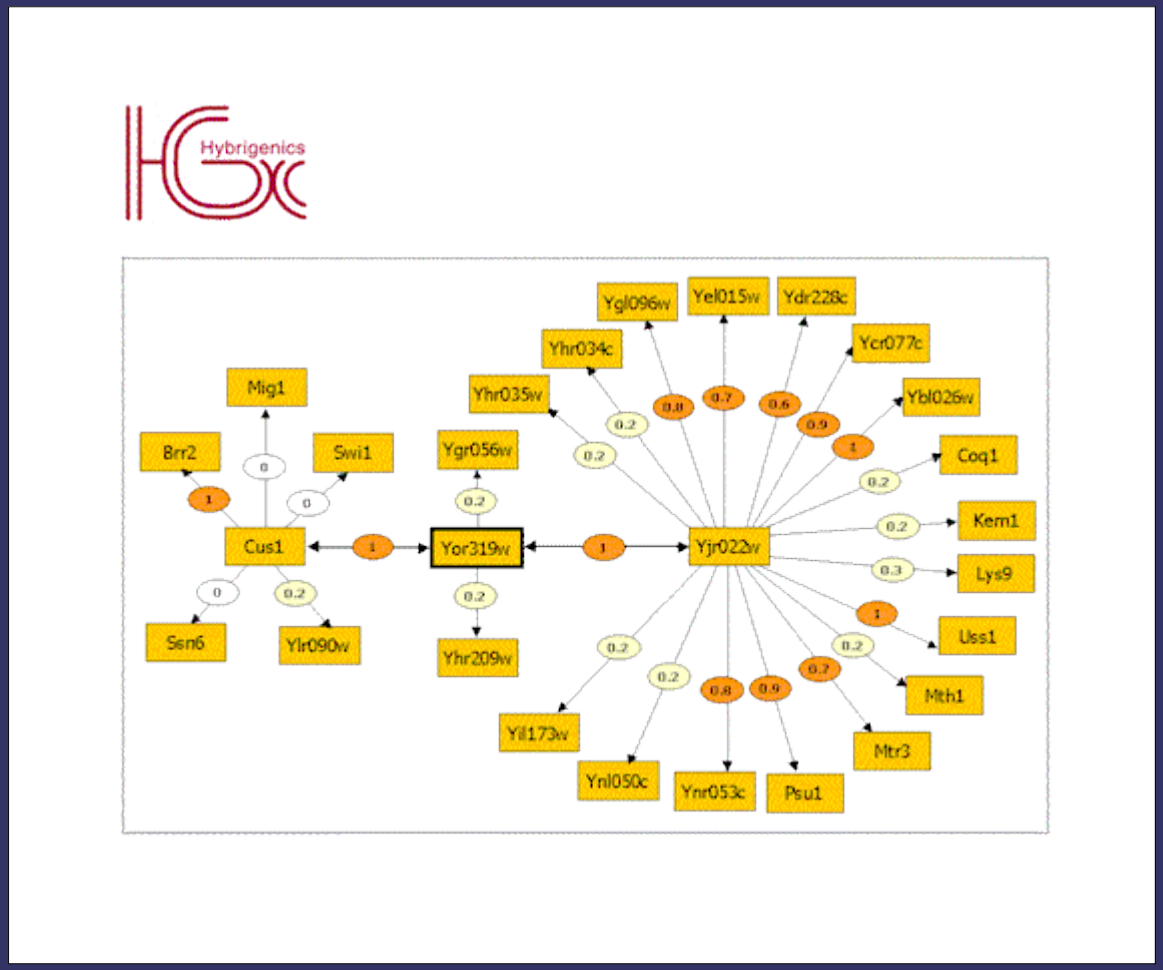
Yeast 2 Hybrid

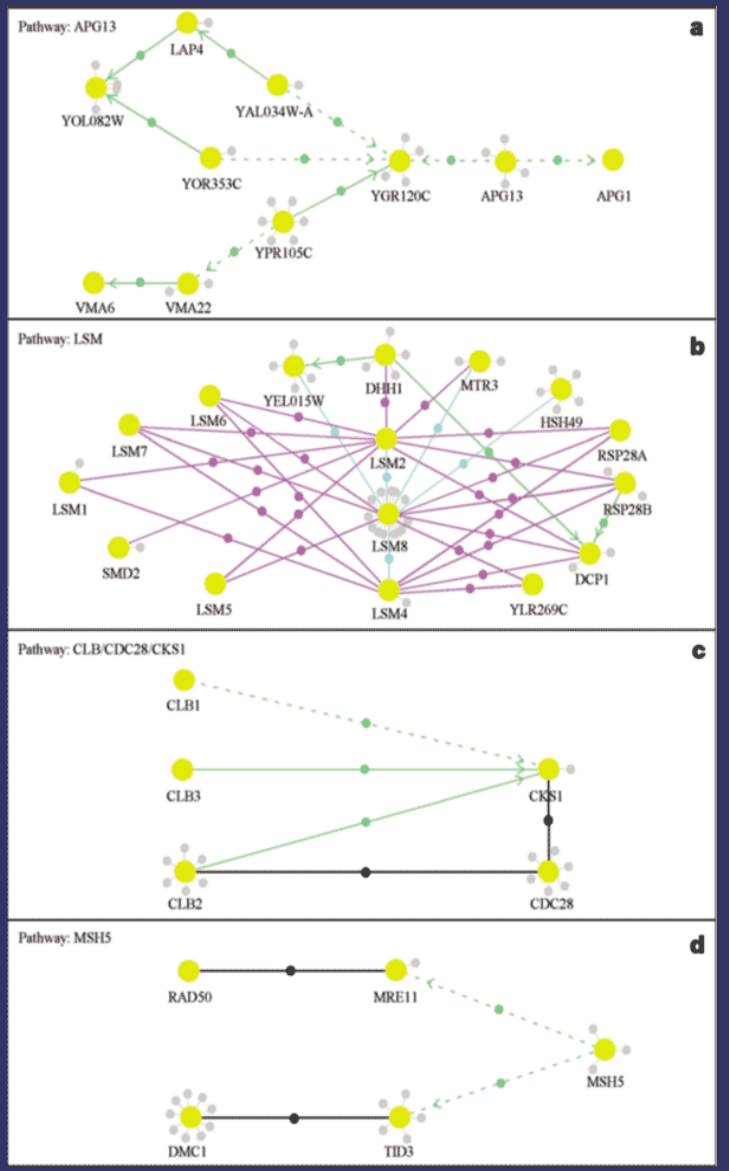
system





Large scale detection of protein interactions by the Yeast 2 hybrid system





C. elegans
Walhout et al. (2000) Science 287:116-122

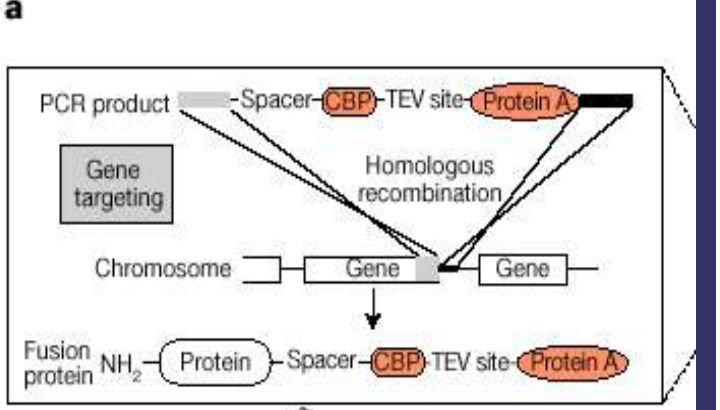
S. cerevisiae
Uetz et al. (2000) Nature 403:623-631
Ito et al. (2000) PNAS 97:1143-1147

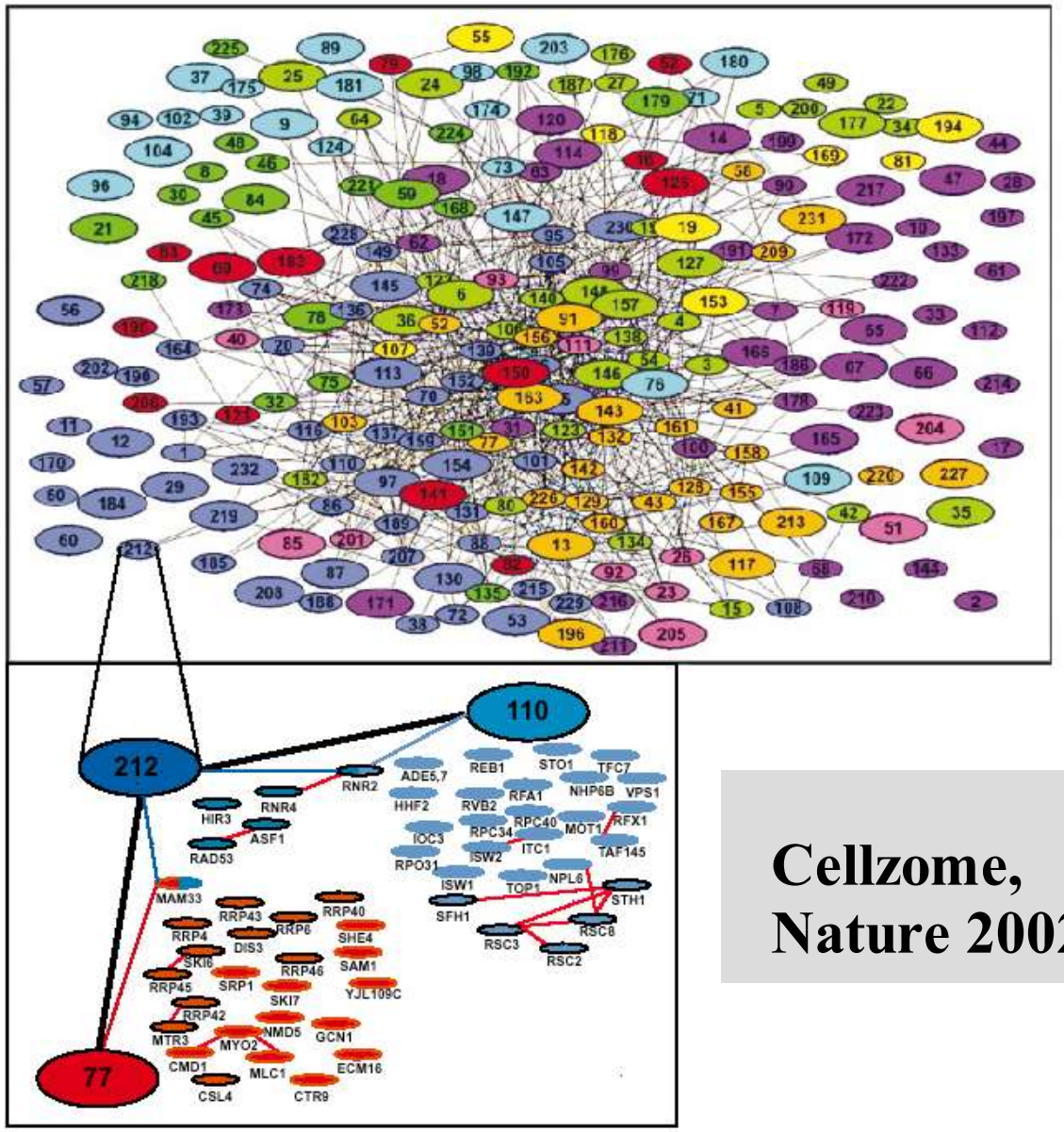
METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

Proteomics:

TAP, Tandem

Affinity Purification





Cellzome, Nature 2002

METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

Text mining (information extraction)

Rules to identify information about interactions in scientific texts.

The current list of rules includes

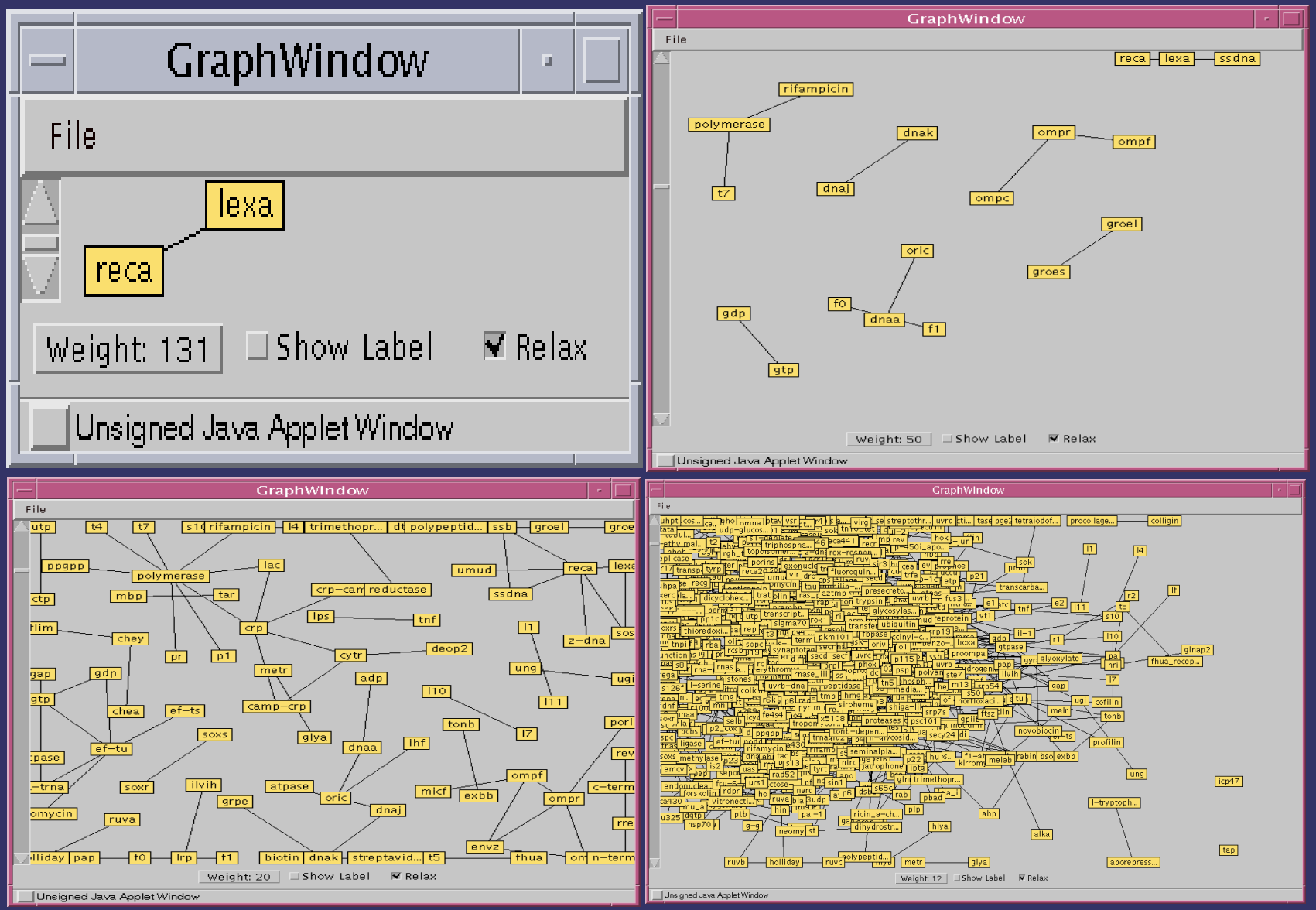
.protein [word]* [verb] [word]* protein
.[verb] of [word]* protein [word]* [by,to] [word]* protein
.[noun] of [word]* protein [word]* [by,with] [word]* protein
.[noun] between [word]* protein [word]* and [word]* protein
.protein [word]* protein [word]* [complex/es, dimer, heterodimer]
.complex formed between [word]* protein [word]* and [word]* protein
.complex/es of [word]* protein [word]* and [word]* protein
.protein [word]* forms a complex with [word]* protein

Sub-rules are used for incorporating particular cases

1.1. protein [word]* [verb] [word]* but not [word]* protein
1.2. protein [word]* cannot [word]* [verb] [word]* protein
1.3. protein [word]* does not [word]* [verb] [word]* protein
1.4. protein [word]* did not [word]* [verb] [word]* protein
1.5. protein [word]* was not [word]* [verb] [word]* protein
1.6. protein [word]* not [word]* [verb] [word]* by [word]* protein
1.7. protein [word]* not required for [word]* [verb] [word]* protein
1.8. protein [word]* failed to [word]* [verb] [word]* protein

E. coli INTERACTIONS DATABASE PROTOTYPE

Interaction networks identified by extracting information from the literature.



by C. Blaschke

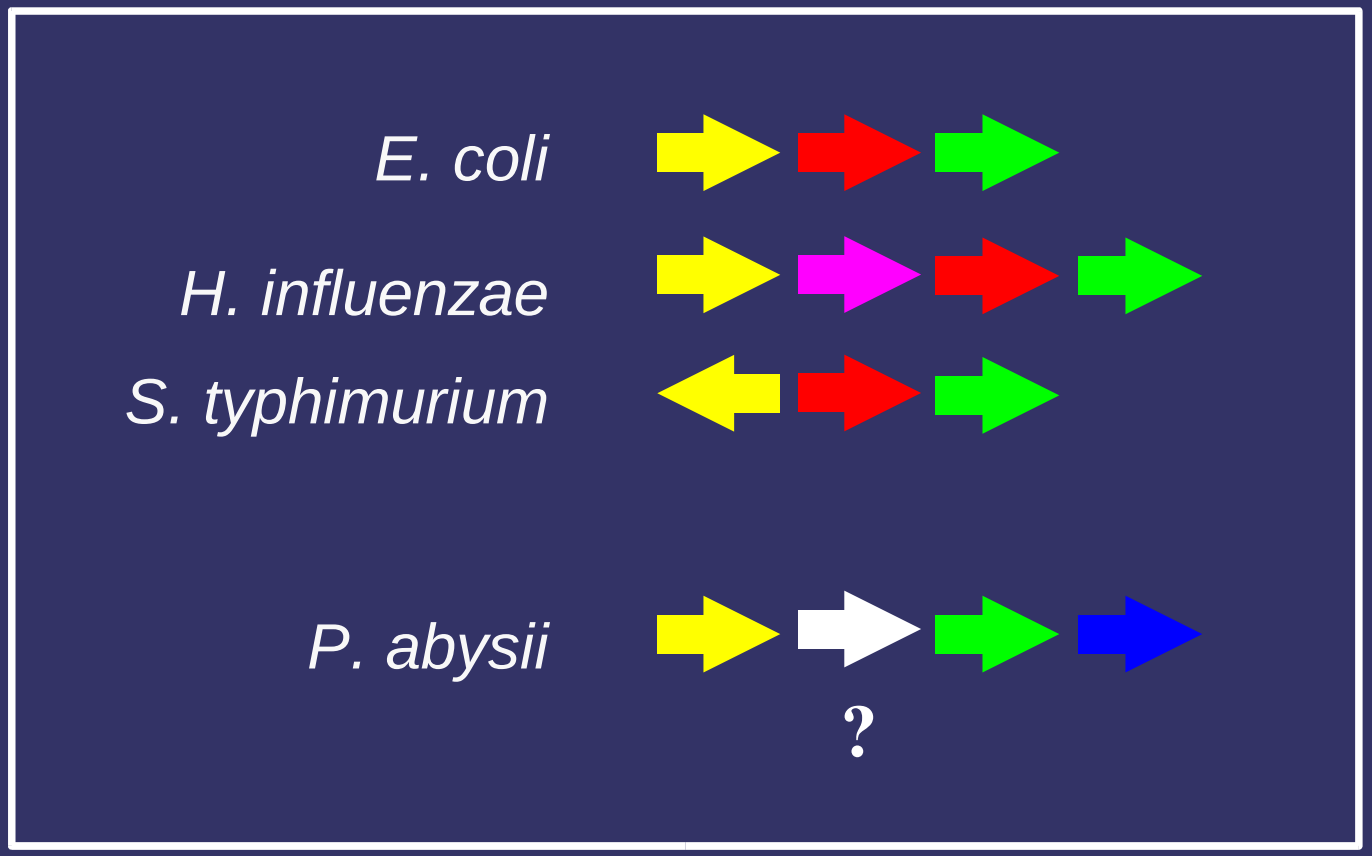
METHODS BASED ON THE DETECTION OF FUNCTIONAL INTERACTIONS

Comparative Sequence Genomics

- Methods based on the analysis of evolutionary information stored in sequences.
- There are FIVE methods, so far
  - Conservation of gene context or neighbourhood
  - Detection of gene fusions
  - Similarity of gene trees
  - Identification of correlated mutations
  - Similarity of phylogenetic profiles

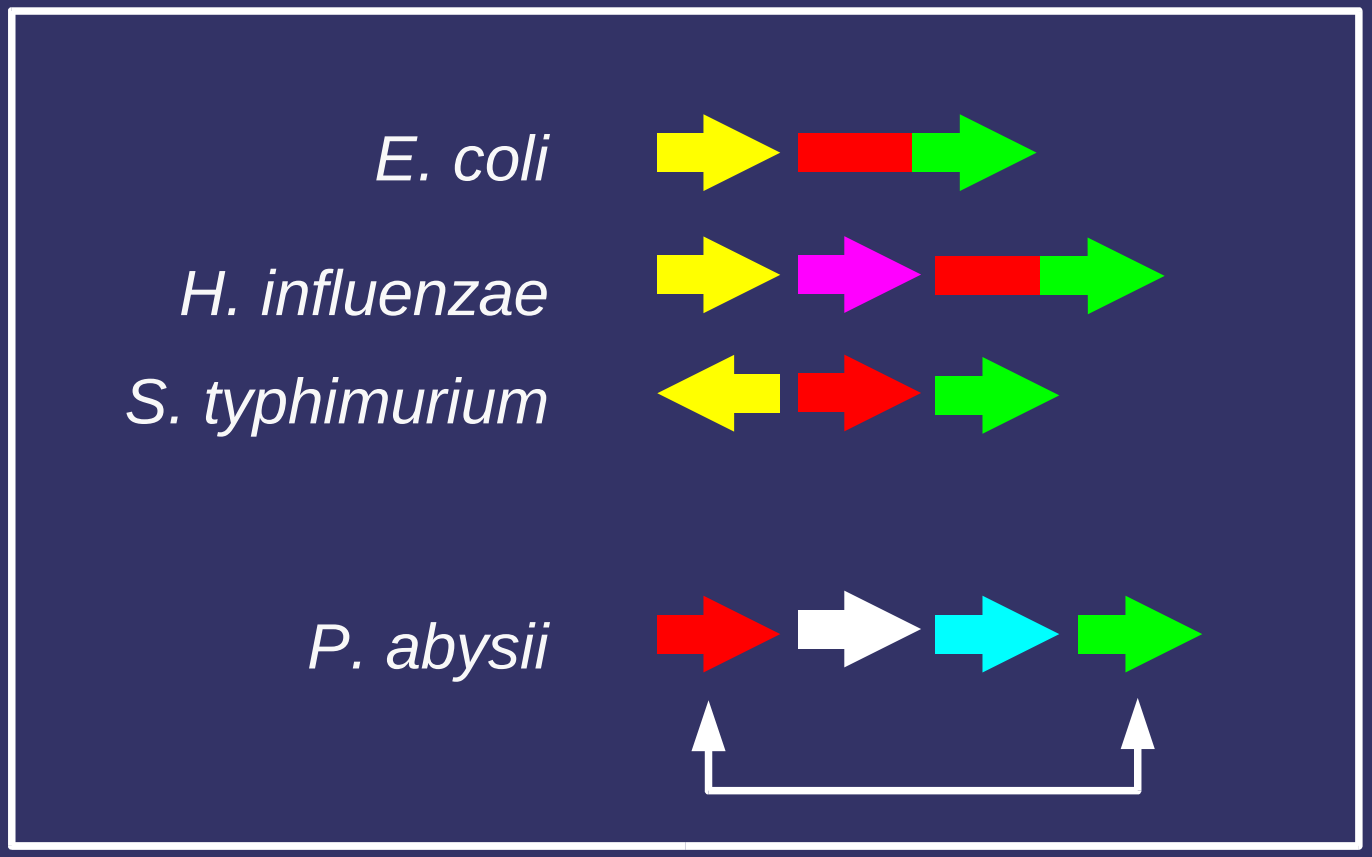
METHODS BASED ON COMPARATIVE SEQUENCE GENOMICS

Conservation of gene context or neighborhood



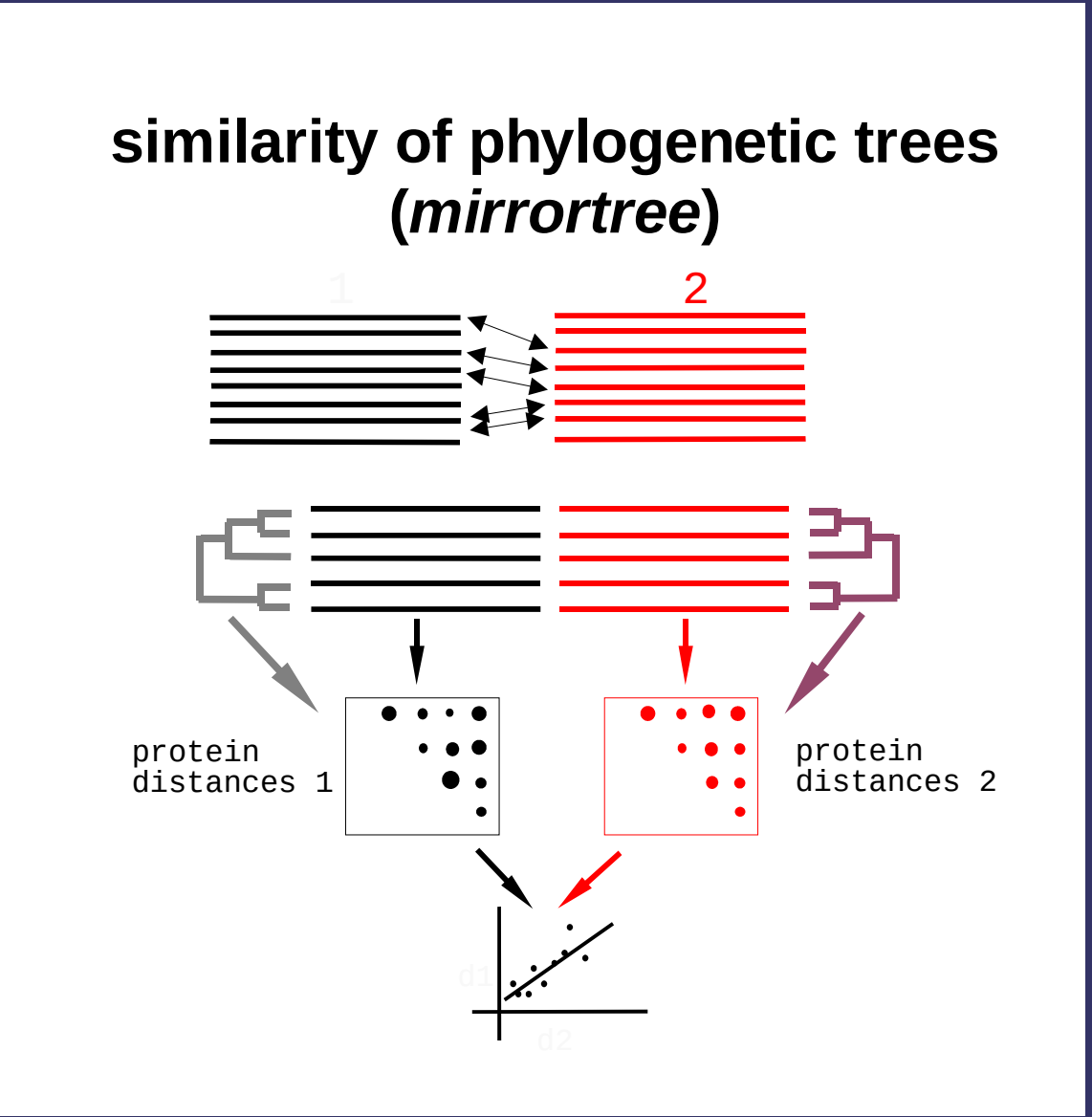
METHODS BASED ON COMPARATIVE SEQUENCE GENOMICS

Detection of gene fussions, to infer physical interactions



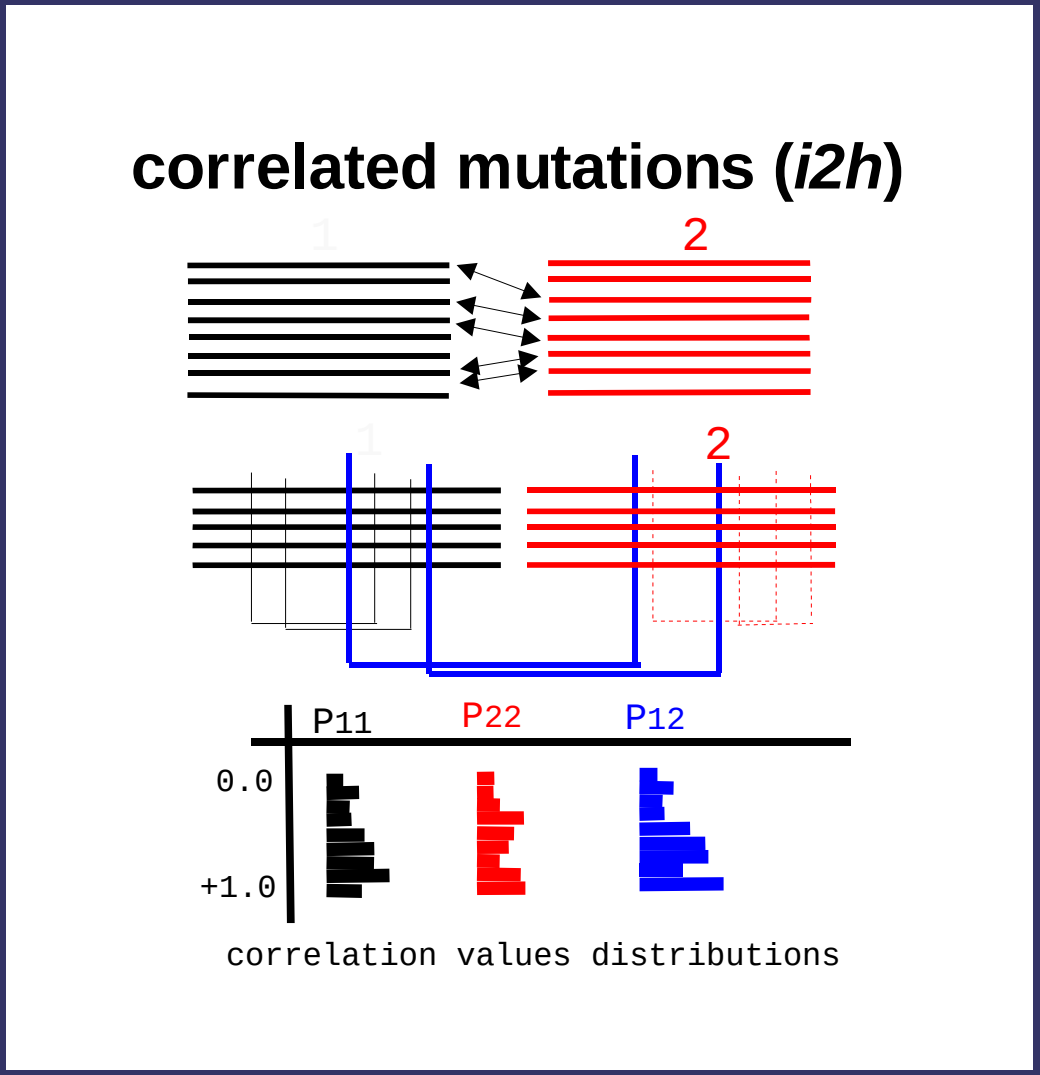
METHODS BASED ON COMPARATIVE SEQUENCE GENOMICS

Similarity of gene trees



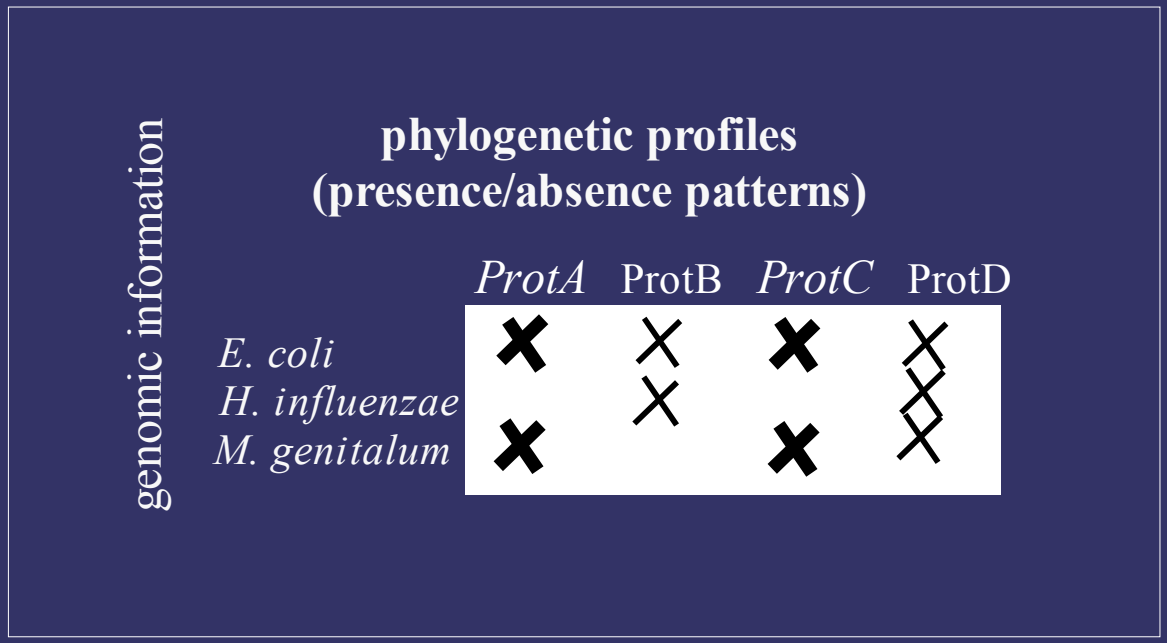
METHODS BASED ON COMPARATIVE SEQUENCE GENOMICS

Detection of correlated mutations



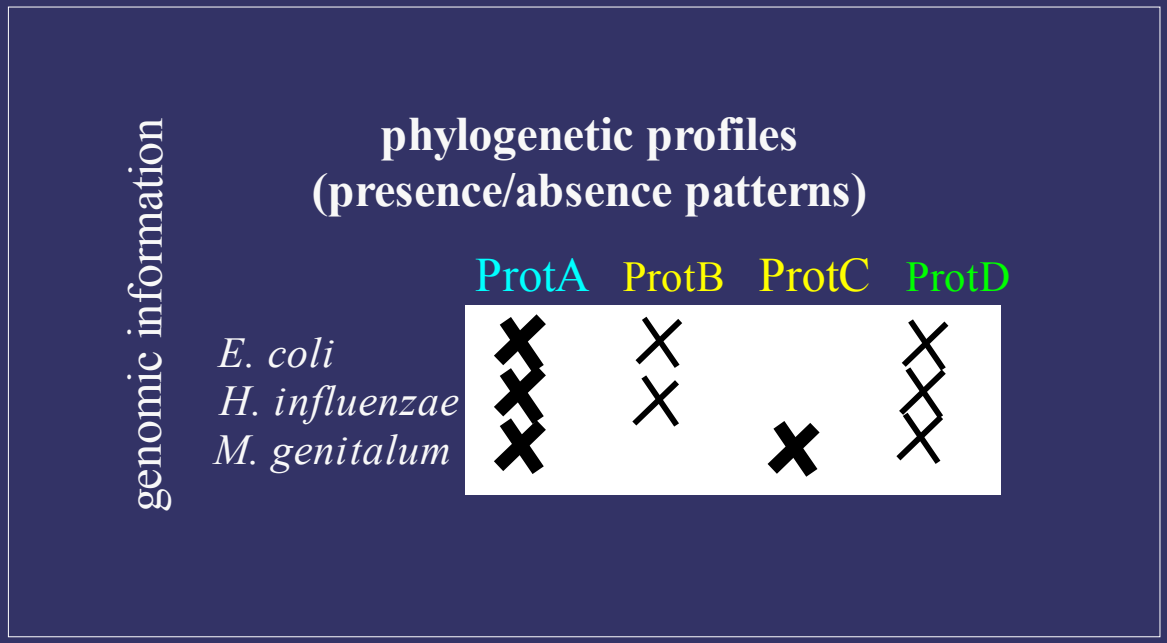
METHODS BASED ON COMPARATIVE SEQUENCE GENOMICS

Similarity of phylogenetic profiles



NON ORTHOLOGOUS GENE DISPLACEMENT

Anti-correlation of phylogenetic profiles



Morett et al, 2003

DATABASES OF PROTEIN INTERACTIONS

Predicted interactions, by comparative genomics (integration of several methods):

- **EcID**
- **Predictome**
- **STRING**
- **Prolinks**

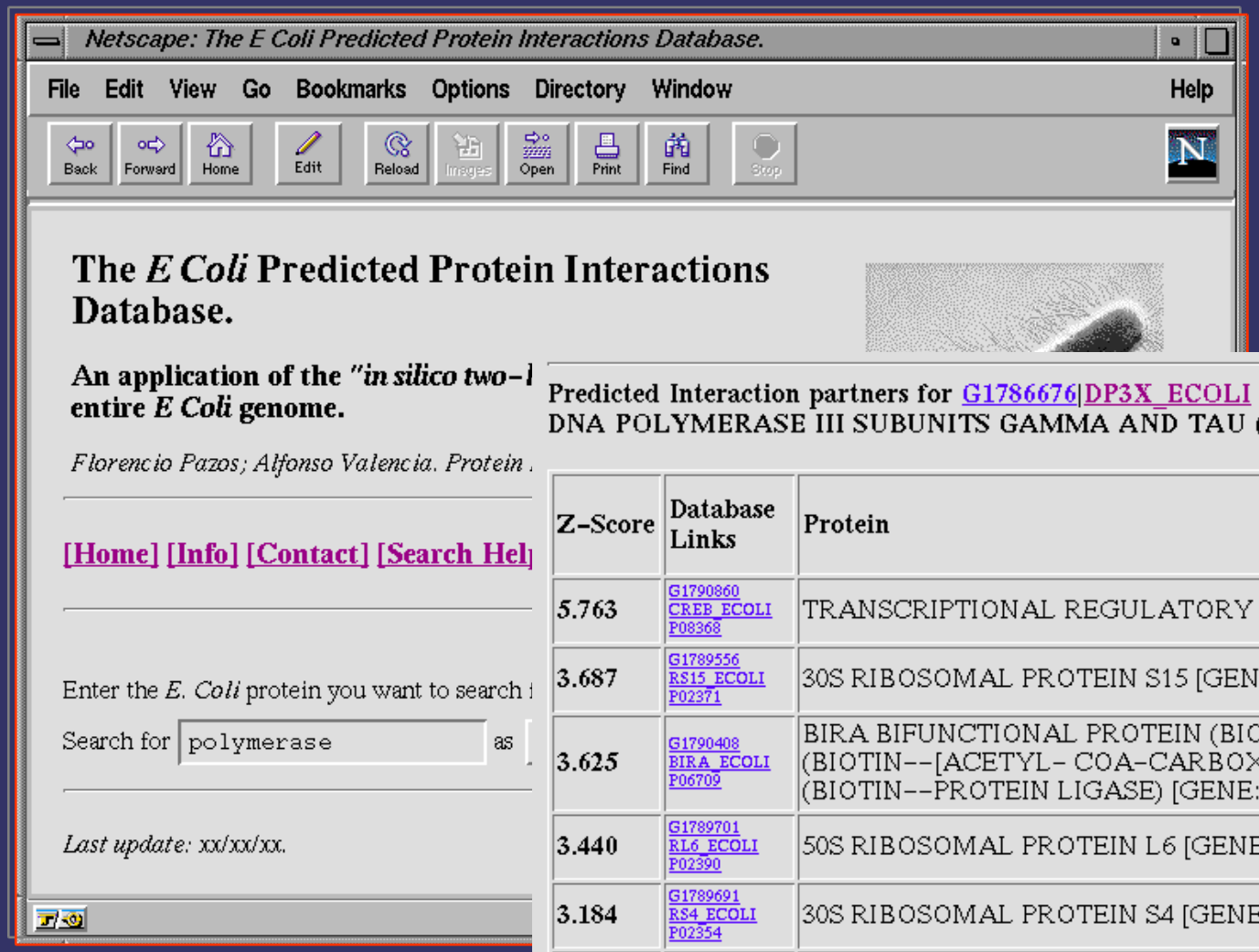
Experimentally detected protein interactions:

- **DIP**
- **BIND**
- **MIPS**
- **PIMrider**

Experimentally detected and predicted interactions:

- **IntAct** (EBI).

E. coli i2h database



Predicted Interaction partners for G1786676|DP3X_ECOLI |P06710
DNA POLYMERASE III SUBUNITS GAMMA AND TAU (EC 2.7.7.7) [GENE: DNAX AND DNAZ]

Z-Score	Database Links	Protein	Run ECID for it
5.763	G1790860 CREB_ECOLI P08368	TRANSCRIPTIONAL REGULATORY PROTEIN CREB [GENE: CREB]	>>
3.687	G1789556 RS15_ECOLI P02371	30S RIBOSOMAL PROTEIN S15 [GENE: RPSO OR SECC]	>>
3.625	G1790408 BIRA_ECOLI P06709	BIRA BIFUNCTIONAL PROTEIN (BIOTIN OPERON REPRESSOR) (BIOTIN--[ACETYL- COA-CARBOXYLASE] SYNTHETASE) (EC 6.3.4.15) (BIOTIN--PROTEIN LIGASE) [GENE: BIRA OR BIOR OR DHBB]	>>
3.440	G1789701 RL6_ECOLI P02390	50S RIBOSOMAL PROTEIN L6 [GENE: RPLF]	>>
3.184	G1789691 RS4_ECOLI P02354	30S RIBOSOMAL PROTEIN S4 [GENE: RPSD OR RAMA]	>>
3.119	G2367268 THDF_ECOLI P25522	THIOPHENE AND FURAN OXIDATION PROTEIN THDF [GENE: THDF OR TRME]	>>
2.900	G2367200 TRUB_ECOLI P09171	TRNA PSEUDOURIDINE SYNTHASE B (EC 4.2.1.70) (TRNA PSEUDOURIDINE 55 SYNTHASE) (PSI55 SYNTHASE) (PSEUDOURIDYLATE SYNTHASE) (URACIL HYDROLYASE) (P35 PROTEIN) [GENE: TRUB OR P35]	>>
2.608	G1787357 MFD_ECOLI P30958	TRANSCRIPTION-REPAIR COUPLING FACTOR (TRCF) [GENE: MFD]	>>

by F. Pazos, 1999

STRING: a database of predicted functional associations between proteins.

Von Mering et al. (2003) Nucleic Acids Research, 31 (1):258.

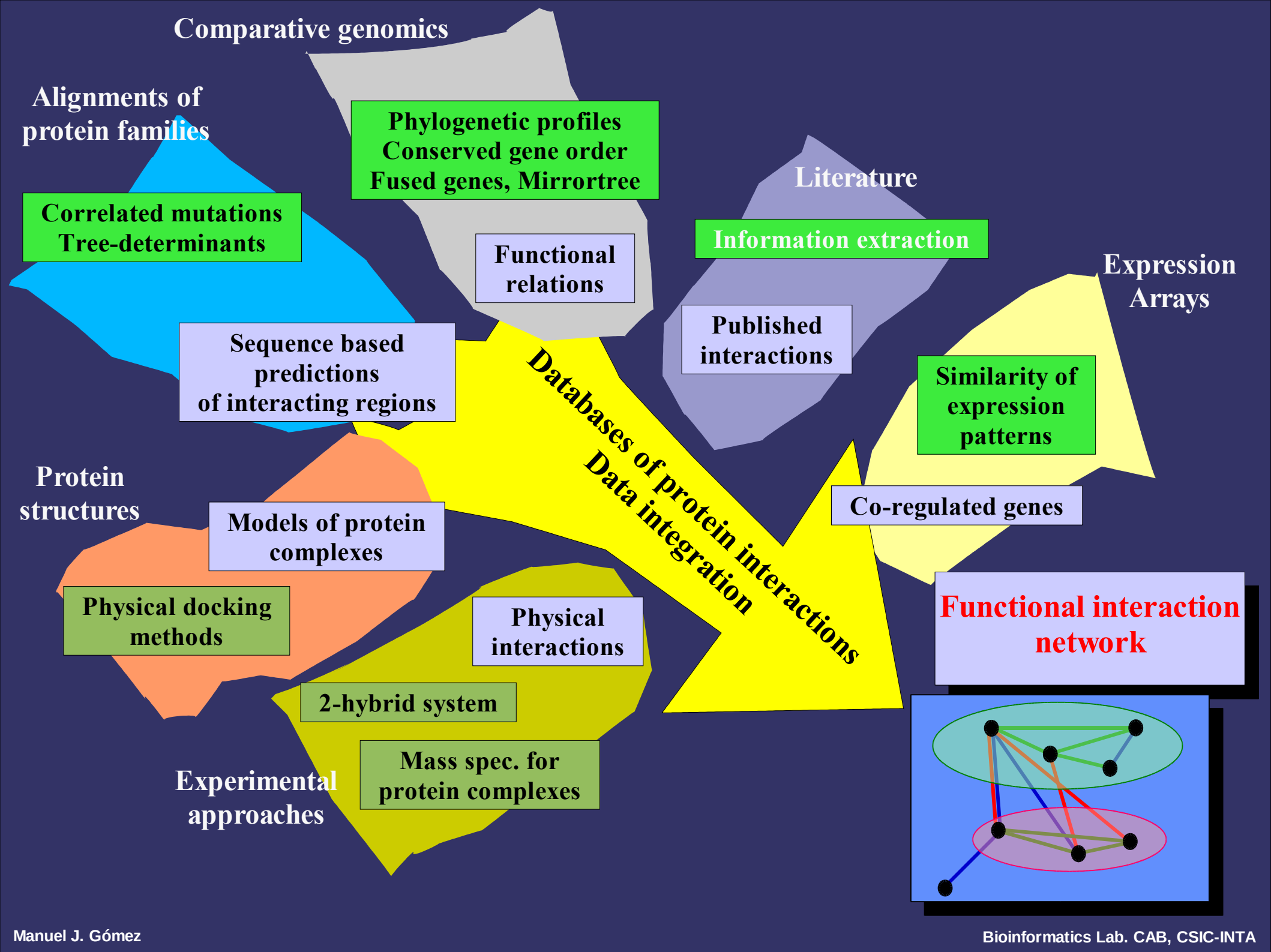
Integrates information about:

- Phylogenetic profiles
- Conservation of gene context
- Existence of gene fusions

With that information, functional associations between COGs are established. The functional associations may be interpreted as physical or regulatory interactions or participation in the same metabolic pathway.

The network of interactions is huge and may include all COGs, depending of the stringency used to calculate it.

In any case, that network can be broken down in modules that may represent individual metabolic pathways, regulatory circuits or protein complexes.



Manuel J. Gómez

Bioinformatics Lab. CAB, CSIC-INTA

This presentation contains material from:

Christian Blaschke, BioAlma
Florencio Pazos, CNB
Alfonso Valencia, CNB
Manuel J Gómez, CAB