Bioinformatics 3

V 3 –

Data for Building Networks

Fri, Oct 25, 2013

Graph Layout I

Requirements:

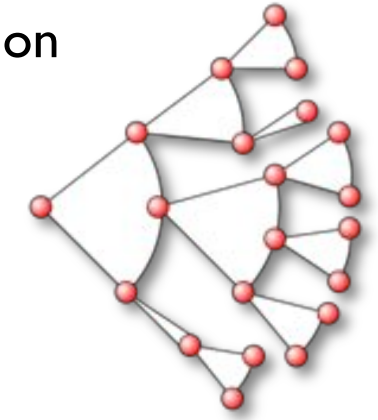
- fast and stable
- nice graphs
- visualize relations
- symmetry
- interactive exploration
- …

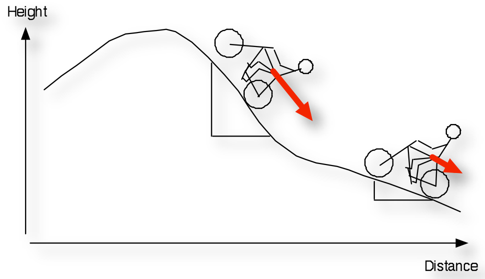
Force-directed Layout:

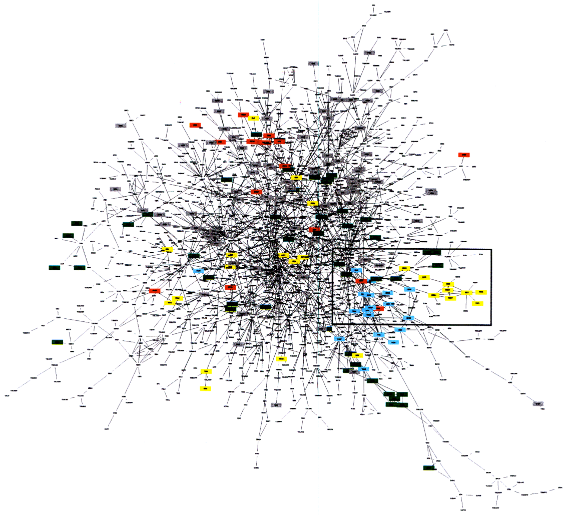
based on energy minimization

→ runtime

→ mapping into 2D





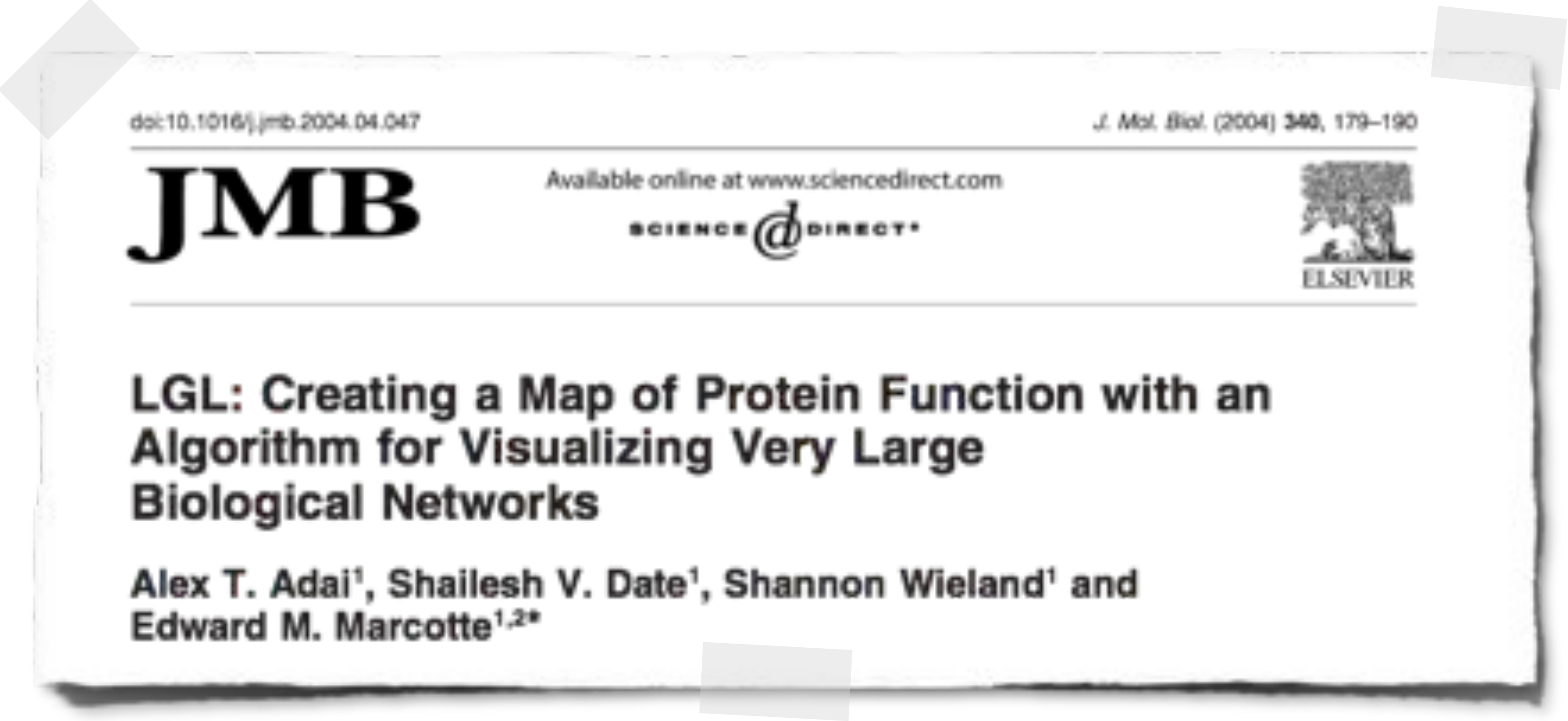


H3: for hierarchic graphs

→ MST-based cone layout

→ hyperbolic space

→ efficient layout for **biological data???**



doi:10.1016/j.jmb.2004.04.047

J. Mol. Biol. (2004) 340, 179–190

JMB

Available online at www.sciencedirect.com

SCIENCE @ DIRECT•

ELSEVIER

LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks

Alex T. Adai[1], Shailesh V. Date[1], Shannon Wieland[1] and Edward M. Marcotte[1,2*]

Aim: analyze and visualize **homologies** within the **protein universe**
50 genomes, 145579 proteins, $21 \cdot 10^9$ BLASTP pairwise sequence comparisons

Expectations:

- homologs will be close together
- **fusion** proteins („Rosetta Stone proteins") will **link** proteins of related function.

→ need to visualize an extremely large network!

→ develop a **stepwise scheme**

LGL: stepwise scheme

(0) **create network** from BLAST E-score

145'579 proteins

E < 10^{-12} → 1'912'684 links , 30737 proteins in the largest cluster

(1) **separate** original network into **connected sets**

11517 connected components, 33975 proteins w/out links

(2) force directed **layout** of each **component independently**,
based on a MST

(3) integrate connected sets into one coordinate system
via a **funnel process**, starting from the largest set

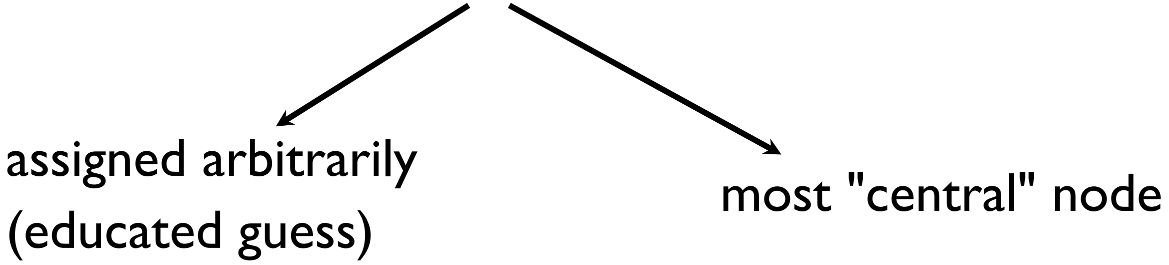
The first connected set is placed at the bottom of a potential funnel.
Other sets are placed one at a time on the rim of the potential funnel and allowed to fall towards the bottom where they are frozen in space upon collision with the previous sets.

Adai et al. J. Mol. Biol. 340, 179 (2004)

Component layout I

For each component independently:

→ start from the **root node** of the MST



assigned arbitrarily
(educated guess)

most "central" node

Centrality: minimize total distance to all other nodes in the component

$$v_{root} = \min\left(\sum_{(v,u)\in V} d(v,u)\right)$$

Level *n*-nodes: nodes that are *n* links away from the root in the MST

Layout → place **root** at the **center**

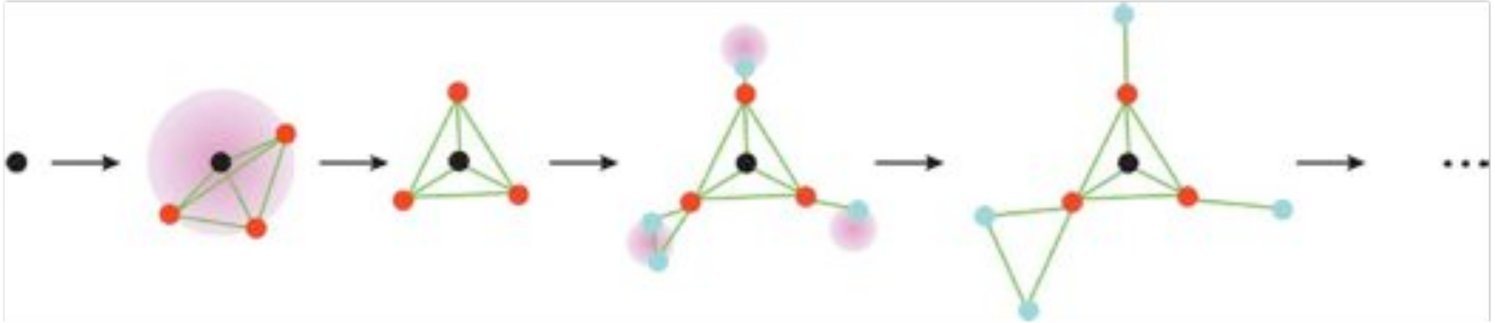
Component Layout II

- start with root node of the MST

- place level-1 nodes on circle (sphere) around root,
  add all links,
  relax springs (+ short-range repulsion)

- place level-2 nodes on circles (sphere) outside their level-1 descendants,
  add all links,
  relax springs

- place level-3 nodes on circles (sphere) outside their level-2 descendants,
  :

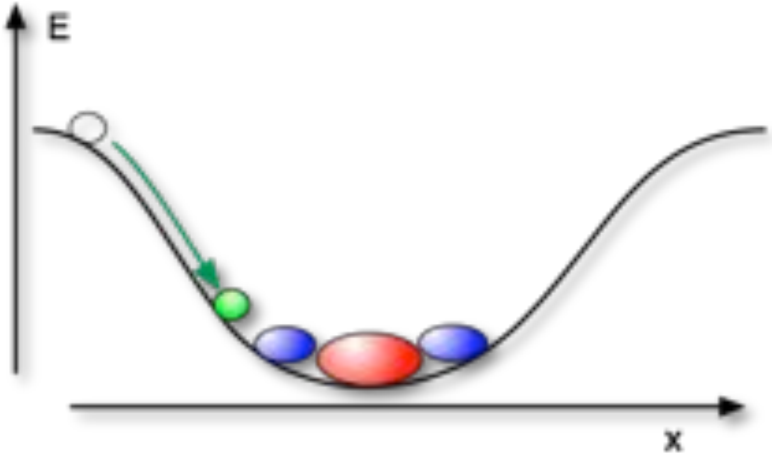


Adai et al. J. Mol. Biol. 340, 179 (2004)

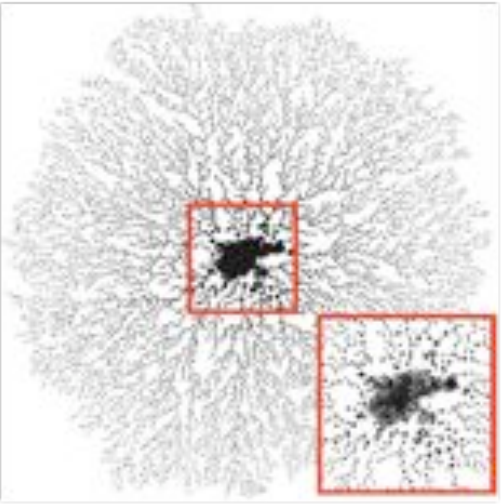
Combining the Components

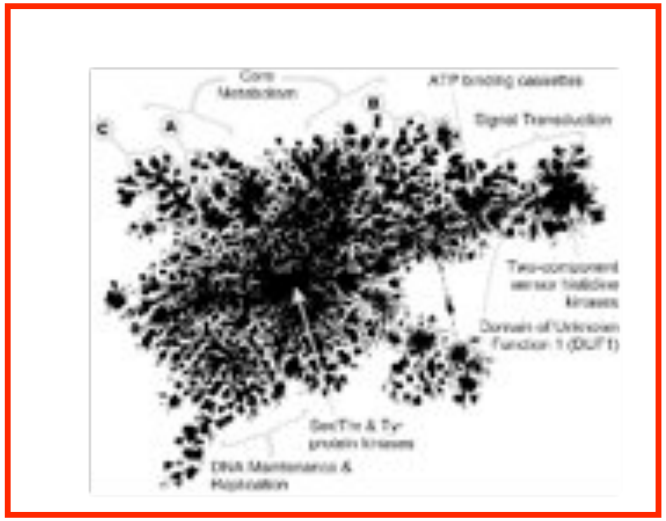
When the components are finished
→ **assemble** using energy **funnel**

- place largest component at bottom
- place next smaller one somewhere on the rim, let it slide down
→ freeze upon contact

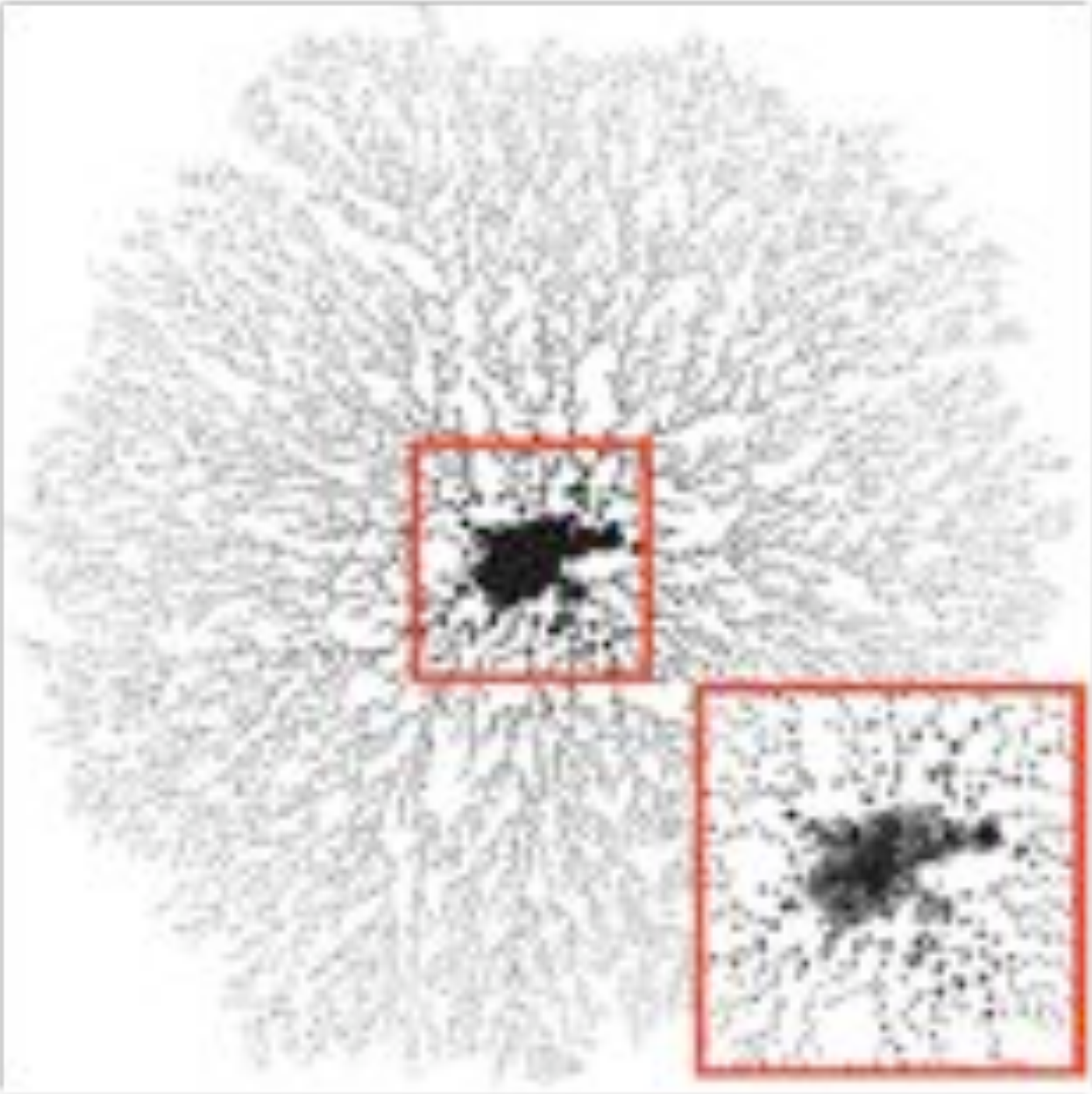


No information in the relative positions of the components!!!



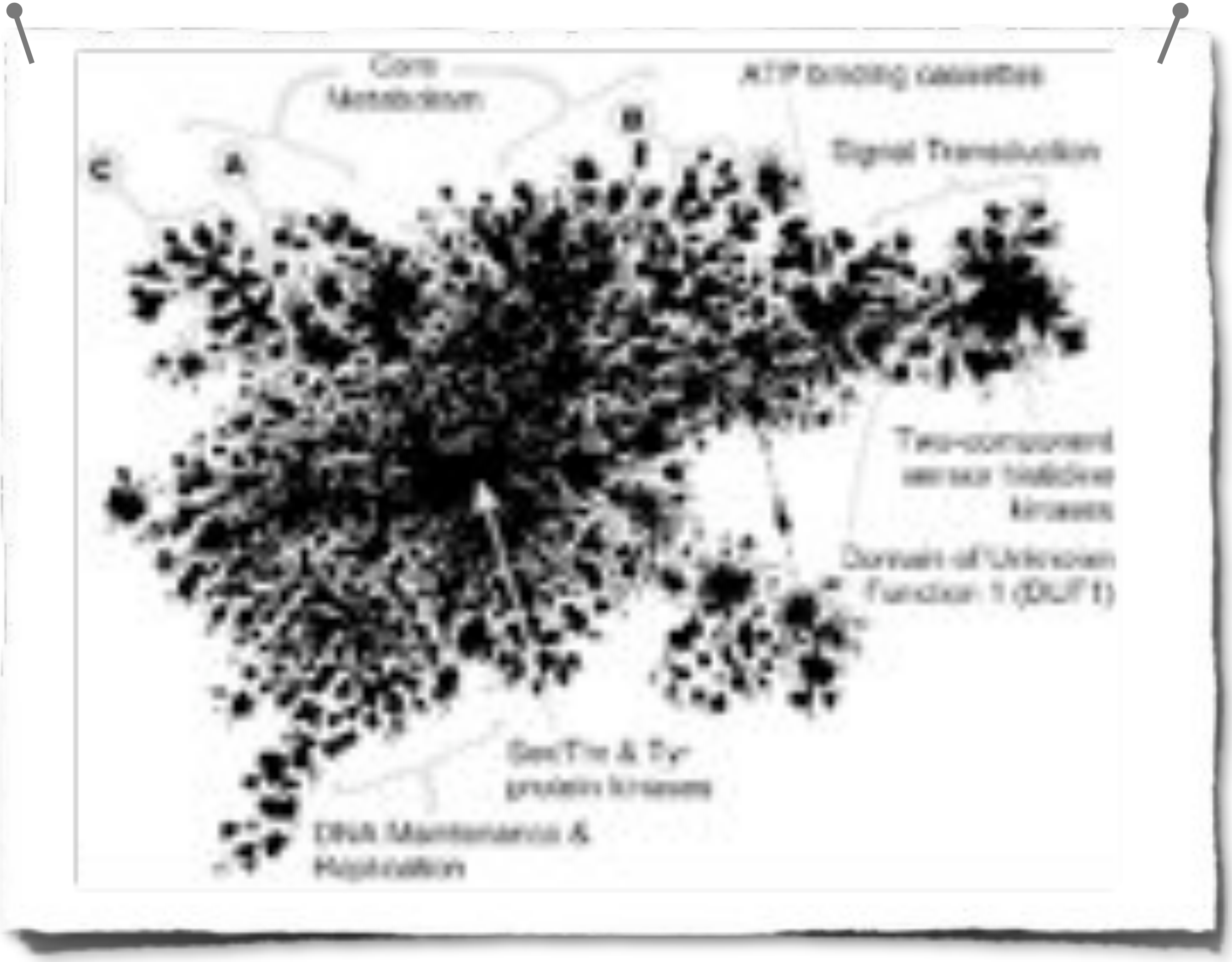

Core Metabolism
ATP binding cassettes
B
C
A
Signal Transduction
Two-component sensor histidine kinases
Domain of Unknown Function 1 (DUF1)
Ser/Thr & Tyr protein kinases
DNA Maintenance & Replication

Adai et al. J. Mol. Biol. 340, 179 (2004)



Adai et al. J. Mol. Biol. 340, 179 (2004)

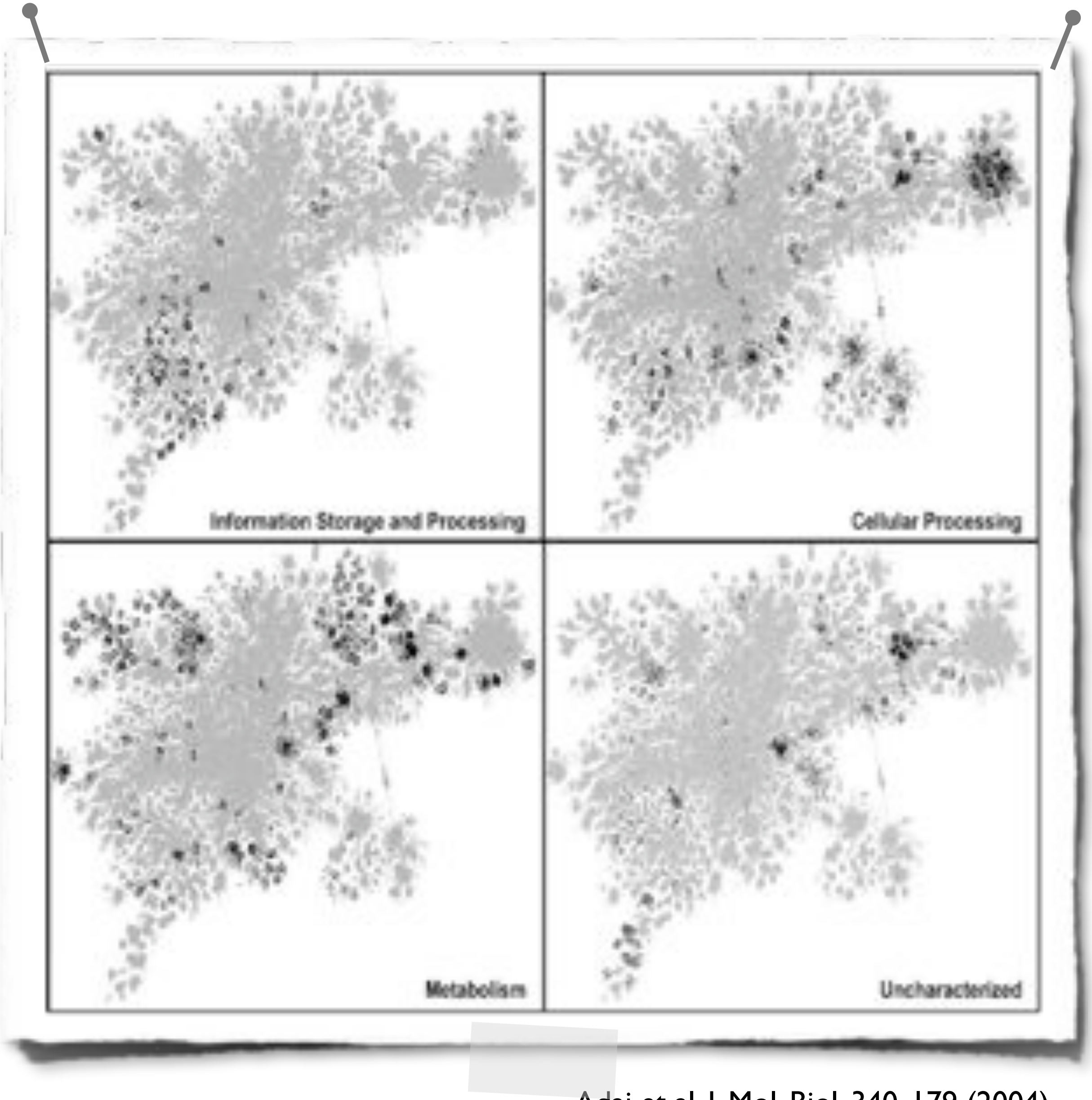
Annotations in the Largest Cluster


Core Metabolism
ATP binding cassettes
B
Signal Transduction
C
A
Two-component sensor histidine kinases
Domain of Unknown Function 1 (DUF1)
Ser/Thr & Tyr protein kinases
DNA Maintenance & Replication

Related functions in the same regions of the cluster → predictions

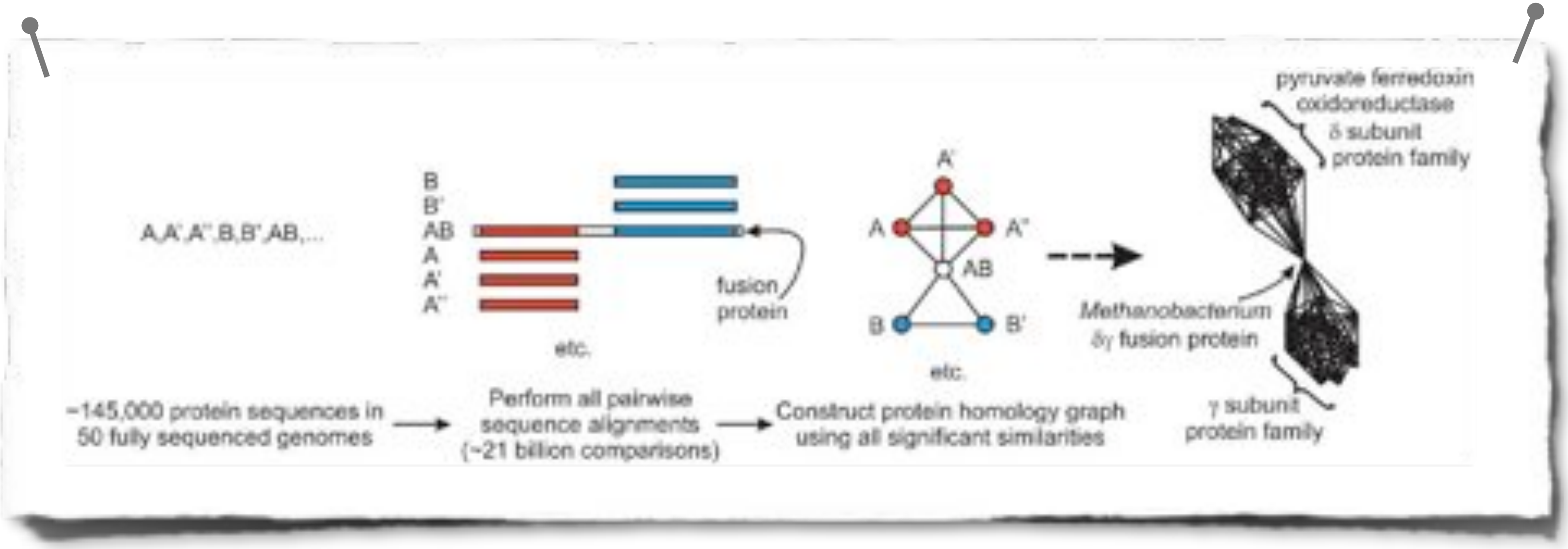
Adai et al. J. Mol. Biol. 340, 179 (2004)

Clustering of Functional Classes



Adai et al. J. Mol. Biol. 340, 179 (2004)

Fusion Proteins


A,A',A",B,B',AB,...
B
B'
AB
A
A'
A"
fusion protein
etc.
A'
A
A"
AB
B
B'
etc.
pyruvate ferredoxin oxidoreductase δ subunit protein family
Methanobacterium δγ fusion protein
γ subunit protein family
~145,000 protein sequences in 50 fully sequenced genomes
Perform all pairwise sequence alignments (~21 billion comparisons)
Construct protein homology graph using all significant similarities

Fusion proteins **connect** two protein homology **families**

A, A', A", AB and B, B', AB

→ historic genetic **events**: fusion, fission, duplications, …

Also **in the network**:

homologies <=> edges

remote homologies <=> in the same cluster

non-homologous functional relations <=> adjacent, linked clusters

Adai et al. J. Mol. Biol. 340, 179 (2004)

Functional Relations between Gene Families

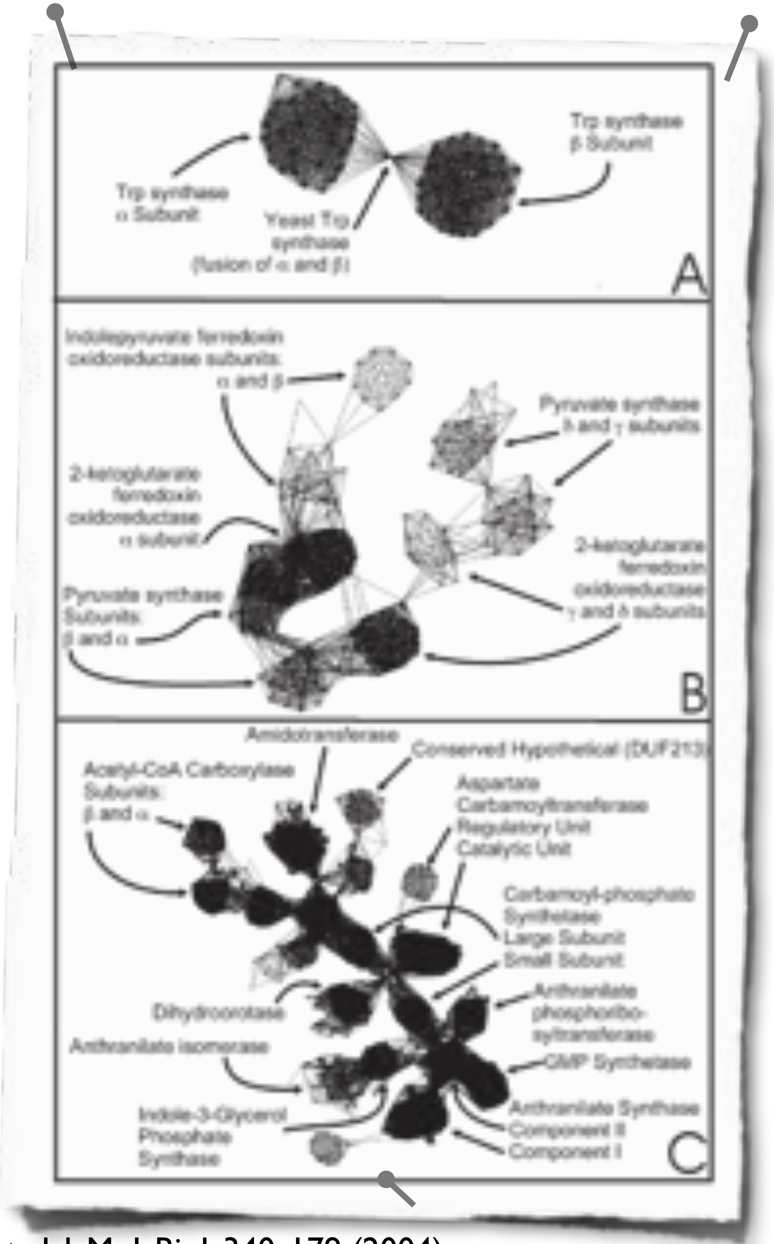
Examples of spatial localization of protein function in the map

A: the linkage of the tryptophan synthase α family to the functionally coupled but non-homologous β family by the yeast tryptophan synthase $\alpha\beta$ fusion protein,

B: protein subunits of the pyruvate synthase and alpha-ketoglutarate ferredexin oxidoreductase complexes

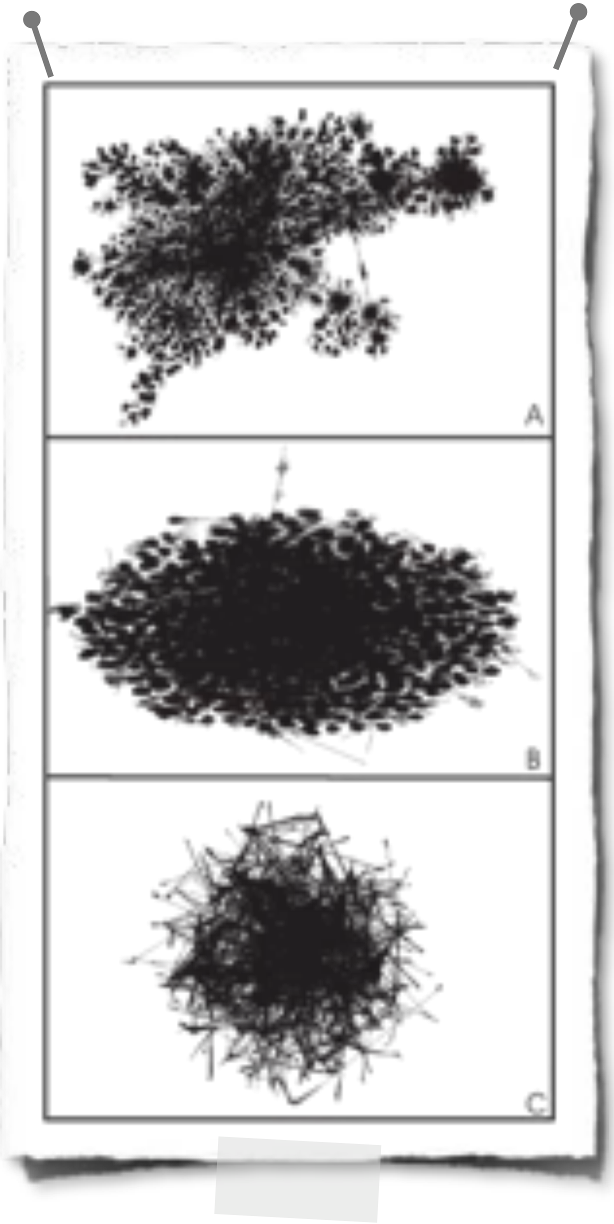
C: metabolic enzymes, particularly those of acetyl CoA and amino acid metabolism

→ DUF213 likely has metabolic function!



Adai et al. J. Mol. Biol. 340, 179 (2004)

And the Winner iiiis…



Compare the layouts from

A: **LGL** – **hierarchic** force-directed layout according to MST
→ structure from homology

B: **global force**-directed layout without MST
→ no structure, no components visible

C: **InterViewer** – collapses similar nodes
→ reduced complexity

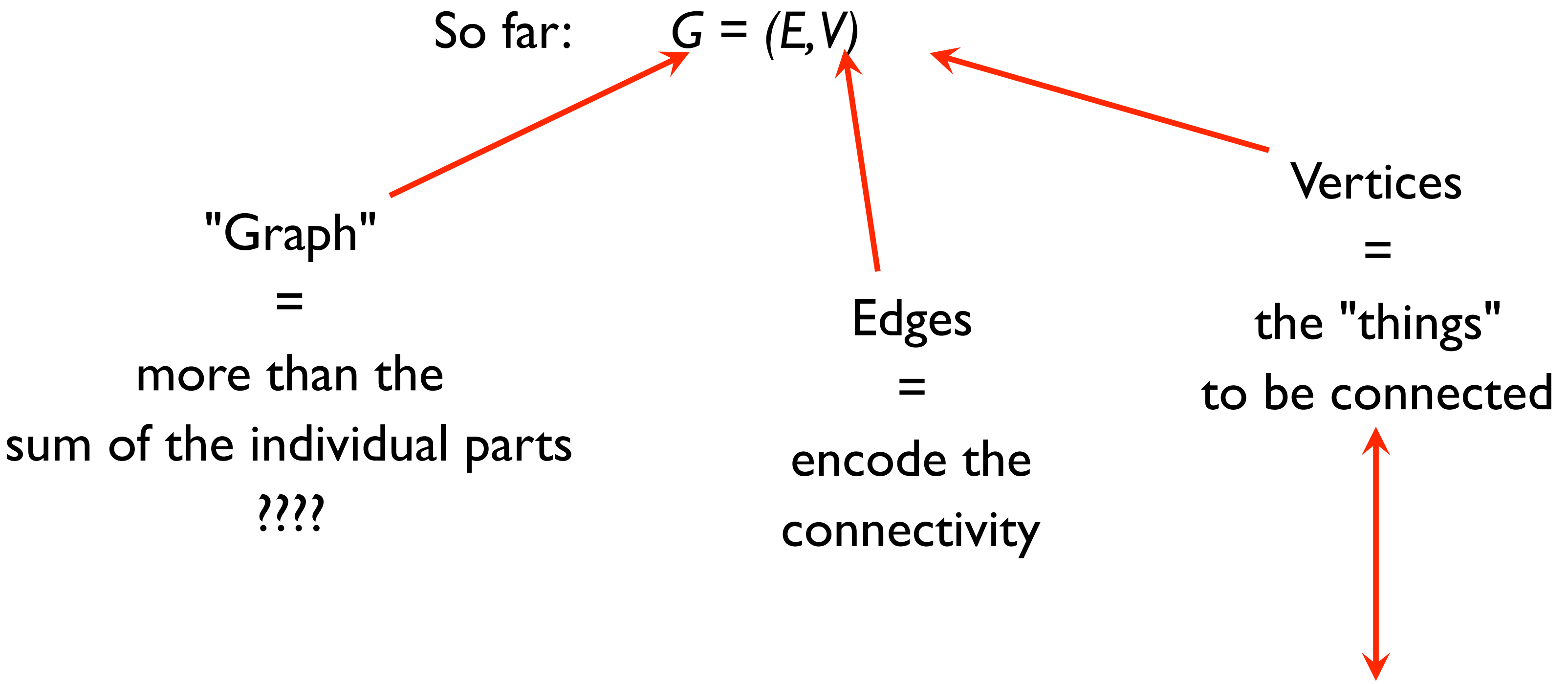
Adai et al. J. Mol. Biol. 340, 179 (2004)

Graph Layout: Summary

Approach	**Idea**
Force-directed spring model	relax energy, springs of appropriate lengths
Force-directed spring-electric model	relax energy, springs for links, Coulomb repulsion between all nodes
H3	spanning tree in hyperbolic space
LGL	hierarchic, force-directed algorithm for modules

the same physical concept, different implementations!

A "Network"



So far: $G = (E,V)$

"Graph"
=
more than the
sum of the individual parts
????

Edges
=
encode the
connectivity

Vertices
=
the "things"
to be connected

→ what are interesting biological "things"?

→ how are they connected?

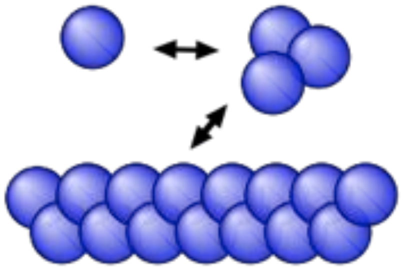
→ are the informations accessible/reliable?

Classified by:

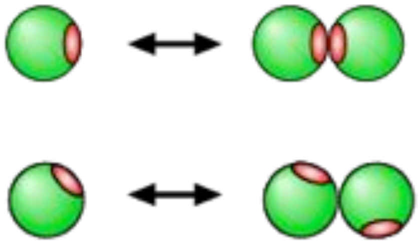
- degree distribution
- clustering
- connected components
- …

Protein Complexes

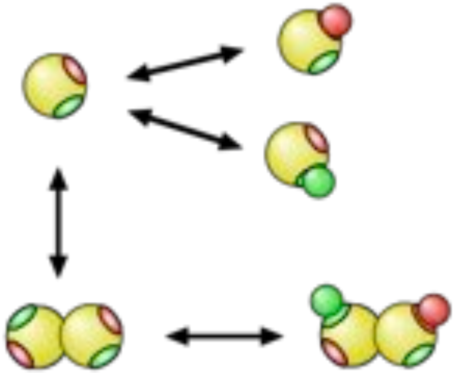
Assembly of structures



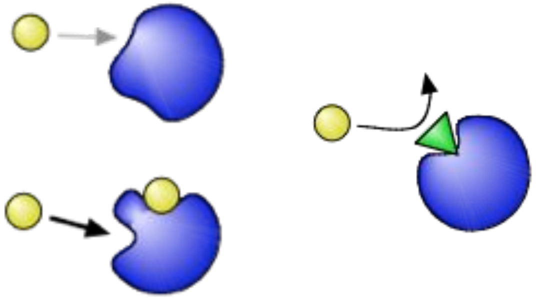
Complex formation may lead to modification of the active site



protein machinery built from parts via dimerization and oligomerization



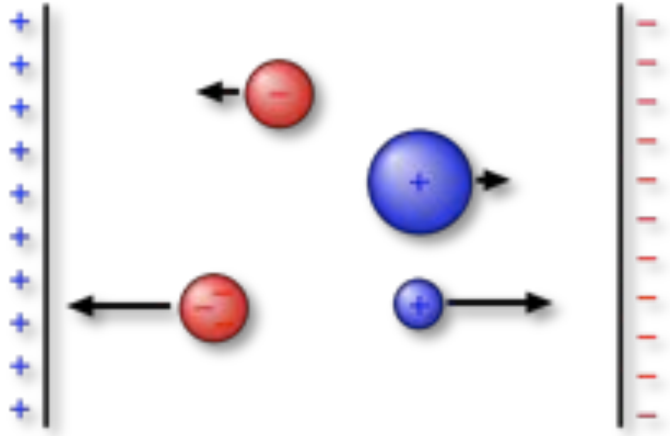
Complex formation may lead to increased diversity

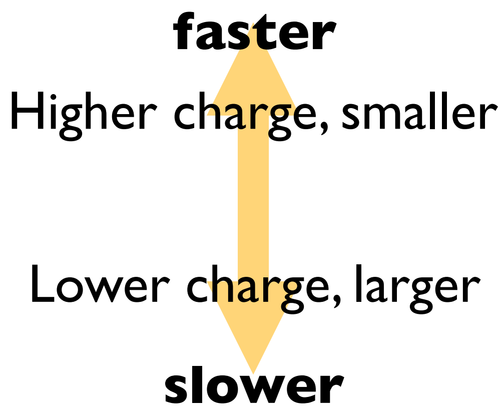


Cooperation and allostery

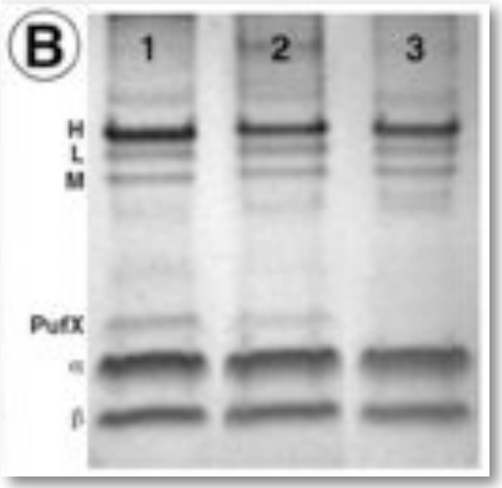
Gel Electrophoresis

Electrophoresis: directed diffusion of charged particles in an electric field





Put proteins in a spot on a gel-like matrix,
apply electric field
→ separation according to size (mass) and charge
→ identify constituents of a complex



Nasty details: protein charge vs. pH, cloud of counter ions,
protein shape, denaturation, …

SDS-PAGE

For better control: denature proteins with detergent

Often used: sodium dodecyl sulfate (**SDS**)
→ denatures and coats the proteins with a negative charge
→ charge proportional to mass
→ traveled distance per time

$$x \propto \frac{1}{\log(M)}$$

→ **SDS-p**oly**a**crylamide **g**el **e**lectrophoresis

After the run: **staining** to make proteins visible

For "quantitative" analysis: compare to **marker** (set of proteins with known masses)

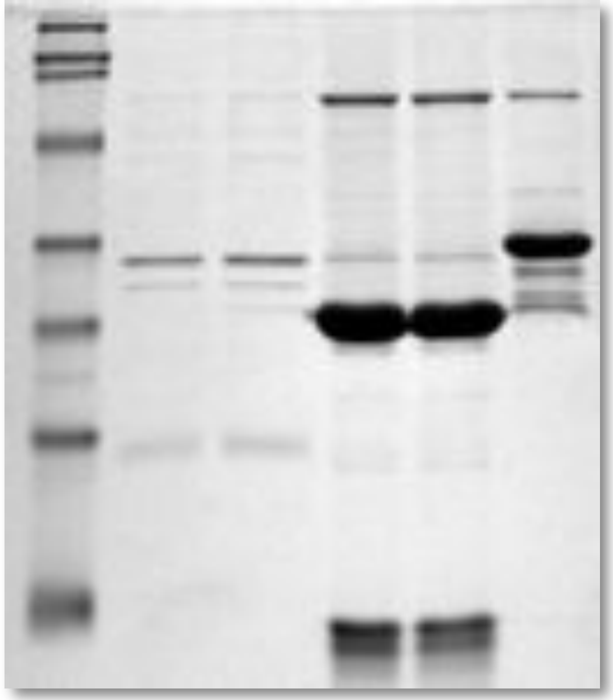


Image from Wikipedia, marker on the left lane

Protein Charge?

$$Protein\ charge\ at\ pH=7$$
$$\cong \sum Lys + \sum Arg - \sum Asp - \sum Glu + \sum co-factors$$

Main source for charge differences: pH-dependent protonation states

<=> Equilibrium between

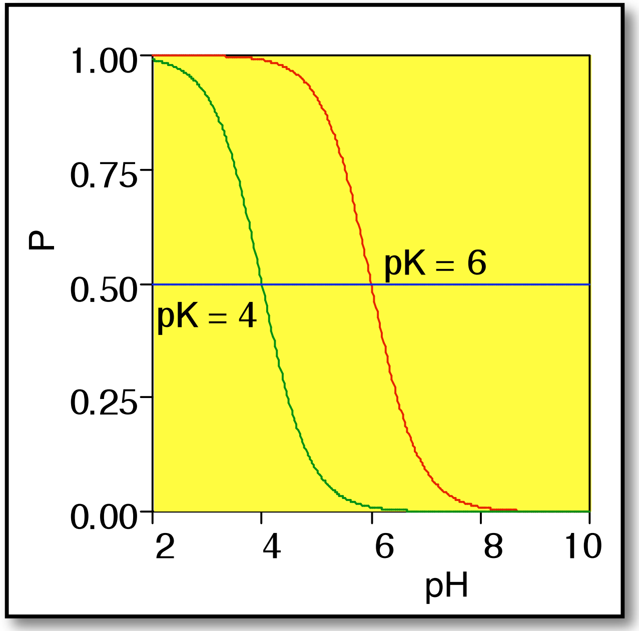
- density (pH) dependent H^+-binding and
- density independent H^+-dissociation

Probability to have a proton:

$$P = \frac{1}{1 + 10^{pH - pK}}$$

pKa = pH value for 50% protonation

Asp 3.7–4.0 … His 6.7–7.1 … Lys 9.3-9.5



Each H^+ has a +1e charge

→ **Isoelectric point**: pH at which the protein is **uncharged**

→ protonation state cancels permanent charges

2D Gel Electrophoresis

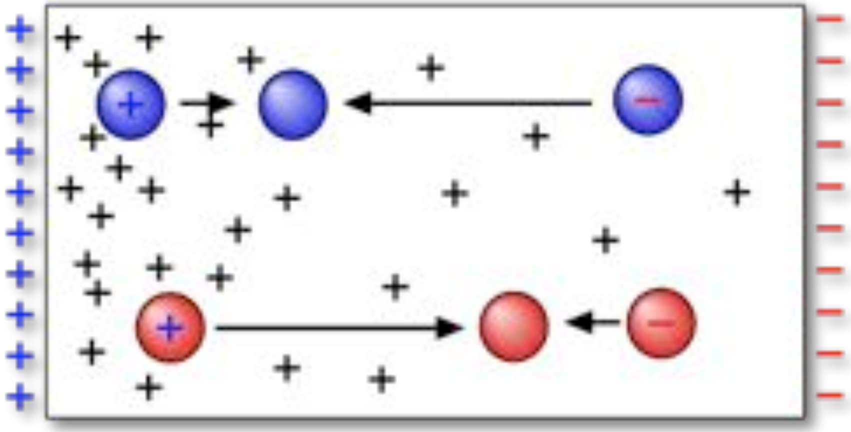
Two steps: i) separation **by isoelectric** point via pH-gradient
ii) separation **by mass** with SDS-PAGE

Step 1:

low pH → high pH

protonated
=> pos. charge

unprotonated
=> neg. charge

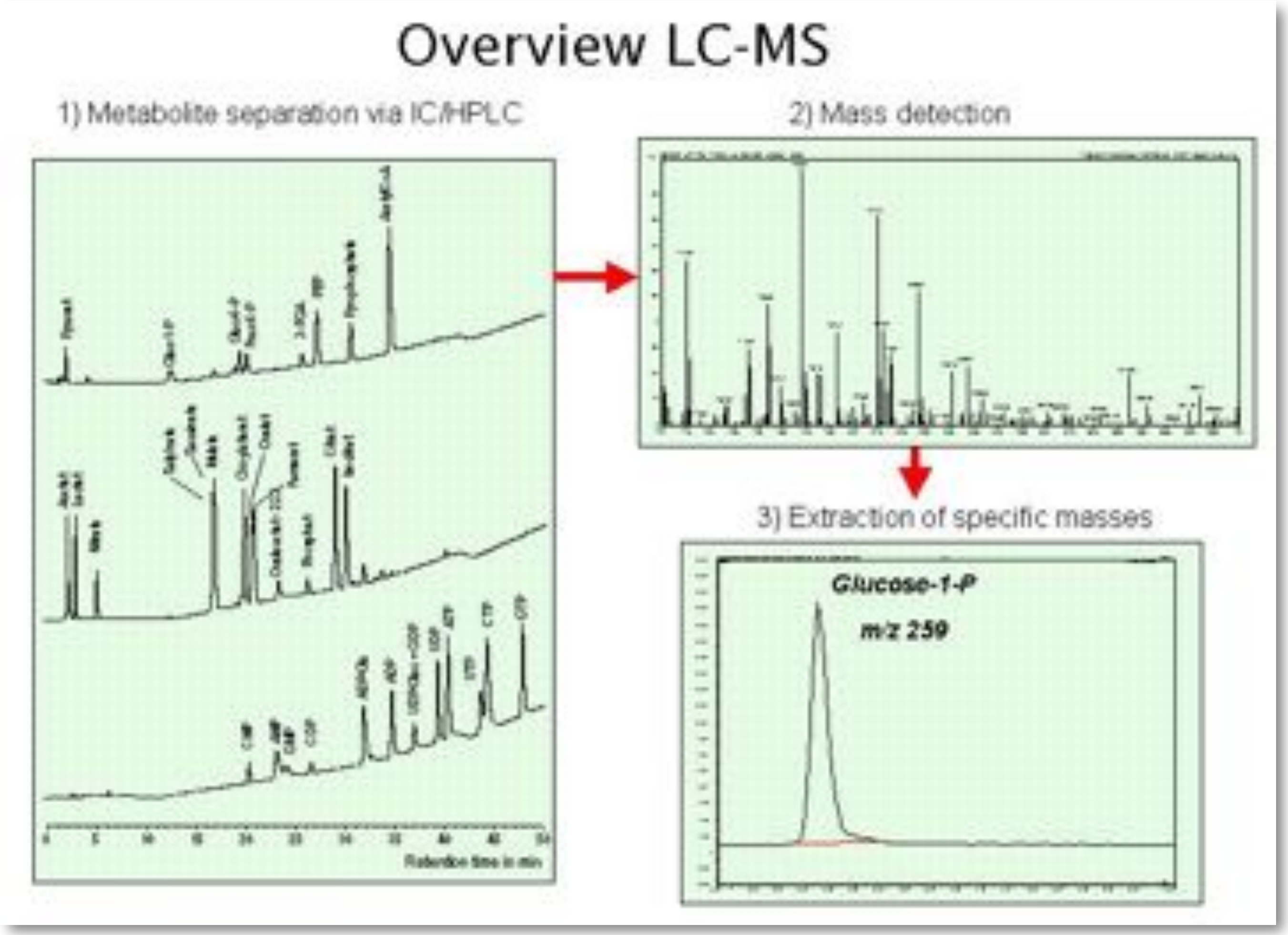


Step 2: SDS-Page

→ Most proteins differ in mass and isoelectric point (pI)

Mass Spectrometry

Identify constituents of a (fragmented) complex via their mass patterns, detect by pattern recognition with machine learning techniques.

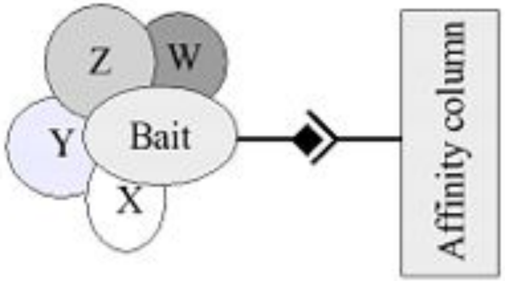


http://gene-exp.ipk-gatersleben.de/body_methods.html

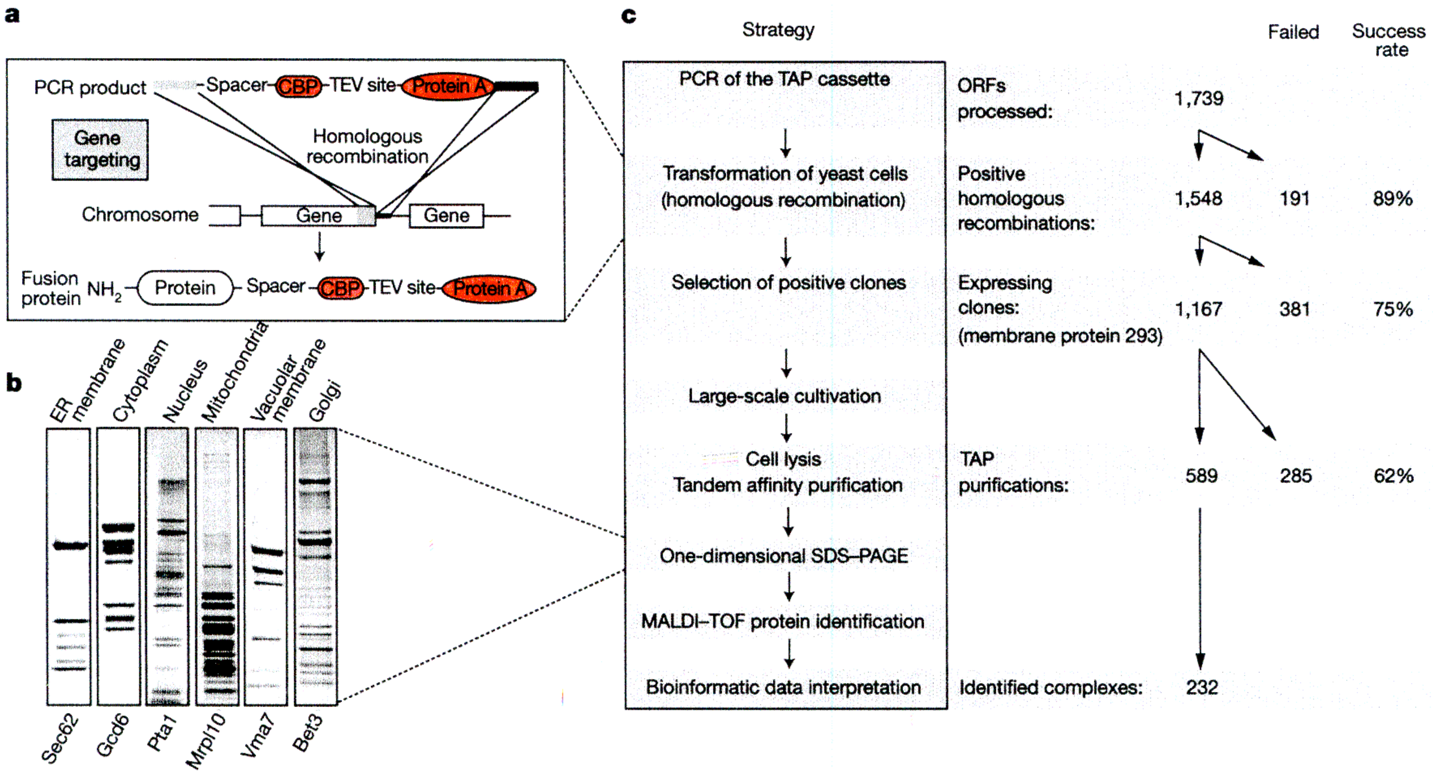
Tandem affinity purification

Yeast 2-Hybrid-method can only identify binary complexes.

In affinity purification, a protein of interest (bait) is tagged with a molecular label (dark route in the middle of the figure) to allow easy purification. The tagged protein is then co-purified together with its interacting partners (W–Z). This strategy can also be applied on a genome scale.



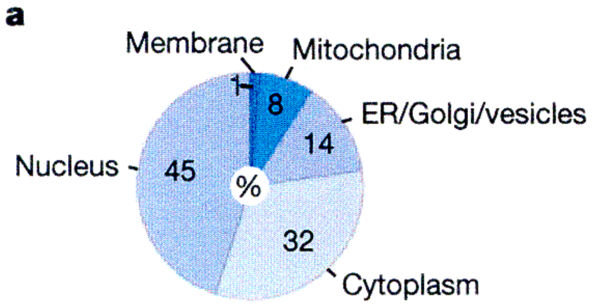
Identify proteins by mass spectrometry (MALDI-TOF).



Gavin *et al. Nature* 415, 141 (2002)

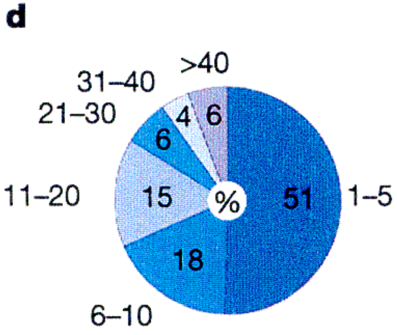
TAP analysis of yeast PP complexes

Identify proteins by scanning yeast protein database for protein composed of fragments of suitable mass.

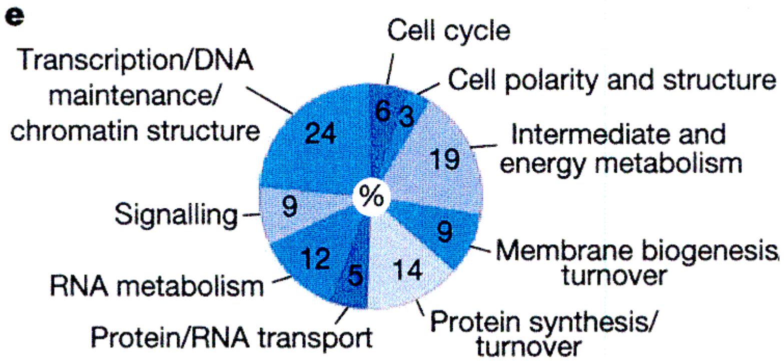


Subcellular localization of identified proteins

Here, the identified proteins are listed according to their localization (a).
(b) lists the number of proteins per complex.



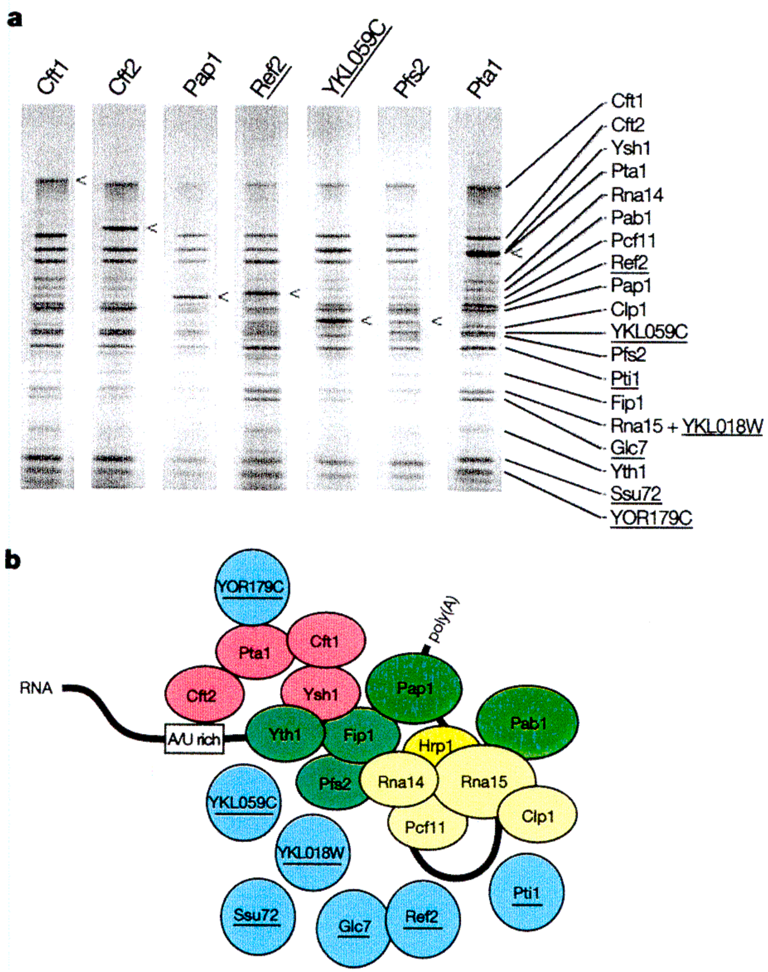
Number of proteins per complex



Distribution of complexes according to function

Gavin *et al. Nature* 415, 141 (2002)

Validation of TAP methodology



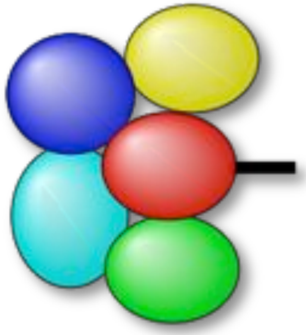
Check of the method:
can the same complex be obtained for different choices of attachment point (tag protein attached to different coponents of complex)?

Yes, more or less (see gel in (a)).

Gavin *et al. Nature* 415, 141 (2002)

Pros and Cons

Advantages:



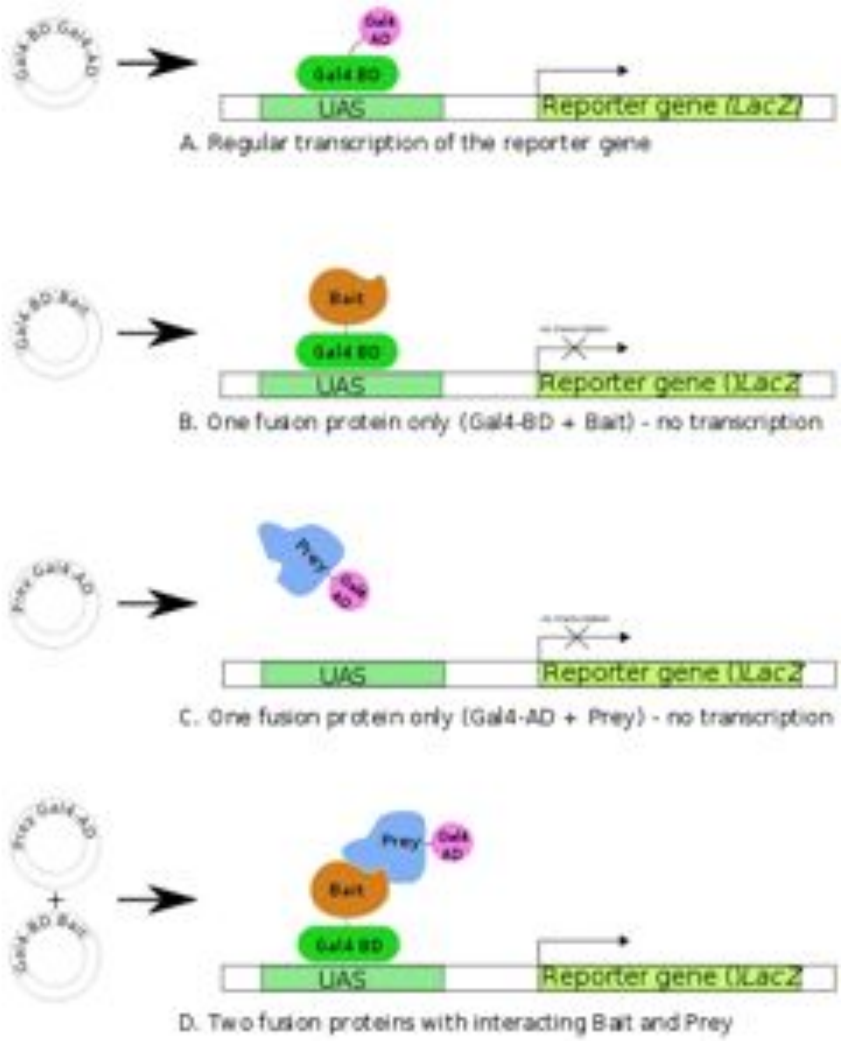
- **quantitative** determination of complex partners ***in vivo*** without prior knowledge
- simple, high yield, high throughput

Difficulties:

- tag may **prevent** binding of the interaction partners
- tag may change (relative) **expression** levels
- tag may be **buried** between interaction partners
  → no binding to beads

Yeast Two-Hybrid Screening

Discover binary protein-protein interactions via physical interaction



A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey

complex of
binding domain (BD) +
activator domain (AD)

fuse bait to BD,
prey to AD
→ expression only when
bait:prey-complex

Performance of Y2H

Advantages:

- *in vivo* test for interactions
- cheap + robust → large scale tests

Problems:

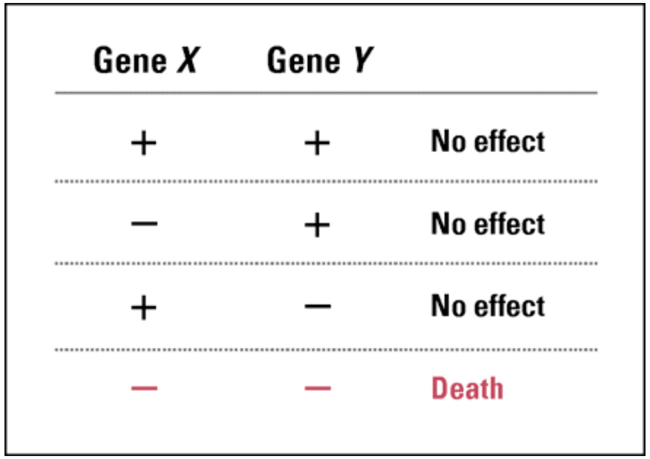
- investigate the interaction between
  (i) overexpressed
  (ii) fusion proteins in the
  (iii) yeast
  (iv) nucleus
- spurious interactions via third protein

→ many false positives
(up to 50% errors)

Synthetic Lethality

Apply two mutations that are viable on their own, but lethal when combined.

In cancer therapy, this effect implies that inhibiting one of these genes in a context where the other is defective should be selectively lethal to the tumor cells but not toxic to the normal cells, potentially leading to a large therapeutic window.



Gene *X*	Gene *Y*	
+	+	No effect
−	+	No effect
+	−	No effect
−	−	Death

http://jco.ascopubs.org/

Synthetic lethality may point to:

- physical interaction (building blocks of a complex)
- both proteins belong to the same pathway
- both proteins have the same function (redundancy)

Gene Coexpression

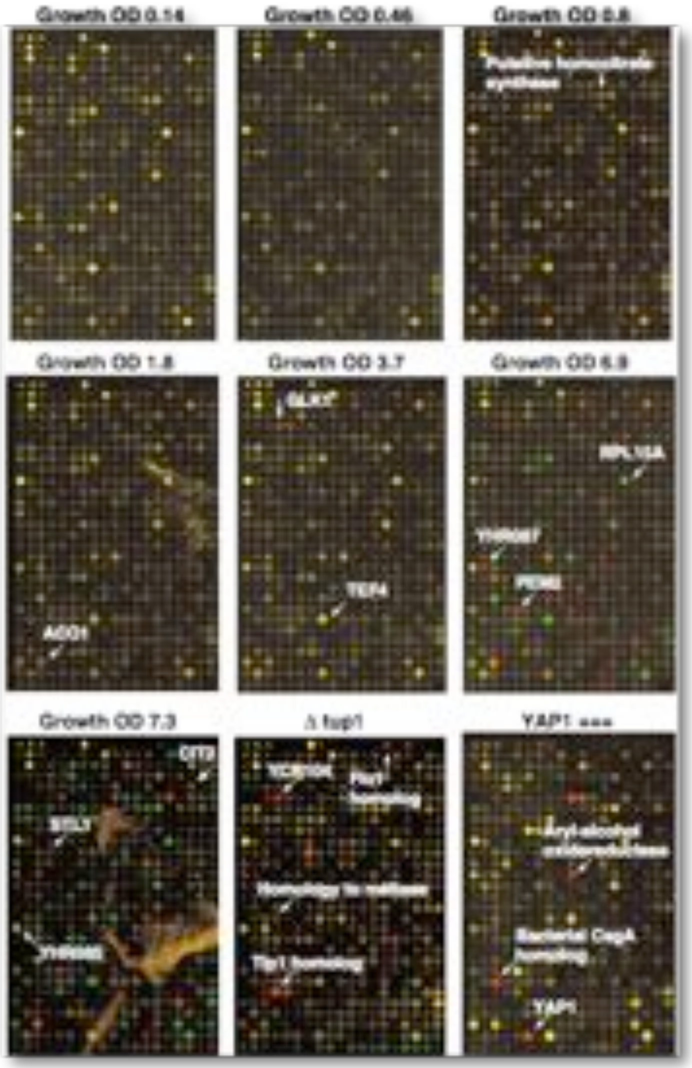
All constituents of a complex should be present at the same point in the cell cycle
→ test for correlated expression

No direct indication for complexes
(too many co-regulated genes),
but useful "filter"-criterion

Standard tool: DNA micro arrays

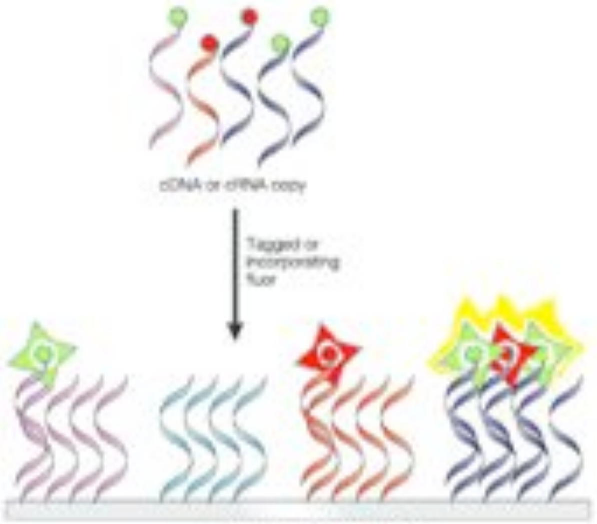
DeRisi, Iyer, Brown, *Science* **278** (1997) 680:

Diauxic shift from fermentation to respiration in *S. cerevisiae*
→ Identify groups of genes with
similar expression profiles

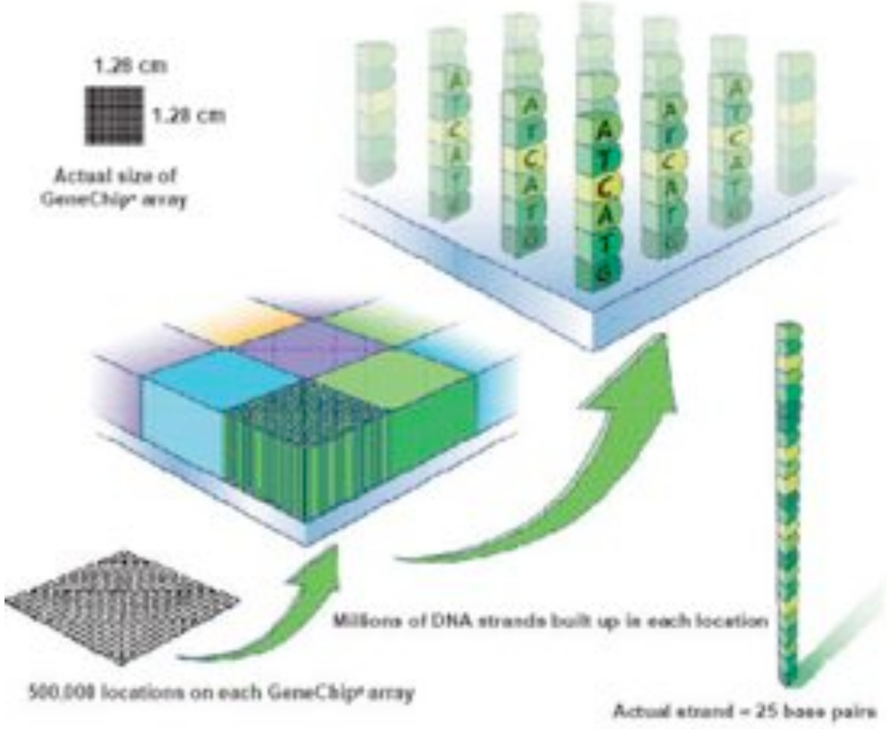


DNA Microarrays

Fluorescence labeled DNA (cDNA) applied to micro arrays
→ hybridization with complementary library strand
→ fluorescence indicates relative cDNA amounts


cDNA or cRNA copy
Tagged or incorporating fluor
cDNA spotted on glass slide or oligonucleotides built on slide

changed from:
A. Butte, Nature Reviews Drug Discovery 1, 951-960, 2002


1.28 cm
1.28 cm
Actual size of GeneChip® array
A
T
C
A
T
G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

http://intmedweb.wfubmc.edu

two labels (red + green) for experiment and control
Usually: red = signal
green = control
→ yellow = "no change"

Diauxic Shift

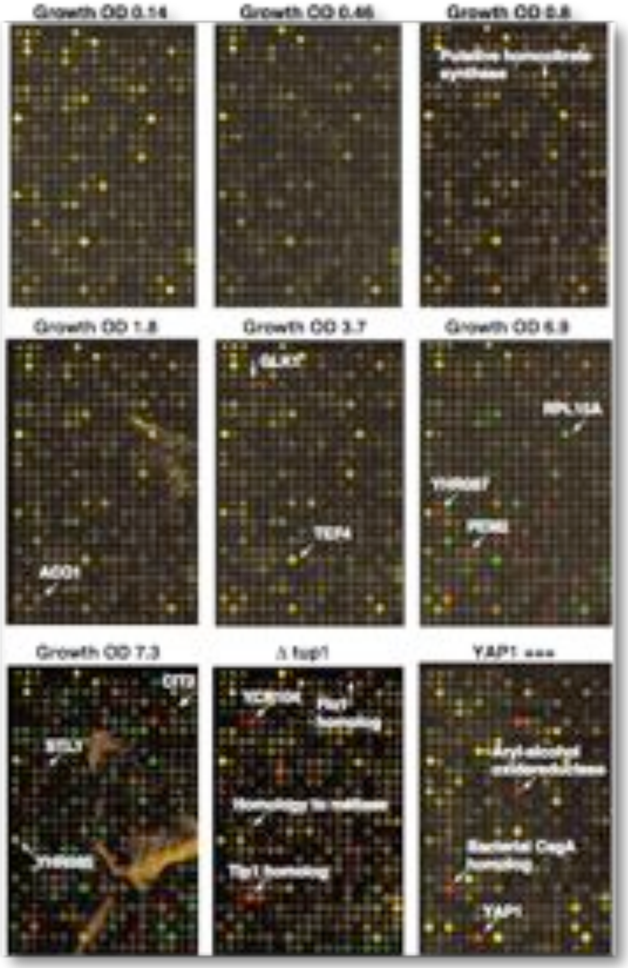
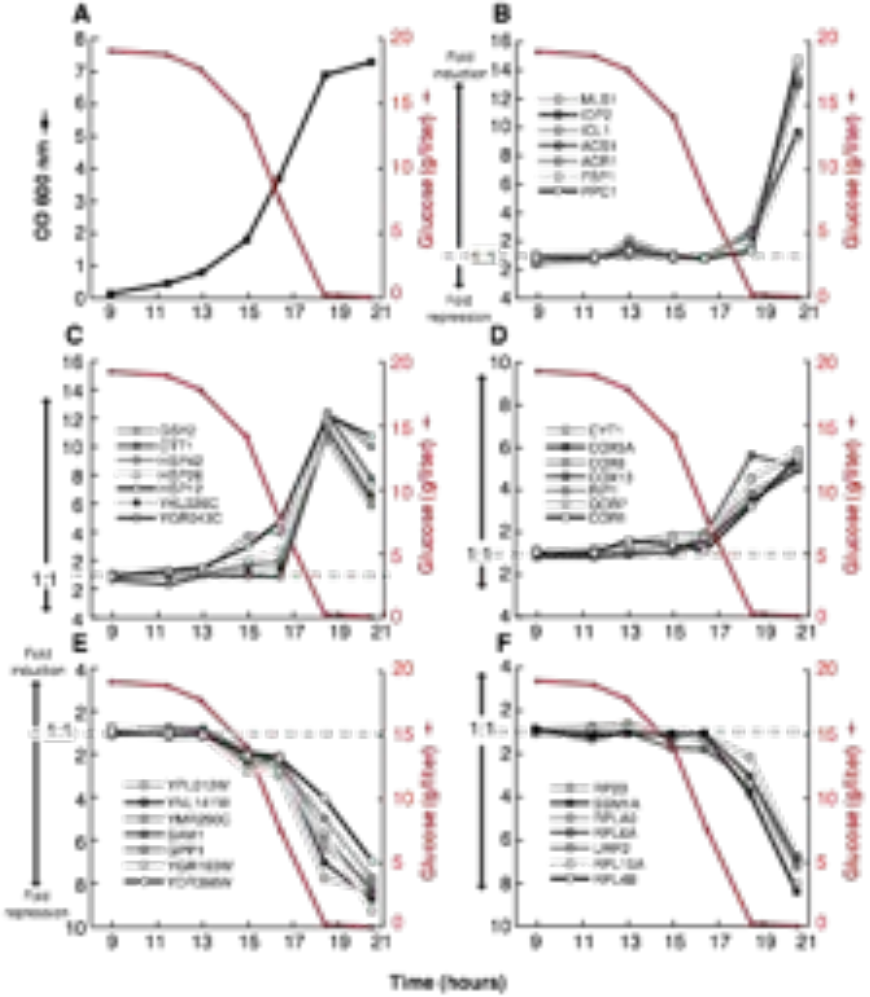


image analysis + clustering



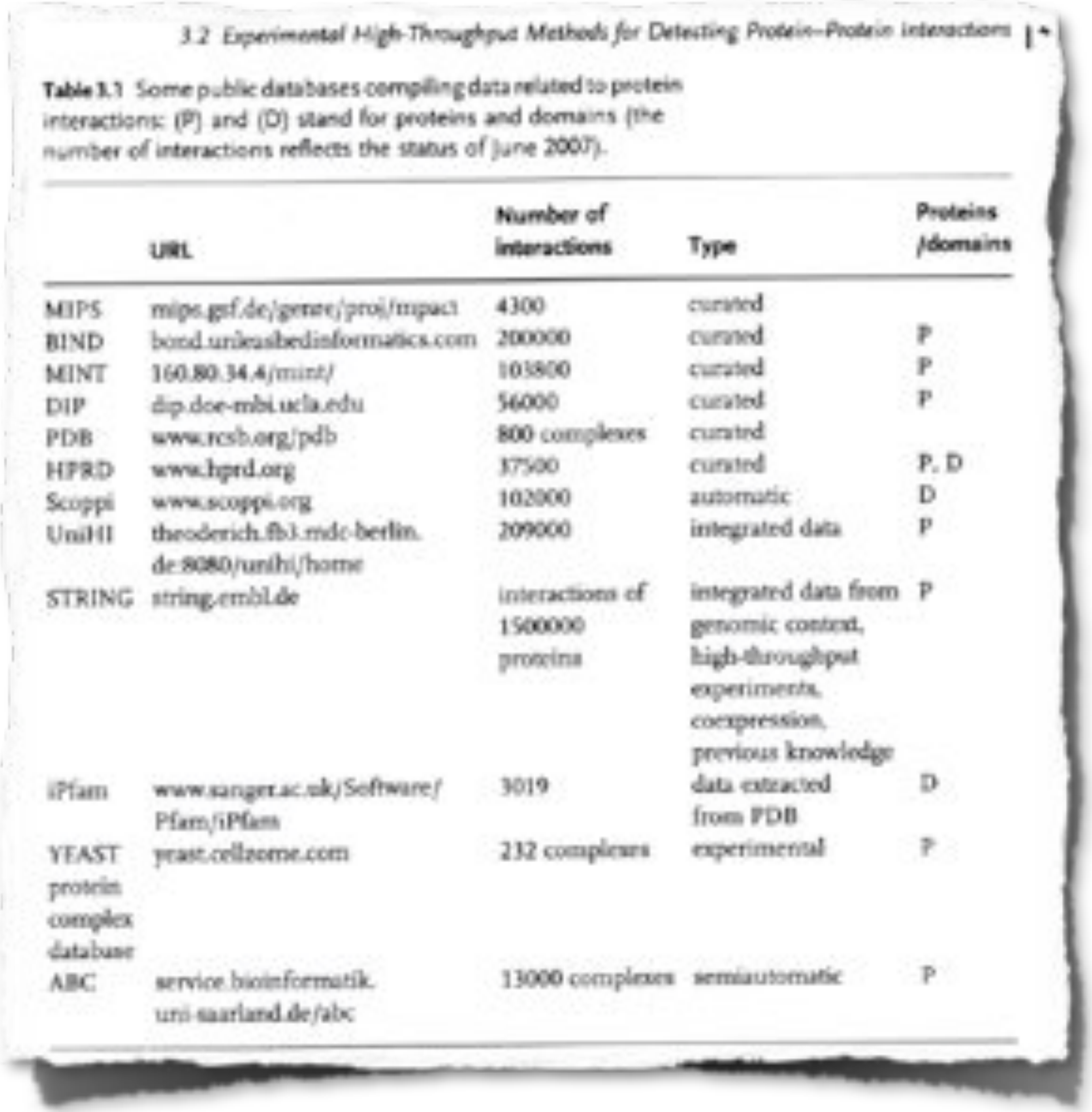
Identify groups of genes with similar time courses = expression profiles

→ **"cause or correlation"?** — biological significance?

DeRisi, Iyer, Brown, *Science* **278** (1997) 680

Interaction Databases

Bioinformatics: make use of existing databases



3.2 Experimental High-Throughput Methods for Detecting Protein–Protein Interactions

Table 3.1 Some public databases compiling data related to protein interactions: (P) and (D) stand for proteins and domains (the number of interactions reflects the status of June 2007).

	URL	Number of interactions	Type	Proteins /domains
MIPS	mips.gsf.de/genre/proj/mpact	4300	curated	
BIND	bond.unleashedinformatics.com	200000	curated	P
MINT	160.80.34.4/mint/	103800	curated	P
DIP	dip.doe-mbi.ucla.edu	56000	curated	P
PDB	www.rcsb.org/pdb	800 complexes	curated	
HPRD	www.hprd.org	37500	curated	P, D
Scoppi	www.scoppi.org	102000	automatic	D
UniHI	theoderich.fb3.mdc-berlin.de:8080/unihi/home	209000	integrated data	P
STRING	string.embl.de	interactions of 1500000 proteins	integrated data from genomic context, high-throughput experiments, coexpression, previous knowledge	P
iPfam	www.sanger.ac.uk/Software/Pfam/iPfam	3019	data extracted from PDB	D
YEAST protein complex database	yeast.cellzome.com	232 complexes	experimental	P
ABC	service.bioinformatik.uni-saarland.de/abc	13000 complexes	semiautomatic	P

(low) Overlap of Results

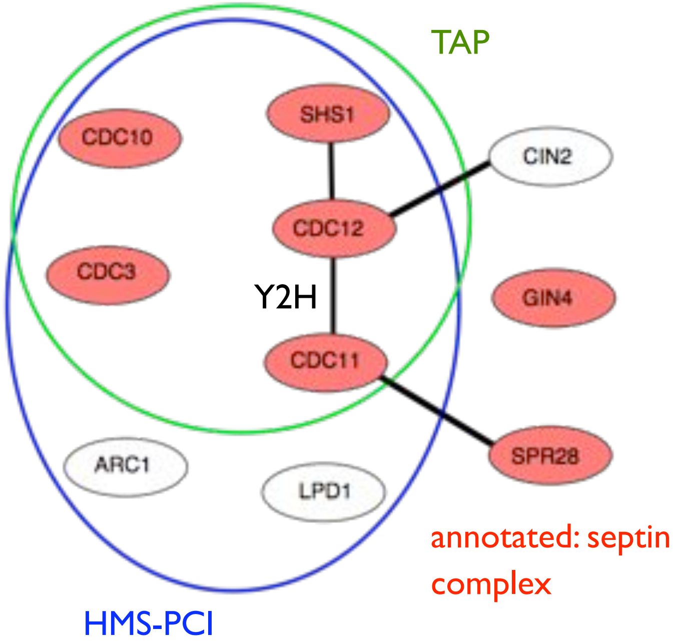
For **yeast**: ~ 6000 proteins => ~18 million potential interactions
rough estimates: ≤ 100000 interactions occur

→ 1 true positive for 200 potential candidates = **0.5%**

→ **decisive** experiment must have **accuracy** << 0.5% false positives

Different experiments detect different interactions

For yeast: 80000 interactions known,
2400 found in > 1 experiment



Problems with experiments:

i) incomplete coverage

ii) (many) false positives

iii) selective to type of interaction and/or compartment

see: von Mering (2002)

Criteria for Reliability

Guiding principles (incomplete list!):

1) **mRNA abundance**:

most experimental techniques are biased towards high-abundance proteins

2) **compartments**:

- most methods have their "preferred compartment"
- proteins from same compartment => more reliable

3) **co-functionality**

complexes have a functional reason (assumption!?)

In-Silico Prediction Methods

Sequence-based:

- gene clustering
- gene neighborhood
- Rosetta stone
- phylogenetic profiling
- coevolution

"Work on the parts list"

→ fast

→ unspecific

→ high-throughput methods for pre-sorting

Structure-based:

- interface propensities
- spatial simulations

"Work on the parts"

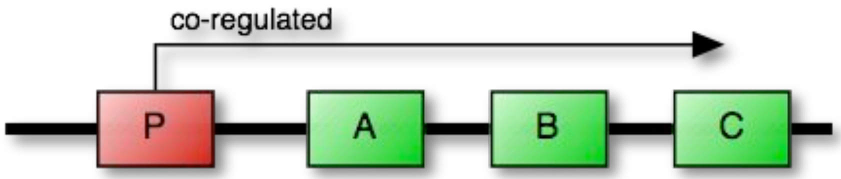
→ specific, detailed

→ expensive

→ accurate

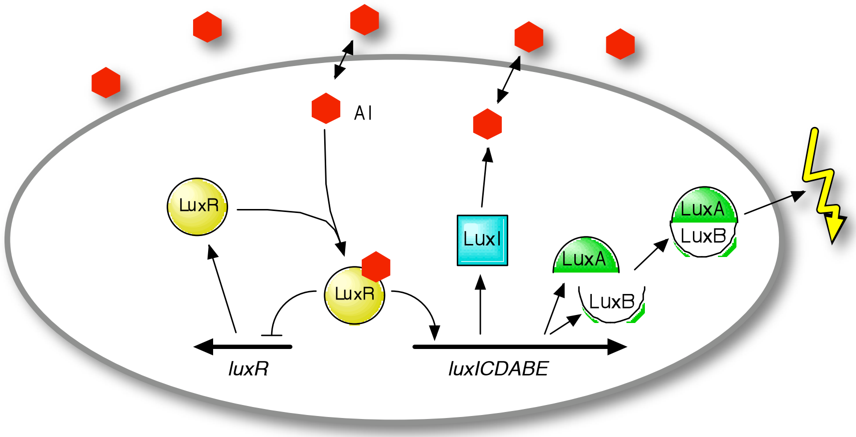
Gene Clustering

Idea: functionally **related** proteins or parts of a complex
are expressed **simultaneously**



Search for genes with a **common promoter**
→ when activated, all are transcribed together as one operand

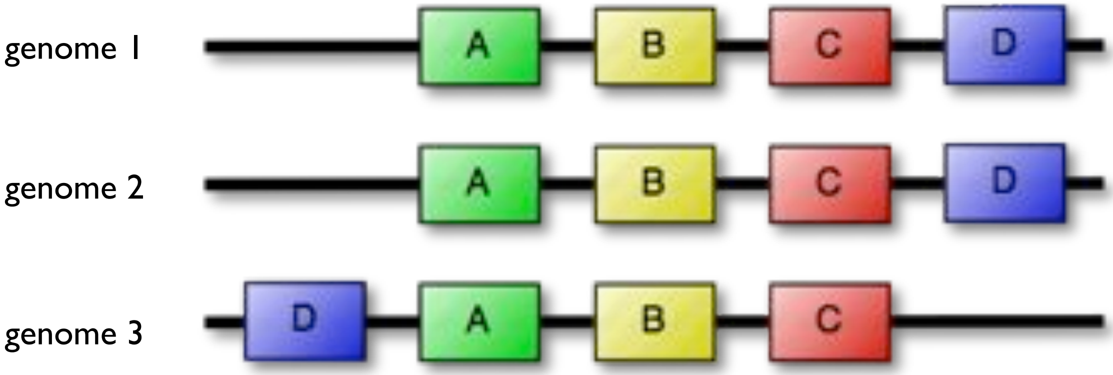
Example:
bioluminescence in *V. fischeri*,
regulated via quorum sensing
→ three proteins: I, AB, CDE



Gene Neighborhood

Hypothesis again: functionally **related** genes are expressed **together**

"functionally" = same {complex | pathway | function | …}



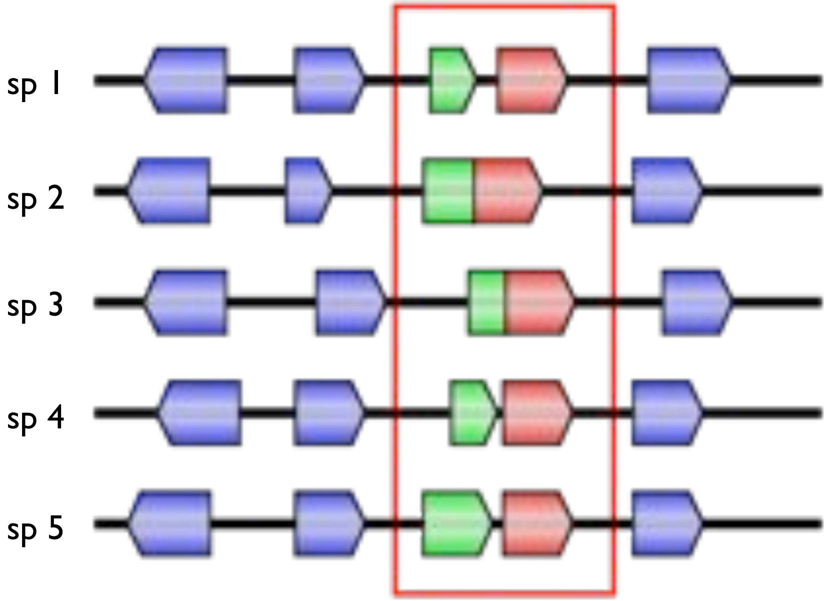
→ Search for **similar sequences** of genes in **different organisms**

(<=> Gene clustering: one species, promoters)

Rosetta Stone Method

Multi-lingual stele from 196 BC,
found by the French in 1799
→ key to deciphering hieroglyphs

Idea: same "**names**" in different genome "**texts**"

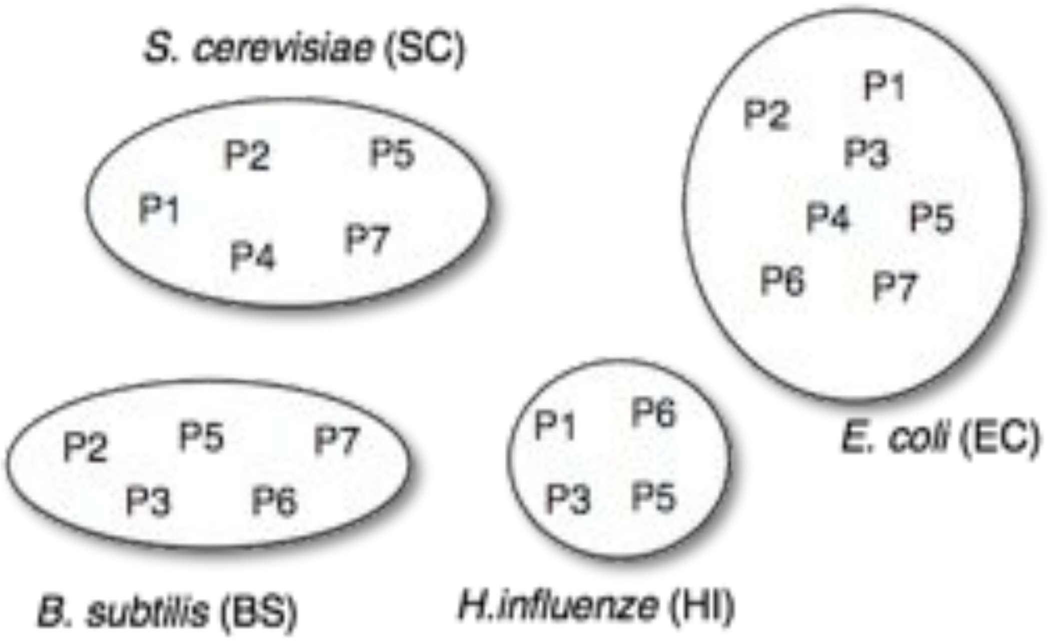


Enright, Ouzounis (2001):
40000 predicted pair-wise interactions
from search across 23 species

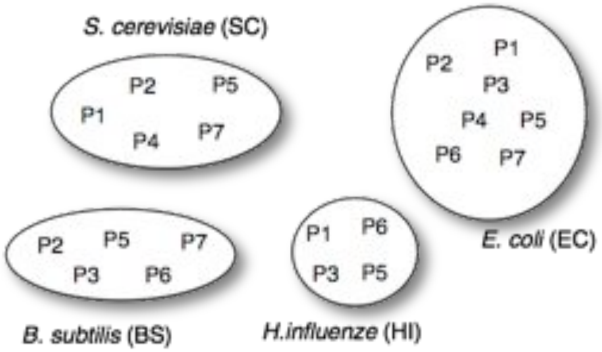
Phylogenetic Profiling

Idea: either **all** or **none** of the proteins of a complex should be **present** in an organism

→ compare presence of protein homologs across species (e.g., via sequence alignment)



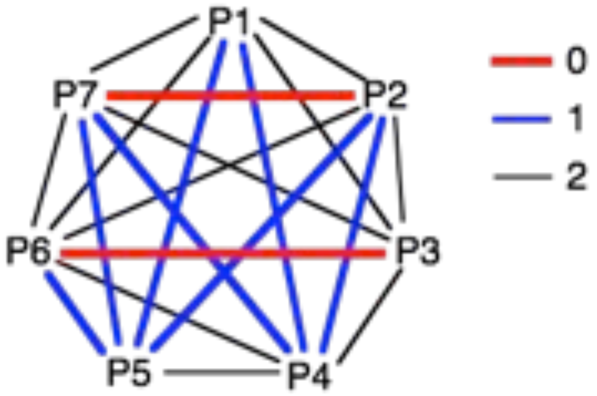
Distances



	EC	SC	BS	HI
P1	1	1	0	1
P2	1	1	1	0
P3	1	0	1	1
P4	1	1	0	0
P5	1	1	1	1
P6	1	0	1	1
P7	1	1	1	0

Hamming distance between species: number of different protein occurrences

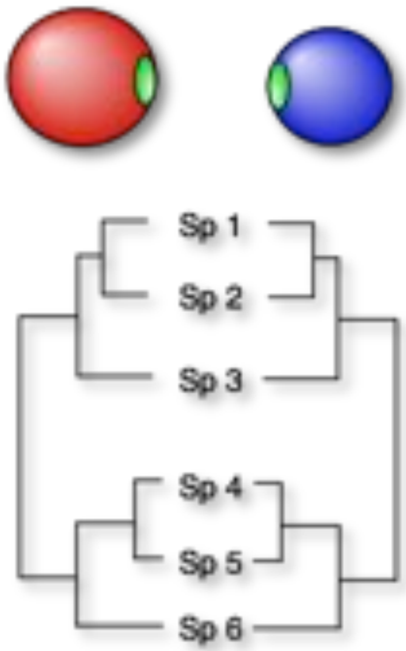
	P1	P2	P3	P4	P5	P6	P7
P1	0	2	2	1	1	2	2
P2		0	2	1	1	2	**0**
P3			0	3	1	**0**	2
P4				0	2	3	1
P5					0	1	1
P6						0	2
P7							0



Two pairs with similar occurrence: P2-P7 and P3-P6

Coevolution

Idea: not only similar static occurence, but similar **dynamic evolution**



Interfaces of complexes are often better conserved than the rest of the protein surfaces.

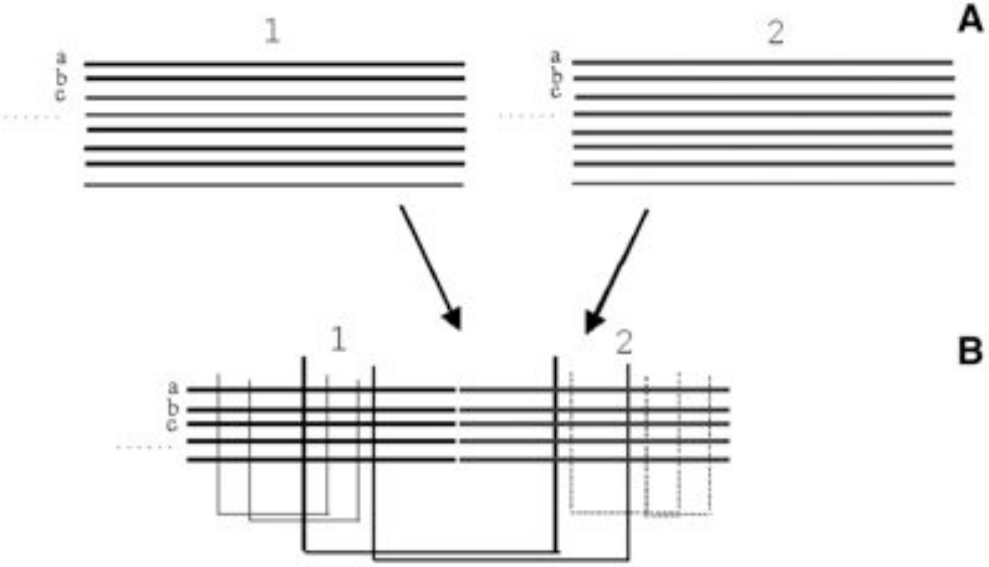
Also: look for potential substitutes
→ anti-correlated
→ missing components of pathways
→ function prediction across species
→ novel interactions

i2h method

Schematic representation of the i2h method.

A: Family alignments are collected for two different proteins, 1 and 2, including corresponding sequences from different species (a, b, c,).

B: A virtual alignment is constructed, concatenating the sequences of the probable orthologous sequences of the two proteins. Correlated mutations are calculated.



Pazos, Valencia, Proteins 47, 219 (2002)

Correlated mutations at interface

Correlated mutations evaluate the similarity in variation patterns between positions in a multiple sequence alignment.

Similarity of those variation patterns is thought to be related to compensatory mutations.

Calculate for each positions *i* and *j* in the sequence a rank correlation coefficient (r_{ij}):

$$r_{ij} = \frac{\sum_{k,l}\left(S_{ikl} - \bar{S}_i\right)\left(S_{jkl} - \bar{S}_j\right)}{\sqrt{\sum_{k,l}\left(S_{ikl} - \bar{S}_i\right)^2}\sqrt{\sum_{k,l}\left(S_{jkl} - \bar{S}_j\right)^2}}$$

where the summations run over every possible pair of proteins *k* and *l* in the multiple sequence alignment.

S_{ikl} is the ranked similarity between residue *i* in protein *k* and residue *i* in protein *l*.

S_{jkl} is the same for residue *j*.

$\bar{S}_i$ and $\bar{S}_j$ are the means of S_{ikl} and S_{jkl}.

Pazos, Valencia, Proteins 47, 219 (2002)

Summary

What you learned **today**: how to get some data on PP interactions

gene clustering

SDS-PAGE TAP

DB

gene neighborhood

MS

micro array

Y2H

Rosetta stone

synthetic lethality

phylogenic profiling

coevolution

type of interaction? — reliability? — sensitivity? — coverage? — …

Next lecture: Mon, Oct. 28, 2013

- combining weak indicators: Bayesian analysis
- identifying communities in networks