Bioinformatics 3

V 3 – Data for Building Networks

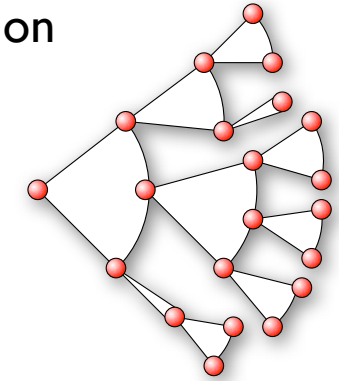
Tue, Oct 25, 2011

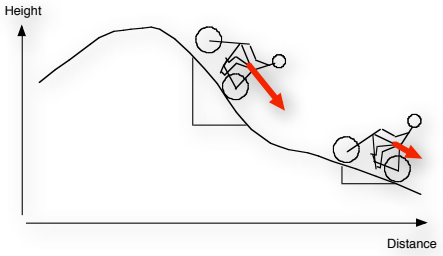
Graph Layout I

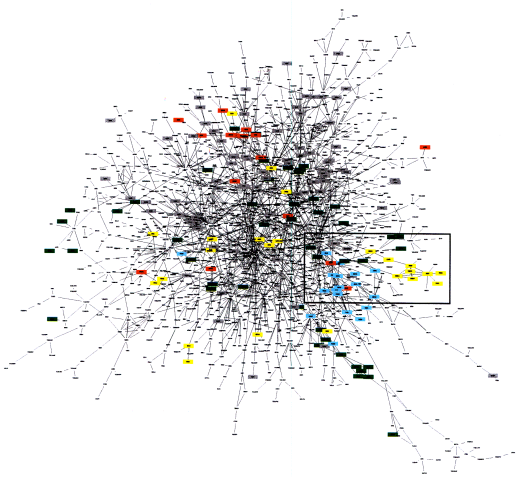
Requirements:

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

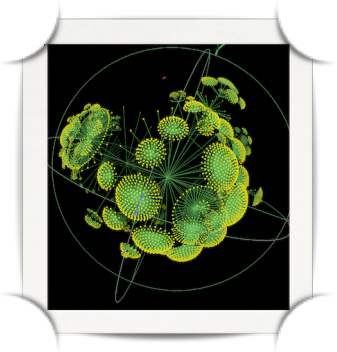
Force-directed Layout:
based on energy minimization
=> runtime
=> mapping into 2D







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 → 21×10^9 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

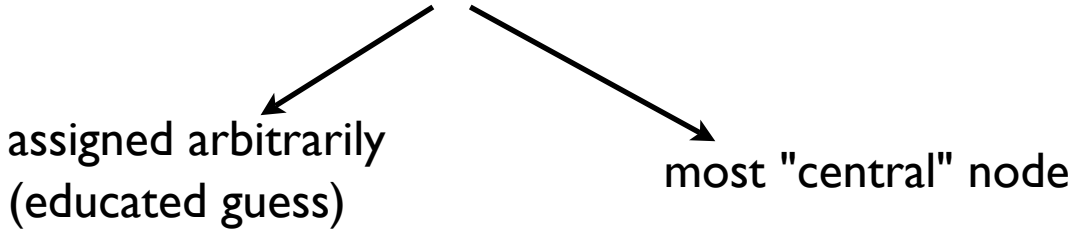
(3) integrate connected sets into one coordinate system via a **funnel process**, starting from the largest set

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



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

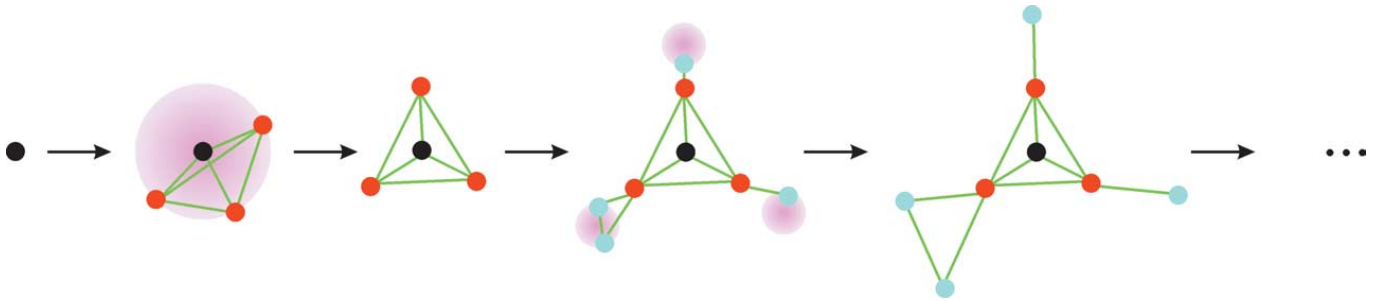
$$v_{root} = \min \left(\sum_{(v,u) \in V} d(v,u) \right)$$

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

Layout => place **root** at the **center**

Component Layout II

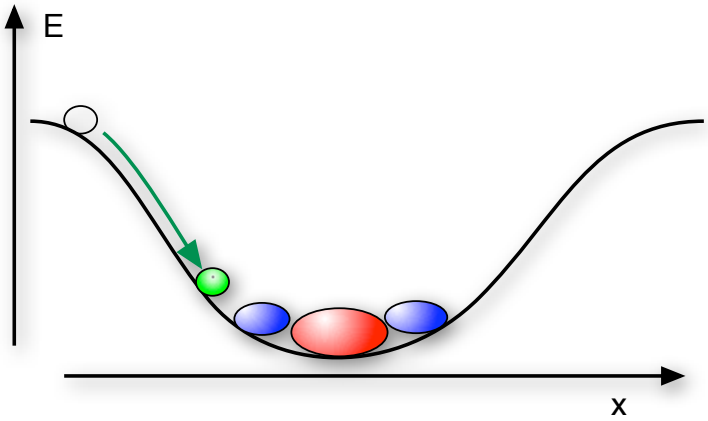
- start with root node of the MST
- place level-1 nodes on circle (sphere) around root,
  add all links,
  relax springs (+ short-range repulsion)
- place level-2 nodes on circles (sphere) outside their level-1 descendants,
  add all links,
  relax springs
- place level-3 nodes on circles (sphere) outside their level-2 descendants,
  :



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

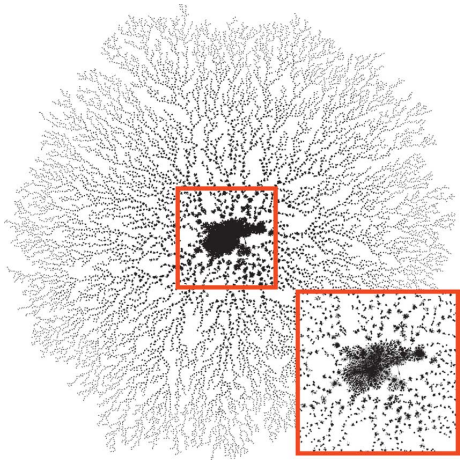
Combining the Components

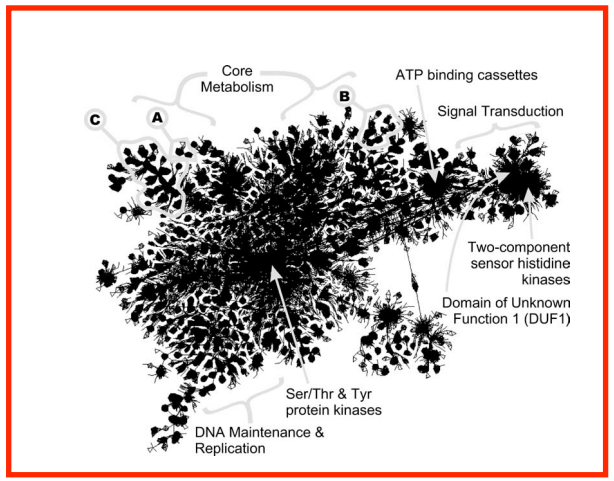
When the components are finished
=> **assemble** with energy **funnel**



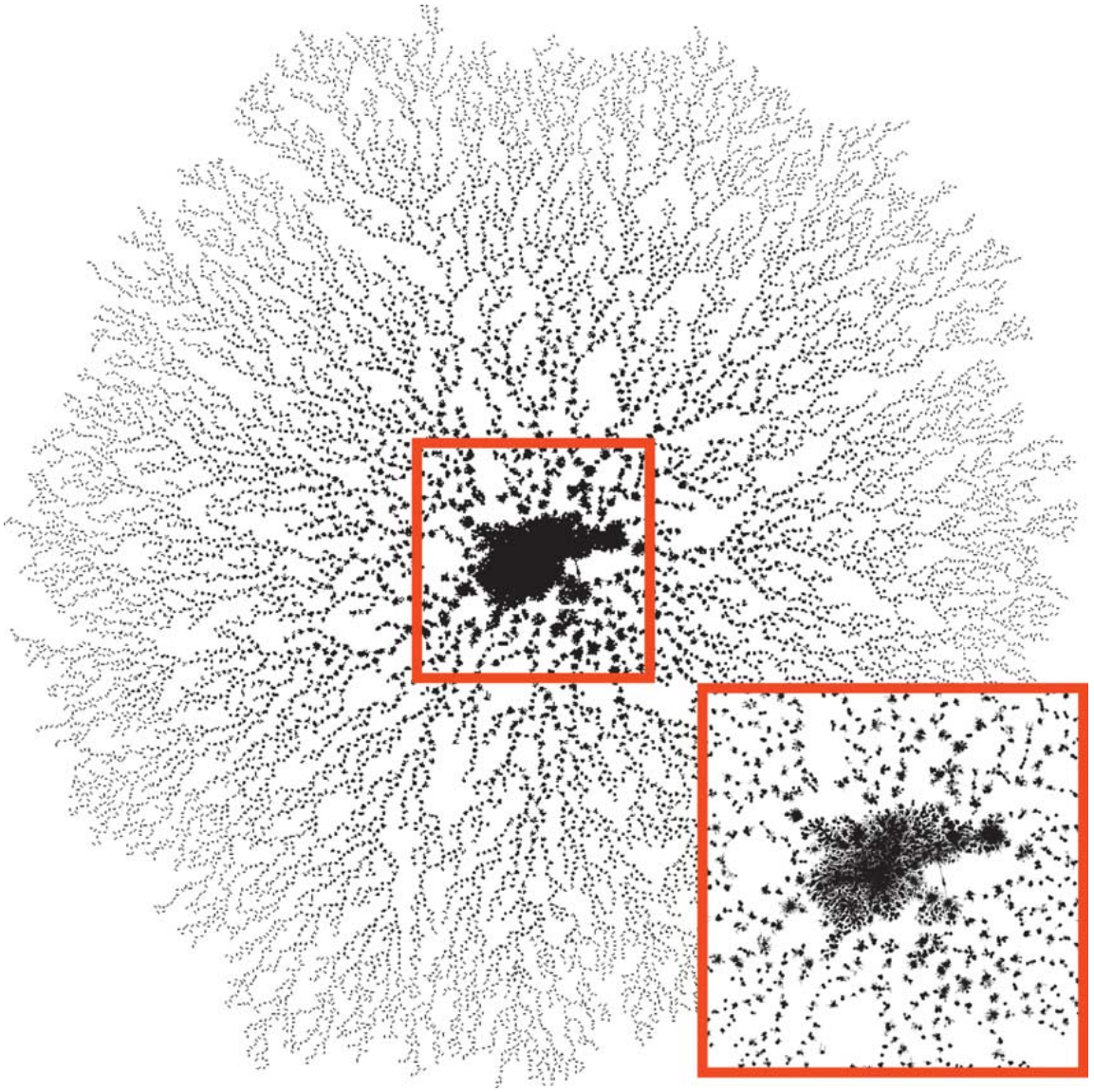
- place largest component at bottom
- place next smaller one somewhere on the rim, let it slide down
  => freeze upon contact

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



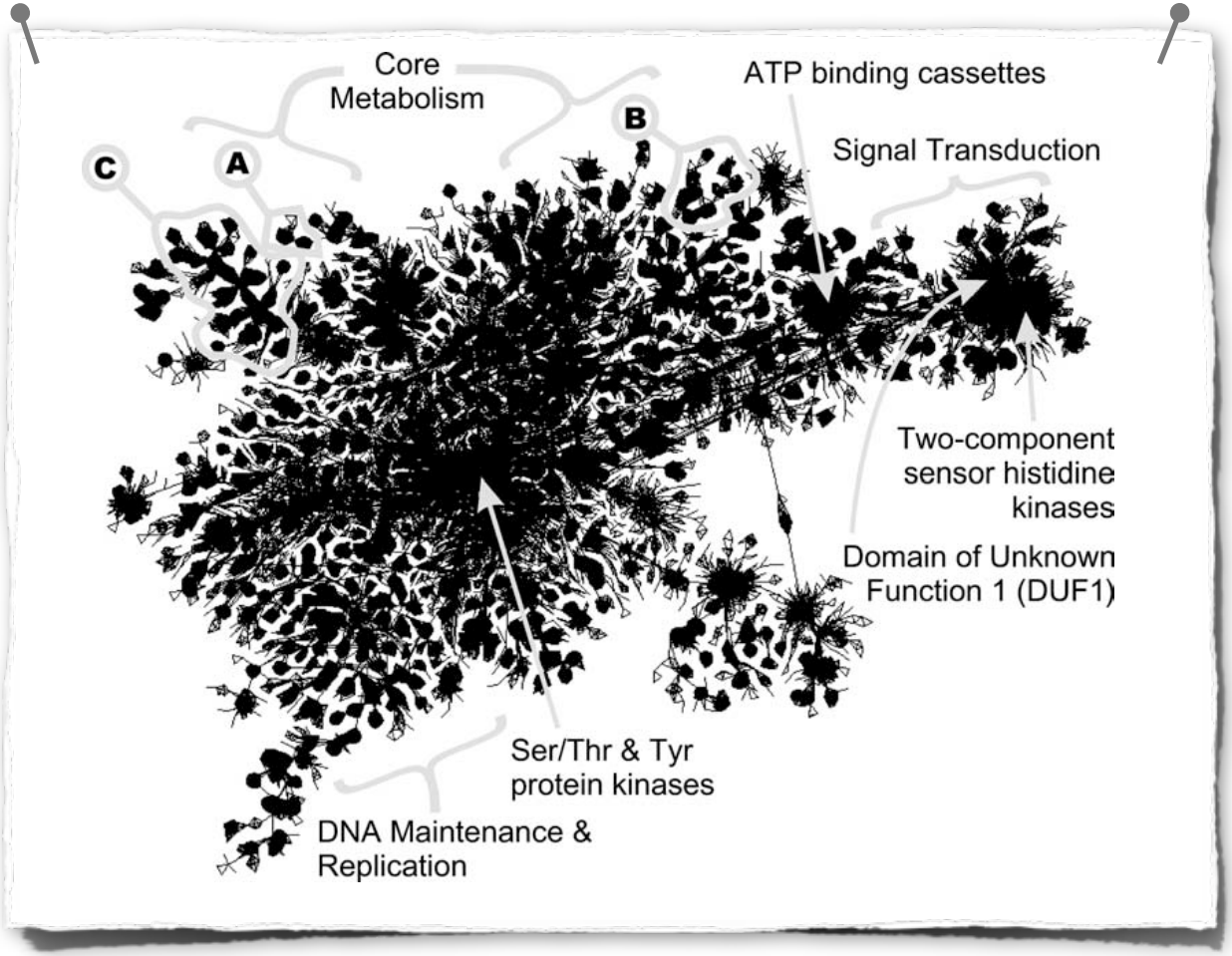


Adai et al. J. Mol. Biol. 340, 179 (2004)



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

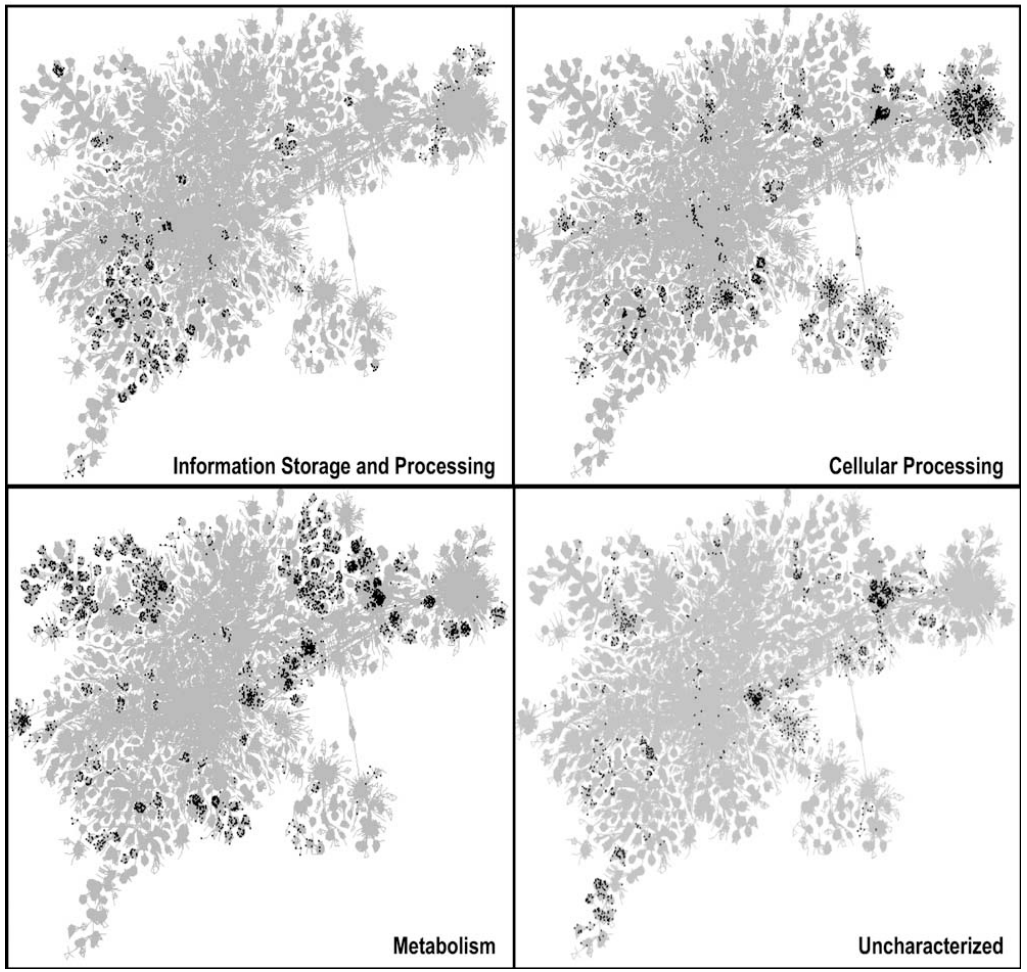
Annotations in the Largest Cluster



Related functions in the same regions of the cluster => predictions

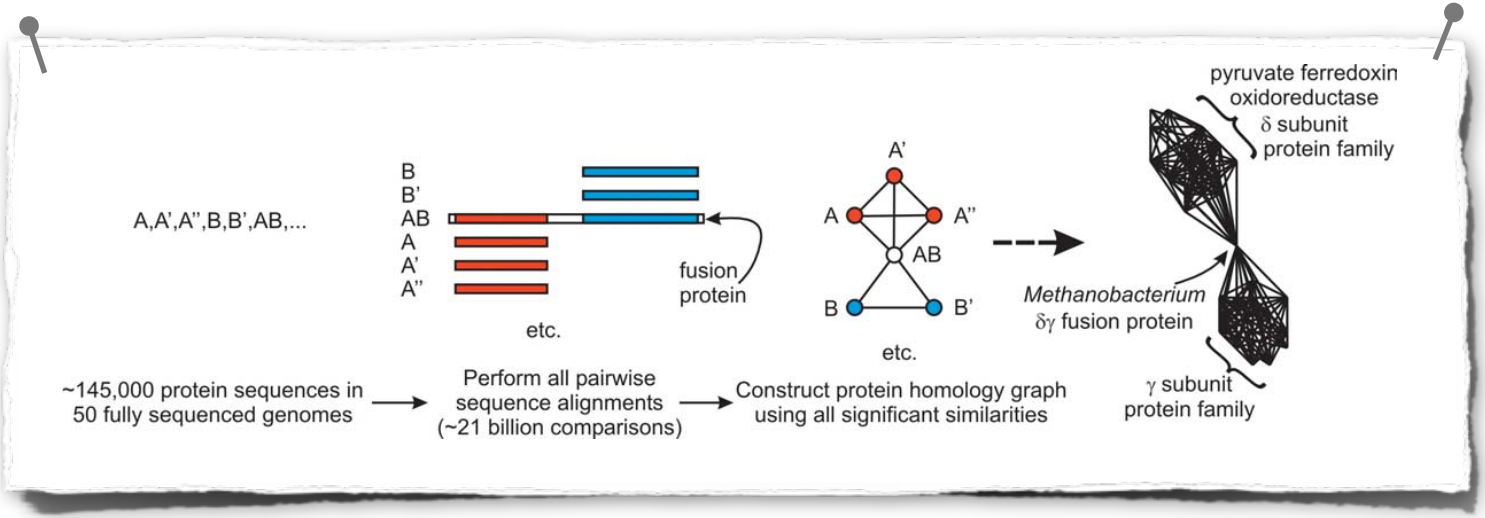
Adai et al. J. Mol. Biol. 340, 179 (2004)

Clustering of Functional Classes



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

Fusion Proteins



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)

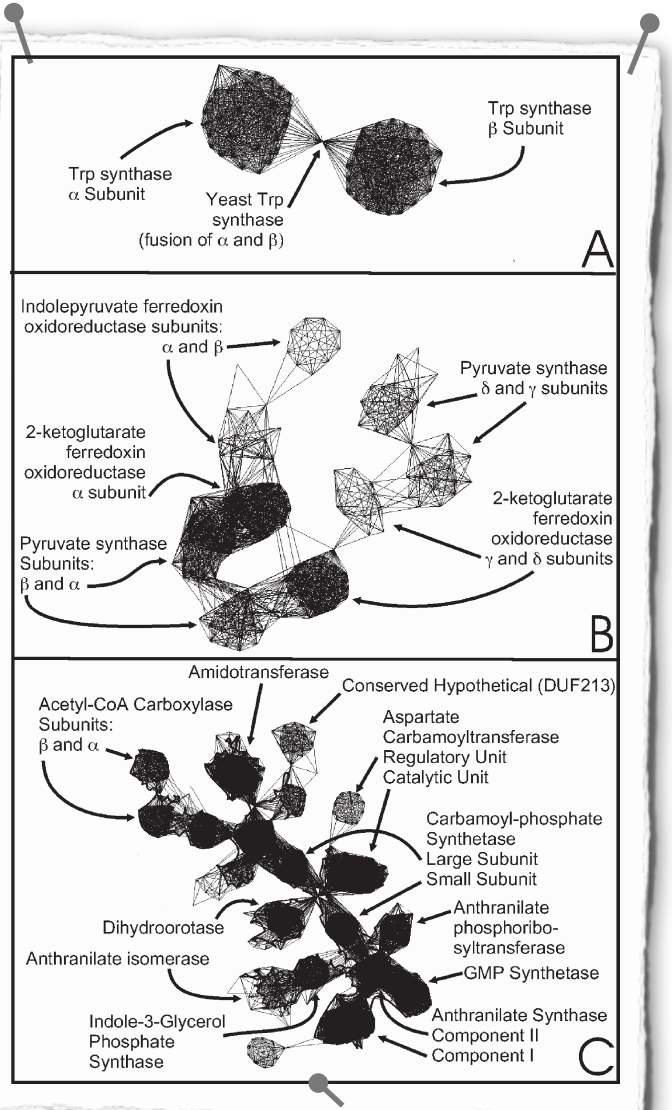
Functional Relations between Gene Families

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 αβ fusion protein,

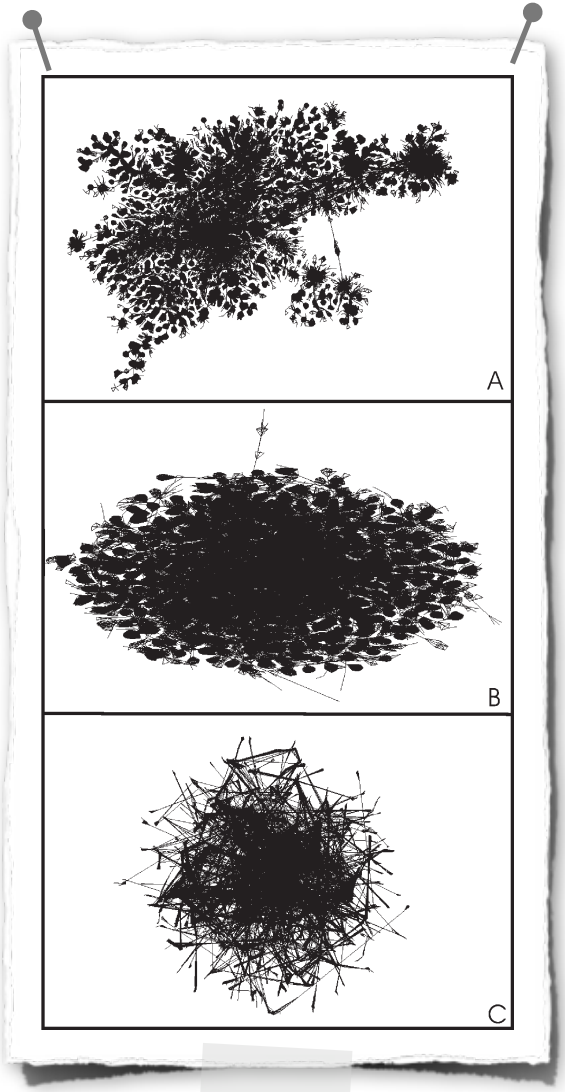
B: protein subunits of the pyruvate synthase and alpha-ketoglutarate ferredexin oxidoreductase complexes

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…



Comparison of the layouts from

A: **LGL** – **hierarchic** force-directed layout
acc. to MST
=> structure from homology

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

C: **InterViewer** – collapses similar nodes
=> reduced complexity

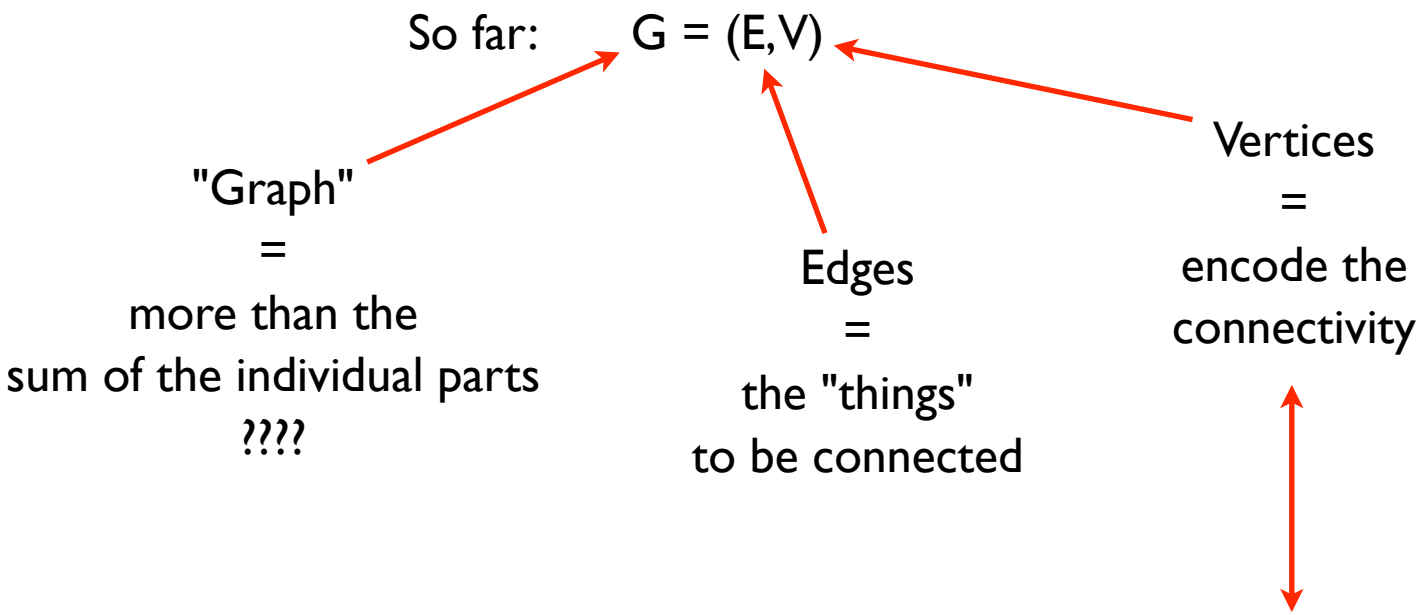
Adai et al. J. Mol. Biol. 340, 179 (2004)

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 alg. for modules

the same physical concept, different implementations!

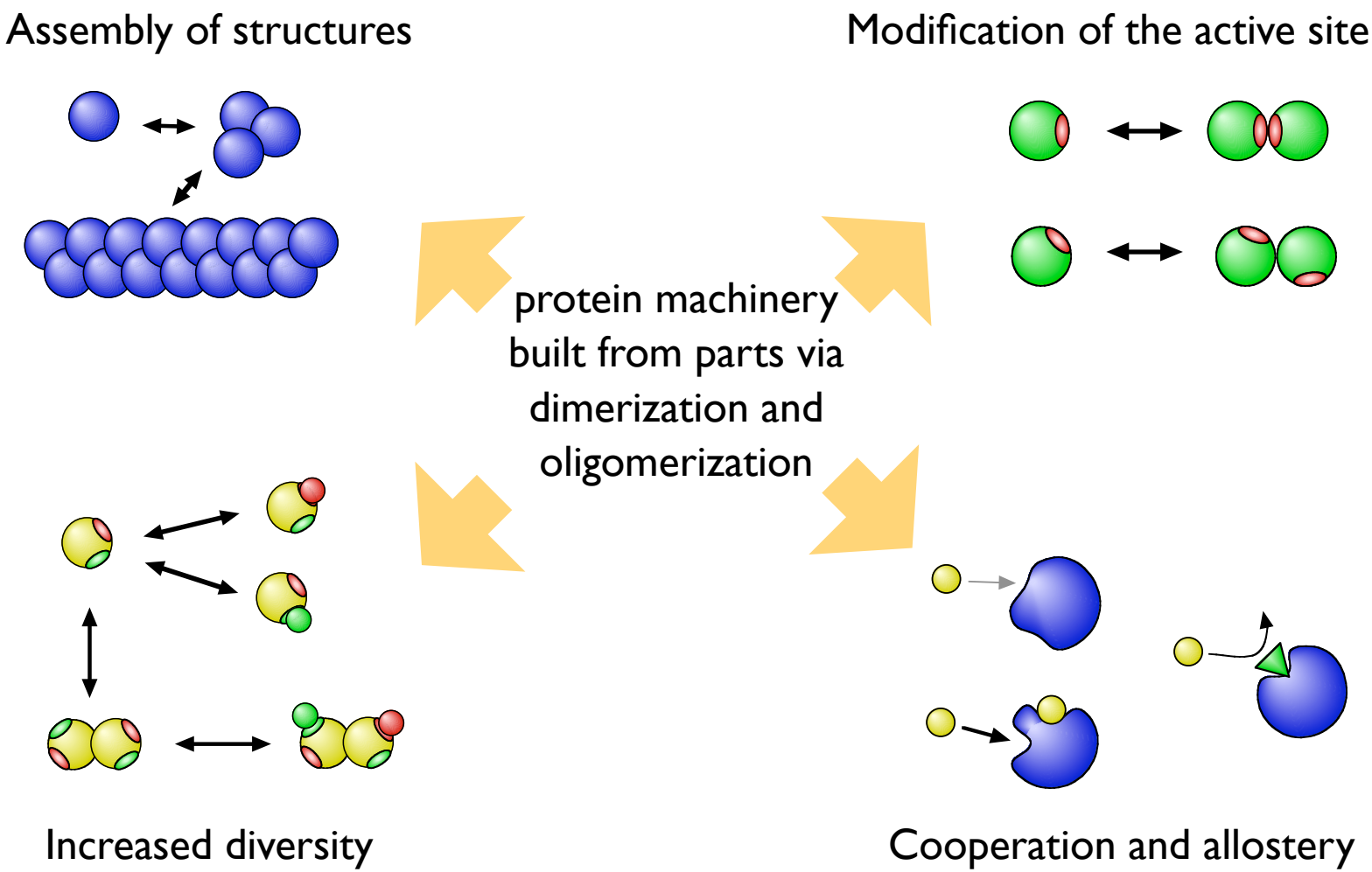
A "Network"



=> what are interesting biological "things"?
=> how are they connected?
=> are the information accessible/reliable?

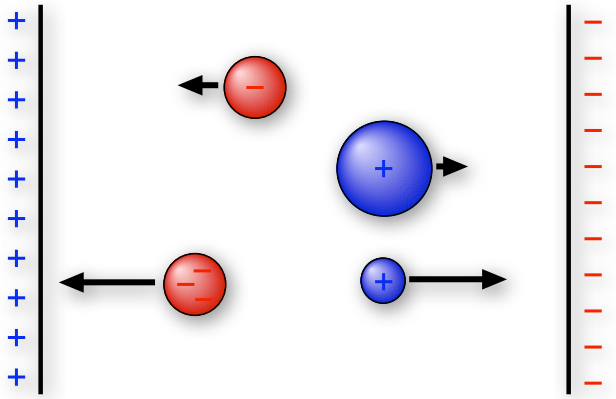
Classified by:
- degree distribution
- clustering
- connected components
- …

Protein Complexes



Gel Electrophoresis

Electrophoresis: directed diffusion of charged particles in an electric field



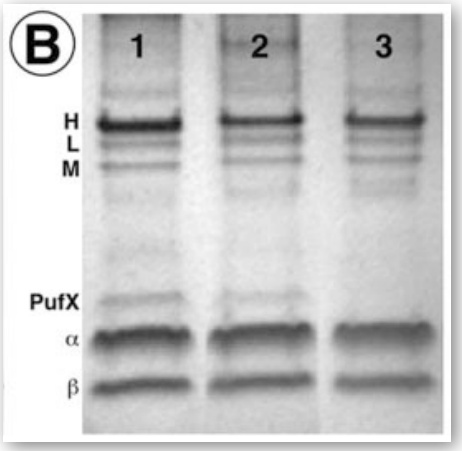
faster

more charge, smaller

less charge, larger

slower

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)

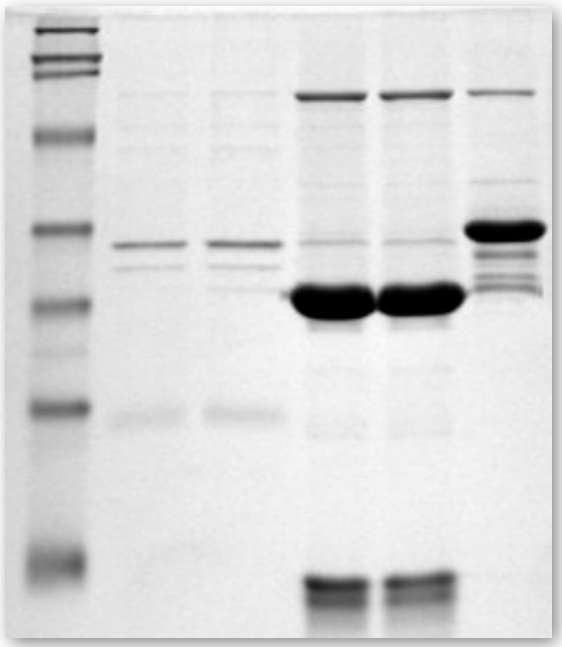


Image from Wikipedia, marker on the left lane

Which Markers???

SDS-Page standard markers

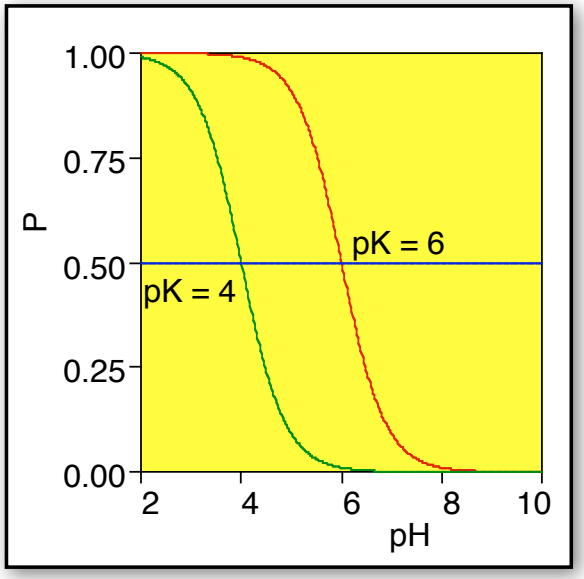
DaJo Posted 12/18/2004 12:38:39 AM	What are the best protein color markers for sds-page and western transfer? I work with proteins ranging from 30-110 kd.
Big E Posted 1/6/2005 1:37:01 AM	**DaJo said:** What are the best protein color markers for sds-page and western transfer? I work with proteins ranging from 30-110 kd. I like the Full-Range Rainbow Molecular Weight Markers from Amersham. The Recombinant proteinsizes are 250,000 160,000 105,000 75,000 50,000 35,000 30,000 25,000 15,000 10,000
protdoc Posted 1/10/2005 8:32:47 PM	**DaJo said:** What are the best protein color markers for sds-page and western transfer? I work with proteins ranging from 30-110 kd. Invitrogen's MultiMark Multi-Colored Standard may be useful, too. It contains 9 markers, ranging from 3 to 185 kDa.
bugme Posted 1/13/2005 9:54:04 PM	I suggest non colored markers for best accuracy. Use Mark12 by invitrogen
smartee Posted 2/6/2005 8:59:03 PM	I would suggest the non-colored. You can do Ponseau staining after transfer to mark the sizes of the bands. As for the brand, I prefer the wide-range marker from Promega. It has easily disctinguishable bands ranging from 15 kDa to 225 kDa, and it worked better for me than Invitrogen's markers. The latter are just too busy and it can be hard to figure out the exact size of the marker band, especially when you're trying to differenciate betwen 70-80-90 kDa.
nin1318 Posted 4/5/2005 8:48:51 PM	i have used the amersham that big e suggested, however, the kaliedoscope precision plus from bio-rad seems to be a better alternative, both in terms of cost, and in quality and reliability. also the color will be visible even after stripping the blots, and is much brighter to begin with. i have done many regressions with bio-rad kaliedoscope markers and they are usually spot on with r2 of .98 or higher...so i don't think that the accuracy of the colors is a problem.
badcell Posted 4/7/2005 11:21:43 AM	I agree with the suggestion tu use kaleidoscope from Bio-Rad. Noncolored markers are suppossed to be more accurate, but do not allow you to follow the course of the electrophoresis, and this can be important, as on occassions the front dye may wash out but the proteins get delayed from some reason, and you only realize after transferring or irreversibly staining the gel. I too have done regressions with the Bio-Rad markers and the accuracy seems to be better than for other brands.
pw_18 Posted 5/5/2005 10:55:19 AM	I've learned to like the Invitrogen Magic-Mark standards, which develop when you add a secondary antibody (anti-IgG-HRP of various species). This way you get your markers directly on the film, and don't have to worry about trying to trace them from the blot. As for accuracy, I'm not convinced that it matters as long as the results are reproducible. SDS-PAGE mobility is influenced by a number of different things (glycosylation, charge distribution, etc.) besides just the base molecular weight.
samm Posted 5/5/2005 12:24:44 PM	PAGE markers are largely relative comparisions. However, I have noticed a distinct difference between prestained and non-stained markers from two companies (Amersham and Fermentas) run side by side, which is not explained bythe companies claimed differences (typically ~1-2kDa) between the two classes of markers. However, as long as your band of interest is at the same relative position to marker, it is fine (across a narrow range of sizes rather than an exact size) - and you can use a series of different percent gels to accurately determine size by PAGE (see the posting in the Analytical chem section).
bwbrian Posted 12/19/2006 2:34:04 AM	Definitely go with the unstained markers. The bands will be much sharper and more accurate, they stain nicely with any protein stain, and you can follow the migration by simply watching the dye front. You can also generally see the unstained bands during migration if you look closely, because they refract the light differently.

A typical discussion on markers for SDS-PAGE, found at www.scientistsolutions.com

Protein Charge?

Main source for different charges: pH-dependent protonation states

<=> Equilibrium between
- density (pH) dependent H^+-binding and
- density independent H^+-dissociation



Probaility 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

H^+ have a +1 charge

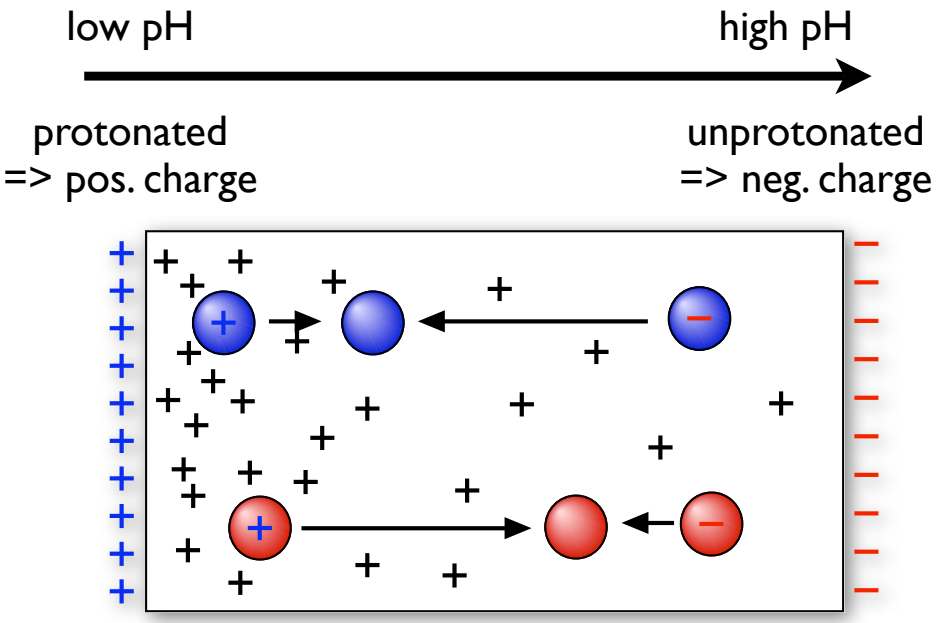
=> **Isoelectric point**: pH at which the protein is **uncharged**

=> 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:

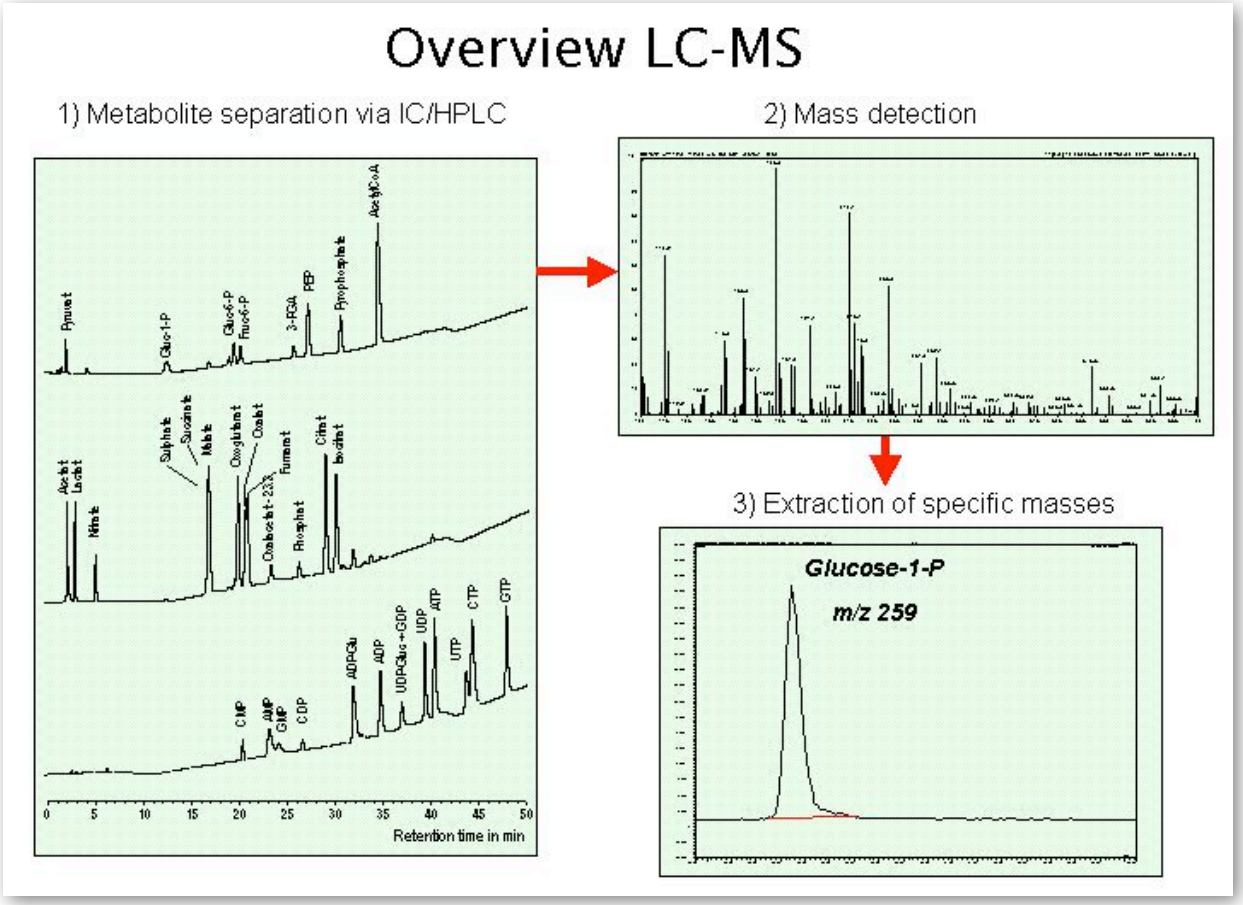


Step 2: SDS-Page

=> Most proteins differ in mass and isoelectric point (pI)

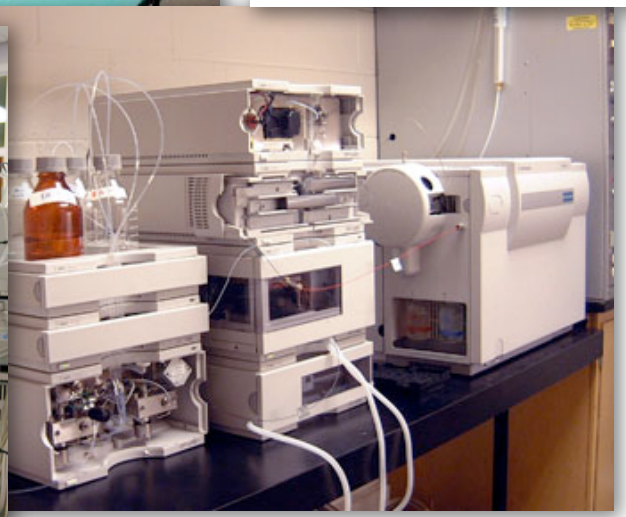
Mass Spectrometry

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







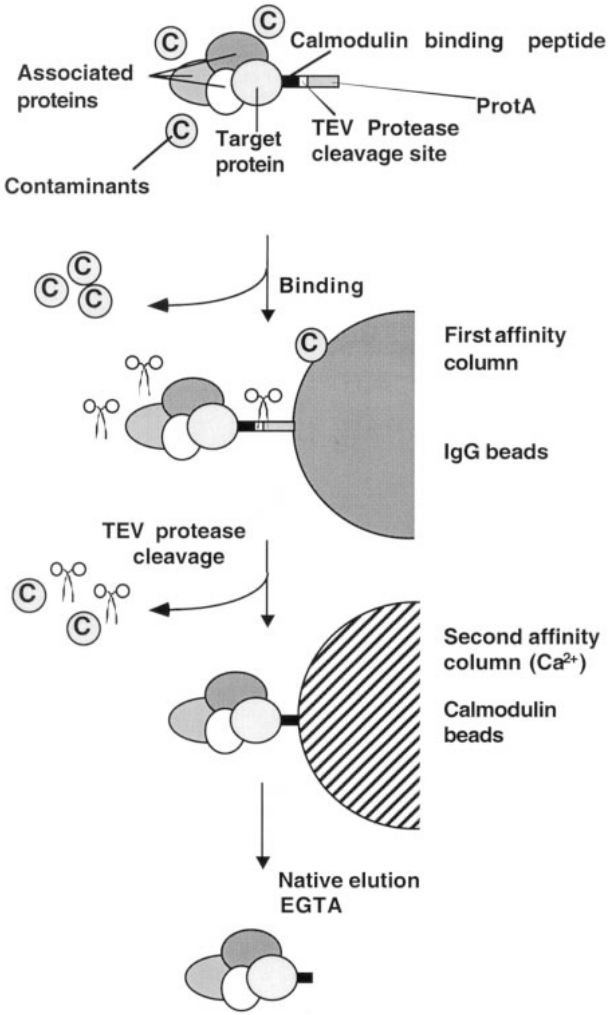


Affinity Chromatography

Puig et al, *Methods* **24** (2001) 218:
Tandem **A**ffinity **P**urification
=> purify target protein in two subsequent binding-elusion steps
=> works on macroscopic amounts
=> identify purified protein with gel, MS, …

Preparation: fuse **TAP tag** to C-terminal
- two IgG binding domains on *Staphylococcus aureus* protein A (ProtA)
- TEV protease cleavage site
- calmodulin binding peptide

=> extract target plus binding partners at near-physiological conditions



Extracted Complexes

Complexes Purified Using the TAP Method

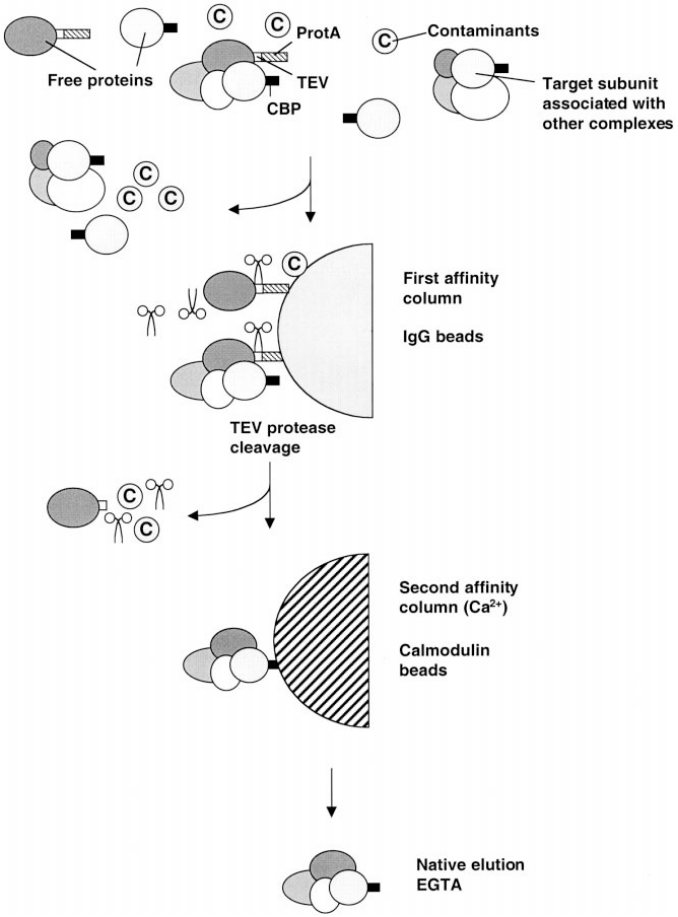
Complex	Function	Protein tagged	Reference
U1 snRNP	Pre-mRNA splicing	Snu71-TAP	(7)
		Luc7p/Snu30p-TAP	(7)
		SmB-TEV-ProtA Nam8p-CBP	This article
		Nam8p-TAP	(23) and this article
U2 snRNP[a]	Pre-mRNA splicing	Lea1p-TAP	(18)
		SmB-TEV-ProtA Lea1p-CBP	(18)
"U6 snRNP"[b]	Pre-mRNA splicing	Lsm8p-TAP	(21)
CBC	Pre-mRNA splicing, nucleocytoplasmic RNA transport	Mud13p-TAP	(7)
BBP-associated	Pre-mRNA splicing, nuclear RNA retention	BBP-TAP	(22)
Mud2p-associated	Pre-mRNA splicing, nuclear RNA retention	Mud2p-TAP	(22)
SF3b	Pre-mRNA splicing	TAP-Rse1p	This article
RNases P/MRP	rRNA and tRNA processing	Pop4-TAP	This article
Dbp5p-associated	Nucleocytoplasmic mRNA transport	Dbp5p-TAP	(19)
Mex67p-associated	Nucleocytoplasmic mRNA transport	Mex67p-TAP	(20)
Mak3/10/31	Protein modification	Mak31-TAP	(7)
Lsm3p-associated	RNA degradation, pre-mRNA splicing	Lsm3p-TAP	(21)
LsmI complex	RNA degradation	Lsm3p-TAP Lsm8p-ProtA	(21)
Xrn1-associated	RNA degradation	Xrn1-TAP	(21)

[a] In this case not all the known components of the complex have been identified.
[b] Contains a mixture of U6, U4/U6, and U4/U6.U5 snRNPs; see Ref. (21).

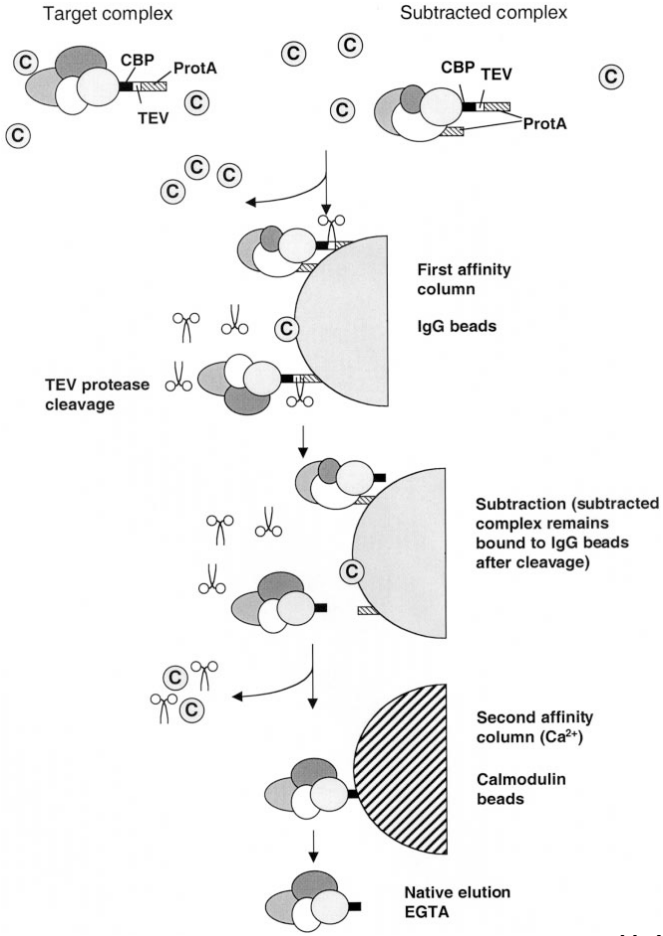
from Puig et al, *Methods* **24** (2001) 218

Variations

Split tag strategy:
=> identify specific pairs



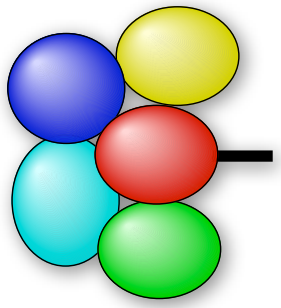
Substraction strategy
=> suppress specific partners



Pros and Cons

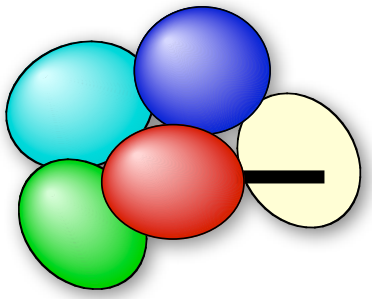
Advantages:

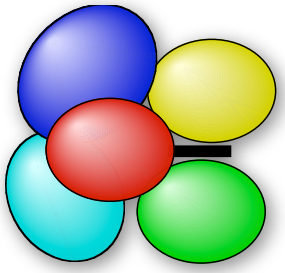
- **quantitative** determination of complex partners ***in vivo*** without prior knowledge
- simple, high yield, high throughput



Difficulties:

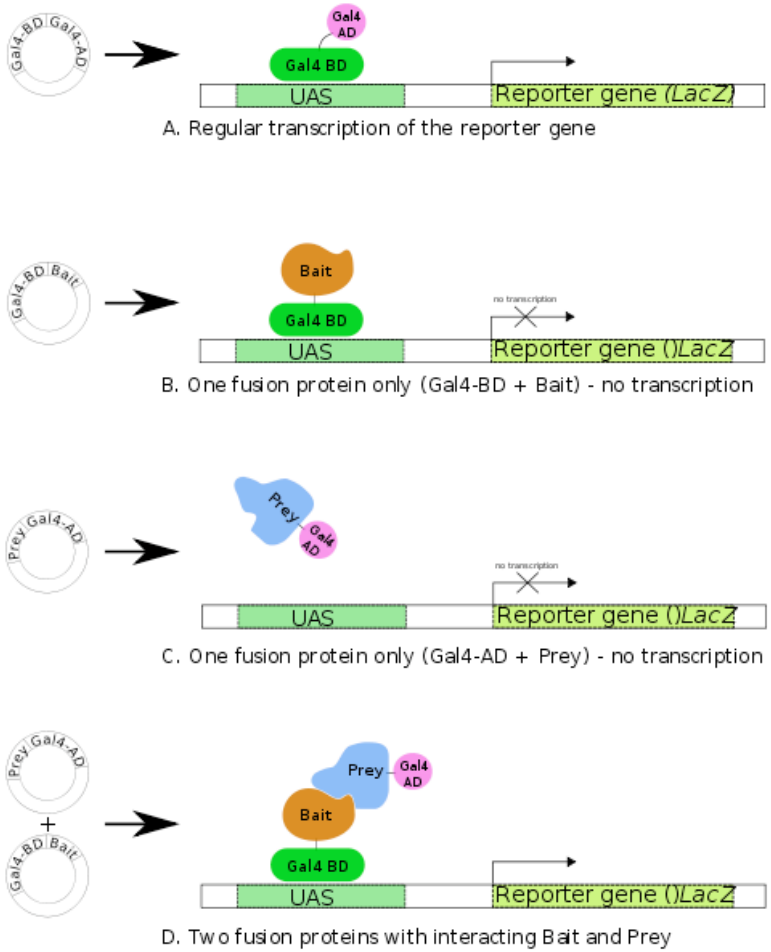
- tag may **prevent** binding of the interaction partners
- tag may change (relative) **expression** levels
- tag may be **buried** between interaction partners
  => no binding to beads





Yeast Two-Hybrid Screening

Discover binary protein-protein interactions via physical interaction



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

May point to:
- physical interaction (building blocks of a complex)
- both proteins in the same pathway
- both proteins have the same function (redundancy)

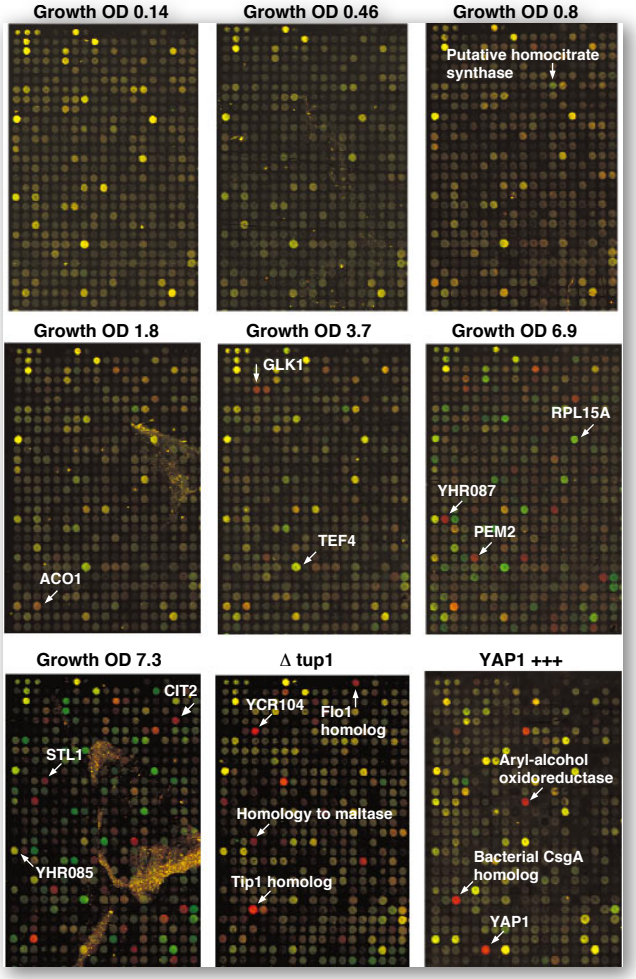
Gene Coexpression

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*
=> groups of genes with similar expression profiles

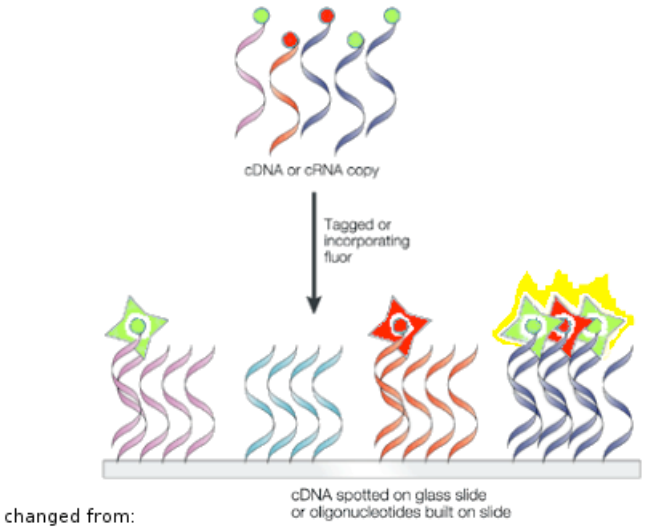


DNA Microarrays

Flourescence labeled DNA (cDNA) applied to micro arrays

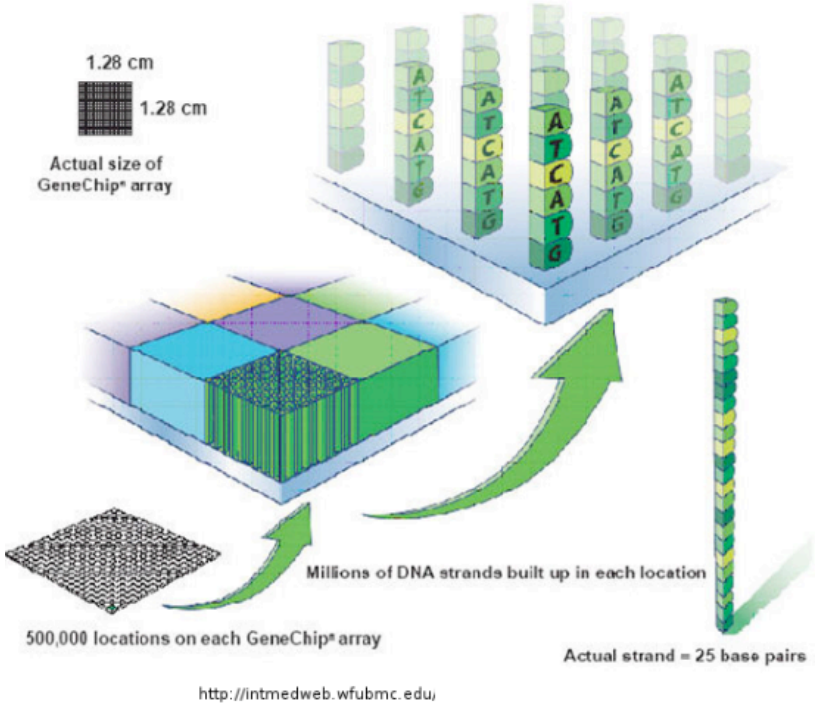
=> hybrization with complementary library strand

=> flourescence indicates relative cDNA amounts



changed from:

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



two labels (red + green) for experiment and control

Usually: red = signal

green = control

=> yellow = "no change"

Diauxic Shift

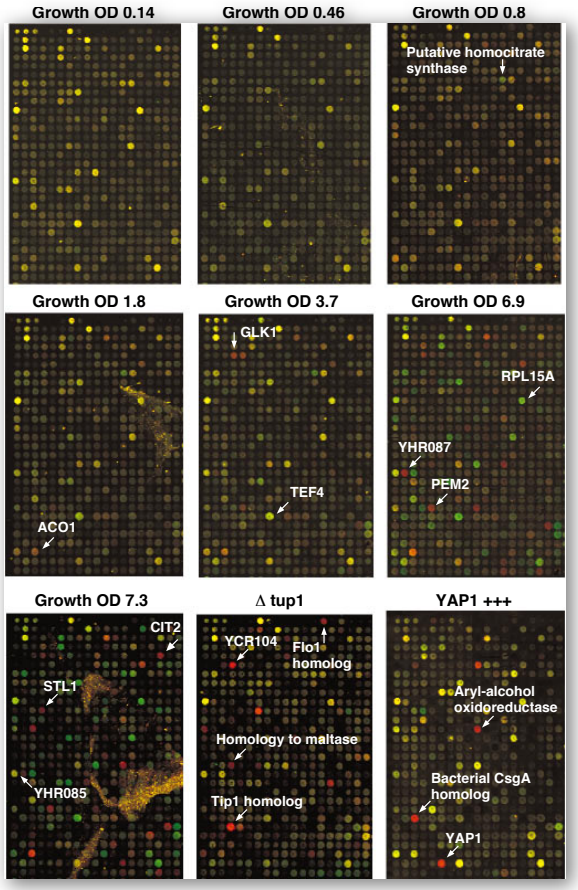
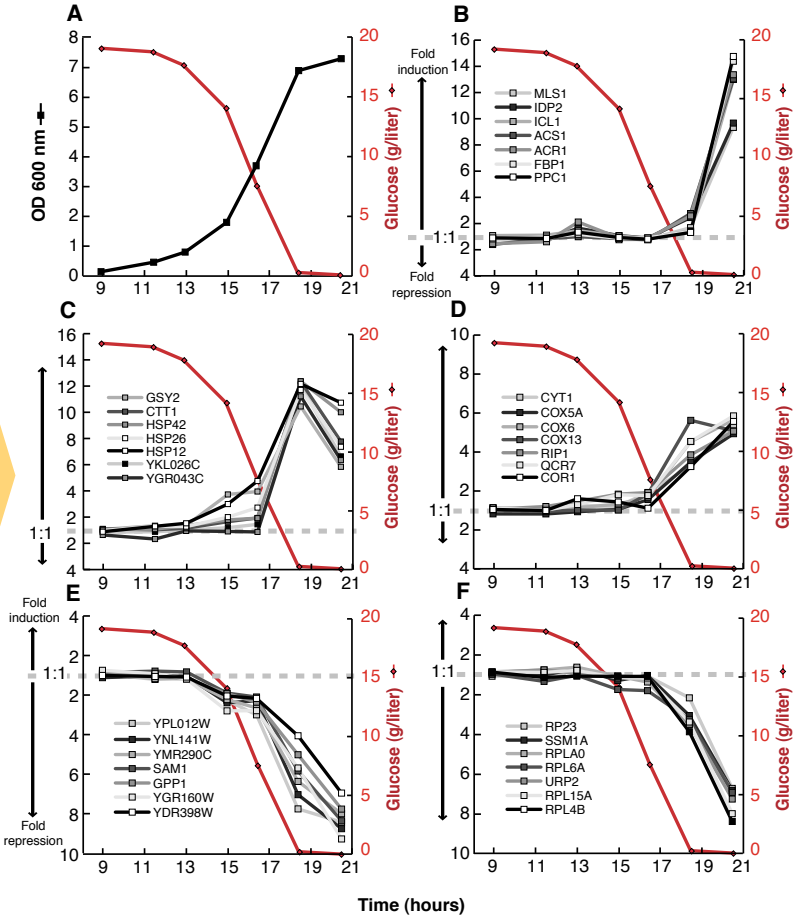


image analysis + clustering



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

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

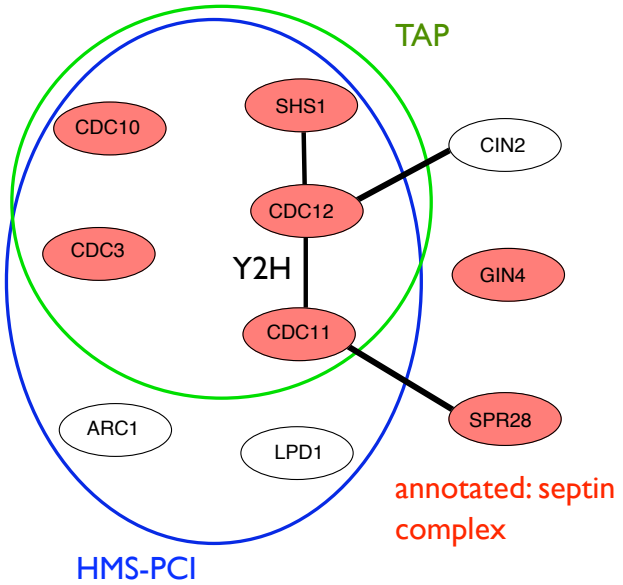
(low) Overlap of Results

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

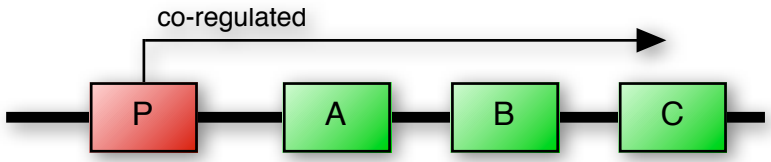
Structure-based:
- interface propensities
- spatial simulations

"Work on the parts list"
=> fast
=> unspecific
=> high-throughput methods
for pre-sorting

"Work on the parts"
=> specific, detailed
=> expensive
=> accurate

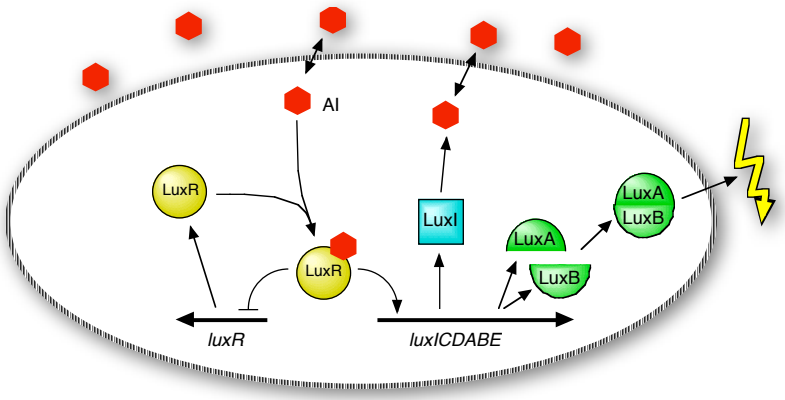
Gene Clustering

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



Search for genes with a **common promoter**
=> when activated, all are transcibed together as one operand

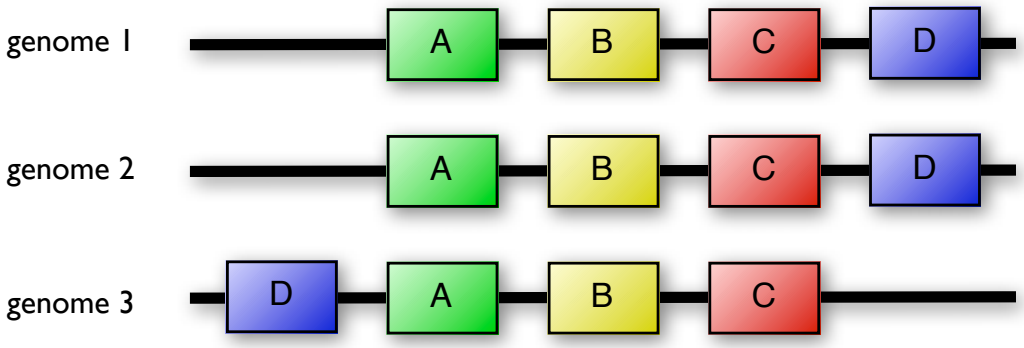
Example:
bioluminescence in *V. fischeri*,
regulated via quorum sensing
=> three proteins: I, AB, CDE



Gene Neighborhood

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

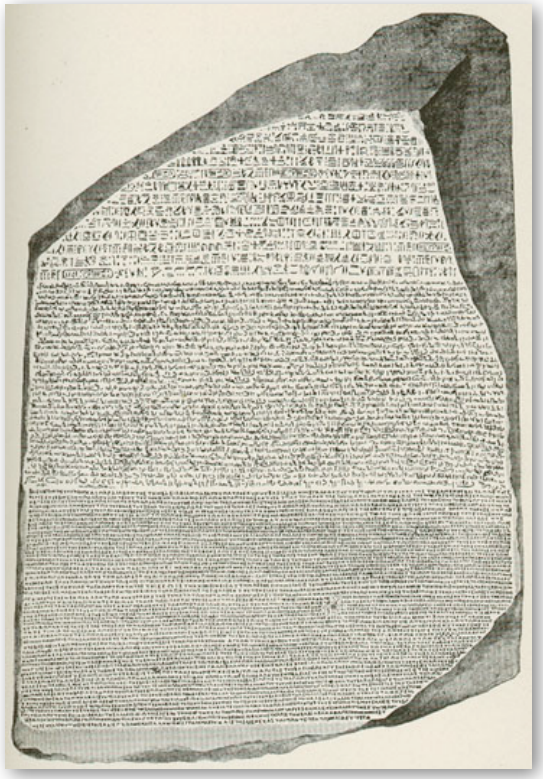
"functionally" = same {complex | pathway | function | …}



=> Search for **similar sequences** of genes in **different organisms**

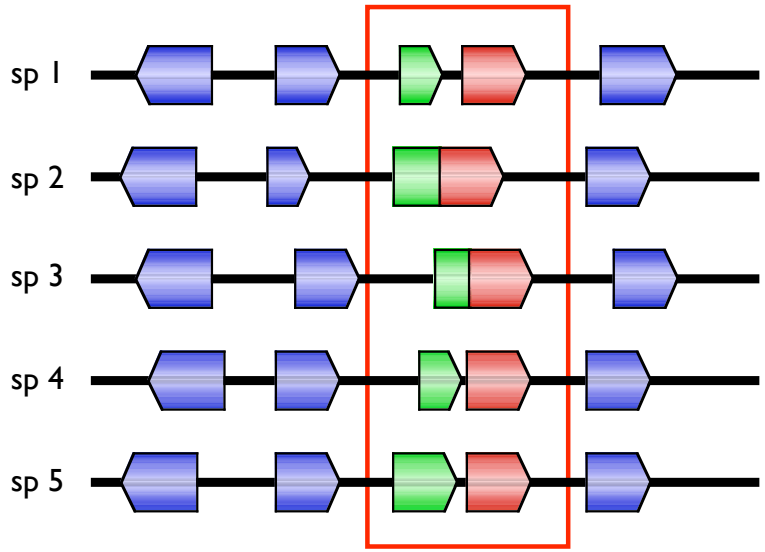
(<=> Gene clustering: one species, promoters)

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**"

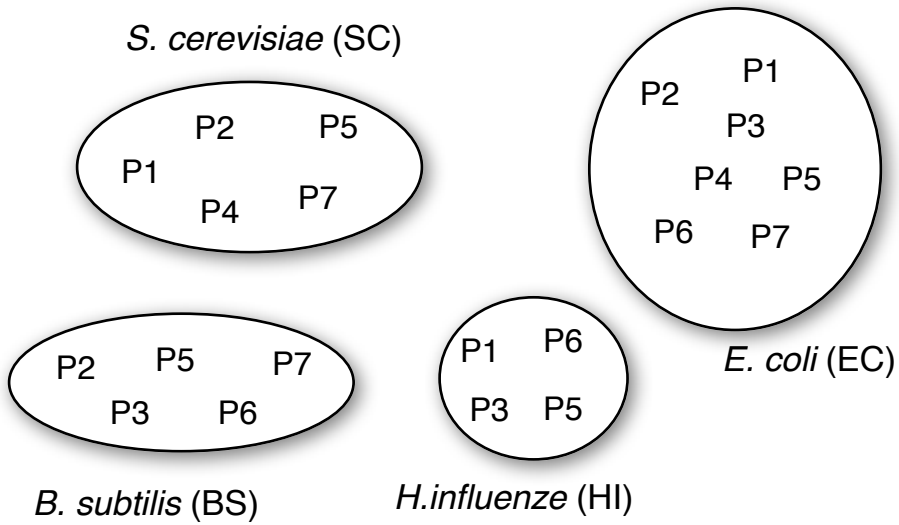


Enright, Ouzounis (2001):
40000 predicted pair-wise interactions
from search across 23 species

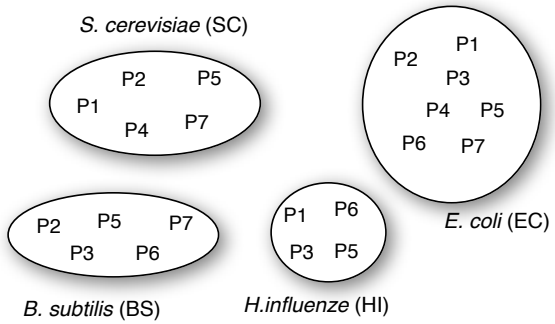
Phylogenetic Profiling

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

=> compare presence of protein homologs across sepcies (e.g., via sequence alignment)



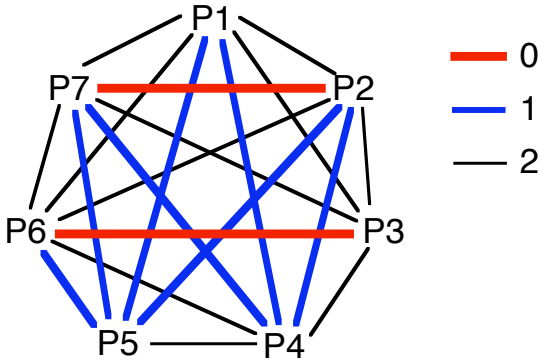
Distances



	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 occurances

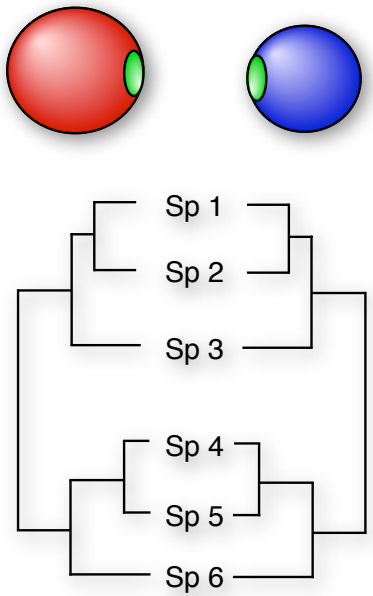
	P1	P2	P3	P4	P5	P6	P7
P1	0	2	2	1	1	2	2
P2		0	2	1	1	2	**0**
P3			0	3	1	**0**	2
P4				0	2	3	1
P5					0	1	1
P6						0	2
P7							0



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

Coevolution

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

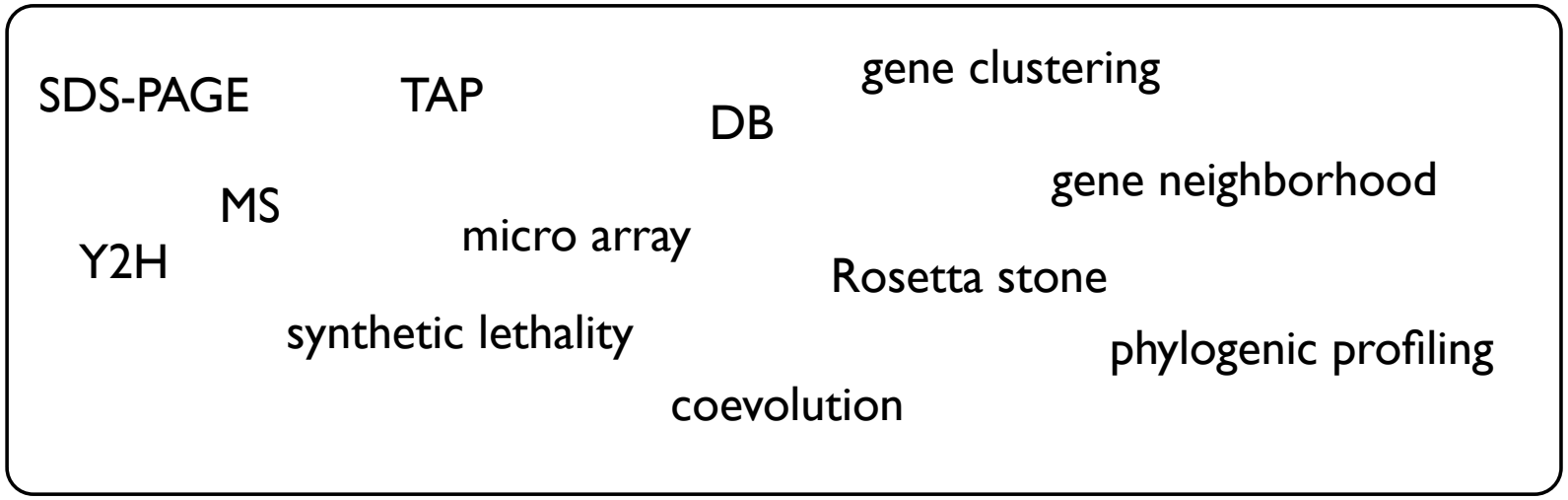


Complexes are often better conserved

Also: look for potential substitutes
=> anti-correlated
=> missing components of pathways
=> function prediction across sepcies
=> novel interactions

Summary

What you learned **today**: how to get some data



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

Next lecture: Fri, Oct. 28, 2010

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

Tutorial: Wed, Nov. 2, 2010, 12:30