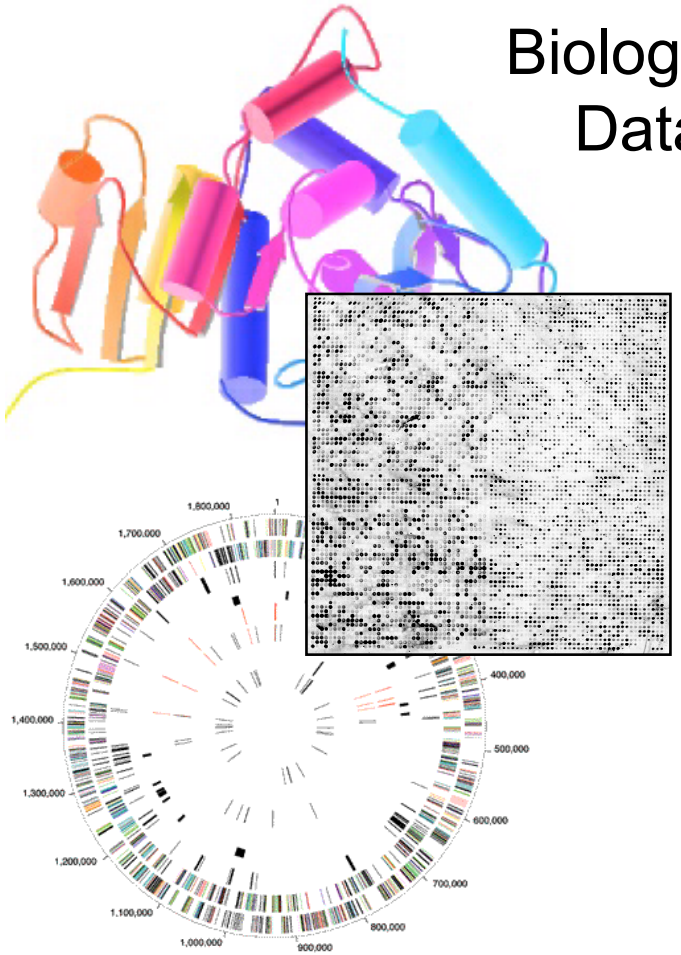
~25' Overview of the GersteinLab in 2008

CS Visiting Students Day,
2008.03.26 15:00-15:30

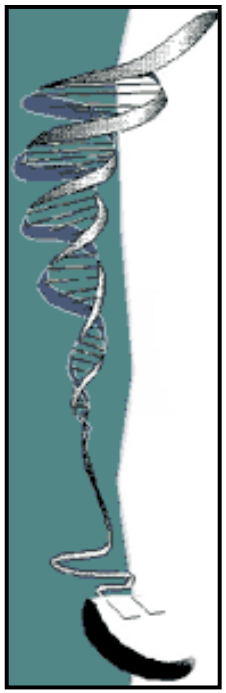
Bioinformatics



Biological Data

+

Computer Calculations



Data Mining

- Importance of knowing the Data
- Best approaches often require detailed domain knowledge – non anonymized aol data, netflix challenge

Bioinformatics represents one of the biggest "open" areas for mining

- Genomics & Astronomy
- Finance, marketing, credit-card fraud
- Security and Intelligence

- Relation to experimental sciences

General Intro. & background on bioinformatics

What is Bioinformatics?

Cor

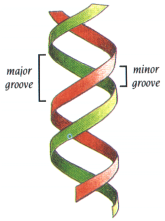
- *(Molecular)* **Bio - informatics**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

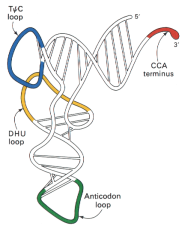
What is the Information?
Molecular Biology as an Information Science

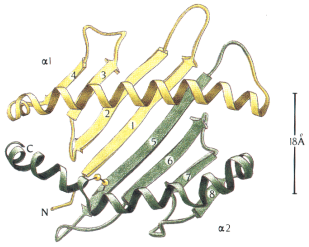
- Central Dogma of Molecular Biology

```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

- Molecules
  - ◊ Sequence, Structure, Function
- Processes
  - ◊ Mechanism, Specificity, Regulation







•Genetic material

•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity

- Central Paradigm for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Protein Function
     -> Phenotype
```

- Large Amounts of Information
  - ◊ Standardized
  - ◊ Statistical

•Most cellular functions are performed or facilitated by proteins.
•Primary biocatalyst
•Cofactor transport/storage
•Mechanical motion/support
•Immune protection
•Control of growth/differentiation

(idea from D Brutlag, Stanford, graphics from S Strobel)

Molecular Biology Information - DNA

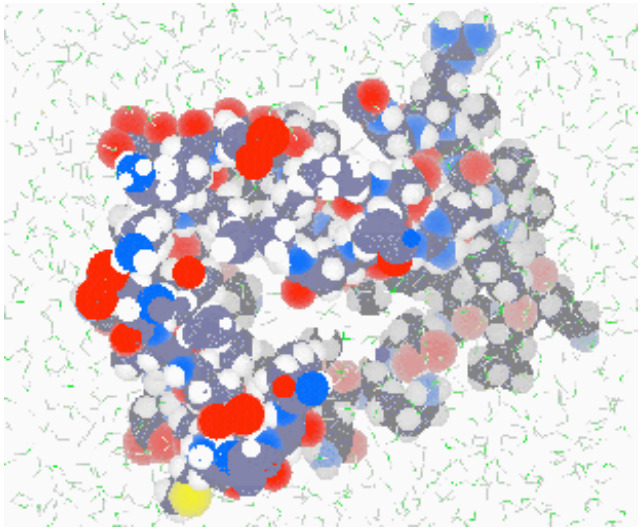
- Raw DNA Sequence
  - ◊ Coding or Not?
  - ◊ Parse into genes?
  - ◊ 4 bases: AGCT
  - ◊ ~1 K in a gene, ~2 M in genome
  - ◊ ~3 Gb Human

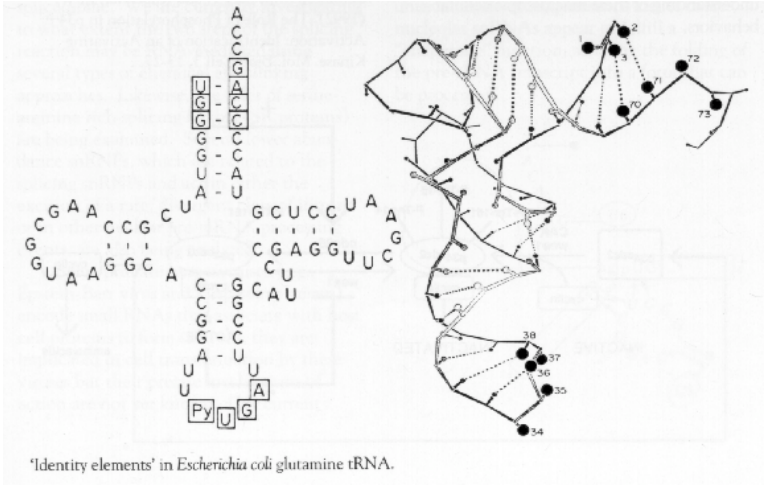
```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctccttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .


. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccatttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Macromolecular Structure

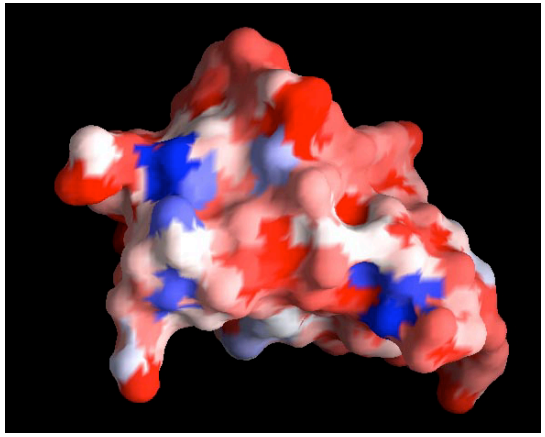
- DNA/RNA/Protein
  - ◊ Almost all protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)





'Identity elements' in *Escherichia coli* glutamine tRNA.



1995

Bacteria, 1.6 Mb, ~1600 genes [*Science* **269**: 496]

1997

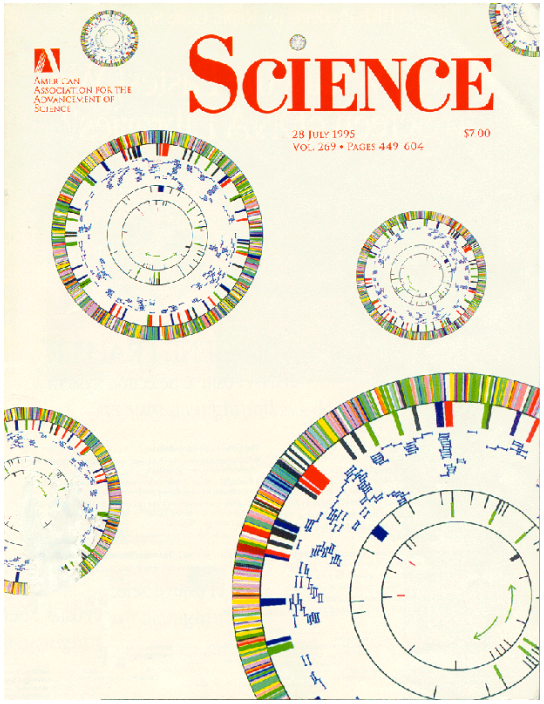
Eukaryote, 13 Mb, ~6K genes [Nature **387**: 1]

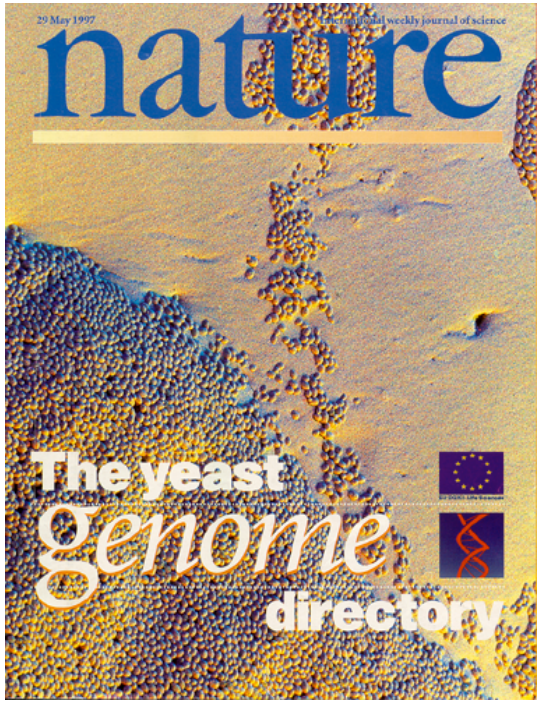
1998

Animal, ~100 Mb, ~20K genes [*Science* **282**: 1945]

2000?

Human, ~3 Gb, ~25K genes





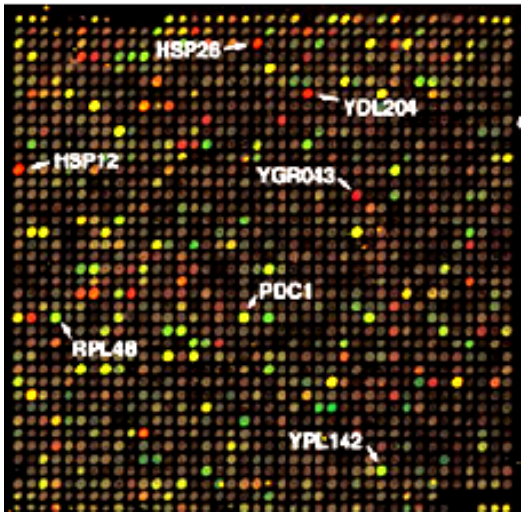
Genomes highlight the **Finiteness** of the "Parts" in Biology

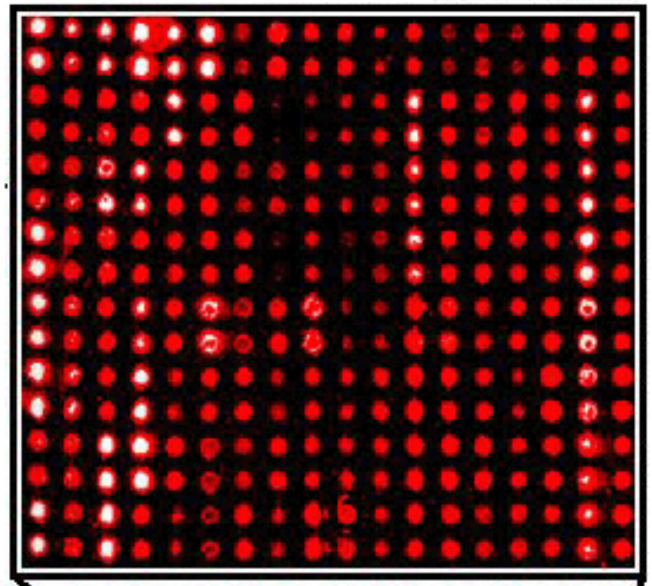
'98 spoof

real thing, Apr '00

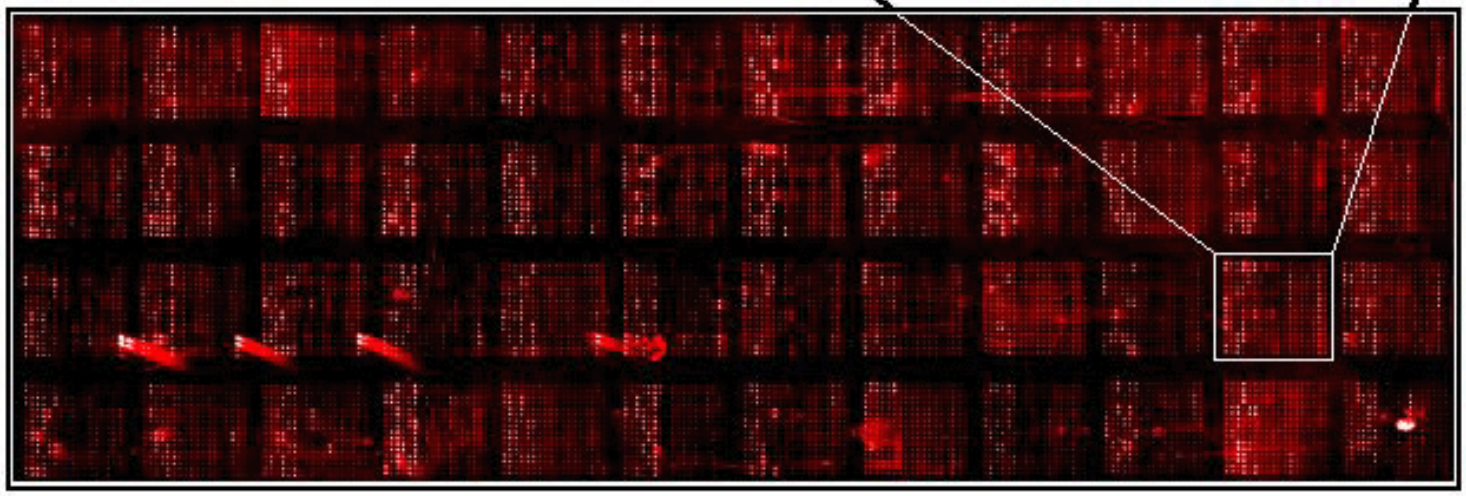
The recent advent and subsequent onslaught of microarray data

1st generation, Expression Arrays (Brown)





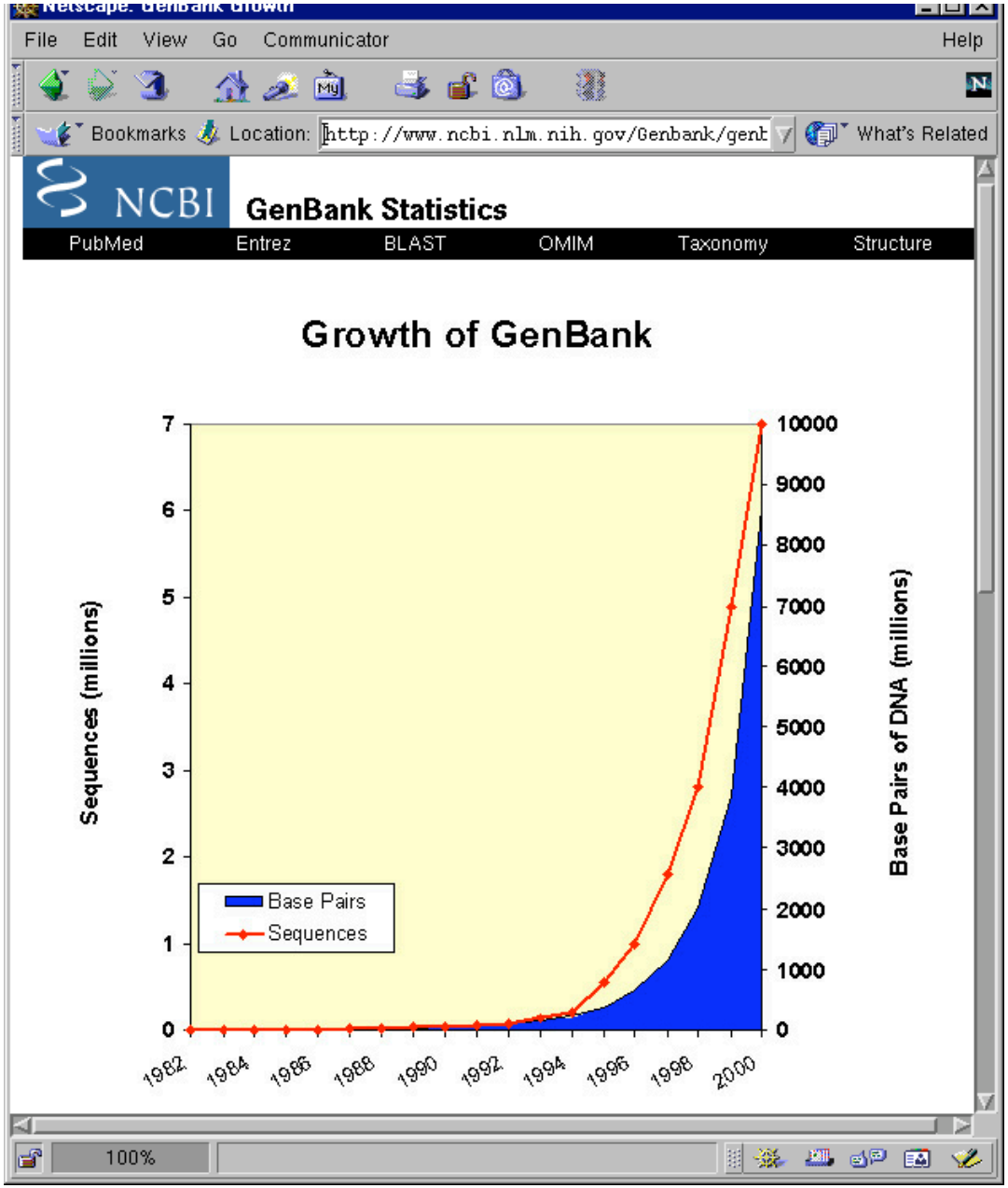
2nd gen., Proteome Chips (Snyder)



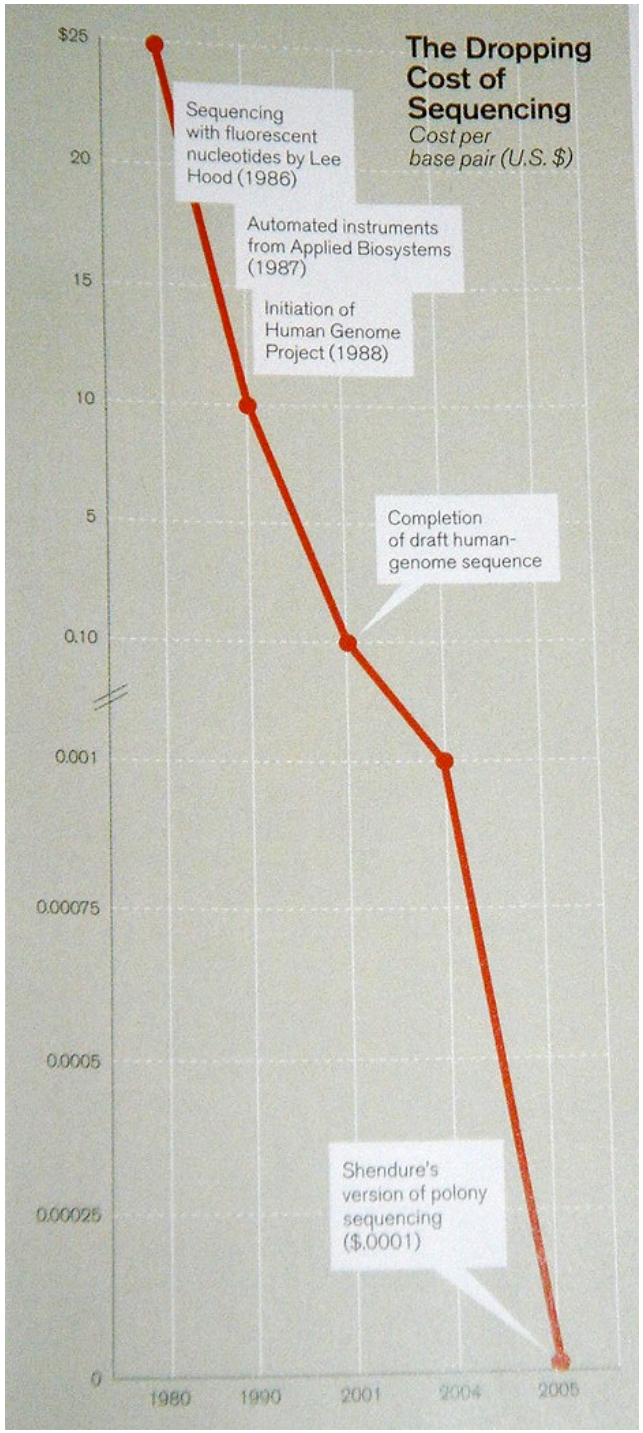
What is Bioinformatics?

- *(Molecular)* **Bio - `informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

Large-scale Information: GenBank Growth



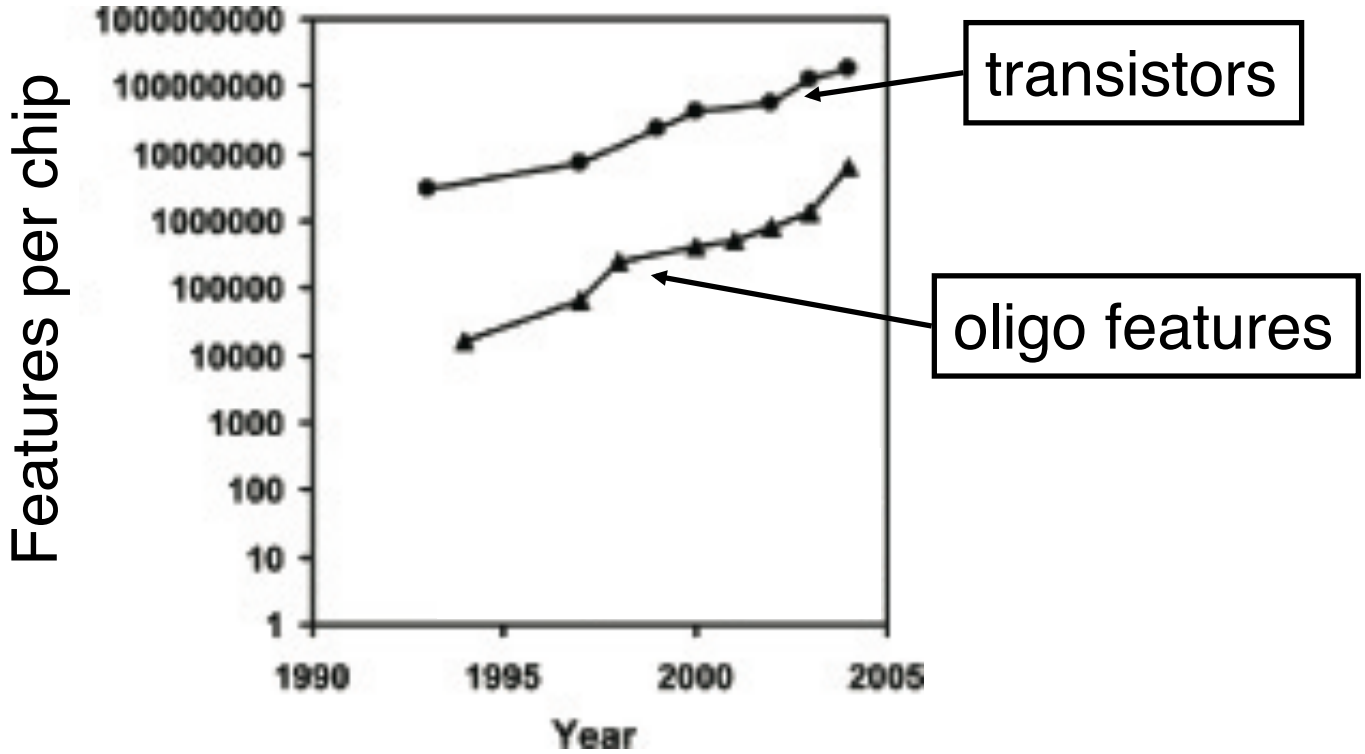
GenBank Data		
Year	**Base Pairs**	**Sequences**
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	8604221980	7077491



The Dropping Cost of Sequencing

- Adapted from Technology Review (Sept./Oct. 2006)

Features per Slide



General Types of "Informatics" techniques in Bioinformatics

- Databases
  - ◊ Building, Querying
  - ◊ Dealing with Complex data
- Text String Comparison
  - ◊ Text Search
  - ◊ 1D Alignment
  - ◊ Significance Statistics
- Dealing with 3D Objects
  - ◊ Graphics (Surfaces, Volumes)
  - ◊ Comparison and 3D Matching (Vision, recognition)
- Physical Simulation
  - ◊ Newtonian Mechanics
  - ◊ Electrostatics
  - ◊ Numerical Algorithms
  - ◊ Simulation
- Finding Patterns
  - ◊ Machine Learning
  - ◊ Clustering
  - ◊ Datamining
  - ◊ Information integration and fusion
    - o Dealing with heterogeneous data
  - ◊ Dimensionality Reduction
    - o (PCA etc)

Research @ GersteinLab.org

- Human Genome Annotation (pseudogenes)
  - ◊ Characterizing the function of non-coding regions, focusing on protein fossils and novel transcriptionally active regions (Pseudogene.org + Tiling.GersteinLab.org)
- Molecular Networks
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe protein function on a large-scale (Networks.GersteinLab.org)
- Macromolecular motions
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)

Research @ GersteinLab.org

- Human Genome Annotation (pseudogenes)
  - ◊ Characterizing the function of non-coding regions, focusing on protein fossils and novel transcriptionally active regions (Pseudogene.org + Tiling.GersteinLab.org)
- Molecular Networks
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe protein function on a large-scale (Networks.GersteinLab.org)
- Macromolecular motions
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)

2001: Most of the genome is not coding (only ~1.2% exon). It consists of elements such as repeats, regulatory regions, non-coding RNAs, origins of replication, pseudogenes, segmental duplications....What do these elements do? How should they be annotated?

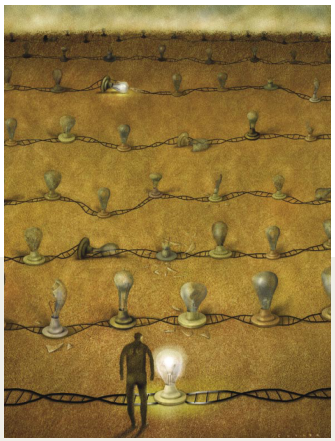
[IHGSC, ***Nature*** 409, 2001]
[Venter et al. ***Science*** 29, 2001]

2007 : Pilot results from ENCODE Consortium on decoding what the bases do

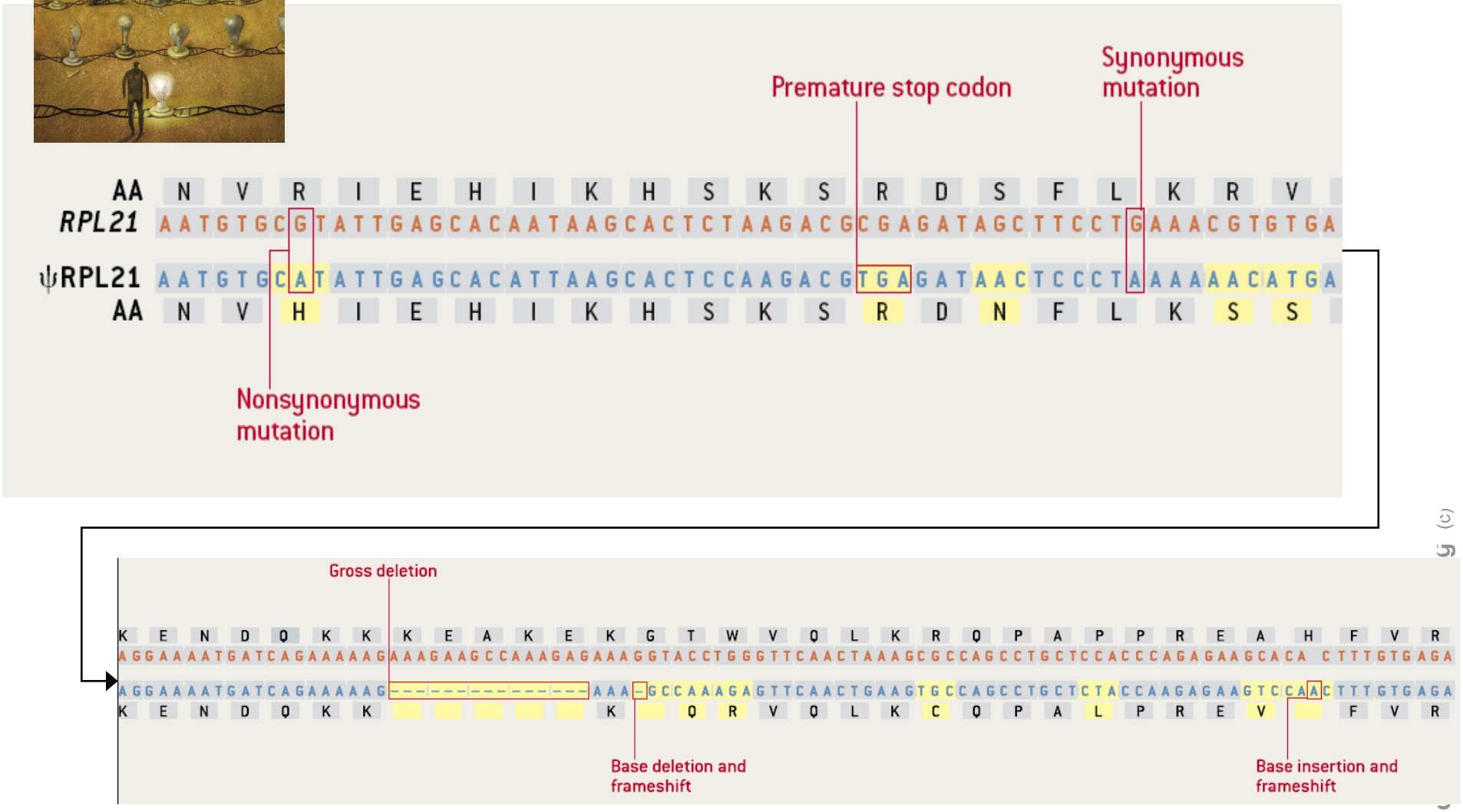
- 1% of Genome (30 Mb in 44 regions)
- Tiling Arrays to assay Transcription & Binding
- Multi-organism sequencing and alignment
- Careful Annotation
- Variation Data

Snyder
Weissman
Miller

[IHGSC, ***Nature*** 409, 2001]
[ENCODE Consortium, ***Nature*** 447, 2007]

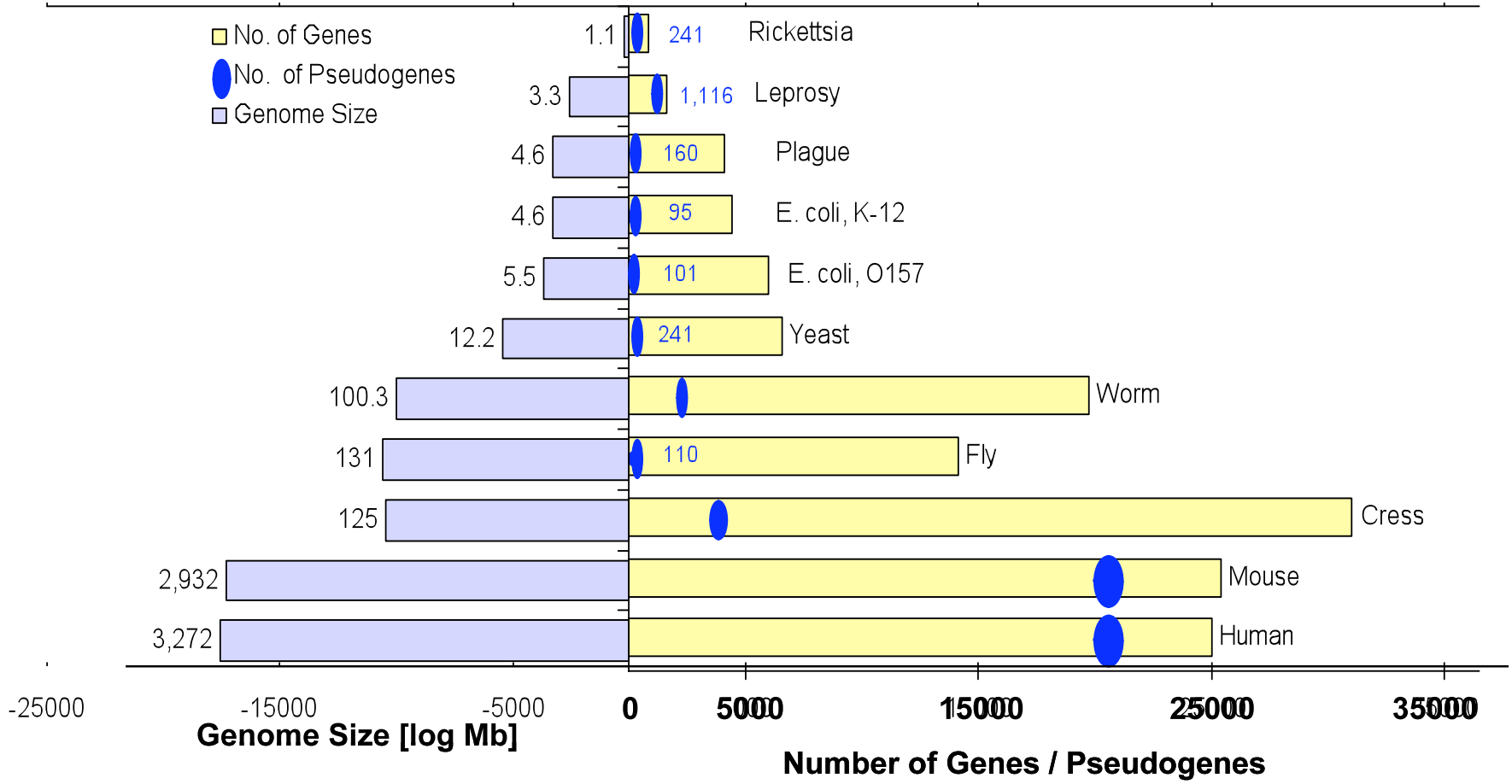


Identifiable Features of a Pseudogene (ψRPL21)



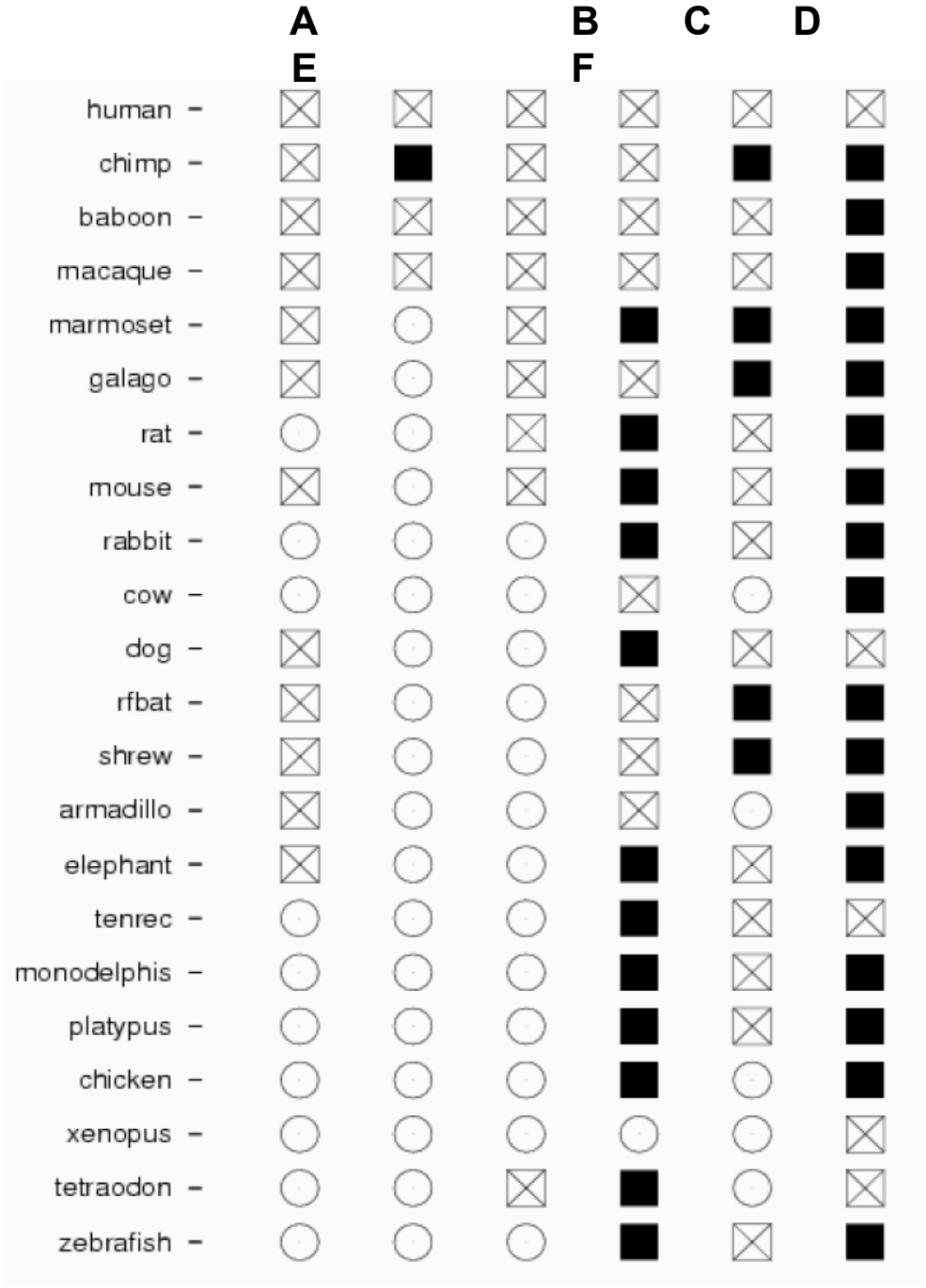
Gerstein & Zheng. Sci Am 295: 48 (2006).

Very Different Distribution of Genes and Pseudogenes in Different Organisms



Zhang & Gerstein (2004) *Curr Opin Genet Dev* 14: 328 + Harrison & Gerstein (2002) J Mol Biol 318: 1155

representative pseudogenes drawn from 201 total



History of Pseudogene Preservation

Based on alignment from ENCODE MSA group

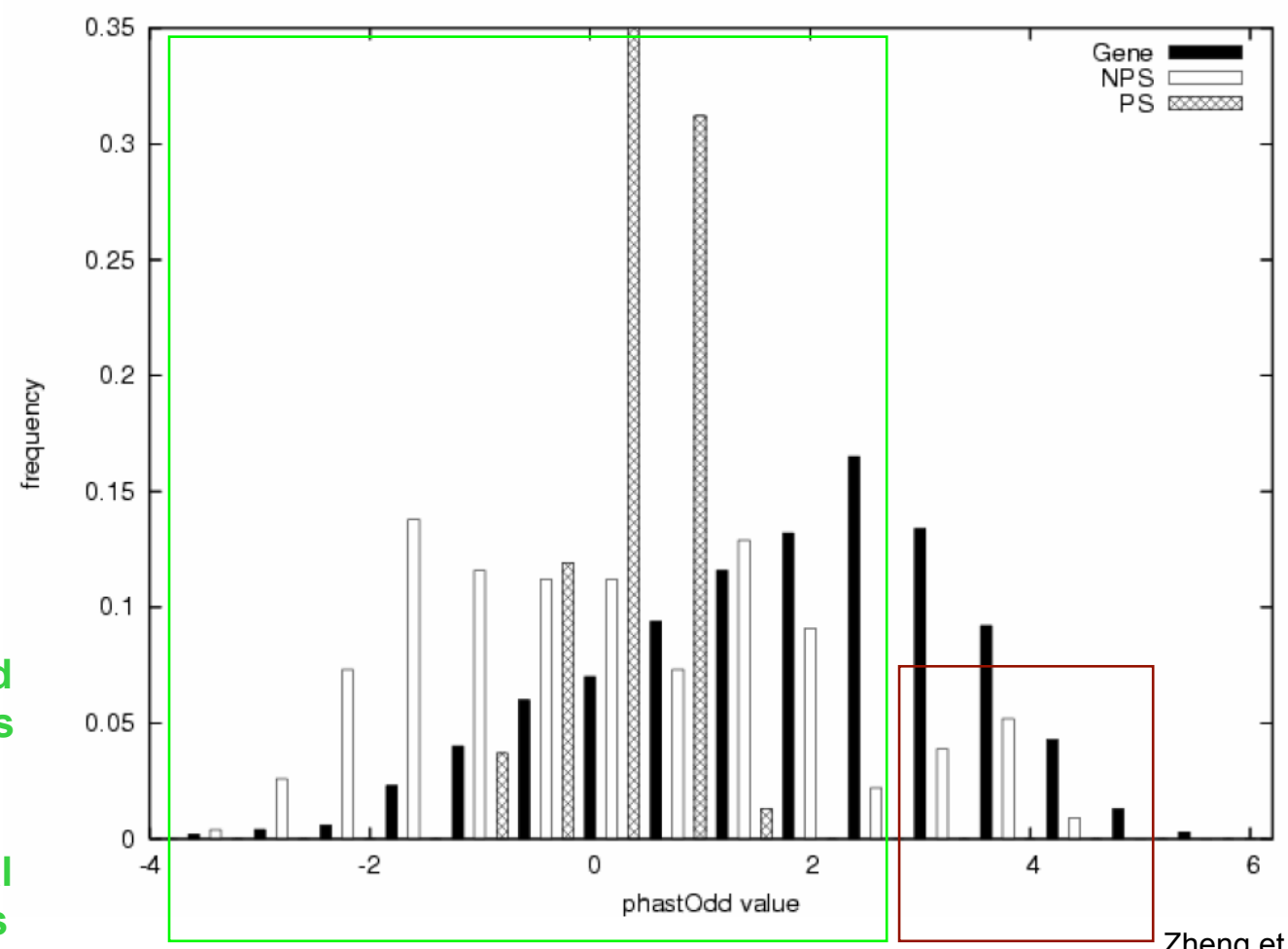
Zheng et al. (2007) Gen. Res.

Absent

Present with Disablement

Present without Disablement

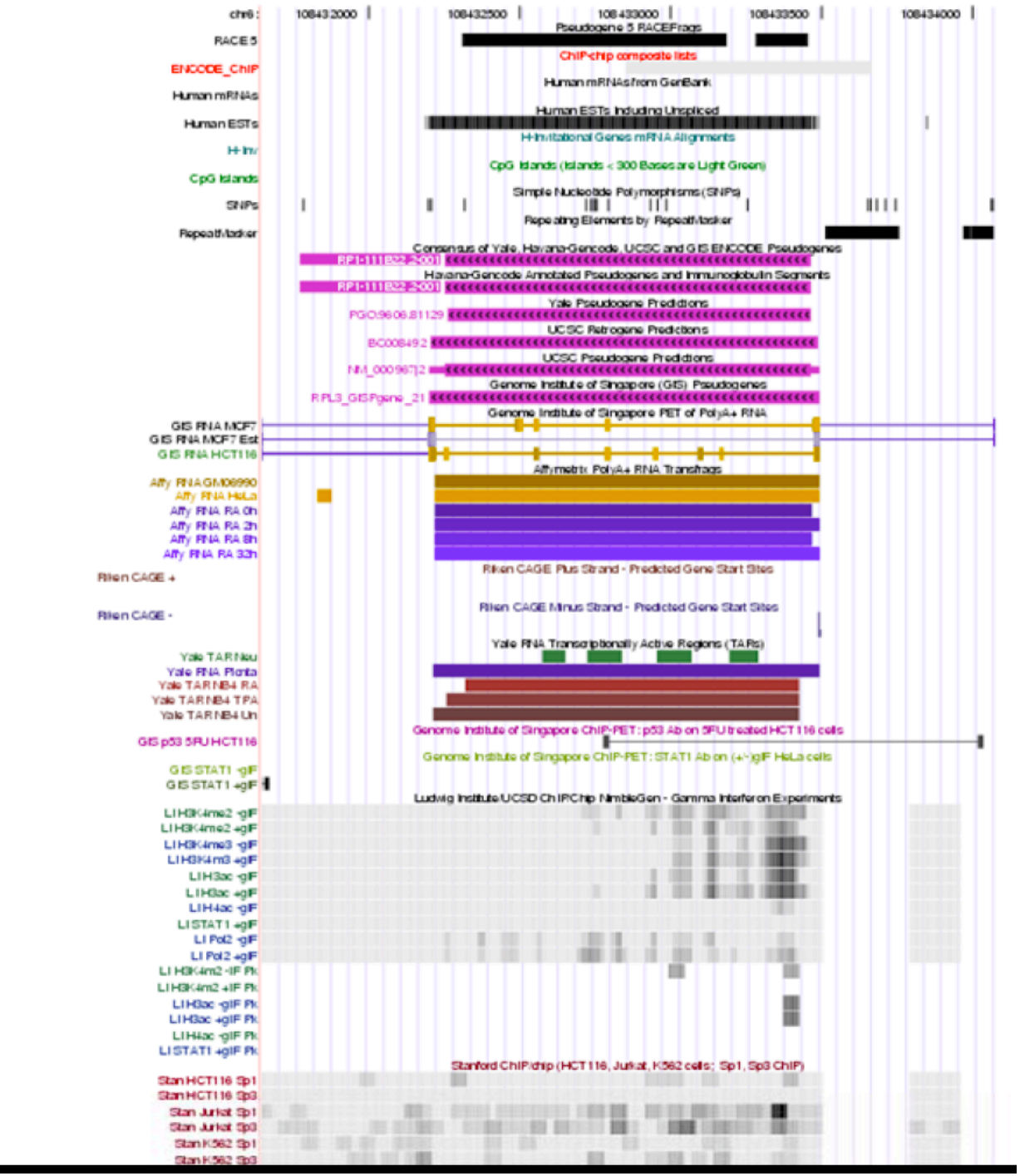
Using phastOdd value to examine neutral evolution of pseudogenes



most good candidates for studying mutational processes

a few non-proc. ψG under constraint

Zheng et al. (2007) Gen. Res.



Composite ChIP hit

Special ψG tracks in browser

diTAG

CAGE

TARs

ChIP-chip

Ex. Pseudogene

Intersecting Transcript-ional Evidence

Zheng et al. (2007) Gen. Res.

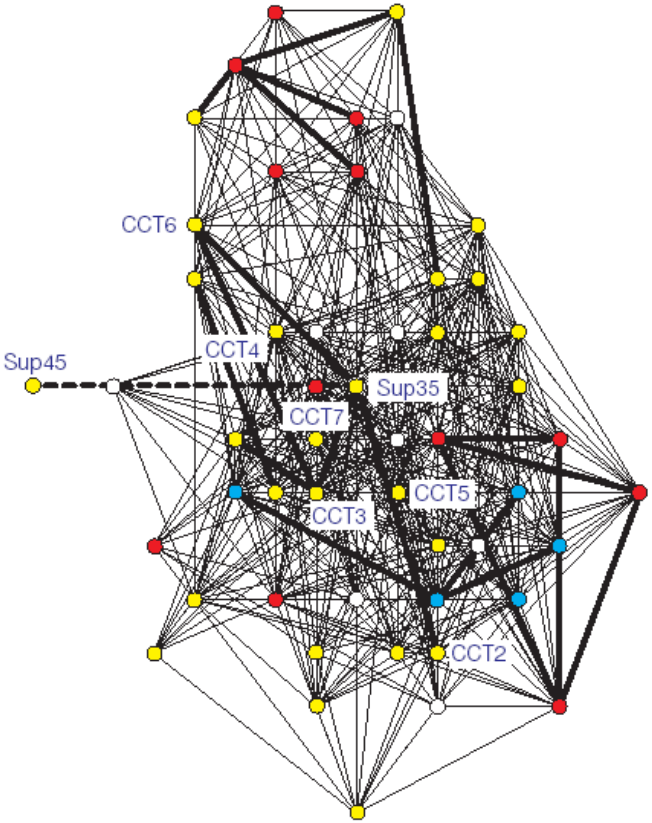
Computational Ideas Used in Genome Annotation

- Mining to find genes, pseudogenes, regulatory regions (e.g. with HMMs)
- Representing annotation in federation of distributed information resources
- How to align and compare "strings" sequences on many different scales
  - ◊ Genome alignment
  - ◊ Genome assembly
- How to statistically quantify variation on many levels (e.g. population, between organism, &c)

Research @ GersteinLab.org

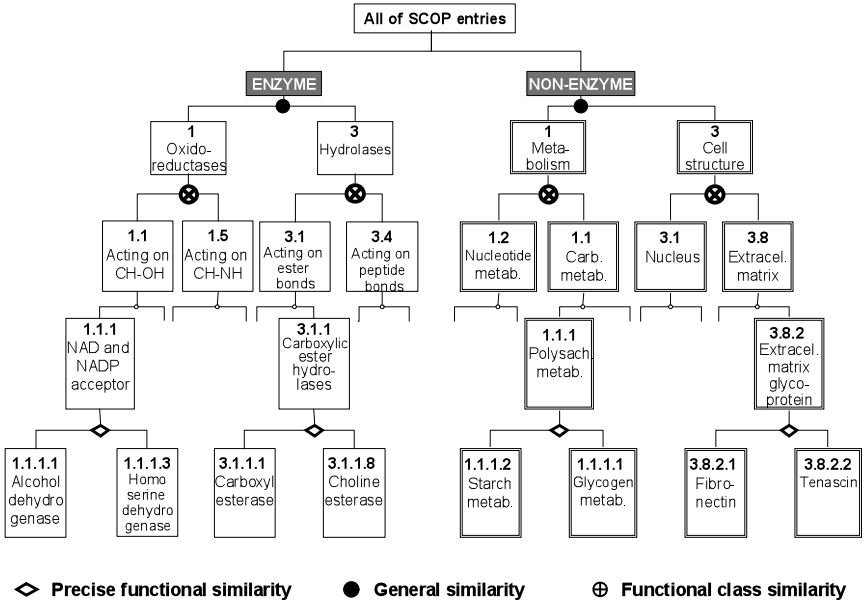
- Human Genome Annotation (pseudogenes)
  - ◊ Characterizing the function of non-coding regions, focusing on protein fossils and novel transcriptionally active regions (Pseudogene.org + Tiling.GersteinLab.org)
- Molecular Networks
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe protein function on a large-scale (Networks.GersteinLab.org)
- Macromolecular motions
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)

Toward Systematic Ontologies for Function, using Networks



General Networks

[Eisenberg et al.]



Hierarchies & DAGs

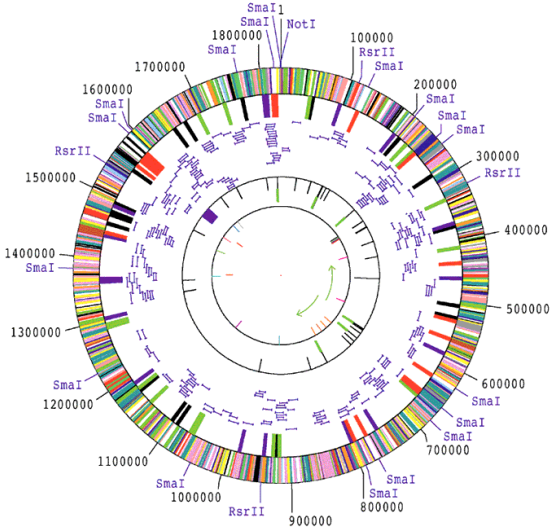
[Enzyme, Bairoch; GO, Ashburner; MIPS, Mewes, Frishman]

nucleic acids | small molecules | proteins

	DNA	RNA	ATP	Metal	CoA	NAD	.....	G protein	CDC28	Calmodulin	.....
protein 1	1.0	0	0	0	0	0	.....	0	0	0	.....
protein 2	0	0.9	0	0	0	0	.....	0	0	0	.....
protein 3	1.0	0	1.0	0	0	0	.....	0	0	0	.....
protein 4	0	0	0	0	0.8	0	.....	0	0	1.0	.....
protein 5	1.0	0	0	0	0	0	.....	0	0.9	0	.....
protein 6	0.9	0					.....				.....
protein 7	0	0.8					.....				.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

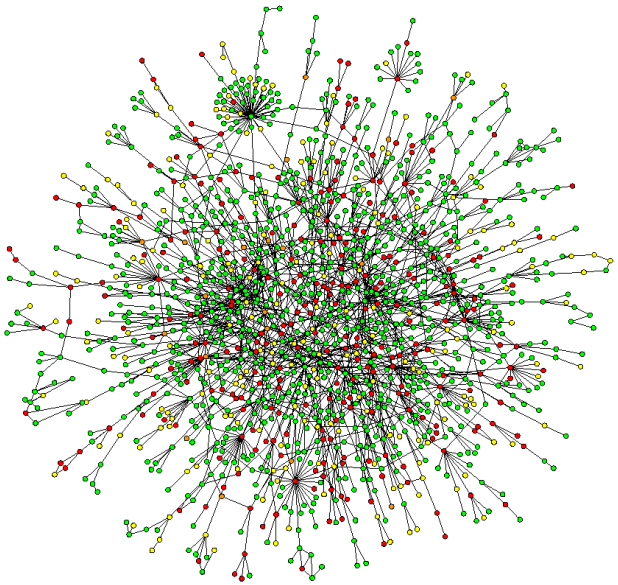
Interaction Vectors [Lan et al, IEEE 90:1848]

Networks occupy a midway point in terms of level of understanding



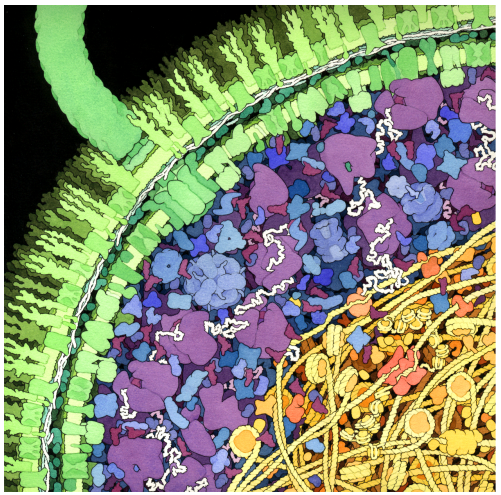
1D: Complete Genetic Partslist

[Fleischmann et al., Science, 269 :496]



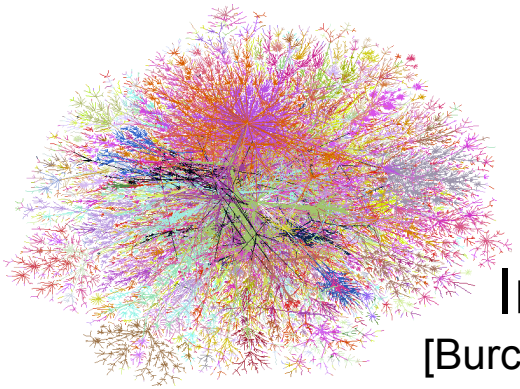
~2D: Bio-molecular Network Wiring Diagram

[Jeong et al. Nature, 41:411]

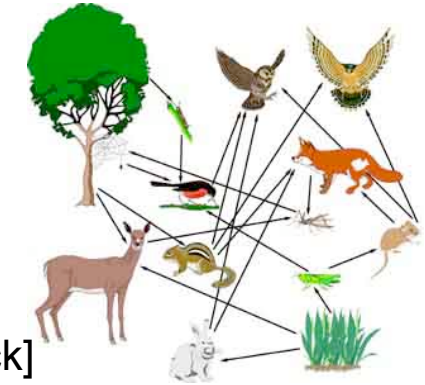


3D: Detailed structural understanding of cellular machinery

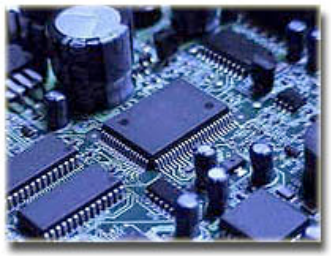
Networks as a universal language



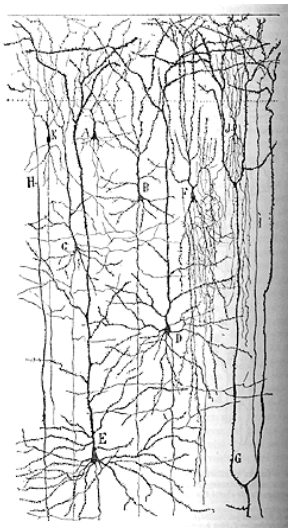
Internet
[Burch & Cheswick]



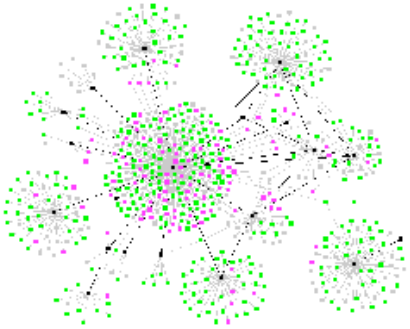
Food Web



Electronic Circuit

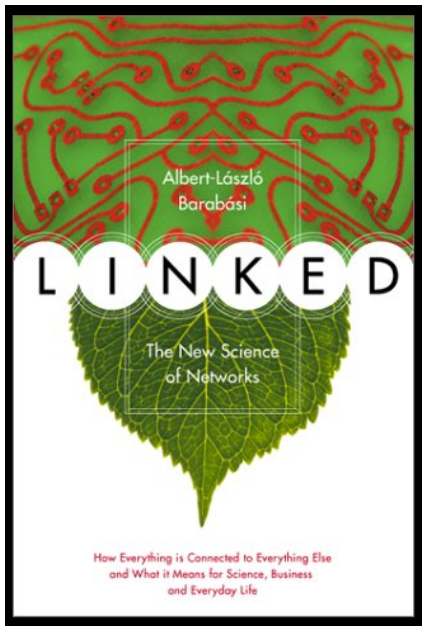


Neural Network
[Cajal]



Disease Spread
[Krebs]

Protein Interactions
[Barabasi]

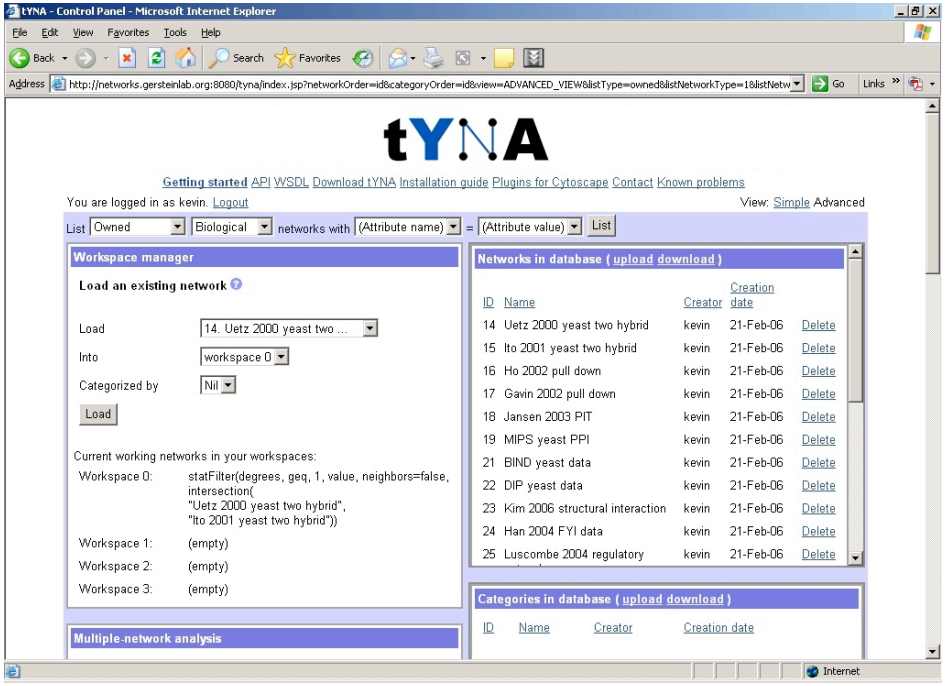


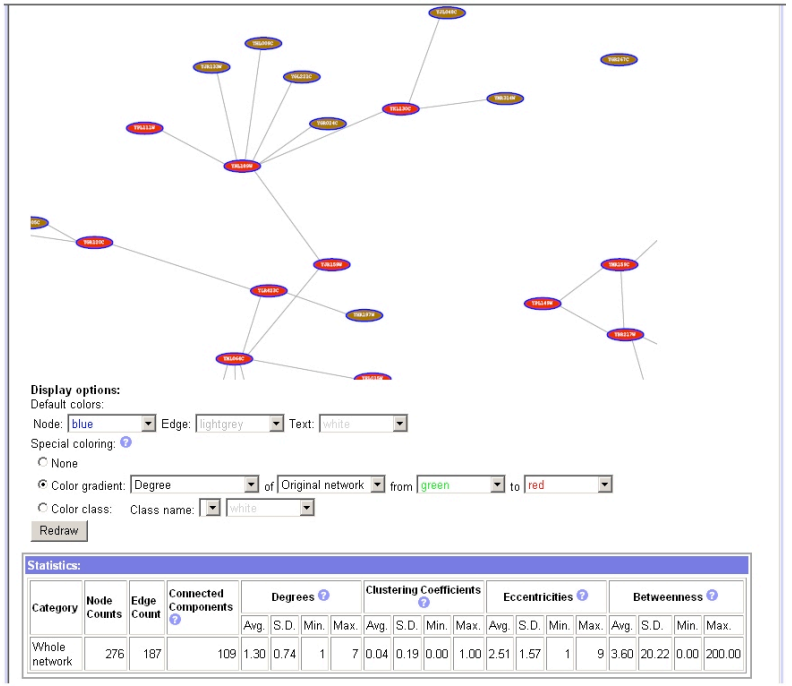
Social Network

- an automated web tool

tYNA

(vers. 2 : "TopNet-like Yale Network Analyzer")

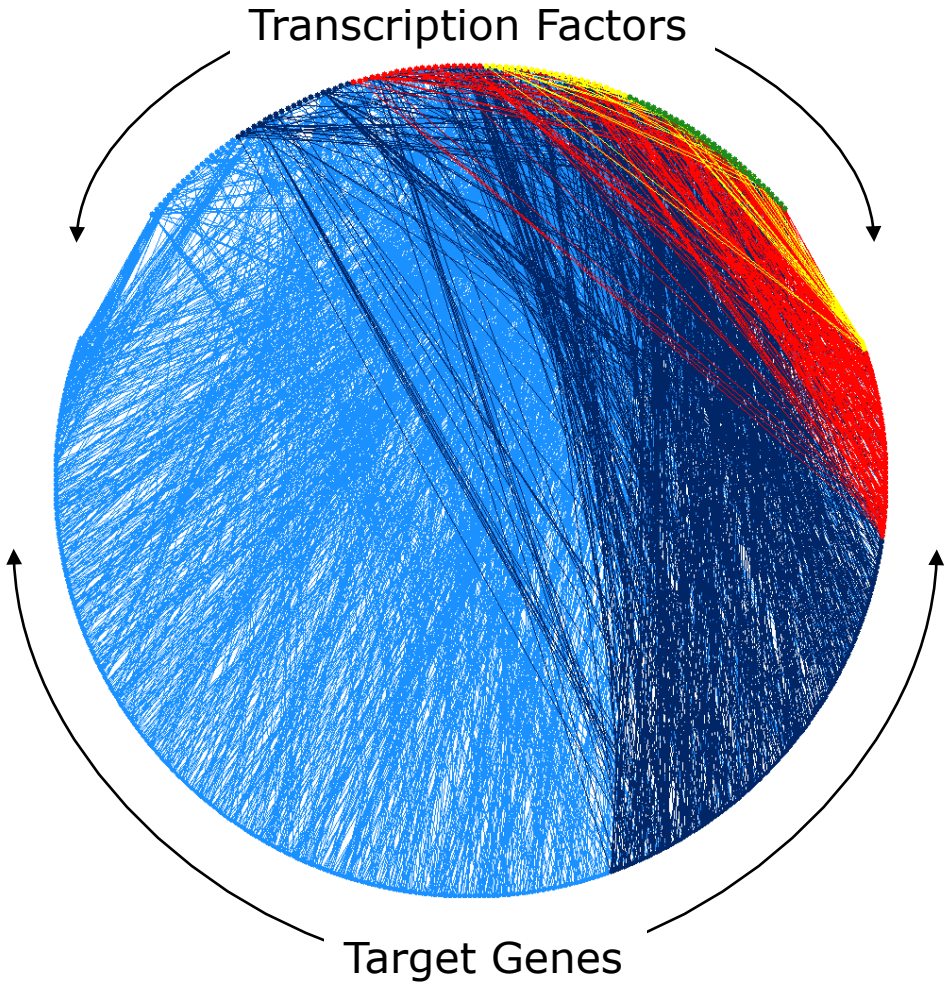




Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]

Yeast Regulatory Network: a platform for integration



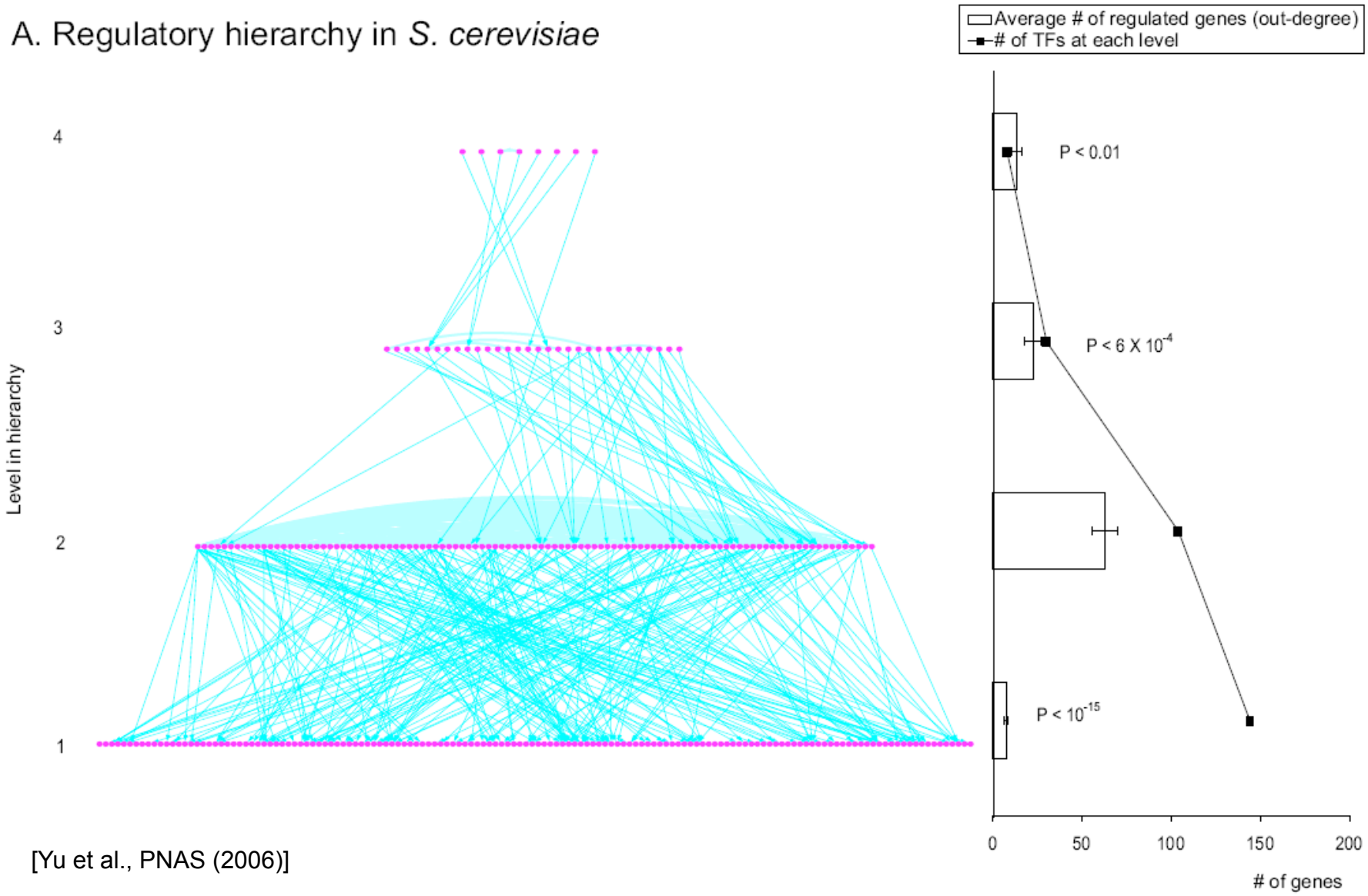
142 transcription factors

3,420 target genes

7,074 regulatory interactions

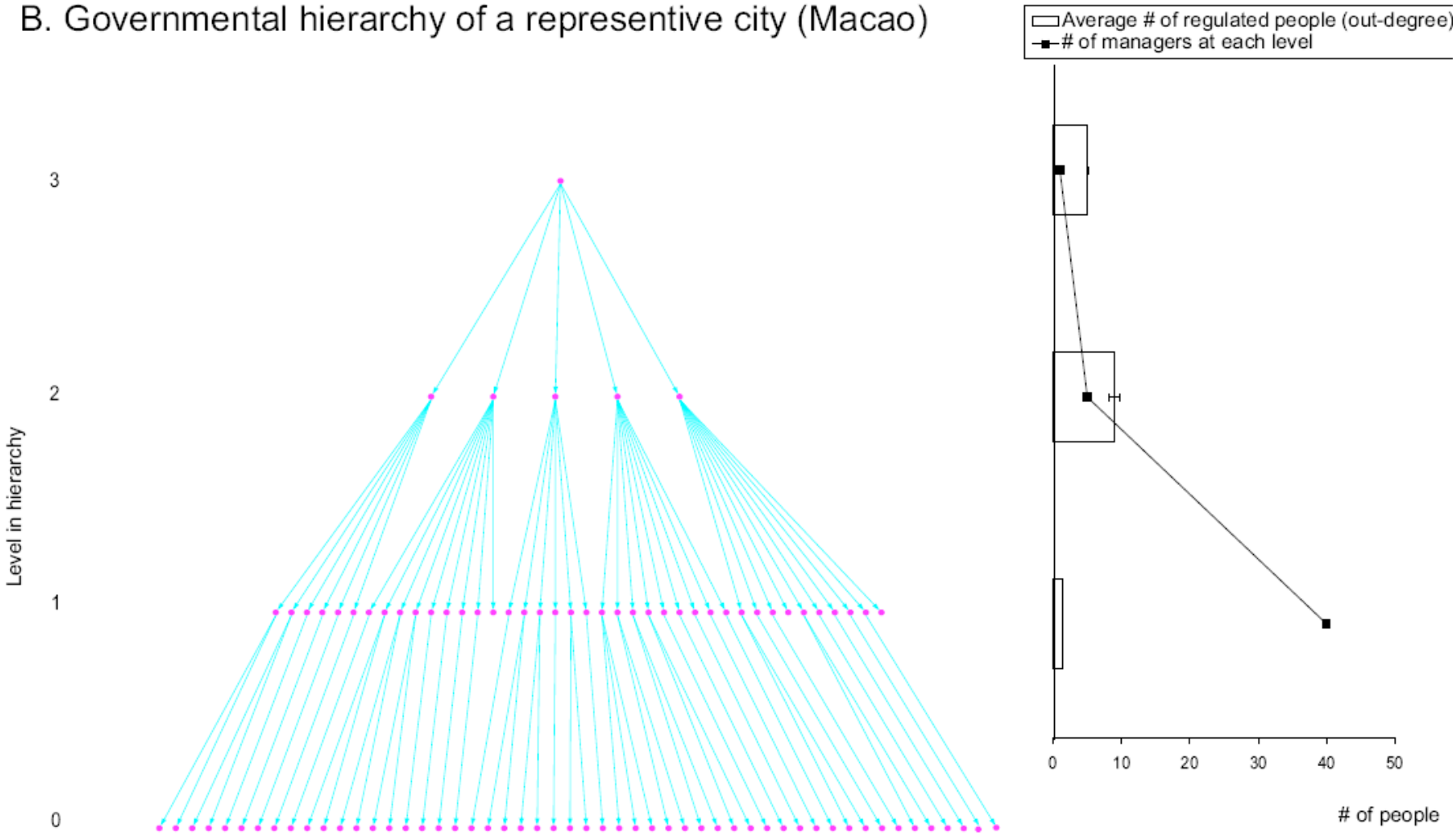
From integrating data from **Snyder et al...** TRANSFAC

Yeast Regulatory Hierarchy: the Middle-managers Rule

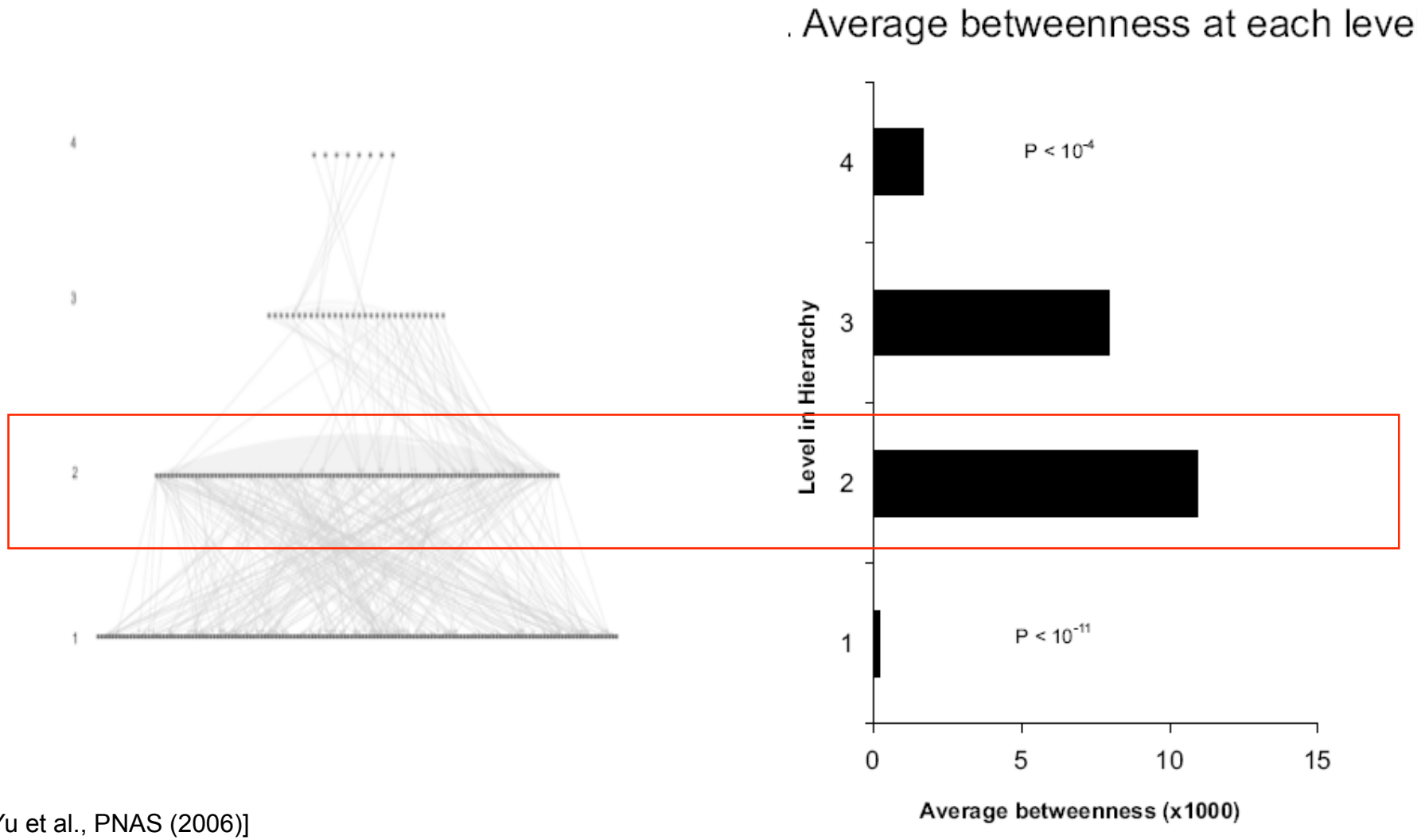


[Yu et al., PNAS (2006)]

Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers



Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



[Yu et al., PNAS (2006)]

Computational Ideas Used in Network Analysis

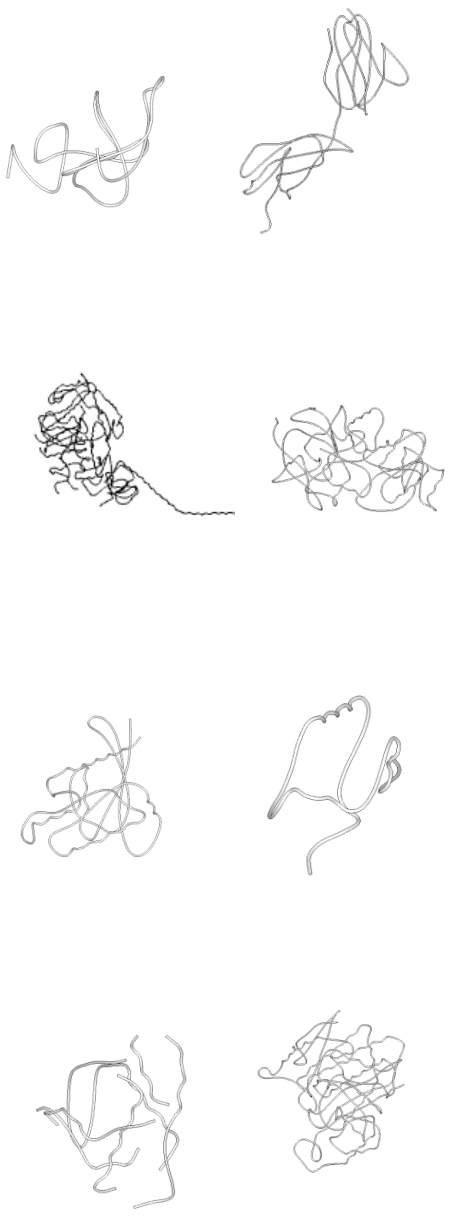
- Mining to find new connections (e.g. with Bayesian approaches) and to fuse together new types of data
- Interesting statistics that can be calculated on networks
- Ways of simplifying and visualizing complex "hairballs" of relationships (e.g. spectral methods such as SVD or CCA)

Research @ GersteinLab.org

- Human Genome Annotation (pseudogenes)
  - ◊ Characterizing the function of non-coding regions, focusing on protein fossils and novel transcriptionally active regions (Pseudogene.org + Tiling.GersteinLab.org)
- Molecular Networks
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe protein function on a large-scale (Networks.GersteinLab.org)
- Macromolecular motions
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)

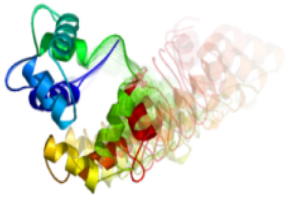
Surveying structural flexibility on a proteomic scale

- **Questions**
  - ◊ How do we describe a wide-range of structural variability in standard terms?
  - ◊ Can we develop simple models to explain constraints on protein flexibility?
  - ◊ What information about flexible hinge location is encoded in sequence?



MolMovDB.org

molmovdb.org



Database of Macromolecular Movements
with Associated Tools for Flexibility and Geometric Analysis

This describes the motions that occur in proteins and other macromolecules, particularly using movies. Associated with it are a variety of free software tools and servers for structural analysis. (Citation info)

Databases

The database is divided into two sections:

Protein Motions: Manually curated descriptions of conformational changes in hundreds of distinct proteins, with references and movie links.

Movies: Thousands of morphs of transitions between PDB files, viewable through a Java applet or as MPEG or GIF movies. Most of these are submissions to the Morph Server by database users.

Search database: Full-text

Servers

Morph Server: A web-based tool for generating and animating chemically realistic interpolations between two conformations. Now supports RNA, DNA, and multi-subunit complexes.

We have also developed new servers for analysis of helix interactions and normal modes of protein domains.

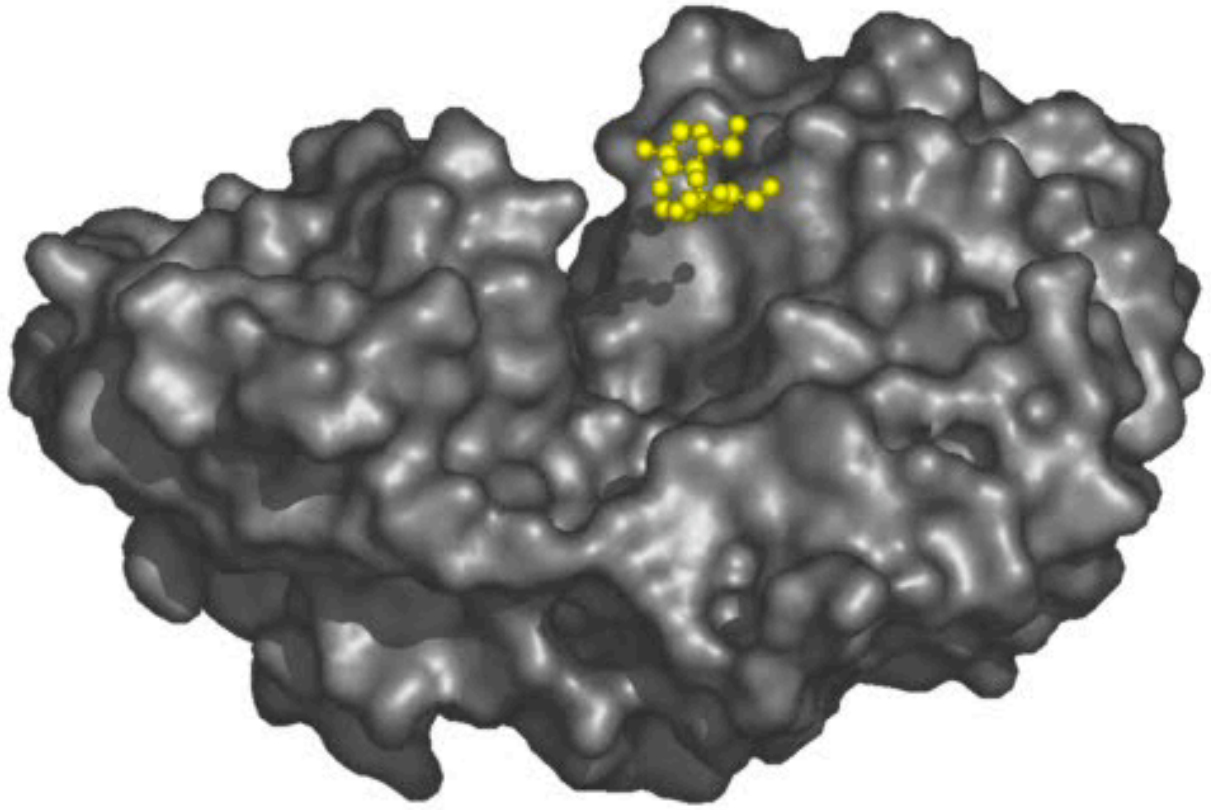
Hinge Prediction: The new HingeMaster server predicts hinge locations in single protein structures. The algorithm combines FlexOracle, TLSMD, StoneHinge and NSHP hinge predictors for maximum accuracy.

Studies and Resources

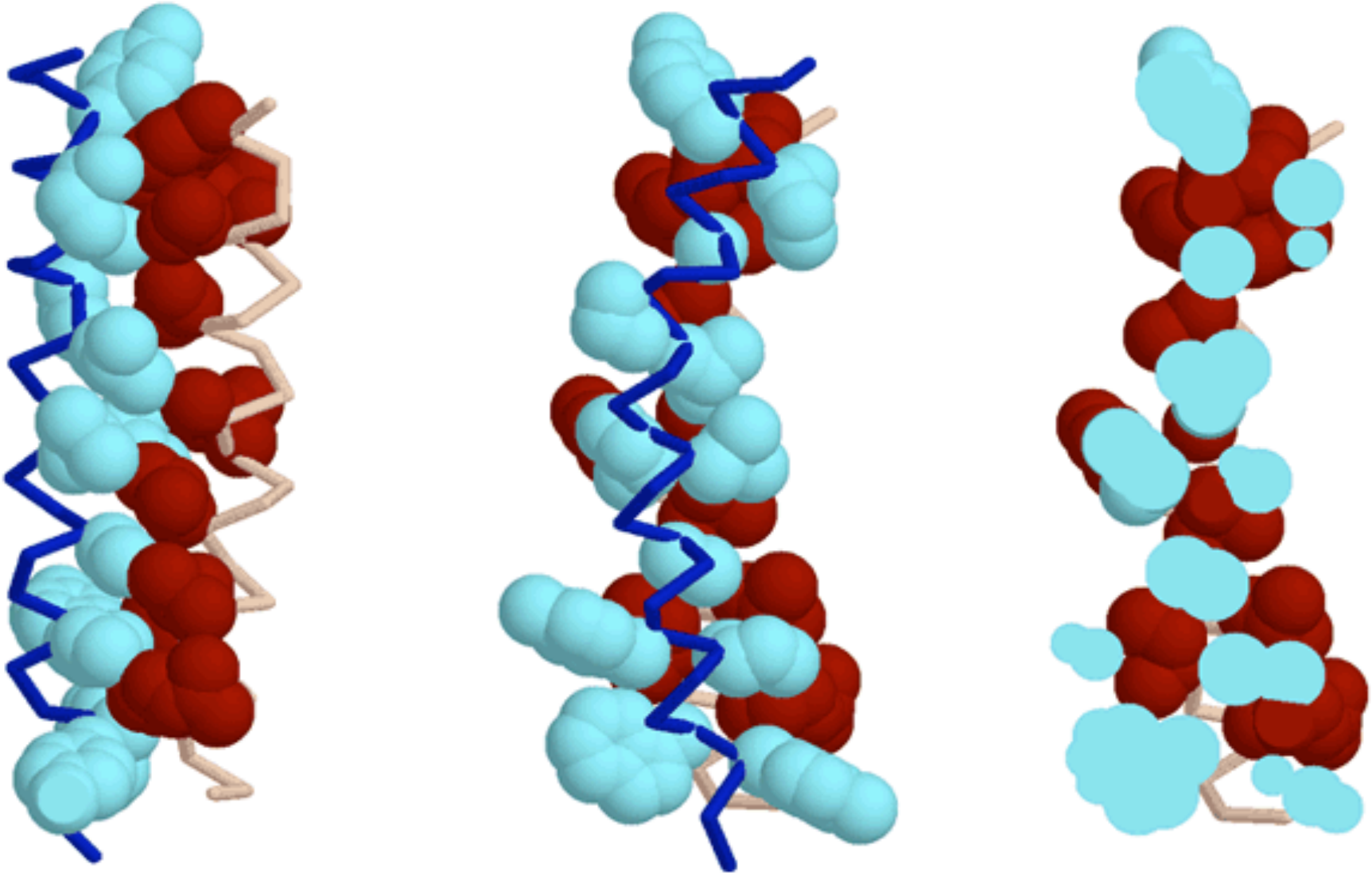
- The help page gives an overview of the database and citation information. There is also a Morph Server FAQ list.
- View a list of Gerstein lab publications describing applications of the database, or related to protein motions in general.
- Samuel Flores's 2005 article in "The Pharma Frontier" explains molecular motions to the public.
- We have individual pages focusing on membrane protein motions, large-scale protein motions, and protein-protein binding motions
-  Other useful programs for structure analysis are available for download.

Example "Morph": MBP

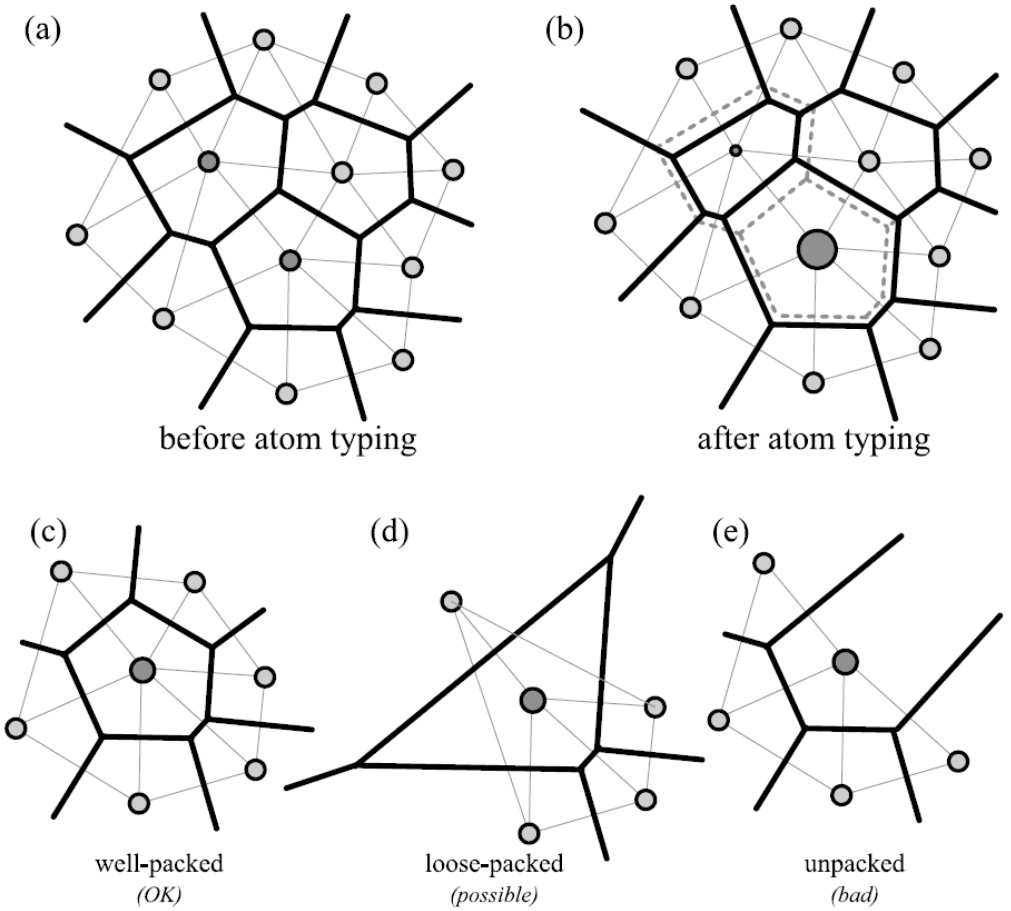
- 2 Known Crystal Structures
  (endpoints, not necessarily same seq.)
- Std. Geometric Stats. (from structure comparison)
- Pathway Interpolation



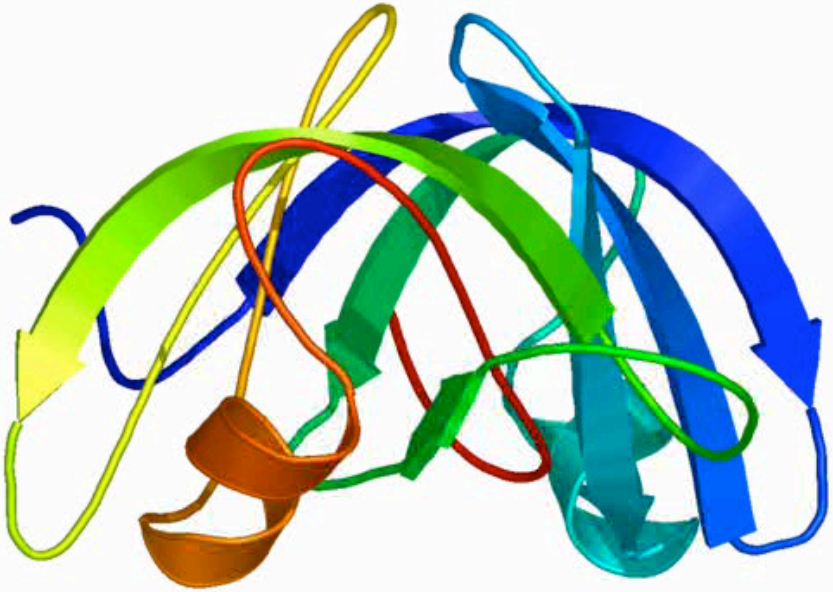
Interdigitating structure of protein interfaces constrains motion



Packing Tools - Voronoi software to calculate packing volumes (geometry.molmovdb.org)



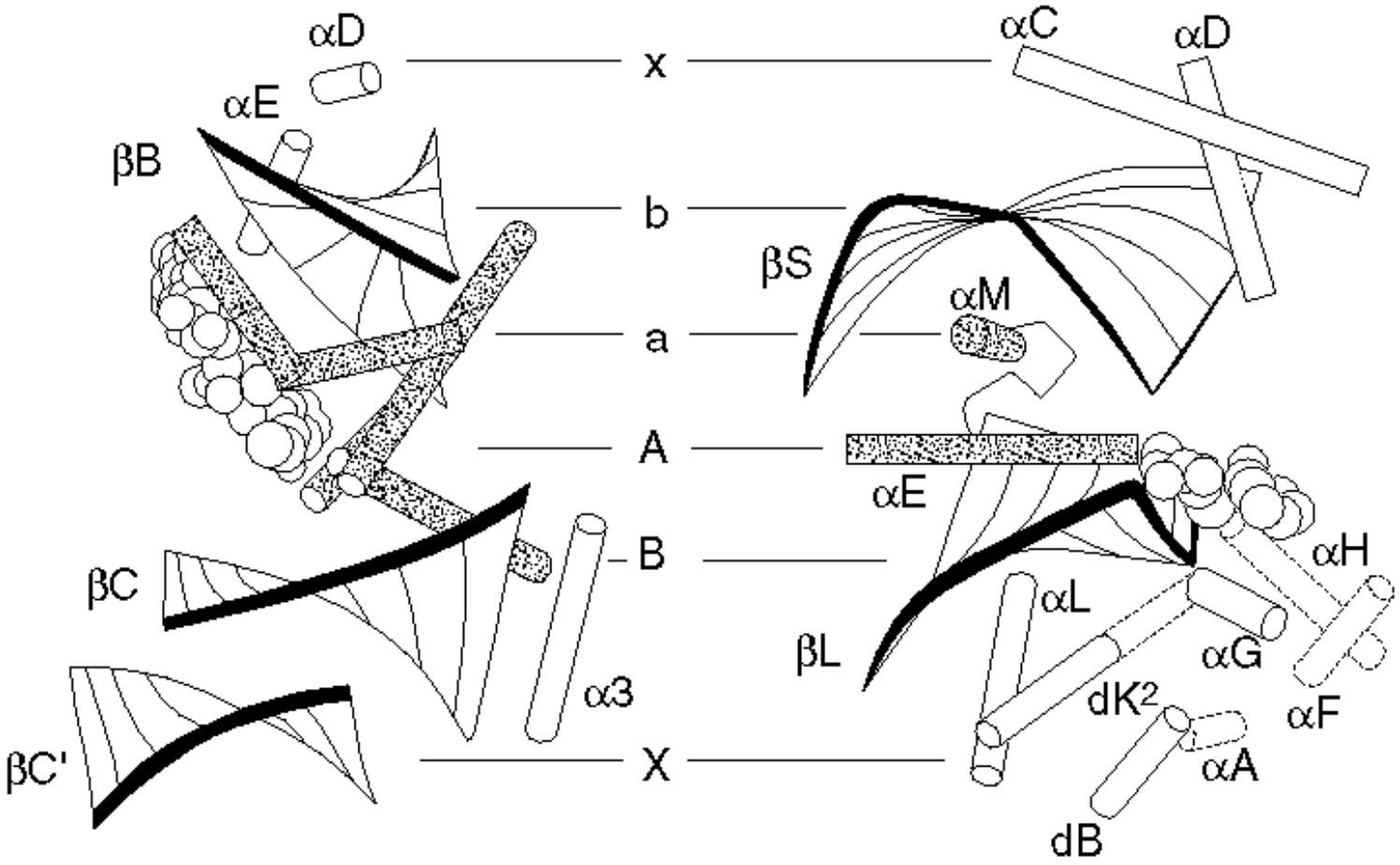
Small Shearing Domain Motions: Molybdenum-binding protein & GAPDH



[Lawson]

[Wonacott]

Proteins With Shear Motions are Often Divided into Layers

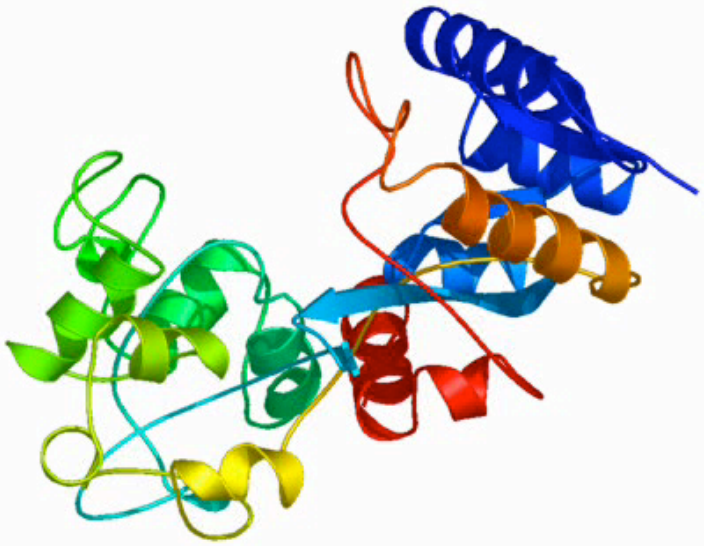


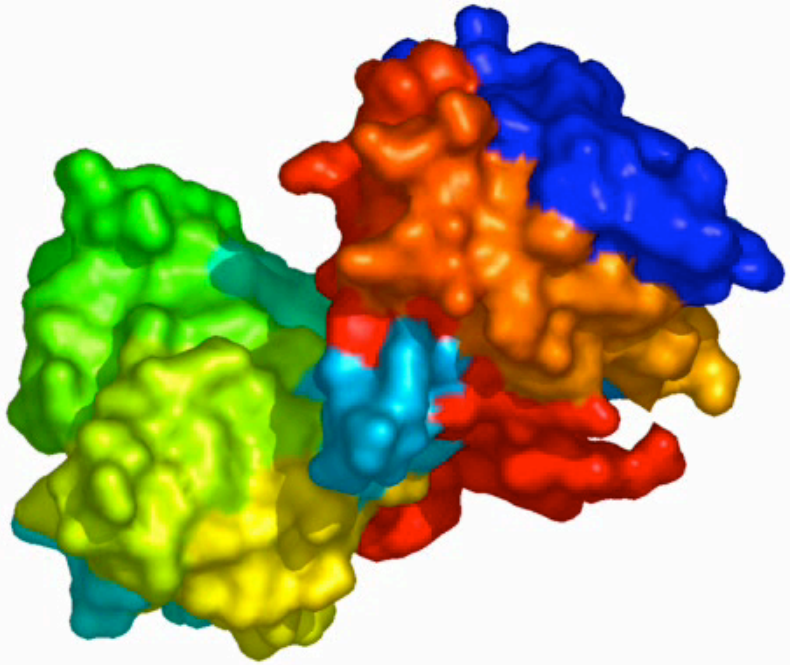
GAPDH

Hexokinase

[Steitz]

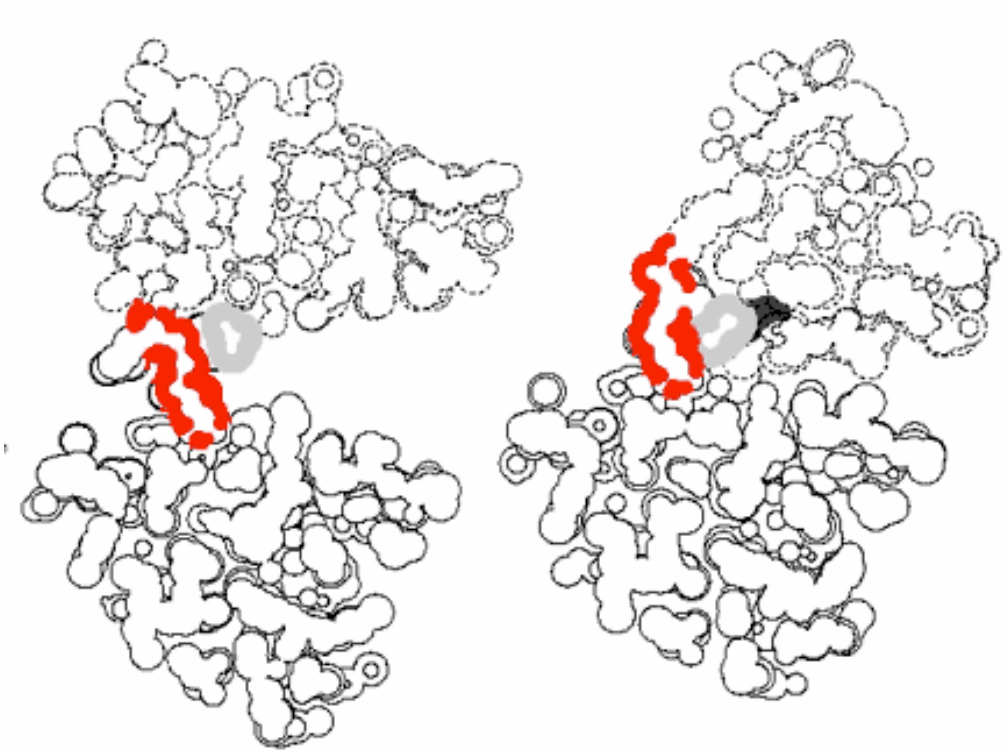
Transferrin: Interdomain Hinges





[Baker]

Transferrin hinge involves absence of steric constraints (continuously maintained interfaces), esp. at hinge



Computational Ideas Used in Molecular Motions

- How to simulate realistic motion of a molecule (efficient computation)
- How to represent surfaces and volumes and calculate quantities such as packing efficiency

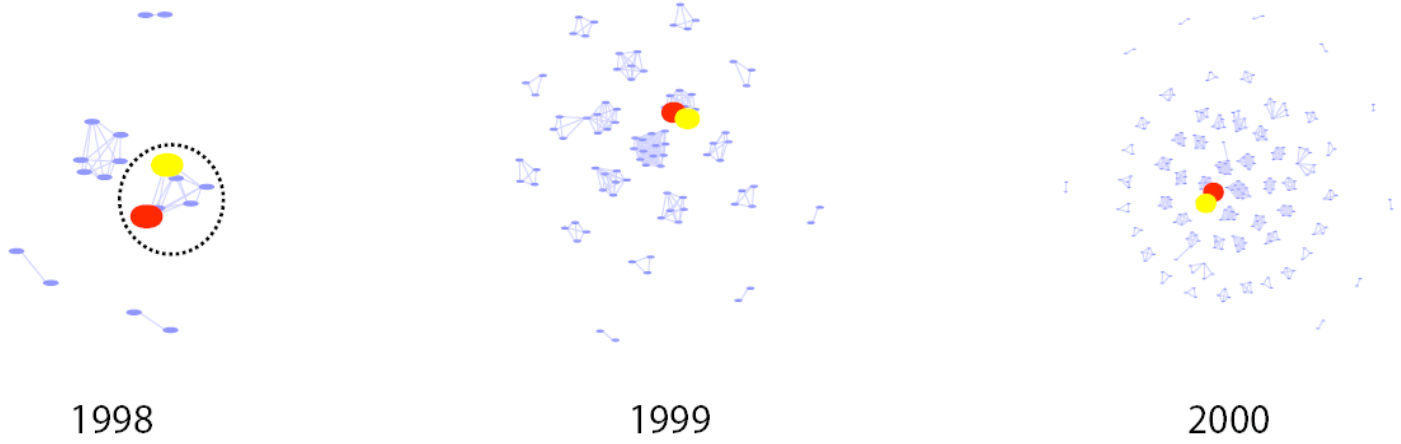
Proteomics Research @ GersteinLab.org

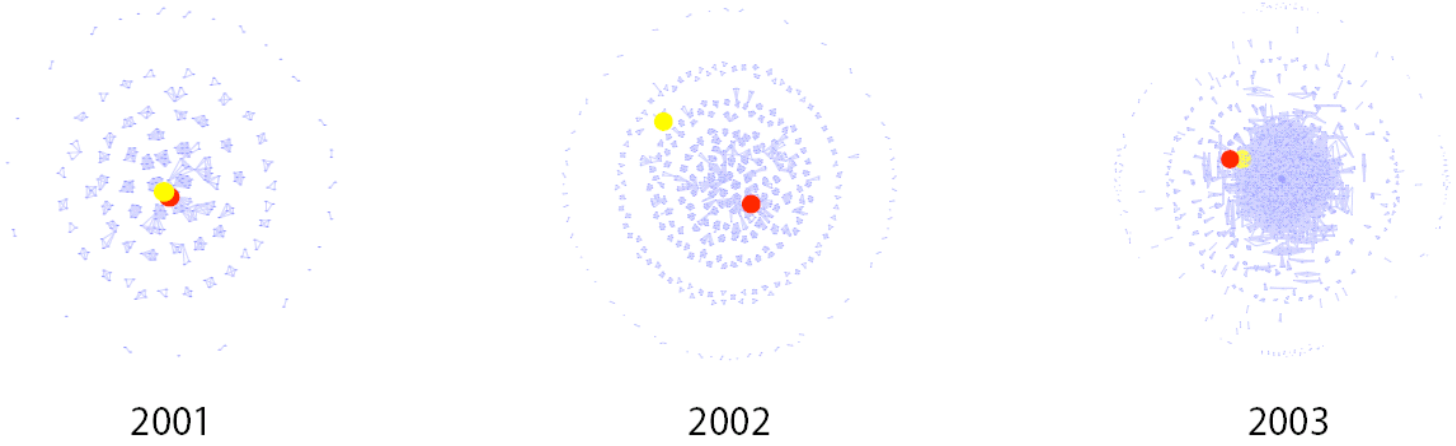
- Macromolecular motions
  - ◊ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)

- Molecular Networks
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe protein function on a large-scale (Networks.GersteinLab.org)

- Human Genome Annotation (protein fossils)
  - ◊ Characterizing the function of non-coding regions, focusing on protein fossils and novel transcriptionally active regions (Pseudogene.org + Tiling.GersteinLab.org)

RNAi: Birth of a Field in the Literature Culmin-ating in the 2006 Nobel





Andrew Fire Craig Mello

Source:
Gerstein & Douglas.
PLoS Comp. Bio. 3:e80
(2007)
PubNet.GersteinLab.org