All the Data and Publications from Science on Web: A Vision for Harnessing this to Study the Structure of Science

Mark B Gerstein
Yale (Comp. Bio. & Bioinformatics)

NSF Workshop on Knowledge Management and Visualization Tools

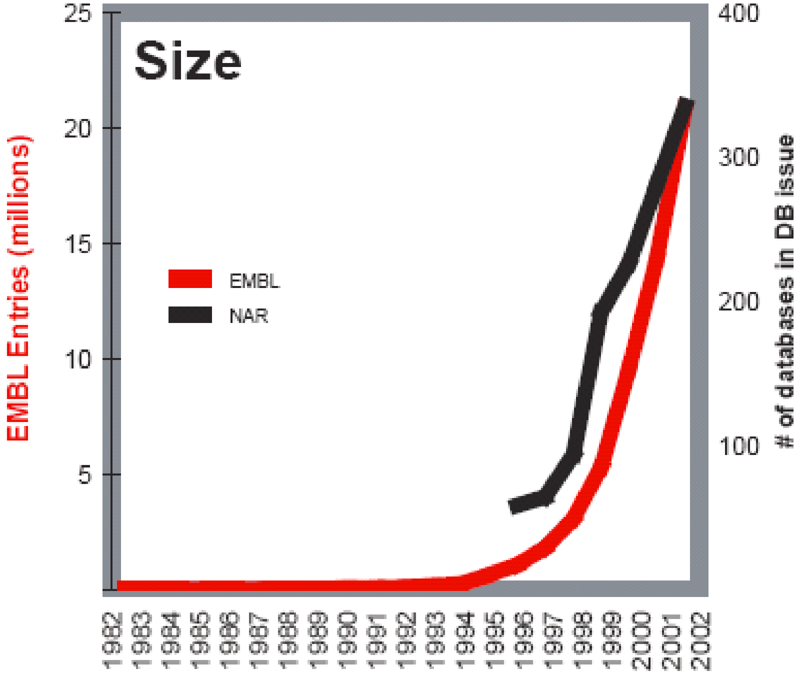
2008.03.11, 09:30-10:00

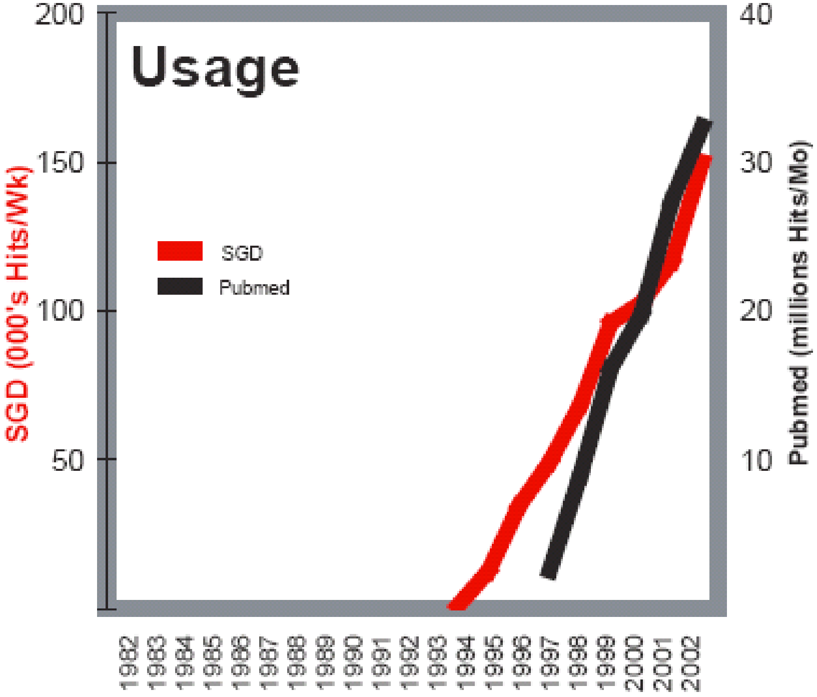
Slides downloadable from Lectures.GersteinLab.org

(Please read permissions statement.)

()

Rapid growth in DBs in biology, changing the landscape





[Greenbaum & Gerstein, Nat. Biotech. ('03)]

Current Situation Facing DBs and Journals: A Changing Landscape

- Distinctions Blurring
  - ◊ Reading Journals via queries
    - Reading DB entries
  - ◊ Towards reading literature with computers
    - Mining text and correlating papers

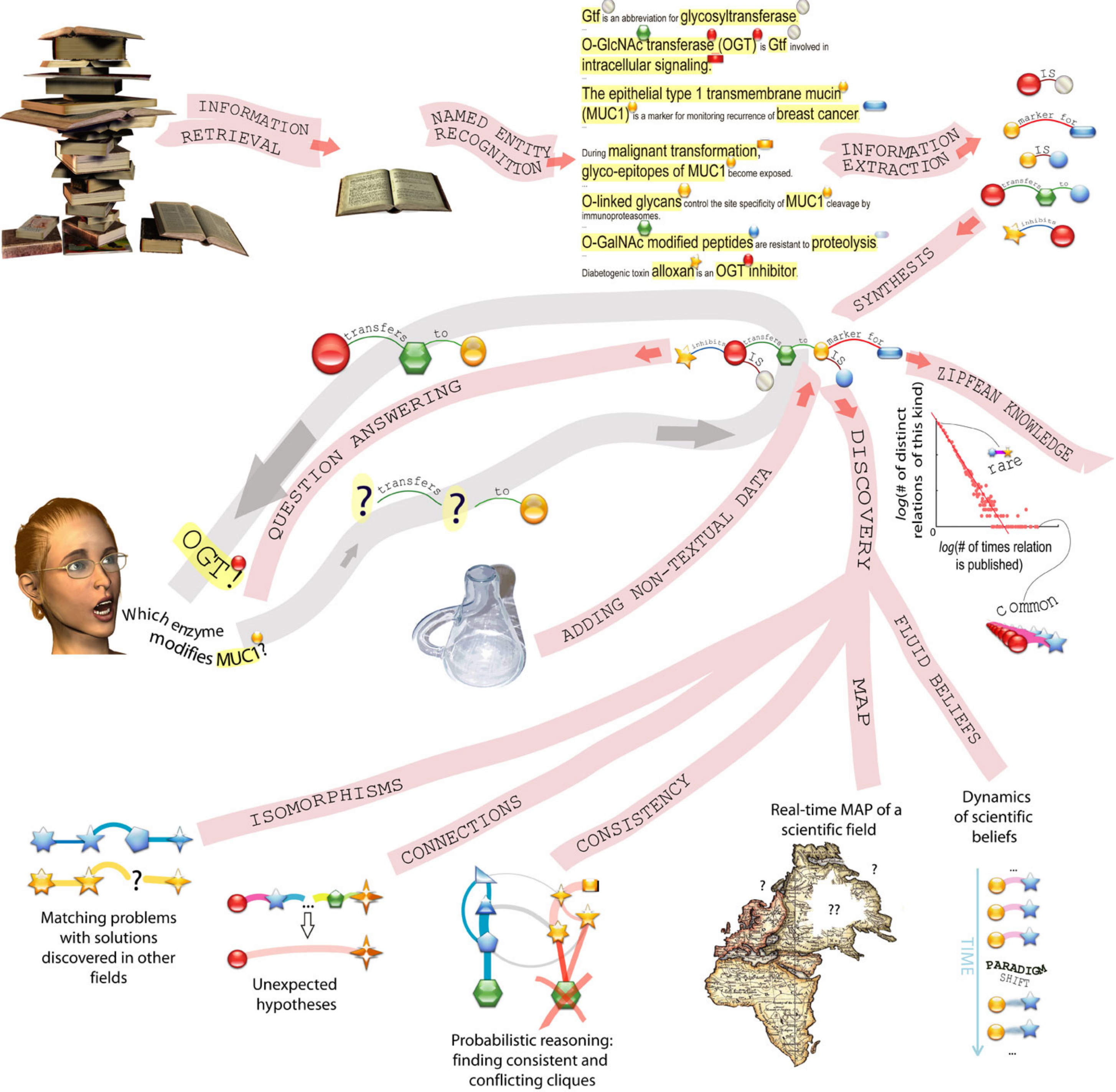
- Biology as a science of heterogeneous *facts*
  - ◊ Well-suited to database storage

[Gerstein, Bioinformatics ('99); Gerstein & Junker. Nature Yearbook ('02)]

The Challenge

- Volume and growth of publications
- Hard to keep up with field
- Missed opportunities in connections between fields
- Harness the power of technology to help scientists share information?
- Deeper use of computer technology in handling scientific texts?
- Discover new scientific relationships

Overall Process of Web Mining



Gtf is an abbreviation for glycosyltransferase

...

O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.

The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer.

During malignant transformation, glyco-epitopes of MUC1 become exposed.

...

O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.

...

O-GalNAc modified peptides are resistant to proteolysis.

Diabetogenic toxin alloxan is an OGT inhibitor.

Which enzyme modifies MUC1?

Matching problems with solutions discovered in other fields

Unexpected hypotheses

Probabilistic reasoning: finding consistent and conflicting cliques

Real-time MAP of a scientific field

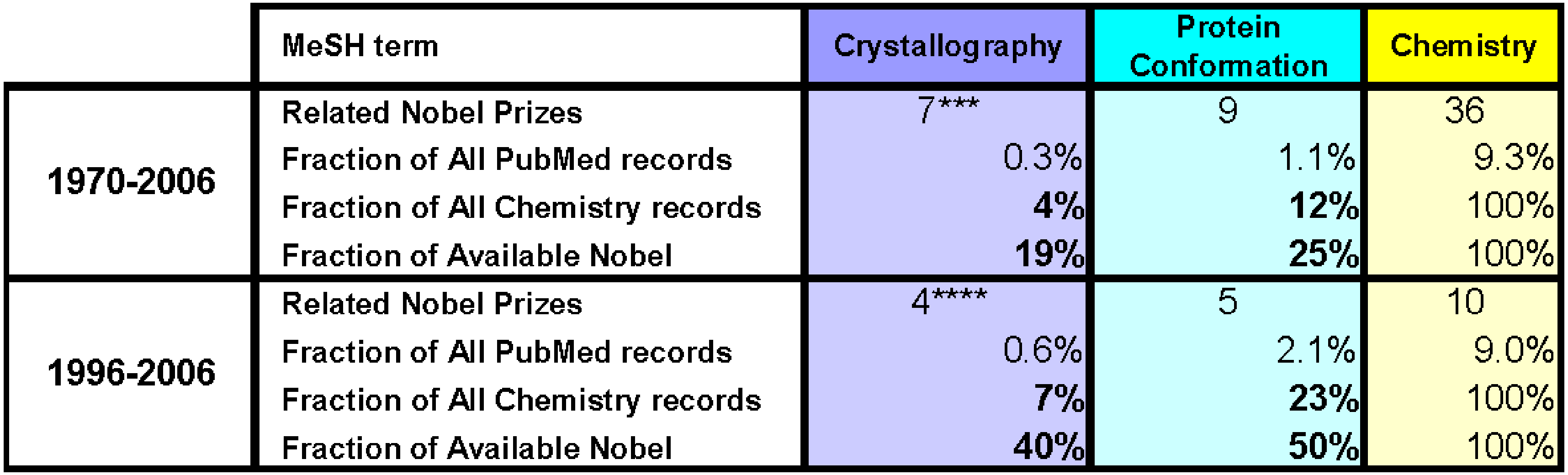
Dynamics of scientific beliefs

[Rzhetsky et al, Cell ('08, submitted)]

Examples Illuminating Current State of Affairs:

Mining Simple Term Occurrence Statistics to Understand and Justify Directions in Science

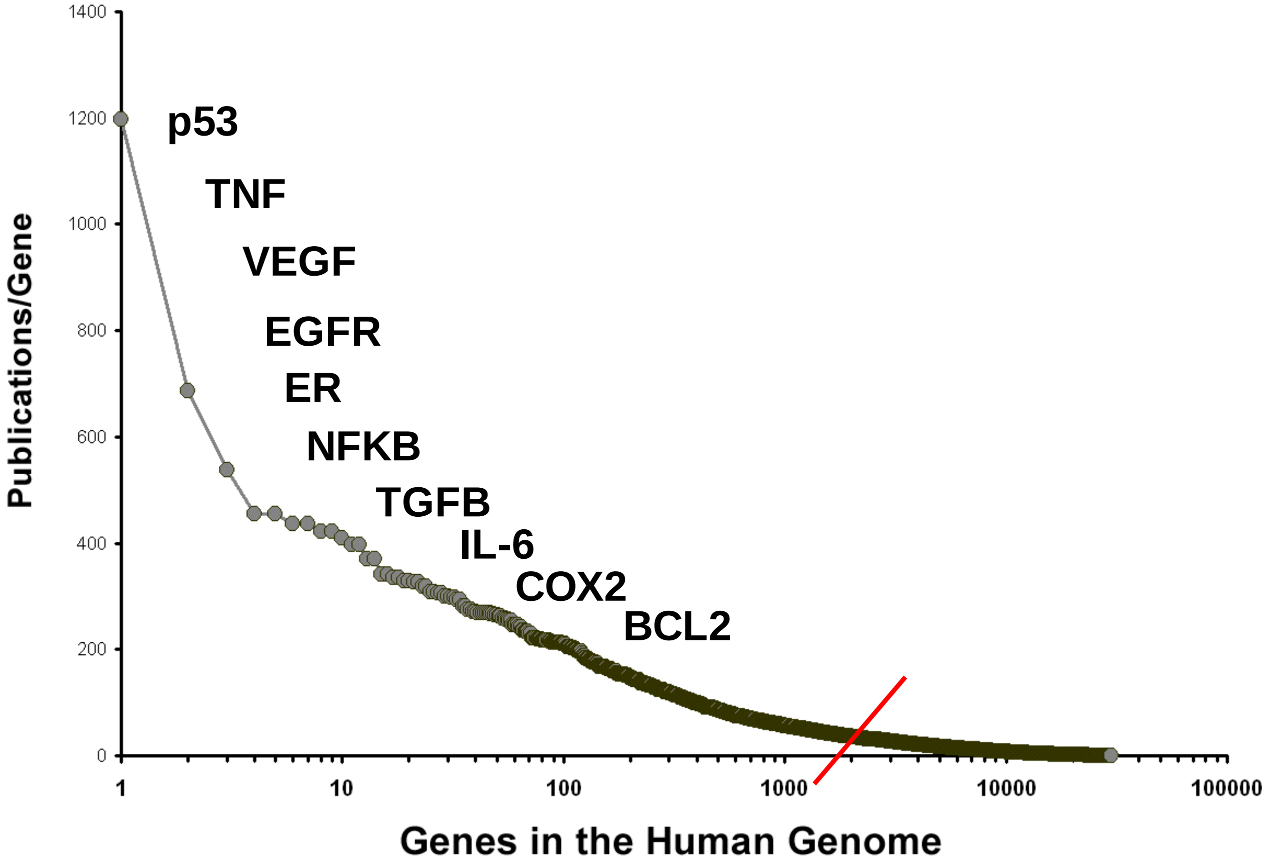
Over-representation of crystallography among the Nobel Prizes, highlighted by the 2006 Nobels



	MeSH term	Crystallography	Protein Conformation	Chemistry
1970-2006	Related Nobel Prizes	7***	9	36
	Fraction of All PubMed records	0.3%	1.1%	9.3%
	Fraction of All Chemistry records	**4%**	**12%**	100%
	Fraction of Available Nobel	**19%**	**25%**	100%
1996-2006	Related Nobel Prizes	4****	5	10
	Fraction of All PubMed records	0.6%	2.1%	9.0%
	Fraction of All Chemistry records	**7%**	**23%**	100%
	Fraction of Available Nobel	**40%**	**50%**	100%

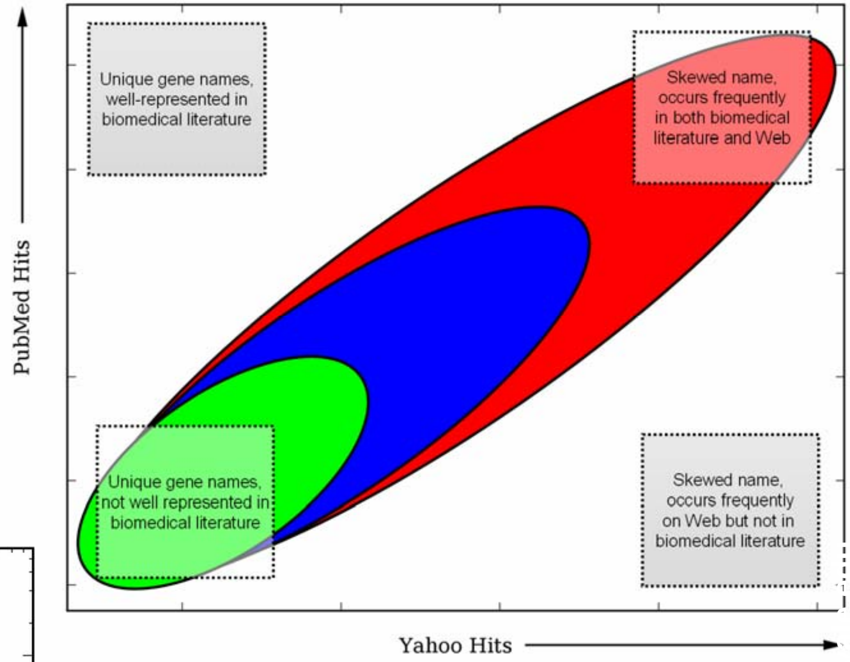
[Seringhaus & Gerstein, Science (2007)]

The current state of mammalian gene annotation: a rationale for data driven research



Adapted from Su and Hogenesch, Genome Biology, 2007

Gene Name Skew

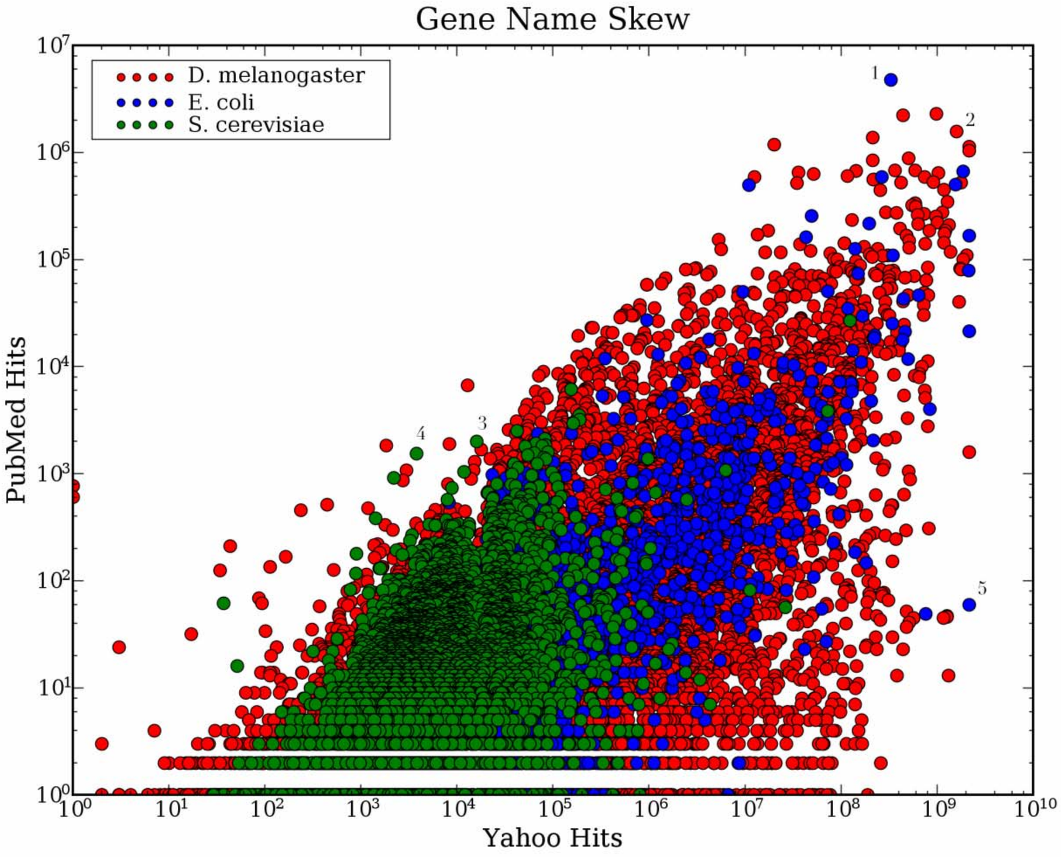


Unique gene names, well-represented in biomedical literature

Skewed name, occurs frequently in both biomedical literature and Web

Unique gene names, not well represented in biomedical literature

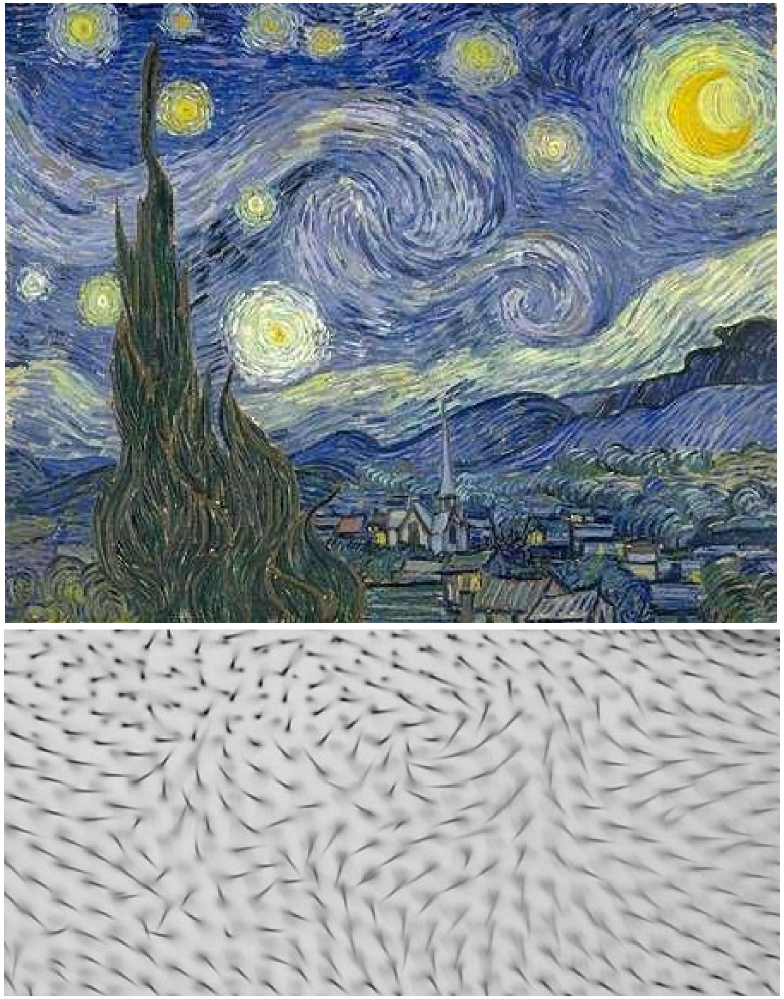
Skewed name, occurs frequently on Web but not in biomedical literature



[Seringhaus et al. GenomeBiology (2008)]

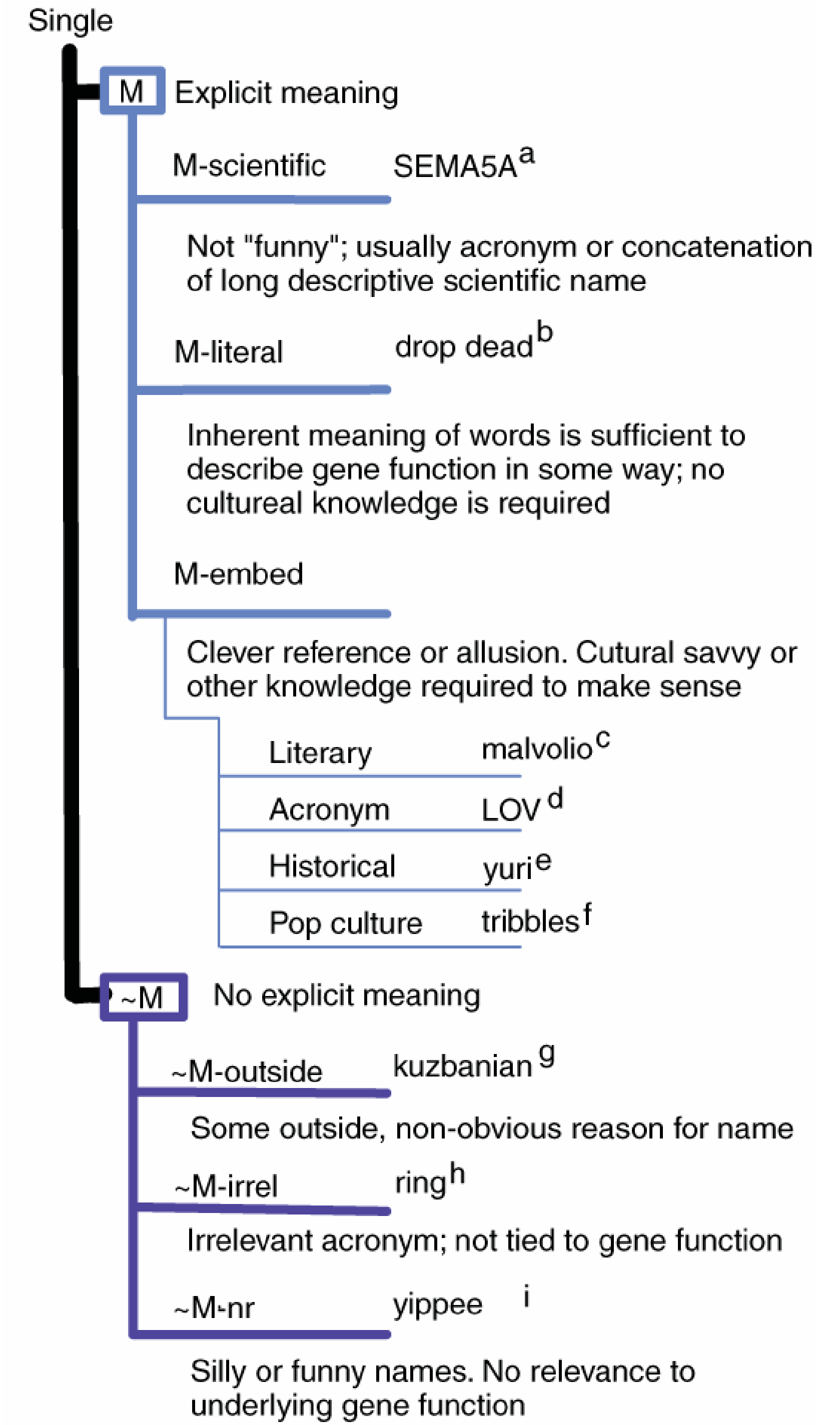
Ex. Naming Issue: Starry Night

- **Starry night** (P Adler, ’94)



[Seringhaus et al. GenomeBiology (2008)]

Naming Pathologies: Related to Single Genes



Single

M Explicit meaning

M-scientific SEMA5A[a]

Not "funny"; usually acronym or concatenation of long descriptive scientific name

M-literal drop dead[b]

Inherent meaning of words is sufficient to describe gene function in some way; no cultureal knowledge is required

M-embed

Clever reference or allusion. Cutural savvy or other knowledge required to make sense

Literary malvolio[c]

Acronym LOV[d]

Historical yuri[e]

Pop culture tribbles[f]

~M No explicit meaning

~M-outside kuzbanian[g]

Some outside, non-obvious reason for name

~M-irrel ring[h]

Irrelevant acronym; not tied to gene function

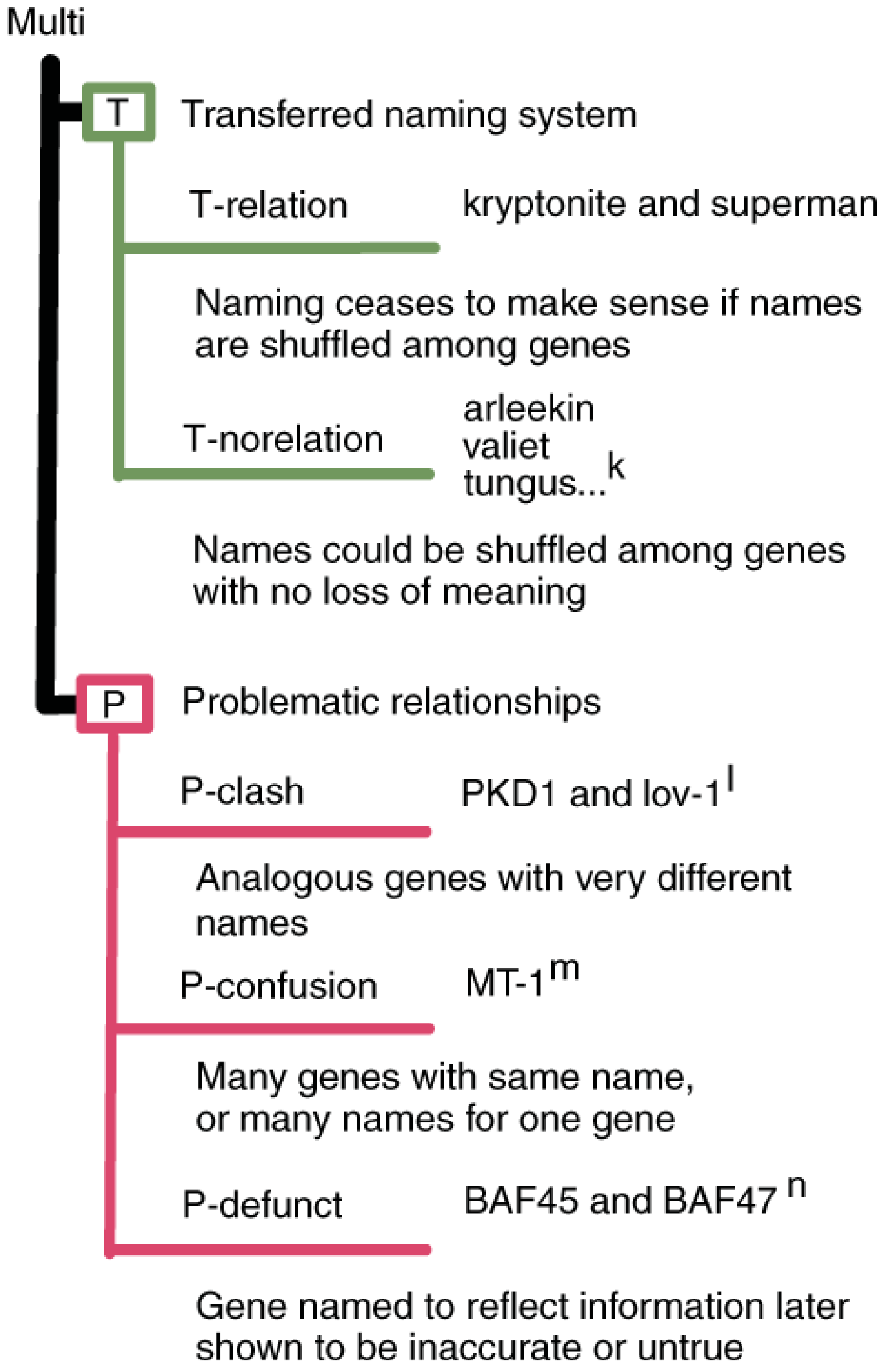
~M-nr yippee[i]

Silly or funny names. No relevance to underlying gene function

(b) drop dead: flies with mutations in drop dead die rapidly after their brain rapidly deteriorates. (c) malvolio: gene needed for normal taste behaviour. Malvolio in Shakespeare's Twelfth Night tasted "with distempered appetite". (d) LOV: light, oxygen, or voltage (LOV) family of blue-light photoreceptor domains. (e) yuri: this gene was discovered on the anniversary of Yuri Gagarin's space flight. Mutants have problems with gravitaxis and cannot stay aloft. (f) tribbles: cells divide uncontrollably, like the eponymous Star Trek characters. (g) kuzbanian: mutants have uncontrollable bristle growth. Koozbanians are alien Muppets with uncontrollable hair growth; spelling was changed to avoid copyright infringement. (h) ring: really interesting new gene. (i) yippee: a graduate student's reaction on cloning the gene

[Seringhaus et al. GenomeBiology (2008)]

Naming Pathologies: Involving Multiple Gene Names



Multi

T Transferred naming system

T-relation kryptonite and superman

Naming ceases to make sense if names are shuffled among genes

T-norelation arleekin valiet tungus...[k]

Names could be shuffled among genes with no loss of meaning

P Problematic relationships

P-clash PKD1 and lov-1[l]

Analogous genes with very different names

P-confusion MT-1[m]

Many genes with same name, or many names for one gene

P-defunct BAF45 and BAF47[n]

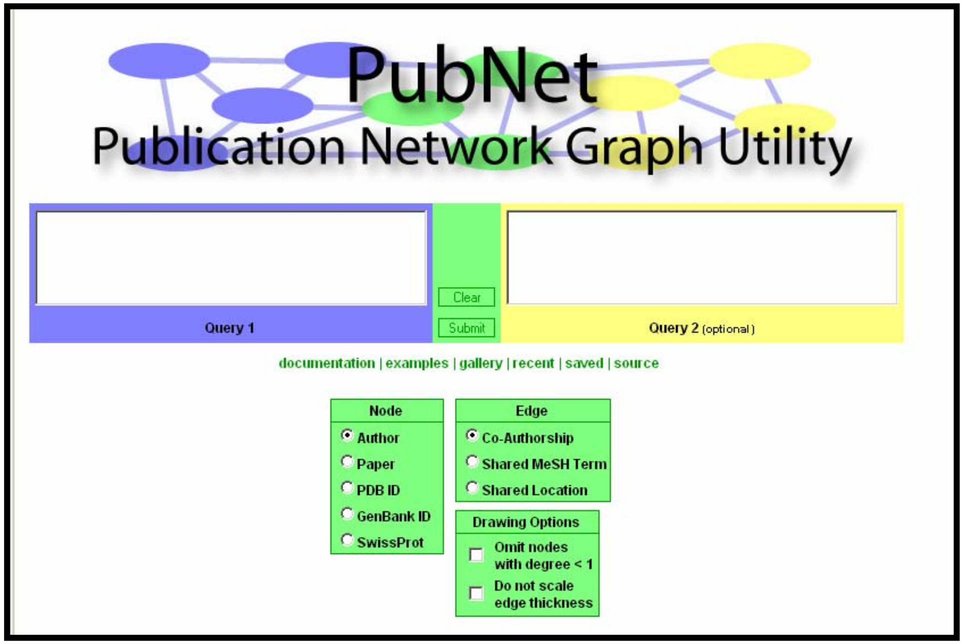
Gene named to reflect information later shown to be inaccurate or untrue

(j) kryptonite and superman: the kryptonite mutation suppresses the function of the SUPERMAN gene. (k) arleekin, valient, tungus: mutations in arleekin, valient, tungus and 29 other genes affect long-term memory. Named after Pavlov's dogs. (l) PKD1 (human) and lov-1 (worm): these are homologs, although their names do not suggest it. (m) MT-1: this label can refer to at least 11 different human genes. (n) BAF45 and BAF47: names for the same gene, reflecting a revision of the molecular weight of product.

[Seringhaus et al. GenomeBiology (2008)]

Examples Illuminating Current State of Affairs:

Using Network Representations to Make Maps of Science -- Studying the Publication Patterns of Genomics Consortia


PubNet
Publication Network Graph Utility
Clear
Submit
Query 1
Query 2 (optional)
documentation | examples | gallery | recent | saved | source

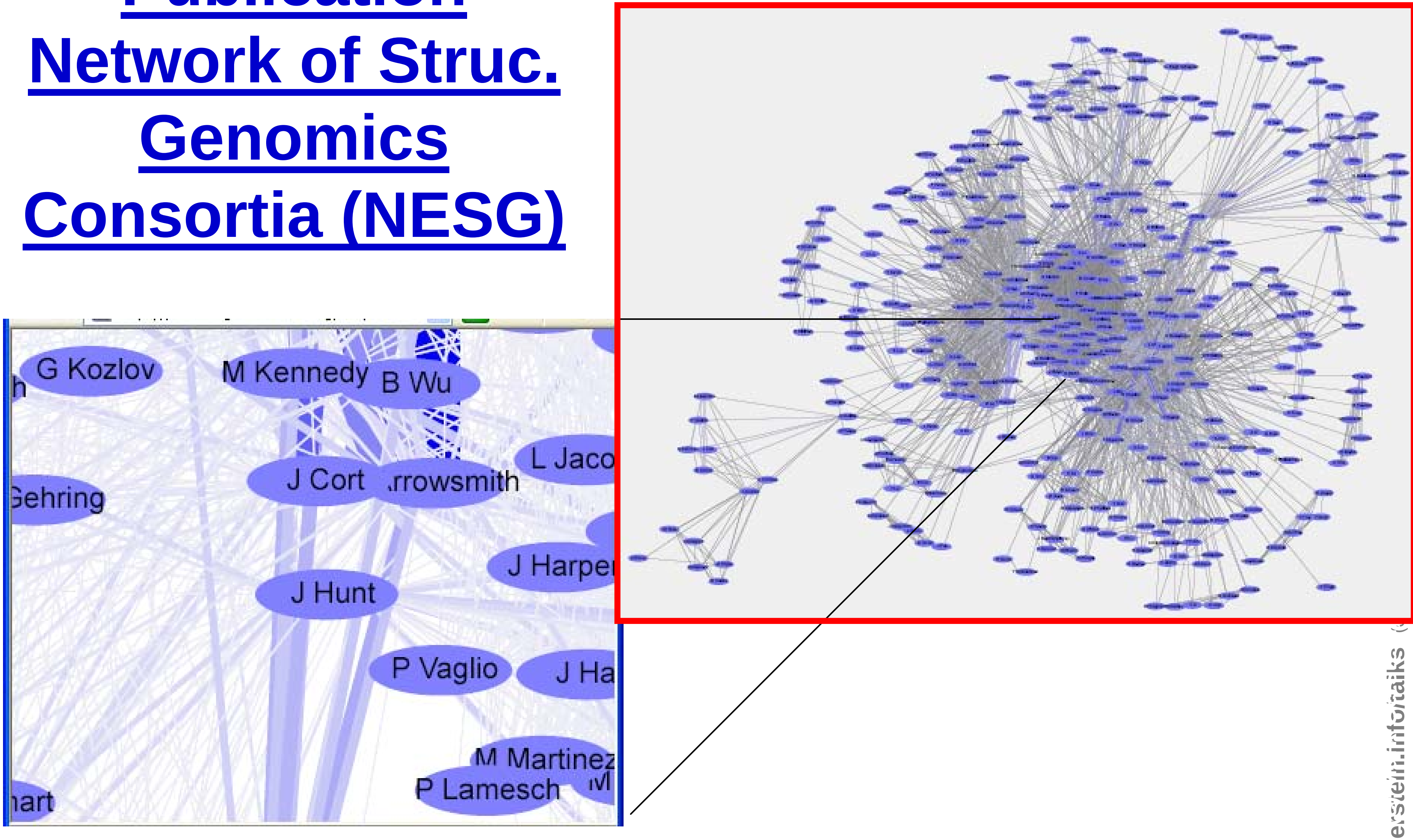
Node
- ● Author
- ○ Paper
- ○ PDB ID
- ○ GenBank ID
- ○ SwissProt

Edge
- ● Co-Authorship
- ○ Shared MeSH Term
- ○ Shared Location

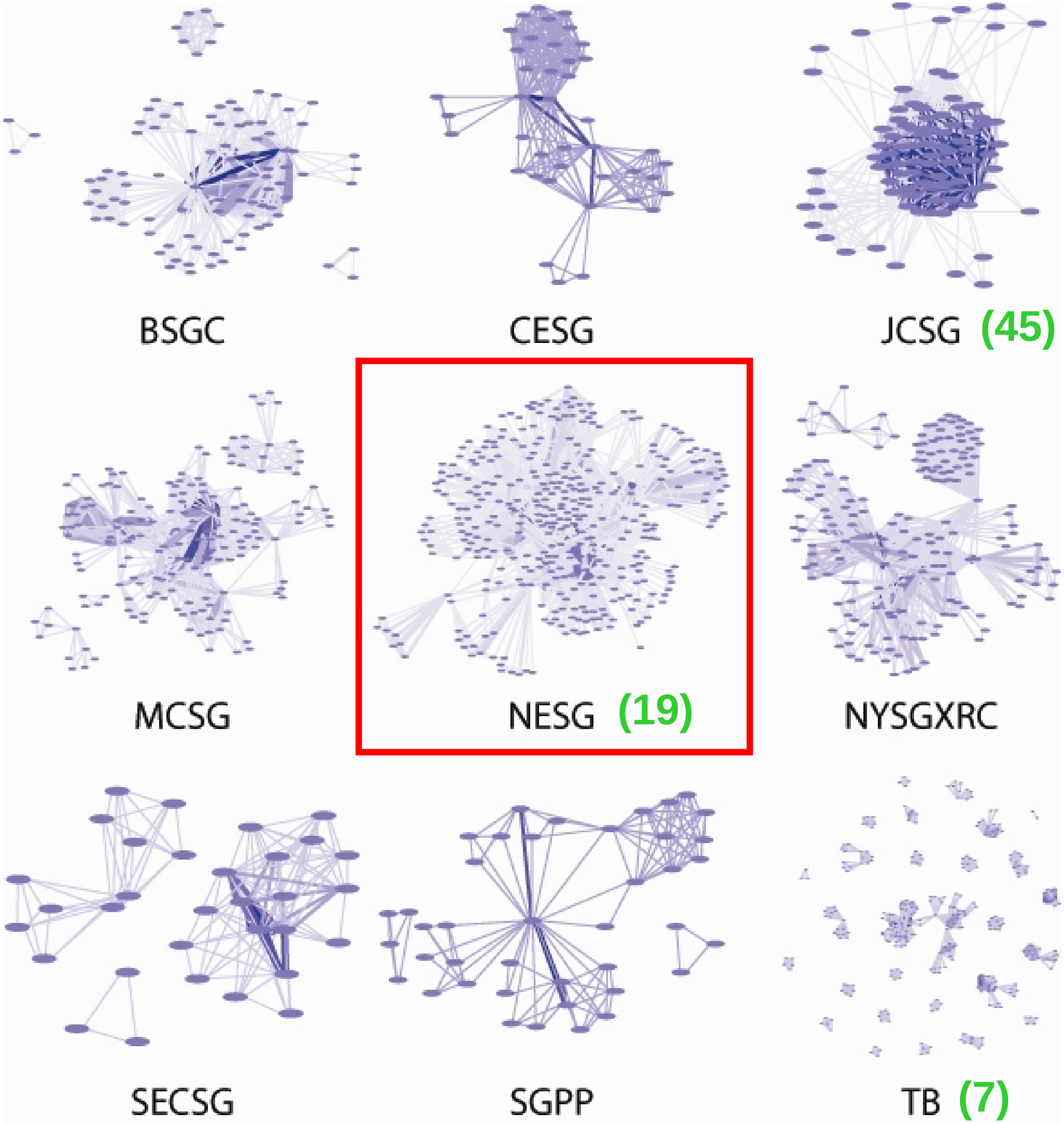
Drawing Options
- □ Omit nodes with degree < 1
- □ Do not scale edge thickness

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

Co-Authorship Publication Network of Struc. Genomics Consortia (NESG)



[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

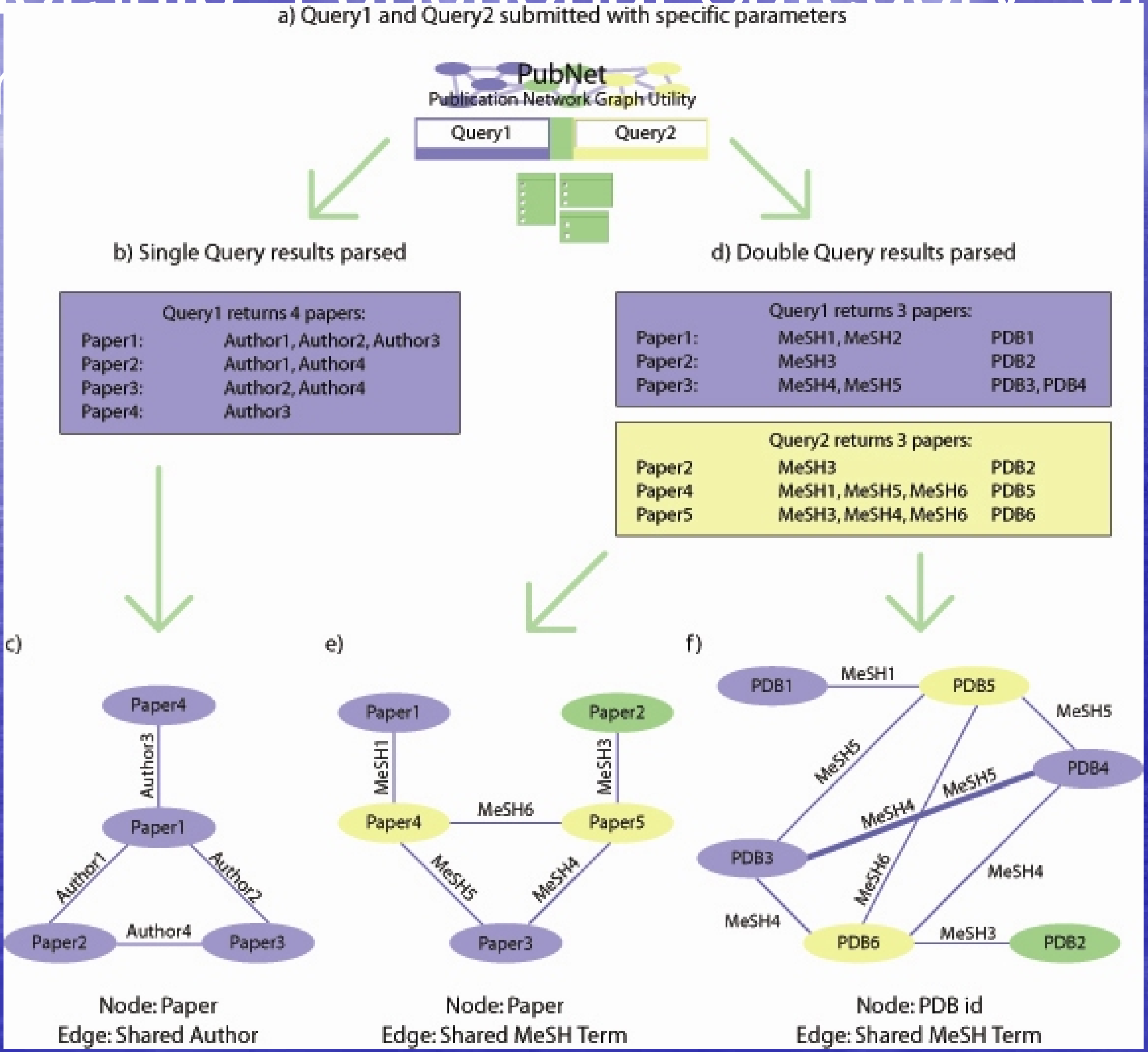


Co-authorship Networks comparing the 9 NIH Structural Genomics Centers

Average Degree

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

More Complex Literature Networks: Linking Proteins, Papers or Papers by Authorship, Functional Category, or Location



a) Query1 and Query2 submitted with specific parameters

b) Single Query results parsed

Query1 returns 4 papers:

Paper1:	Author1, Author2, Author3
Paper2:	Author1, Author4
Paper3:	Author2, Author4
Paper4:	Author3

d) Double Query results parsed

Query1 returns 3 papers:

Paper1:	MeSH1, MeSH2	PDB1
Paper2:	MeSH3	PDB2
Paper3:	MeSH4, MeSH5	PDB3, PDB4

Query2 returns 3 papers:

Paper2	MeSH3	PDB2
Paper4	MeSH1, MeSH5, MeSH6	PDB5
Paper5	MeSH3, MeSH4, MeSH6	PDB6

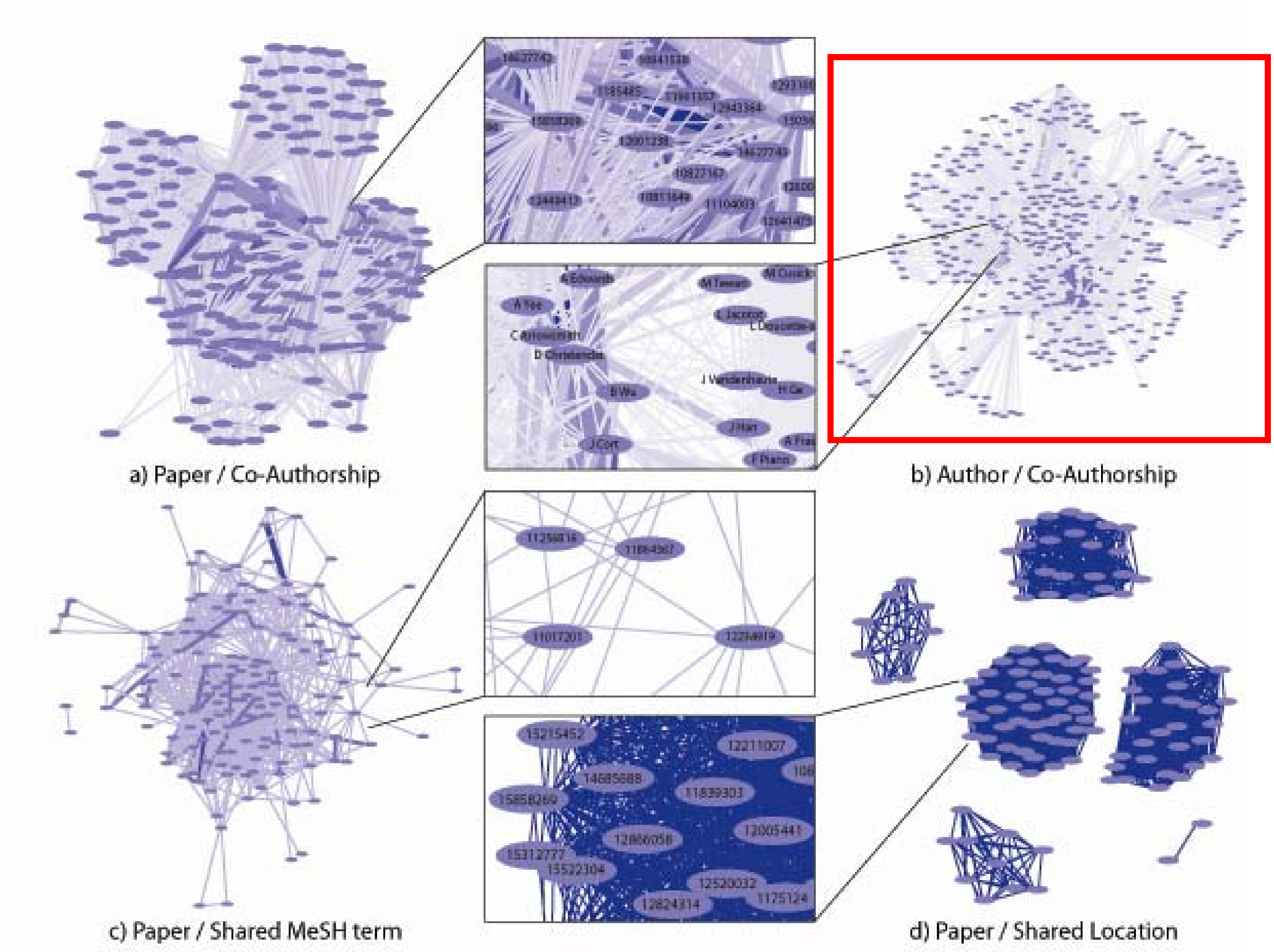
Node: Paper
Edge: Shared Author

Node: Paper
Edge: Shared MeSH Term

Node: PDB id
Edge: Shared MeSH Term

(c) Mark Gerstein, 2002, Yale, bioinfo.mbb.yale.edu

Different Representations of Publication Network of NESG



a) Paper / Co-Authorship

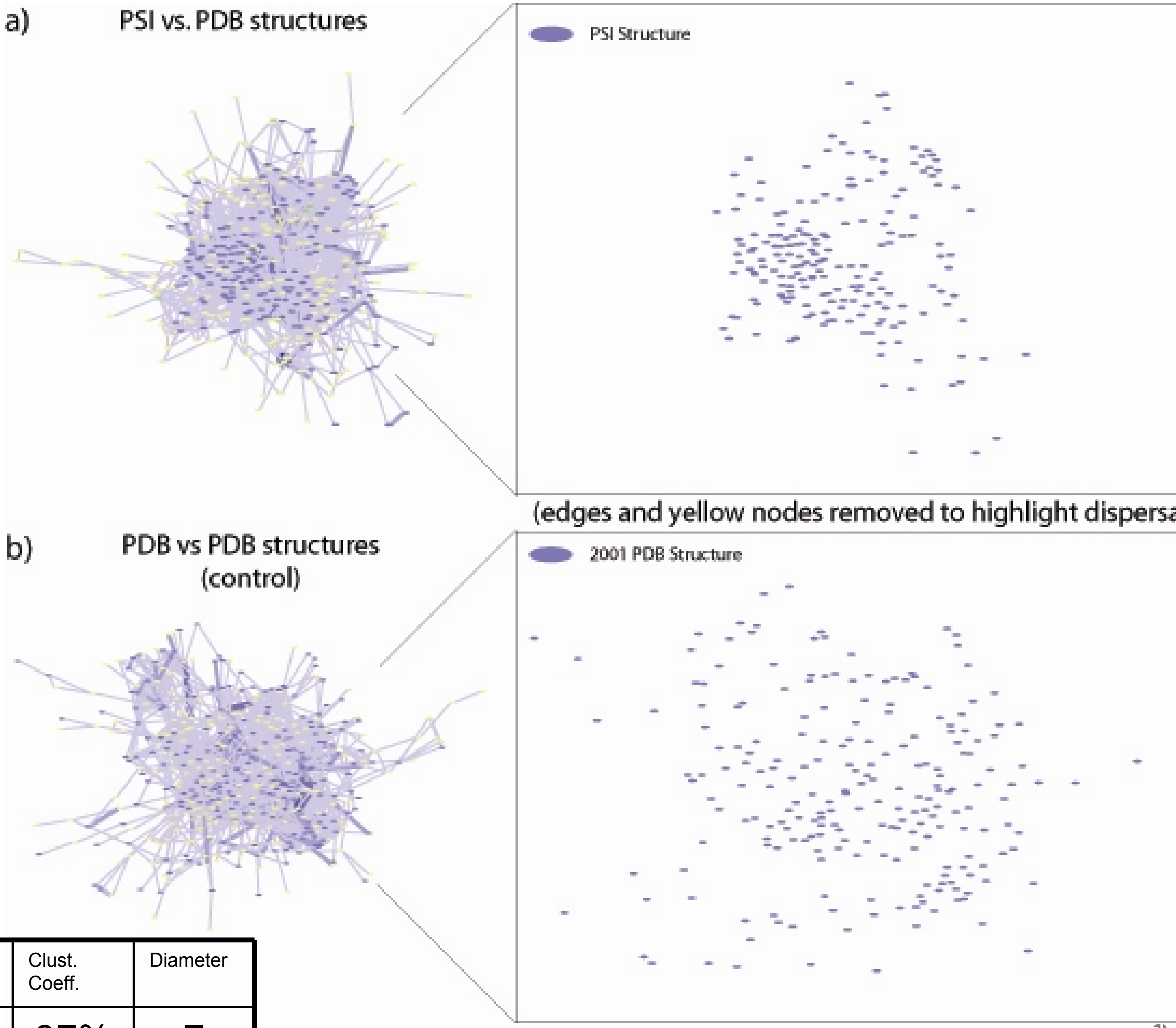
b) Author / Co-Authorship

c) Paper / Shared MeSH term

d) Paper / Shared Location

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

Clustering structures determined by struc. genomics consortia according to functional similarity: Is there a functional bias in consortia structures?

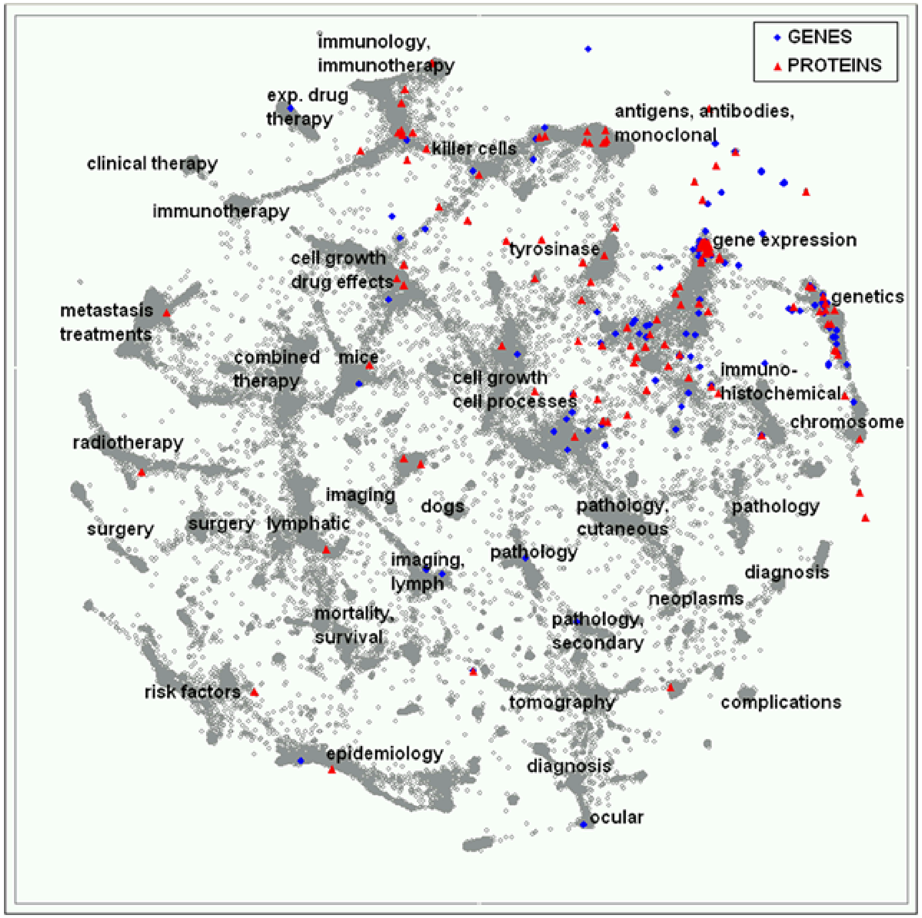


(edges and yellow nodes removed to highlight dispersa

	Avg. Degree	Avg. Path	Clust. Coeff.	Diameter
PSI	24	2.6	37%	7
PDB	6	3.9	31%	9

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

Making Larger Maps: Mapping a whole field

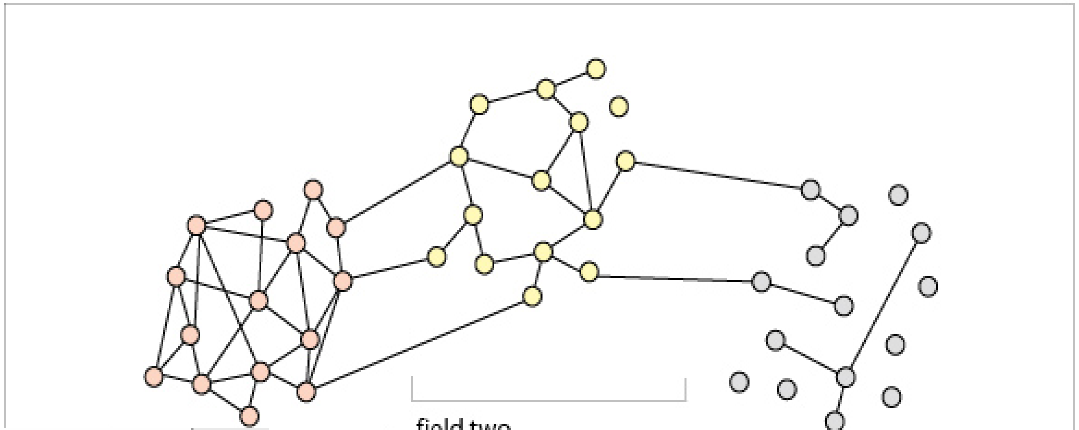


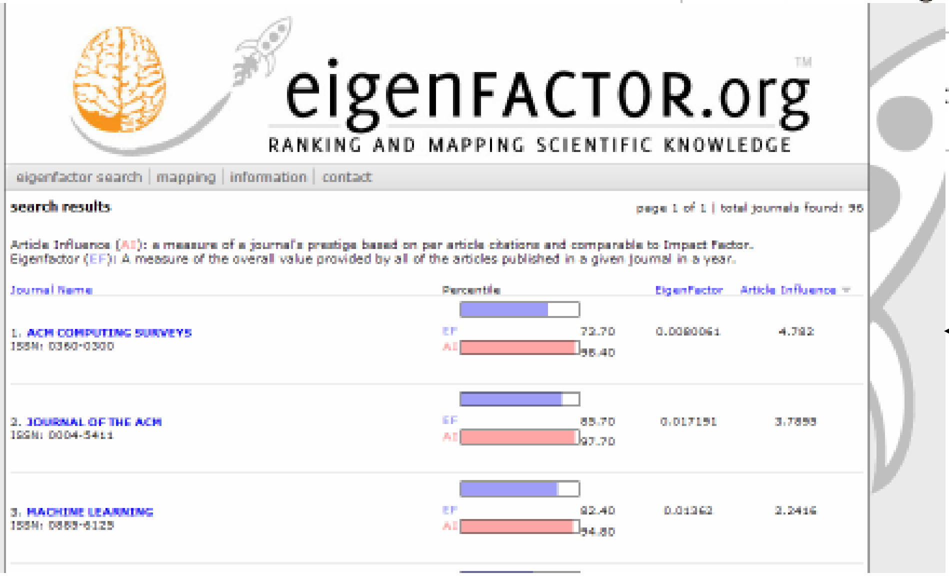
[Boyack, Kevin W., Mane, Ketan and Börner, Katy. (2004). Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. IV2004 Conference, London, UK, pp. 965-971.]

Ranking Journal Influence - Eigenfactor.org

"Ranks journals much as Google ranks websites."

Adjusts for citation differences among → disciplines




eigenFACTOR.org
RANKING AND MAPPING SCIENTIFIC KNOWLEDGE
eigenfactor search | mapping | information | contact

search results

page 1 of 1 | total journals found: 56

Article Influence (AI): a measure of a journal's prestige based on per article citations and comparable to Impact Factor.
Eigenfactor (EF): A measure of the overall value provided by all of the articles published in a given journal in a year.

Journal Name	Percentile	EigenFactor	Article Influence
1. ACM COMPUTING SURVEYS ISSN: 0360-0300	EF 73.70 AI 98.40	0.0080061	4.792
2. JOURNAL OF THE ACM ISSN: 0004-5411	EF 85.70 AI 97.70	0.017151	3.7855
3. MACHINE LEARNING ISSN: 0885-6125	EF 82.40 AI 94.80	0.01362	2.2416

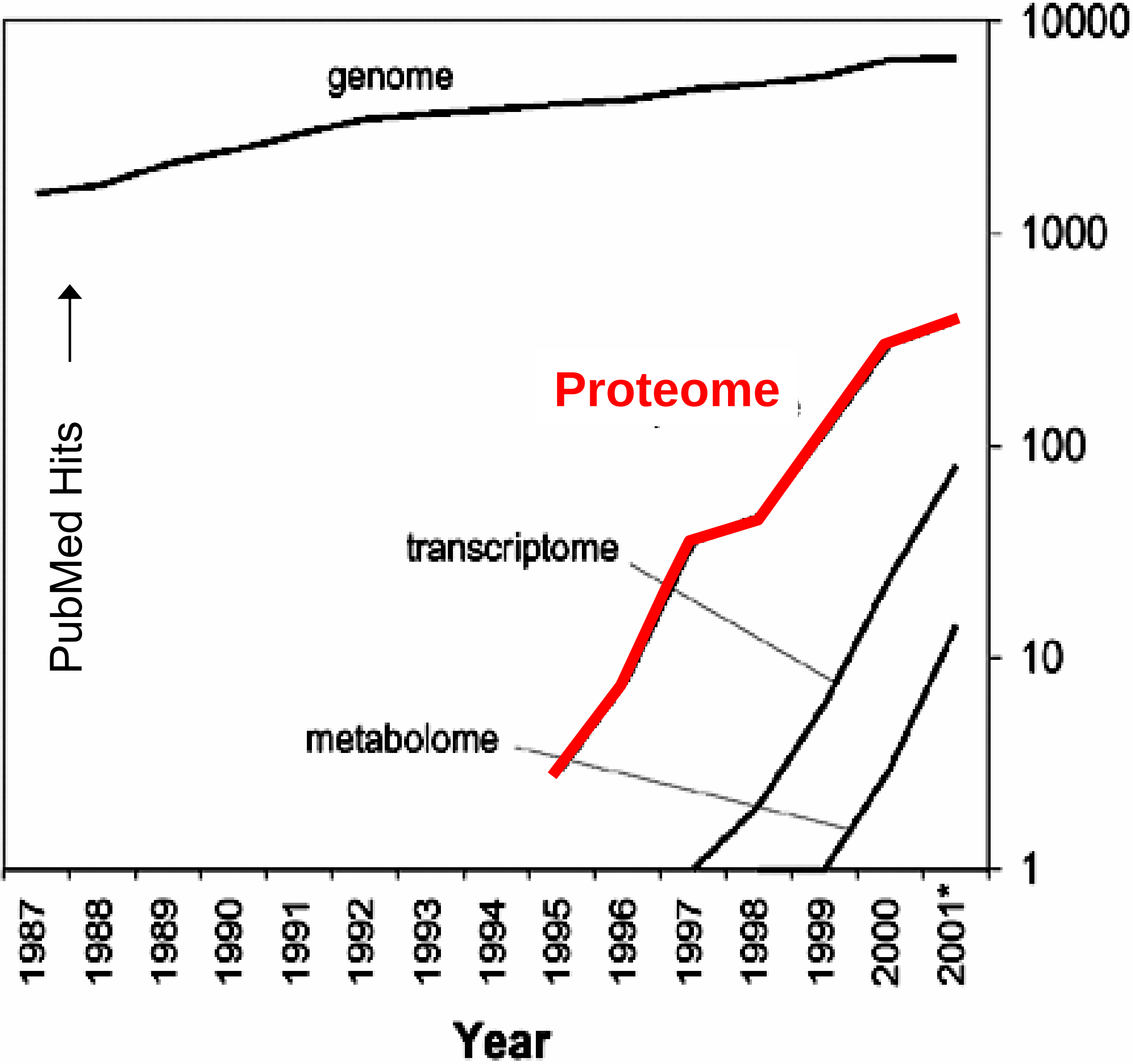
← Ranking of journals in computer science (top of list).

Examples Illuminating Current State of Affairs:

Analyzing the Dynamics of Science

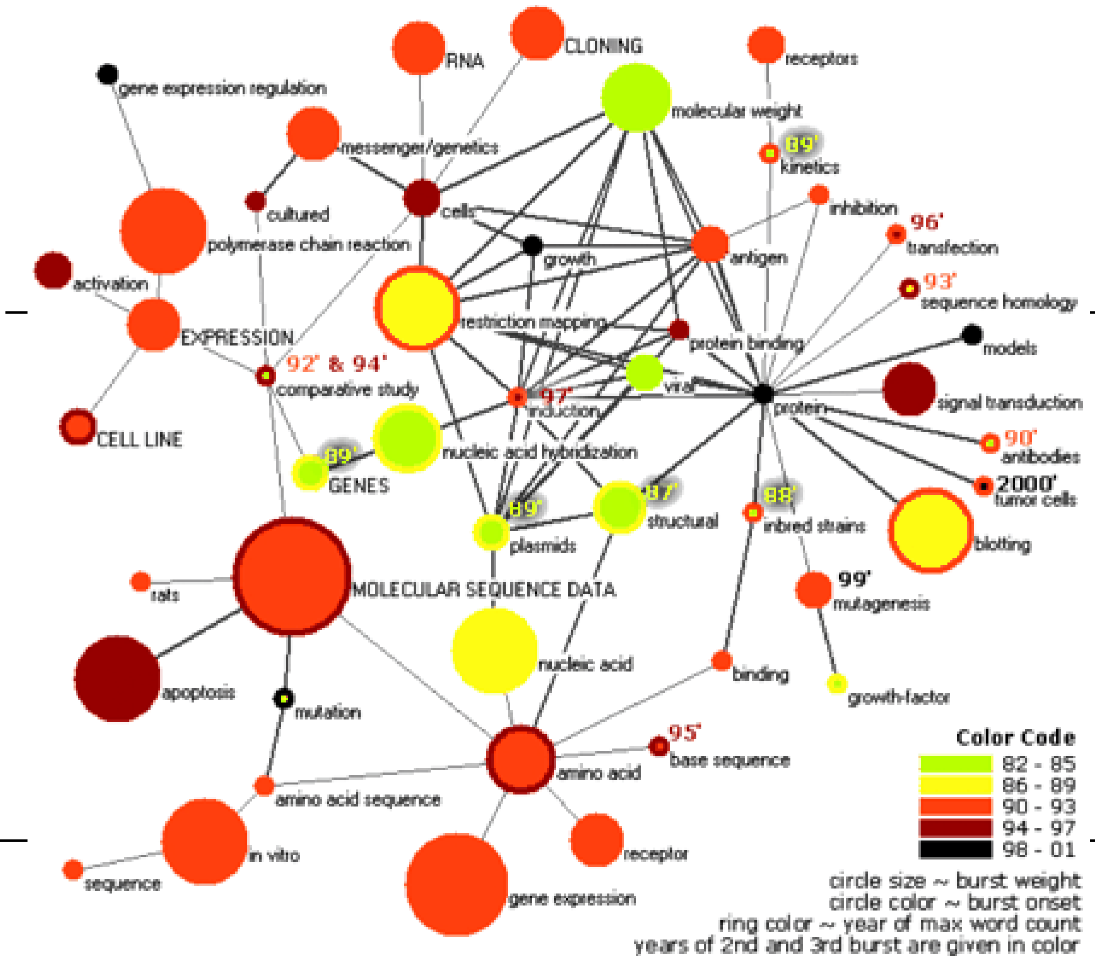
	Google Hits	Pub-Med Hits
Genome	~1880000	66171
Proteome	**~63,000**	**703**
Transcriptome	3520	72
Physiome	2980	15
Metabolome	349	12
Phenome	4980	6
Morphome	238	2
Interactome	56	2
Glycome	46	1
Secretome	21	1
Ribonome	1	1
Orfeome	42	-
Regulome	18	-
Cellome	17	-
Operome	8	-
Transportome	1	-
Functome	1	-

'Omics terms of the years



Evolution of Science

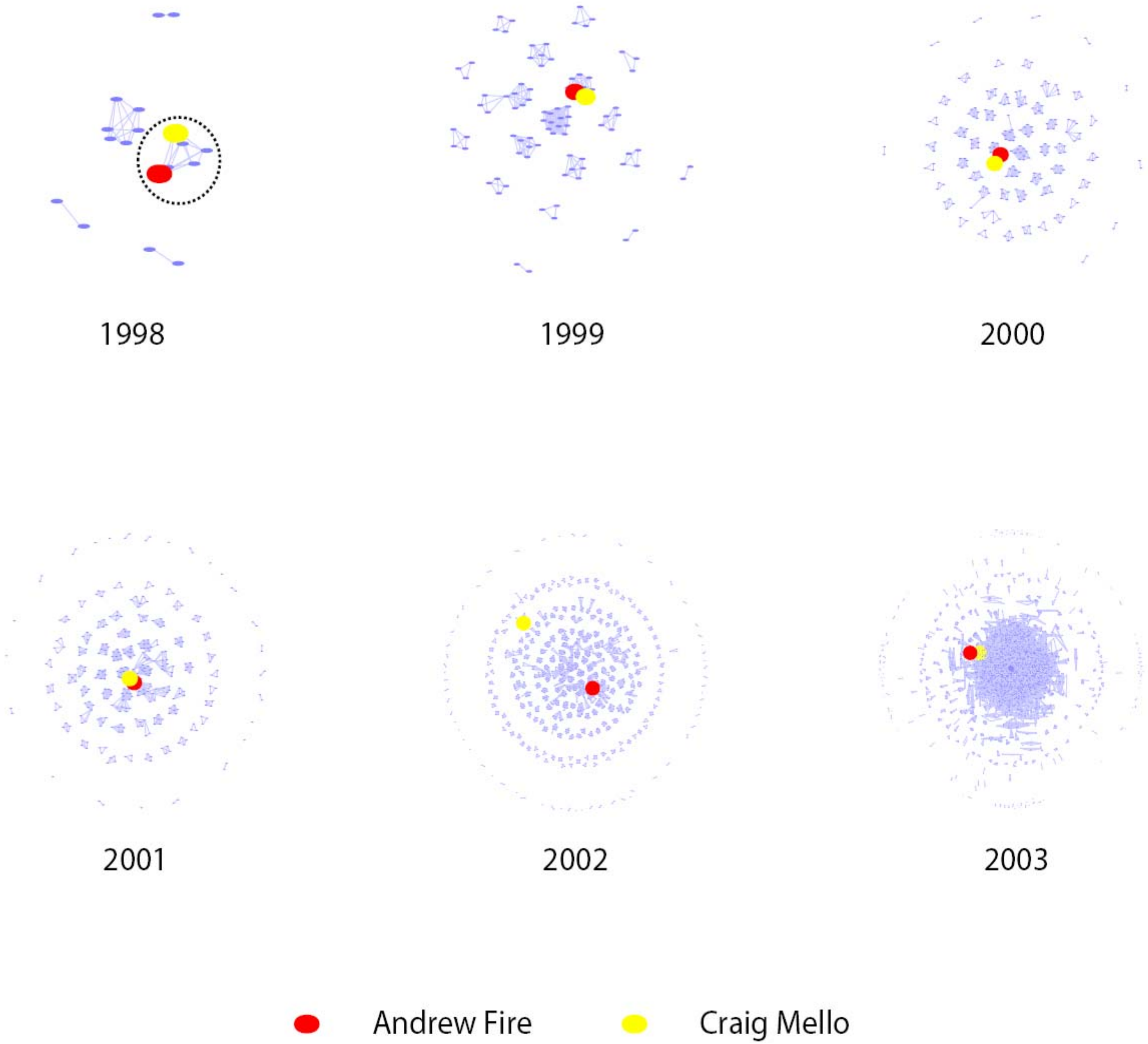
□ K. Mane, K. Börner (2004). Mapping topics and topic bursts in *PNAS*. 101 (Supplement 1): 5287.



Map based on "bursty" words in life sciences publications since 1980. Older fundamental research (center) led to four different areas (subgraphs in corners).

This and other domain maps at http://www.scimaps.org.

RNAi: Birth of a Field in the Literature Culminating in the 2006 Nobel



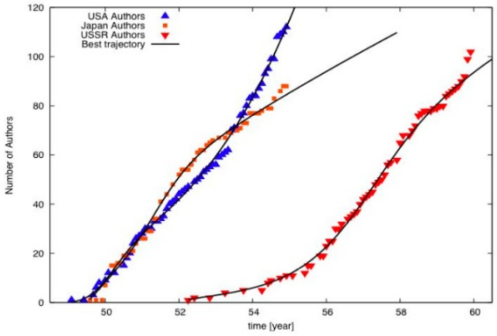
Source:
Gerstein & Douglas.
PLoS Comp. Bio. 3:e80
(2007)
PubNet.GersteinLab.org

The Social Dynamics of Innovation and Scientific Discovery

(a)

$$\begin{cases} \dot{S} = & \Lambda - \beta S \frac{I}{N} - bS \frac{Z}{N} - \mu S \\ \dot{E} = & (1-p)\beta S \frac{I}{N} + (1-l)bS \frac{Z}{N} - \rho E \frac{I}{N} - \varepsilon E - \mu E \\ \dot{I} = & p\beta S \frac{I}{N} + \rho E \frac{I}{N} + \varepsilon E - \delta I - \mu I \\ \dot{Z} = & lbS \frac{Z}{N} - \mu Z \\ R = & N - (S + E + I + Z) \end{cases}$$

(b)



(c)

$S(t)$	Susceptible
$E(t)$	Incubators
$I(t)$	“Infectious”
$Z(t)$	Stiflers
$R(t)$	Scientifically inactive
Λ	Recruitment rate
$1/\mu$	Scientific life expectancy
ε	Natural progression rate
$1/(\mu+\delta)$	“Infectious” period
β	Transmission rate
ρ	Transmission rate
b	Transmission rate
l	Effectiveness
p	Effectiveness

The population dynamics of authors in an emerging field is well described by models similar to those of epidemics, but that take into account contact processes and intentionality characteristic of human social dynamics. Panel (a) shows a SEIRZ model, (b) its best solution applied to the spread of Feynaman diagrams in the USA, Japan and the Soviet Union, and (c) details the model parameter's interpretation. The spread of ideas is characterized by relatively low contact rates (compared to infectious diseases), and very long lifetimes for the idea, as well as intentional strtuctures to promote interaction between individuals during the learning process.

The Patterns of Discovery and the Spread of Ideas as Epidemics.

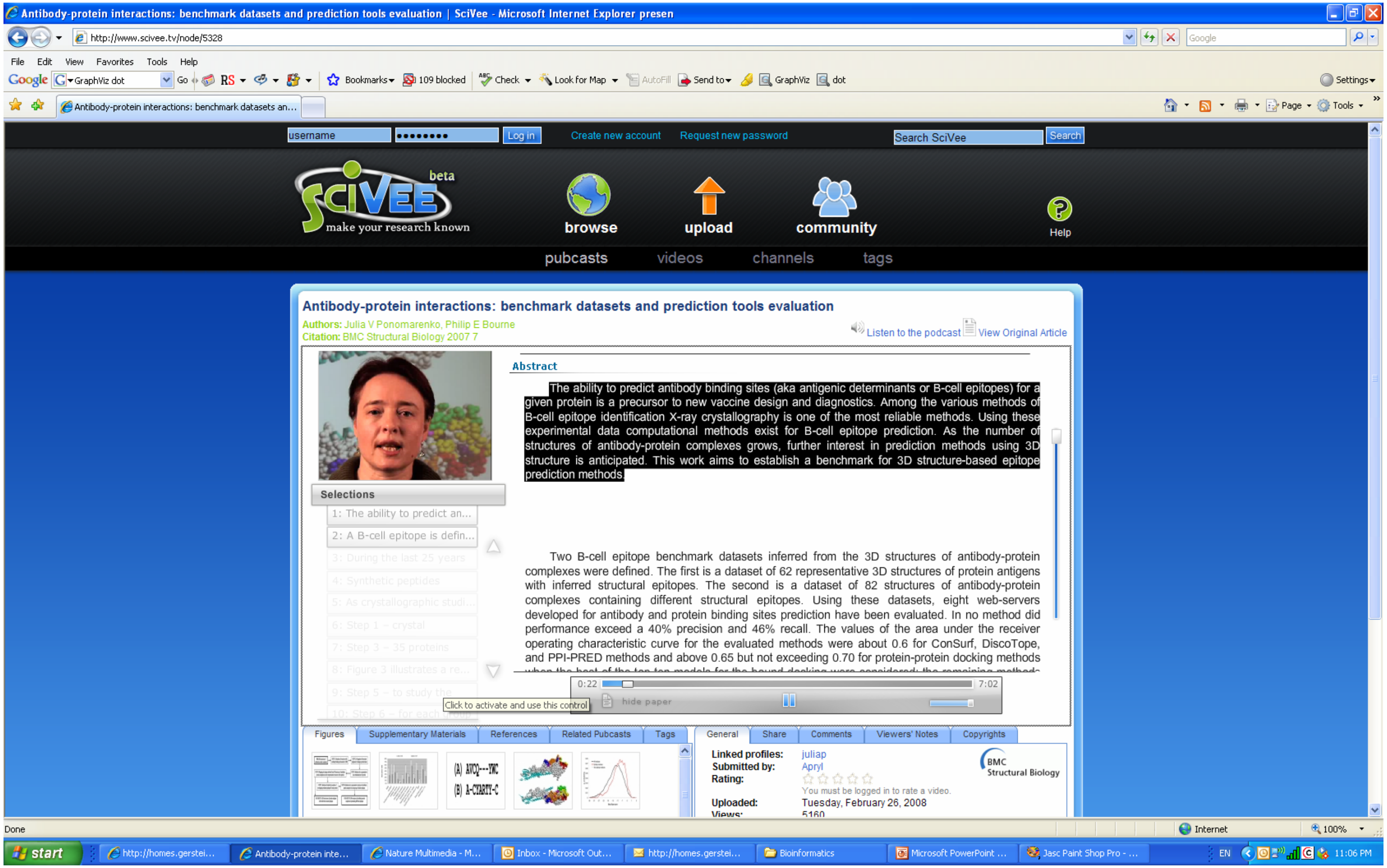
[Bettencourt et al., Physica A (2006)]

Examples Illuminating Current State of Affairs:

Mashing up the Text from Scientific Publications with other information sources to make Science more Understandable

- Mashing up scientific texts with streamed video, genome annotation, protein structure & interactions
- SciVee
  - ◊ http://www.scivee.tv
  - ◊ Partnership: NSF, PloS, San Diego Supercomputing Center
  - ◊ Pubcasts—video correlated with PLoS papers automatically displayed as video runs
  - ◊ Videos—scientists upload their own without papers
- Journal of Visualized Experiments (JoVE)
  - ◊ http://www.jove.com
  - ◊ Monthly issues of theme-related videos
  - ◊ Procedure walk-throughs, interviews
  - ◊ High-quality video and sound

SciVee



Antibody-protein interactions: benchmark datasets and prediction tools evaluation | SciVee - Microsoft Internet Explorer presen
http://www.scivee.tv/node/5328
File Edit View Favorites Tools Help
username
Log in
Create new account
Request new password
Search SciVee
Search
beta
make your research known
browse
upload
community
Help
pubcasts
videos
channels
tags

Antibody-protein interactions: benchmark datasets and prediction tools evaluation

Authors: Julia V Ponomarenko, Philip E Bourne
Citation: BMC Structural Biology 2007 7

Listen to the podcast
View Original Article

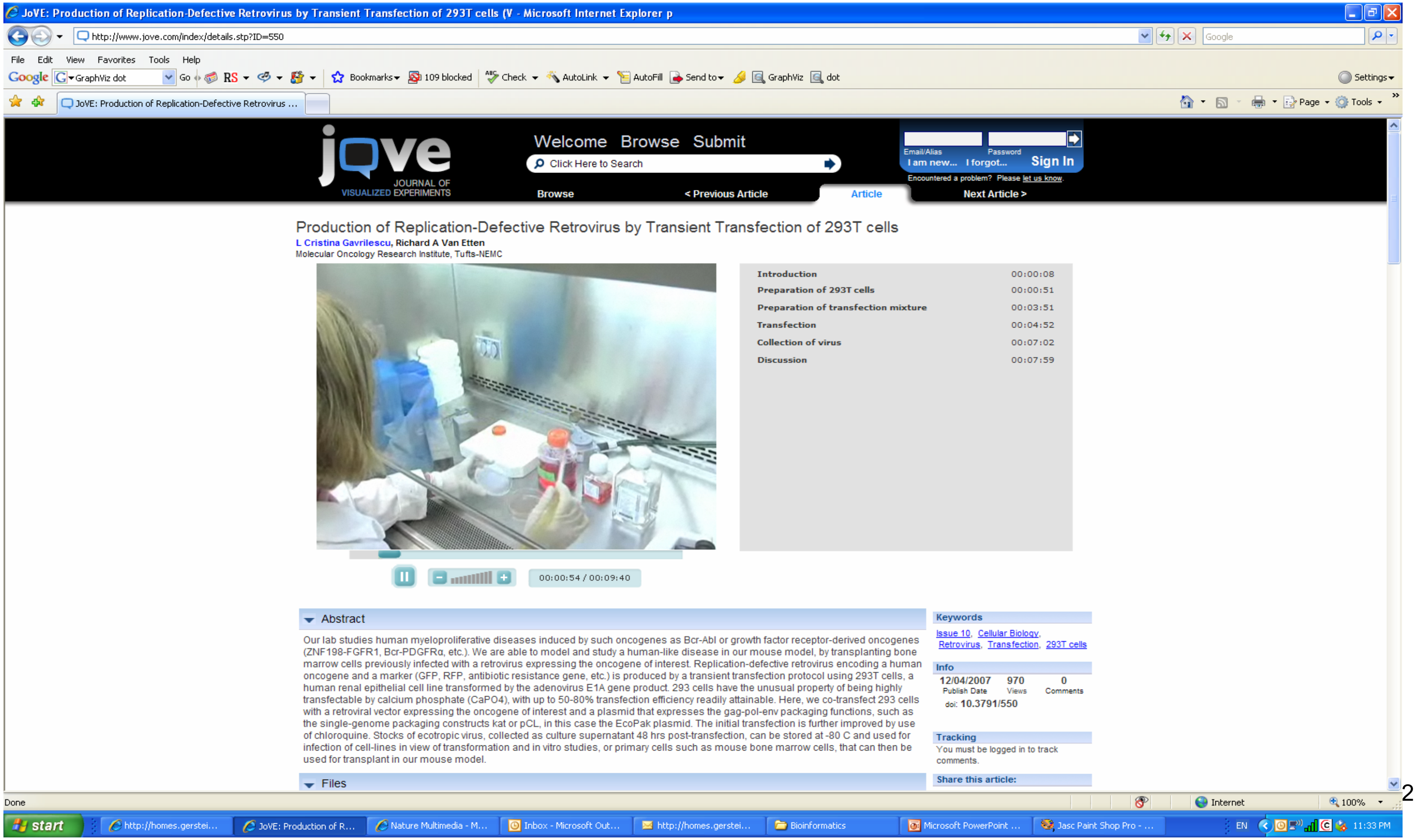
Abstract

The ability to predict antibody binding sites (aka antigenic determinants or B-cell epitopes) for a given protein is a precursor to new vaccine design and diagnostics. Among the various methods of B-cell epitope identification X-ray crystallography is one of the most reliable methods. Using these experimental data computational methods exist for B-cell epitope prediction. As the number of structures of antibody-protein complexes grows, further interest in prediction methods using 3D structure is anticipated. This work aims to establish a benchmark for 3D structure-based epitope prediction methods.

Two B-cell epitope benchmark datasets inferred from the 3D structures of antibody-protein complexes were defined. The first is a dataset of 62 representative 3D structures of protein antigens with inferred structural epitopes. The second is a dataset of 82 structures of antibody-protein complexes containing different structural epitopes. Using these datasets, eight web-servers developed for antibody and protein binding sites prediction have been evaluated. In no method did performance exceed a 40% precision and 46% recall. The values of the area under the receiver operating characteristic curve for the evaluated methods were about 0.6 for ConSurf, DiscoTope, and PPI-PRED methods and above 0.65 but not exceeding 0.70 for protein-protein docking methods when the best of the top ten models for the bound docking were considered; the remaining methods

Linked profiles: juliap
Submitted by: Apryl
Rating: ☆☆☆☆☆ You must be logged in to rate a video.
Uploaded: Tuesday, February 26, 2008
Views: 5160

JoVE



Production of Replication-Defective Retrovirus by Transient Transfection of 293T cells

L Cristina Gavrilescu, Richard A Van Etten
Molecular Oncology Research Institute, Tufts-NEMC

Introduction	00:00:08
Preparation of 293T cells	00:00:51
Preparation of transfection mixture	00:03:51
Transfection	00:04:52
Collection of virus	00:07:02
Discussion	00:07:59

Abstract

Our lab studies human myeloproliferative diseases induced by such oncogenes as Bcr-Abl or growth factor receptor-derived oncogenes (ZNF198-FGFR1, Bcr-PDGFRα, etc.). We are able to model and study a human-like disease in our mouse model, by transplanting bone marrow cells previously infected with a retrovirus expressing the oncogene of interest. Replication-defective retrovirus encoding a human oncogene and a marker (GFP, RFP, antibiotic resistance gene, etc.) is produced by a transient transfection protocol using 293T cells, a human renal epithelial cell line transformed by the adenovirus E1A gene product. 293 cells have the unusual property of being highly transfectable by calcium phosphate (CaPO4), with up to 50-80% transfection efficiency readily attainable. Here, we co-transfect 293 cells with a retroviral vector expressing the oncogene of interest and a plasmid that expresses the gag-pol-env packaging functions, such as the single-genome packaging constructs kat or pCL, in this case the EcoPak plasmid. The initial transfection is further improved by use of chloroquine. Stocks of ecotropic virus, collected as culture supernatant 48 hrs post-transfection, can be stored at -80 C and used for infection of cell-lines in view of transformation and in vitro studies, or primary cells such as mouse bone marrow cells, that can then be used for transplant in our mouse model.

Files

Keywords

Issue 10, Cellular Biology, Retrovirus, Transfection, 293T cells

Info

12/04/2007	970	0
Publish Date	Views	Comments

doi: 10.3791/550

Tracking

You must be logged in to track comments.

Share this article:

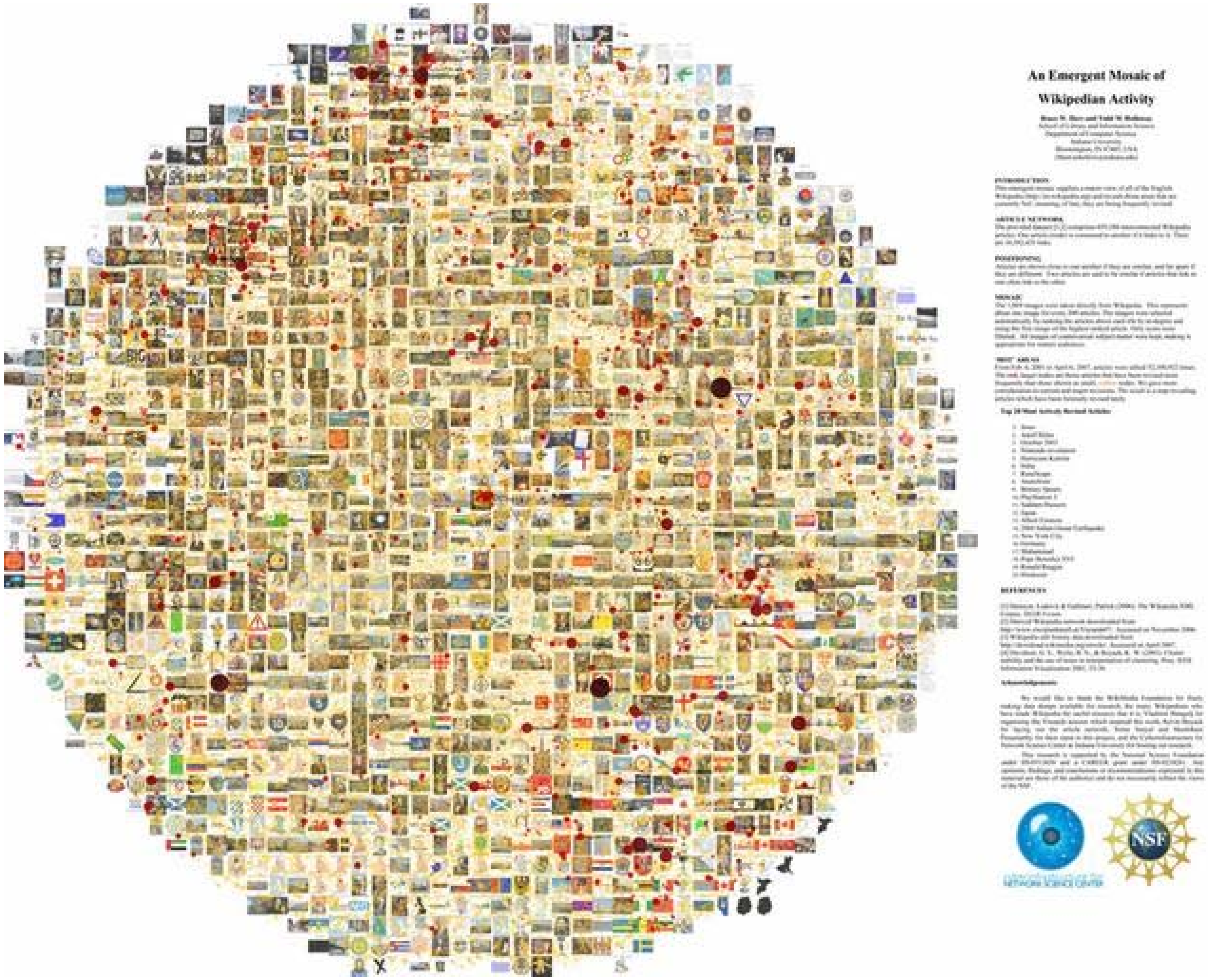
Fusing Data & Papers to Annotate the Genome

- Ideal project for 21st century is annotating every base of the genome
  - ◊ Want to attach all publications and results to the genome
  - ◊ "Fly through Genome" as way to access and understand the literature

- Problem of a good browser....

[Gerstein, Science ('00), Nature ('06)]

We need a Google Earth for the Genome; A Step in this Direction...



An Emergent Mosaic of Wikipedian Activity

Bruce W. Herr and Todd M. Holloway

[illegible]

INTRODUCTION

[illegible]

ARTICLE NETWORK

[illegible]

POSITIONING

[illegible]

MOSAIC

[illegible]

"HOT" AREAS

[illegible]

Top 20 Most Actively Revised Articles

[illegible]

REFERENCES

[illegible]

Acknowledgements

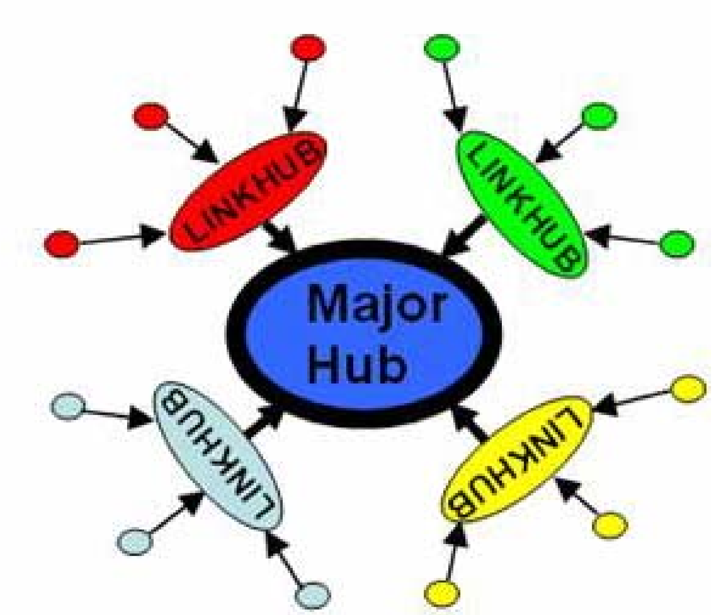
[illegible]

[Herr, Holloway, Börner, Emergent Mosaic of Wikipedian Activity, 2007]

Impediments to the Vision

DB Interoperation & Federated Information Architecture

- Need to perform a distributed query over many sites
  - ◊ Conventional web links
  - ◊ More complex interfaces
- Annotation of the human genome involves a massive federation of interoperating servers
  - ◊ "Administered" by many disparate people and groups



[Smith et al., BMC Bioinfo. ('07)]

Impediment #1: Structuring the Information Correctly for Large-scale Query

Issues with the Current Situtation between DBs & Journals

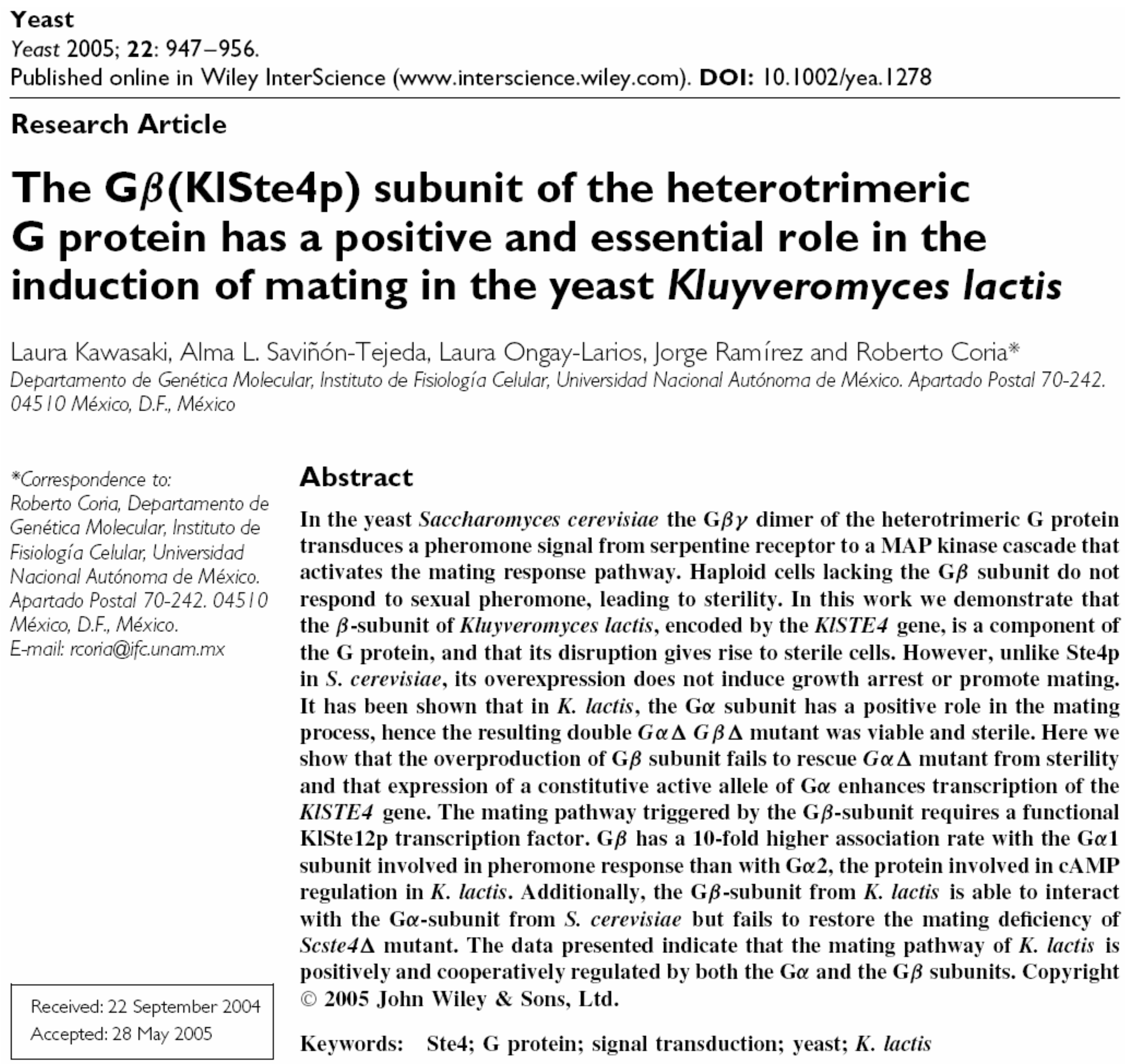
- Not always a clear linkage between papers & DBs
  - ◊ Keeping entries in DB and paper in sync
- Data aliquot
  - ◊ Huge datasets are handled but what of isolated facts
- How to connect key attributes of Journals with DBs
  - ◊ Attribution for credit & accountability
  - ◊ Time stamping of unchanging entries
  - ◊ Citation and history
  - ◊ Well worked out process of QC via refereeing and editing
- Readability of Papers
  - ◊ Detailed data embedded into papers, making text hard to read

Structured Abstract Proposal

- Storing information in papers in machine interpretable fashion
  - for automatic deposition into DBs
  - Abstract + standardized view of all tables
- Cross-referencing it with a specific part of the global genome, proteome, and interactome
  - Article written as annotation from the start
- Done in parallel to submission & revision of normal journal article
  - Refereed & edited normally
  - Capitalizes on peer review & incentives to publish
- Curators vs editors
  - Author is in control and this process
  - But it’s officiated by referees and editors

[Gerstein et al., Nature ('07)]

Ex. Structured Abstract



Yeast
Yeast 2005; **22**: 947–956.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/yea.1278

Research Article

The Gβ(KlSte4p) subunit of the heterotrimeric G protein has a positive and essential role in the induction of mating in the yeast *Kluyveromyces lactis*

Laura Kawasaki, Alma L. Saviñón-Tejeda, Laura Ongay-Larios, Jorge Ramírez and Roberto Coria*
Departamento de Genética Molecular, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México. Apartado Postal 70-242. 04510 México, D.F., México

**Correspondence to:
Roberto Coria, Departamento de Genética Molecular, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México. Apartado Postal 70-242. 04510 México, D.F., México.
E-mail: rcoria@ifc.unam.mx*

Abstract

In the yeast *Saccharomyces cerevisiae* the G$\beta\gamma$ dimer of the heterotrimeric G protein transduces a pheromone signal from serpentine receptor to a MAP kinase cascade that activates the mating response pathway. Haploid cells lacking the Gβ subunit do not respond to sexual pheromone, leading to sterility. In this work we demonstrate that the β-subunit of *Kluyveromyces lactis*, encoded by the *KlSTE4* gene, is a component of the G protein, and that its disruption gives rise to sterile cells. However, unlike Ste4p in *S. cerevisiae*, its overexpression does not induce growth arrest or promote mating. It has been shown that in *K. lactis*, the Gα subunit has a positive role in the mating process, hence the resulting double $G\alpha\Delta$ $G\beta\Delta$ mutant was viable and sterile. Here we show that the overproduction of Gβ subunit fails to rescue $G\alpha\Delta$ mutant from sterility and that expression of a constitutive active allele of Gα enhances transcription of the *KlSTE4* gene. The mating pathway triggered by the Gβ-subunit requires a functional KlSte12p transcription factor. Gβ has a 10-fold higher association rate with the Gα1 subunit involved in pheromone response than with Gα2, the protein involved in cAMP regulation in *K. lactis*. Additionally, the Gβ-subunit from *K. lactis* is able to interact with the Gα-subunit from *S. cerevisiae* but fails to restore the mating deficiency of *Scste4*Δ mutant. The data presented indicate that the mating pathway of *K. lactis* is positively and cooperatively regulated by both the Gα and the Gβ subunits. Copyright © 2005 John Wiley & Sons, Ltd.

Received: 22 September 2004
Accepted: 28 May 2005

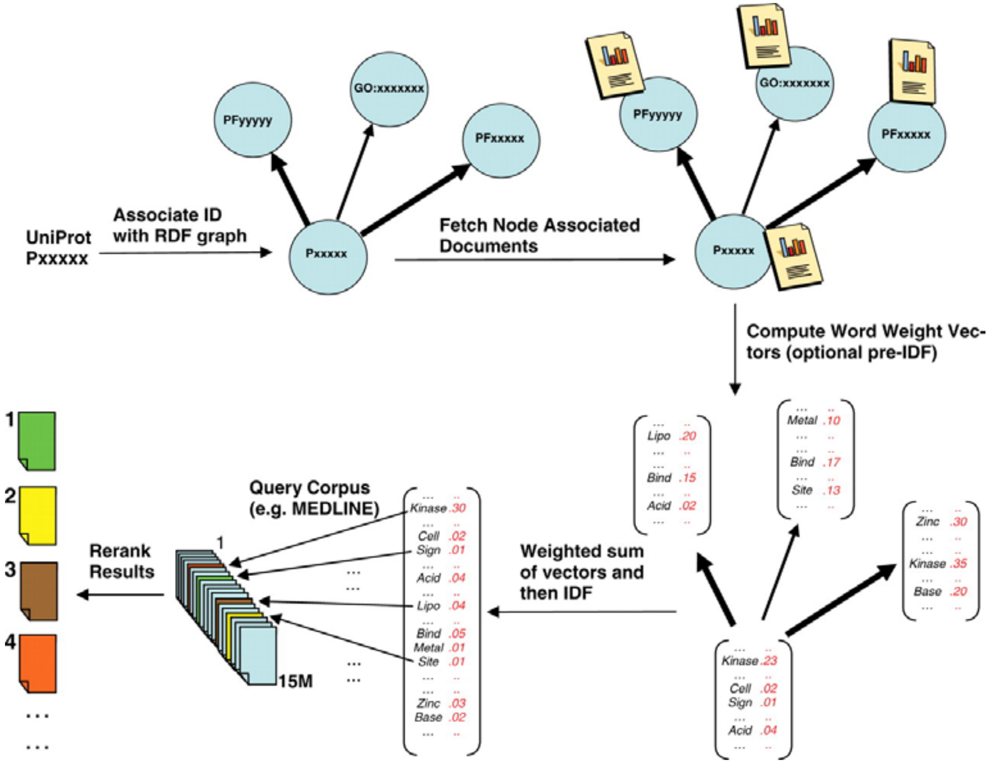
Keywords: Ste4; G protein; signal transduction; yeast; *K. lactis*

Ex. Structured Abstract

- ***K.lactis*** (species)
  - **KlSTE4** (gene)
    - **KlSte4p** (protein)
      - **CLONED**
        - Available at …
      - **SEQUENCED**
        - Sequence ATGTACGCTATAGGC….
      - **MUTANTS**
        - **DELETION**
        - **FUNCTIONAL ASSAYS**
        - Sterile in both MATa and MATα
        - No defect in vegetative growth
        - **STRAIN INFORMATION**
        - Available at….
      - **INTERACTIONS**
        - **TWO-HYBRID**
        - KlGpa1p (10x stronger) = XXX
        - Control (no partner) = XXX
        - KlGpa1p* = XXX
        - KlGpa2p = XXX
        - ScGpa1p = XXX (S. cerevisiae)
      - **COMMENTS**
        - Both KlSte4p and KlGpa1p required to induce mating in *K.lactis*
  - **KlGPA1** (gene)
    - **KlGpa1p** (protein)
      - **INTERACTIONS**
        - **TWO-HYBRID**
        - KlSte4 = XXX
    - **KlGpa1p*** (protein)
      - **INTERACTIONS**
        - **TWO-HYBRID**
        - KlSte4 = XXX
  - **KlGPA2** (gene)
    - **KlGpa2p** (protein)
      - **INTERACTIONS**
        - **TWO-HYBRID**
        - KlSte4 = XXX
- ***S.cerevisiae*** (species)
  - **SCGPA1** (gene)
    - **ScGpa1p** (protein)
      - **INTERACTIONS**
        - **TWO-HYBRID**
        - KlSte4 = XXX

Unsupervised Textmining vs Manually Curated and Structured Documents (e.g. Sem. Web.): Not necessarily a conflict

- Structured abs. might be good training sets for mining



[Smith et al., Bioinformatics ('07); Smith & Gerstein, Science ('06)]

Impediment #2: Access Restrictions Inhibit Large-scale Query

Absence of social framework for protecting "data" on the web

- Researchers unclear on framework
  - ◊ The ambiguity of the present copyright laws governing the protection of databases creates a situation where researchers are (practically) unclear about their rights to extract and combine data
    - Putting articles up on sites, "quoting" annotation
  - ◊ Likewise, researchers are unsure how to get "credit" for combined data ("Mash ups")
    - Disincentive to data integration
- Database owners, unsure of how laws safeguards their information, overprotect their data with licenses and technological mechanisms that impede interoperation.

Technological safeguards to "protect" data

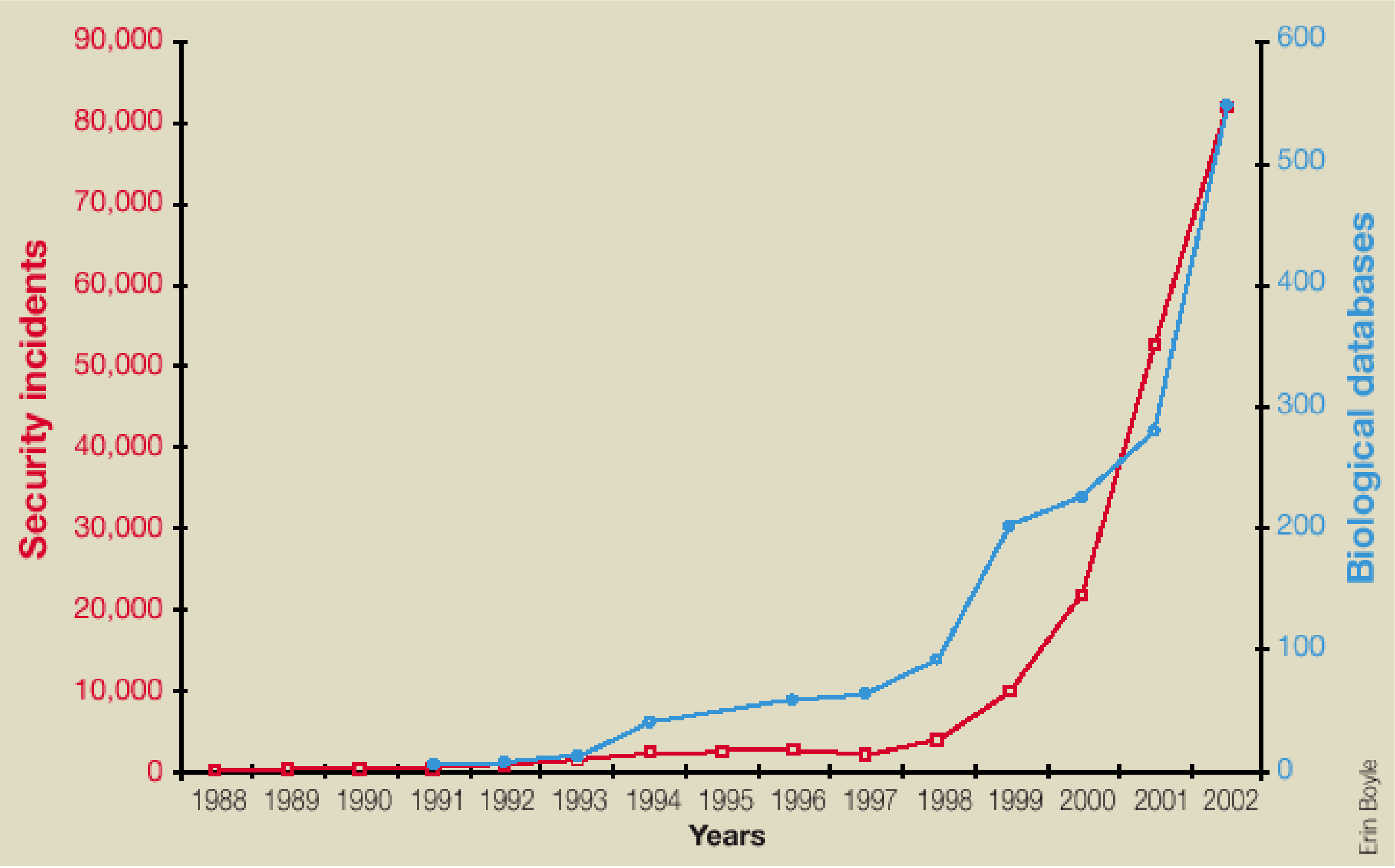
- Limits on Bulk Downloads & Global Analysis
  - ◊ Passwords and IP filtering
    - allow the database owner to limit access to specific users and computers
    - selectively cut off access to researchers performing bulk calculations.
  - ◊ Data can also be presented piecemeal, in response to a specific user query
  - ◊ Examples
    - Incyte Proteome database
    - Cellzome database of interactions.
- Databases can be stored in propriety formats
  - ◊ Extreme is encryption
- Watermarking adds overt or hidden digital fingerprints
  - ◊ Slightly corrupting the data.
  - ◊ Not that common in bio-DBs (but found in British Library).

Free text Issue is Part of this Larger Context

- Different traditions in academic publishing vs DB world
  - Genome sequence is free
  - but have to pay for article about it!
- Many free text initiatives
  - PubMedCentral.NIH.gov & arXiv.org
- Tricky economics of free text
  - potentially efficient
  - but redistributes dollars in world of academic publishing
  - who pays: readers or writers

Impediment #3: Security Considerations Inhibit Large-scale Query

Vast Computer Security Costs in the "Wild West" Internet



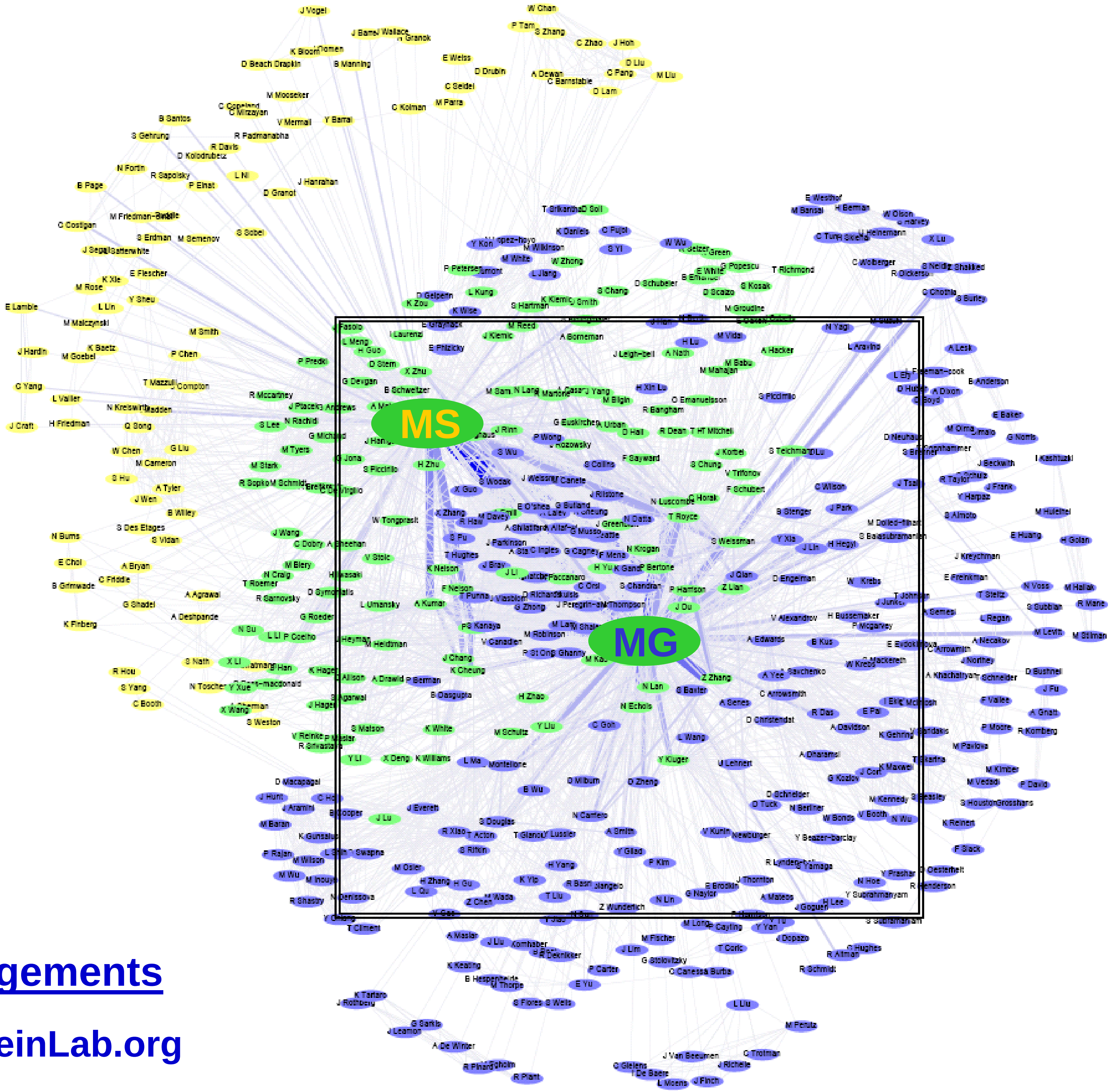
Vast difficulty in securing information servers in academia

- Mundane administration — patches
- Make building intricate systems for interoperation difficult, as researchers have to continually check their interfaces for "holes"
- Unique impact on research (vs business)
  - ◊ Free and broad dissemination of ideas between labs and public is hallmark of research.
  - ◊ Preserving openness precludes standard security practices often employed in a corporate or military environment -- e.g. private networks
  - ◊ Academic computer users exhibit great variability, making effective security procedures more difficult

Vision for Harnessing the Volume of Information on the Web to Study the Structure of Science

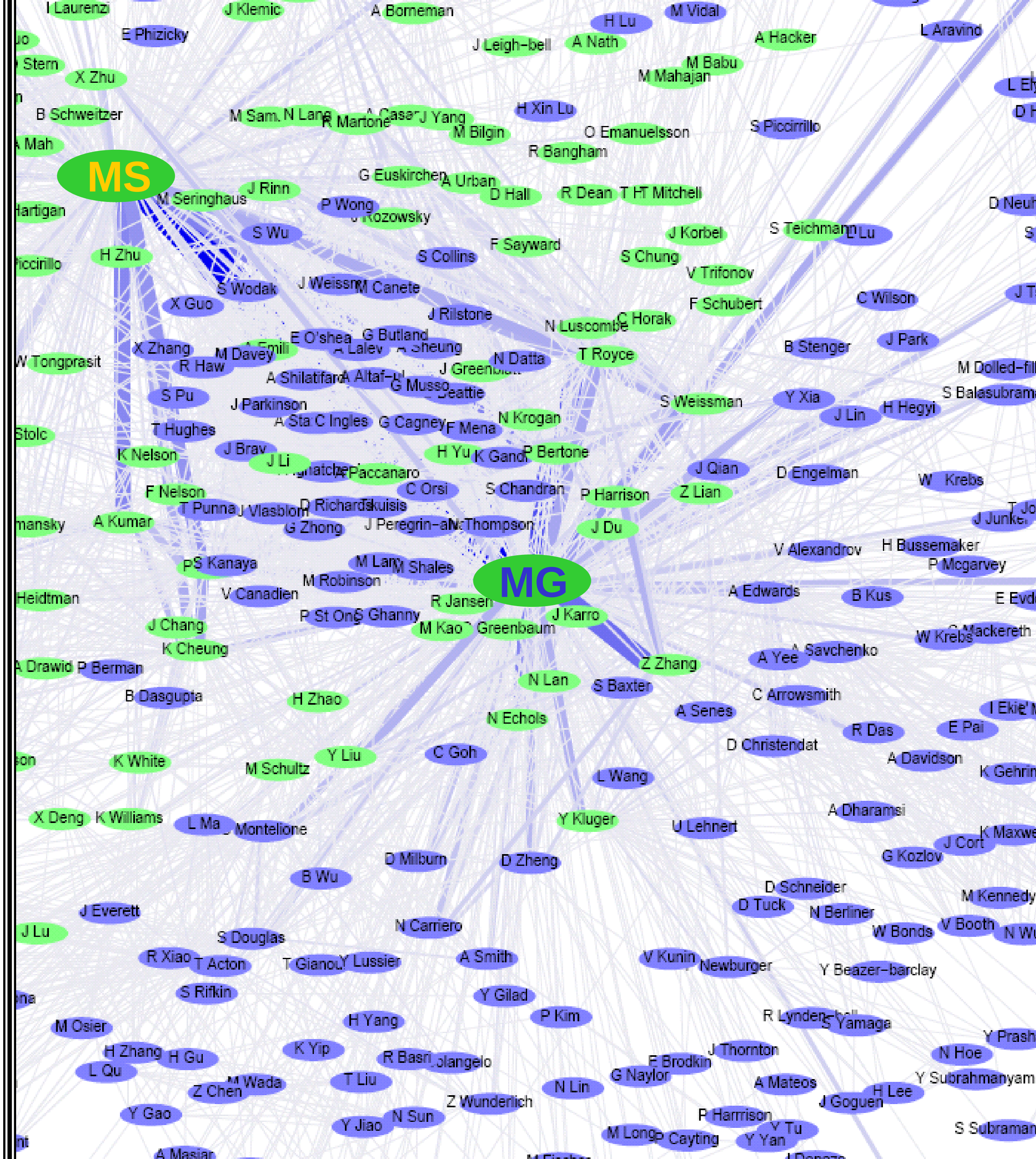
- Main Applications of Large-scale Mining
  - ◊ New Scientific Discoveries (not disc. here)
  - ◊ Understanding Areas of Study through Simple Zipf Stats
    - Crystallography Nobel, Genomics, Gene Naming
  - ◊ Maps of Science
    - Studying a genomics consortia, Bigger Maps to Rank Journals
  - ◊ Dynamics of Science
    - Watching and modeling the appearence of new terms, RNAi ex.

- Impediments Large-scale Mining (as Distributed Query)
  - ◊ (Semi) Structuring the Information in Journals
  - ◊ Overcoming access restrictions
  - ◊ Security Considerations



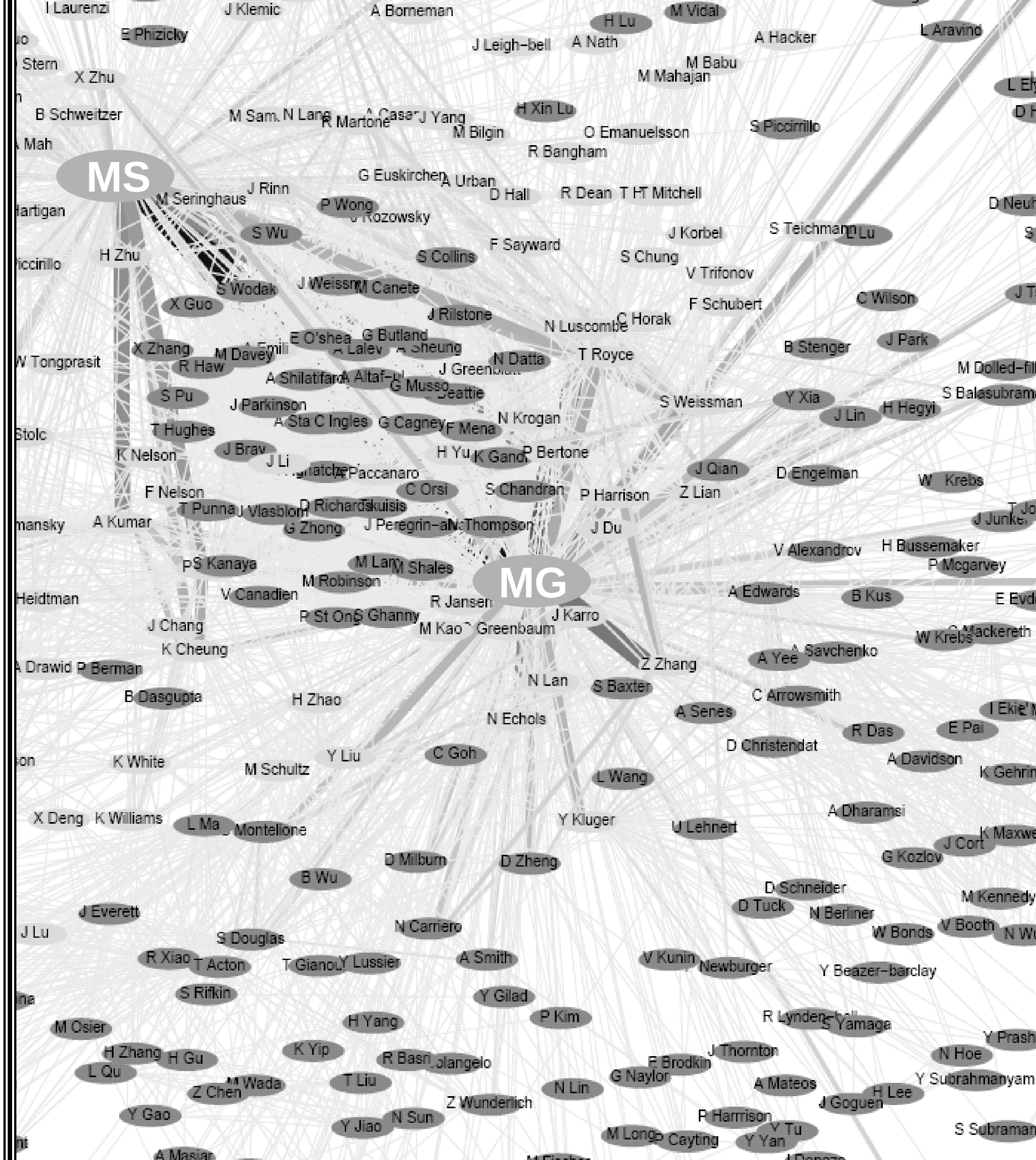
Acknowledgements

TopNet.GersteinLab.org



Acknowledgements

TopNet.GersteinLab.org



Acknowledgements

TopNet.GersteinLab.org

NIH, NSF, Keck

Job opportunities currently for postdocs & students

MS
M Seringhaus
MG
D Greenbaum
K Cheung
M Schultz
G Montelione
S Douglas
A Smith
K Yip
P Cayting
I Laurenzi
J Klemic
A Borneman
H Lu
M Vidal
L Aravind
E Phizicky
J Leigh-bell
A Nath
A Hacker
Stern
X Zhu
M Babu
M Mahajan
B Schweitzer
M Sam
N Lang
R Martone
J Yang
H Xin Lu
M Bilgin
O Emanuelsson
S Piccirillo
Mah
R Bangham
G Euskirchen
A Urban
J Rinn
D Hall
R Dean
T HT Mitchell
M Seringhaus
Hartigan
J Korbel
S Teichmann
F Sayward
S Collins
S Chung
H Zhu
V Trifonov
S Wodak
J Weissm
M Canete
X Guo
C Wilson
F Schubert
J Rilstone
N Luscombe
C Horak
E O'shea
G Butland
X Zhang
M Davey
R Haw
T Royce
N Datta
B Stenger
J Park
W Tongprasit
A Shilatifard
A Altaf-ul
G Musso
J Greenbaum
M Dolled-filhart
S Pu
J Parkinson
S Weissman
Y Xia
S Balasubramanian
J Lin
H Hegyi
A Sta C Ingles
G Cagney
F Mena
N Krogan
T Hughes
Stolc
J Bray
K Nelson
J Li
H Yu
K Gand
P Bertone
J Qian
D Engelman
W Krebs
Paccanaro
C Orsi
S Chandran
P Harrison
Z Lian
F Nelson
T Punna
J Vlasblom
D Richardson
A Kumar
G Zhong
J Peregrin-alvarez
Thompson
J Du
J Junker
V Alexandrov
H Bussemaker
P Mcgarvey
PS Kanaya
M Lam
M Shales
M Robinson
V Canadien
A Edwards
B Kus
E Evdokimova
Heidtman
R Jansen
P St Onge
S Ghanny
J Karro
J Chang
M Kao
C Greenbaum
W Krebs
Mackereth
A Savchenko
A Drawid
P Berman
S Baxter
B Dasgupta
H Zhao
C Arrowsmith
N Echols
A Senes
I Ekiel
R Das
E Pai
D Christendat
C Goh
A Davidson
K White
K Gehring
L Wang
A Dharamsi
X Deng
K Williams
Y Kluger
U Lehnert
K Maxwell
J Cort
D Milburn
D Zheng
G Kozlov
B Wu
D Schneider
D Tuck
N Berliner
M Kennedy
J Everett
W Bonds
V Booth
J Lu
Newburger
Y Beazer-barclay
R Xiao
T Acton
Lussier
A Smith
S Rifkin
Y Gilad
P Kim
R Lynden-bell
S Yamaga
M Osier
Y Prasad
H Zhang
H Gu
R Basri
J Thornton
F Brodkin
N Hoe
L Qu
G Naylor
M Wada
T Liu
N Lin
A Mateos
Y Subrahmanyam
Z Chen
H Lee
Z Wunderlich
J Goguen
Y Gao
Y Jiao
N Sun
S Subramaniam
A Masiar

[Acknowledgements]

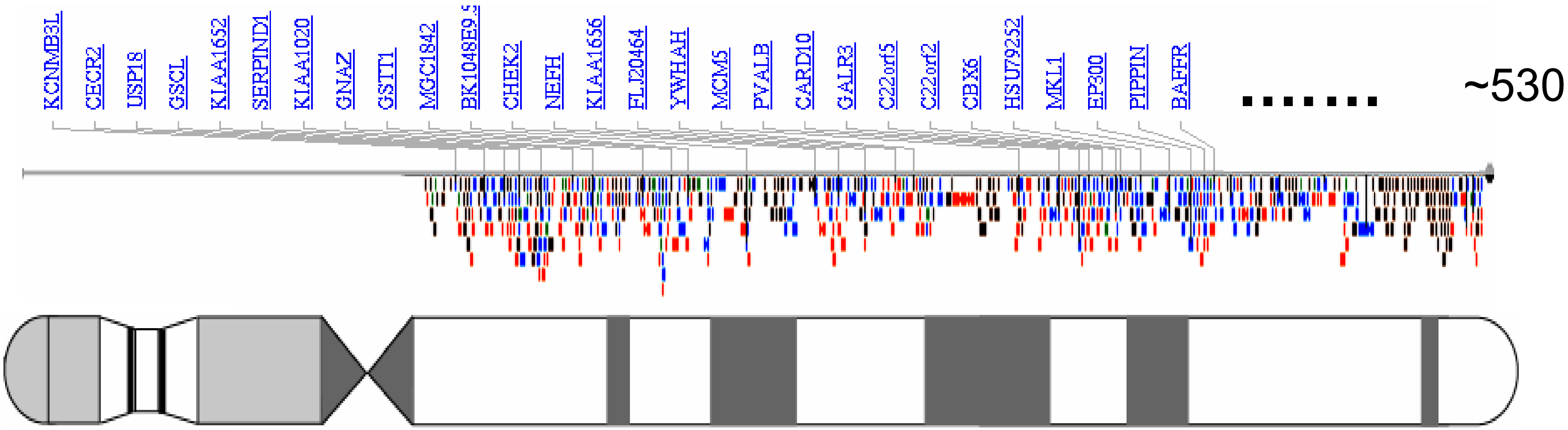
TopNet.GersteinLab.org

Acknowledgements

D Greenbaum
J Junker
S Douglas
A Smith
M Seringhaus

bioinfo.mbb.yale.edu
papers.gersteinlab.org/papers/epublishing

The problem: Grappling with Function on a Genome Scale?



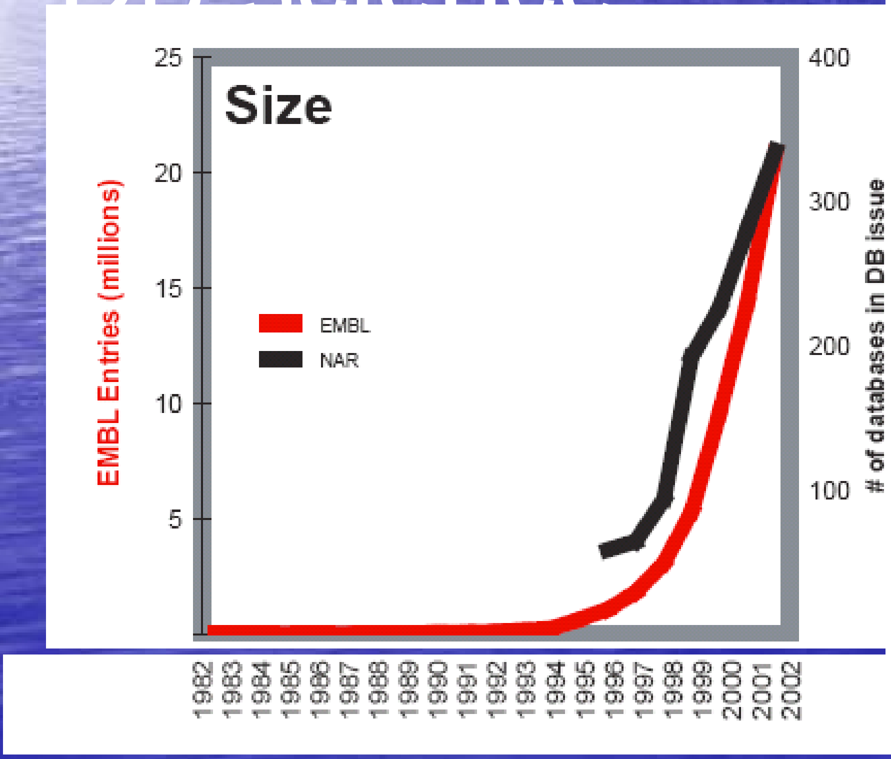
- 250 of ~530
  originally characterized on chr. 22
  [Dunham et al. Nature (1999)]

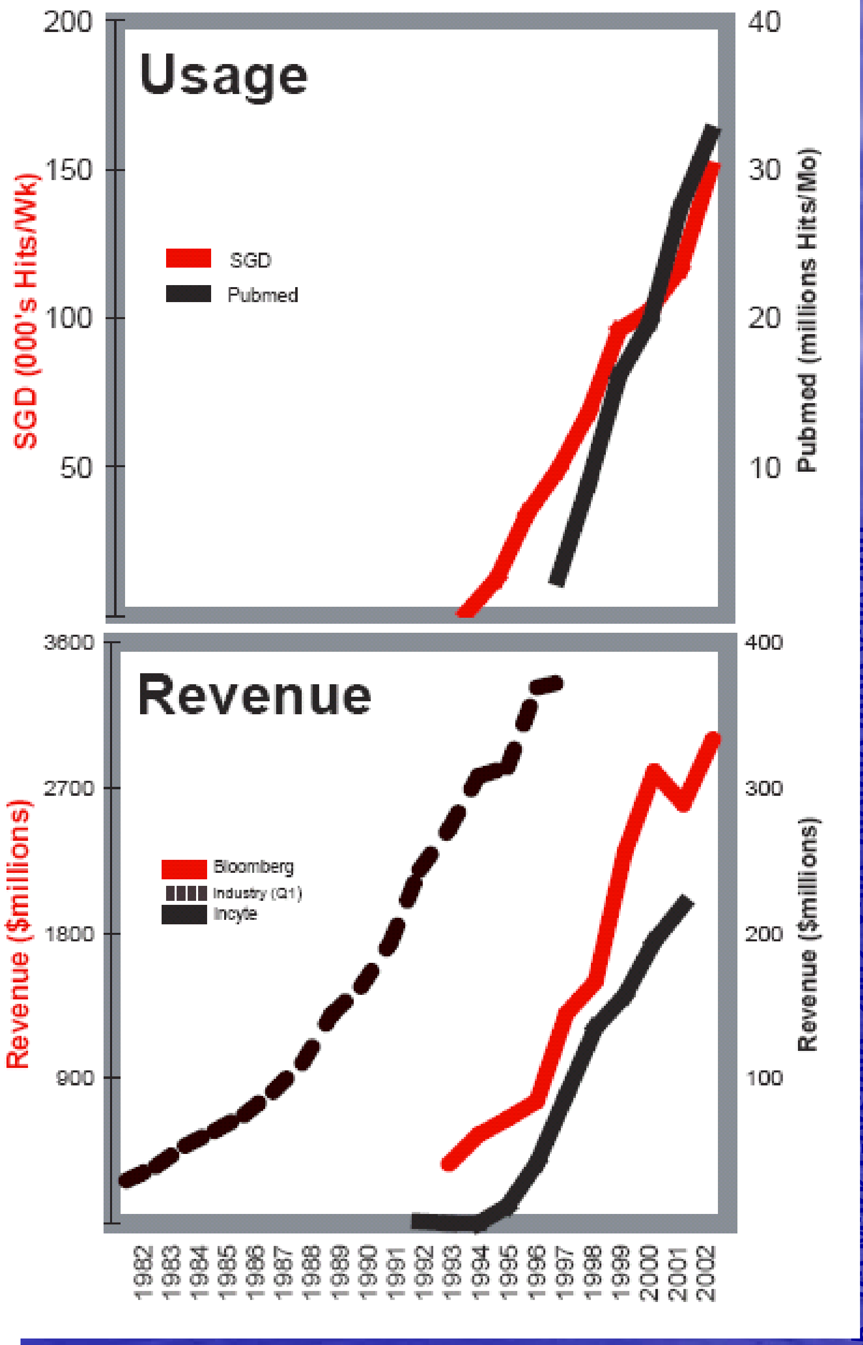
- >25K Proteins in Entire Human Genome
  (with alt. splicing)

EXTRA

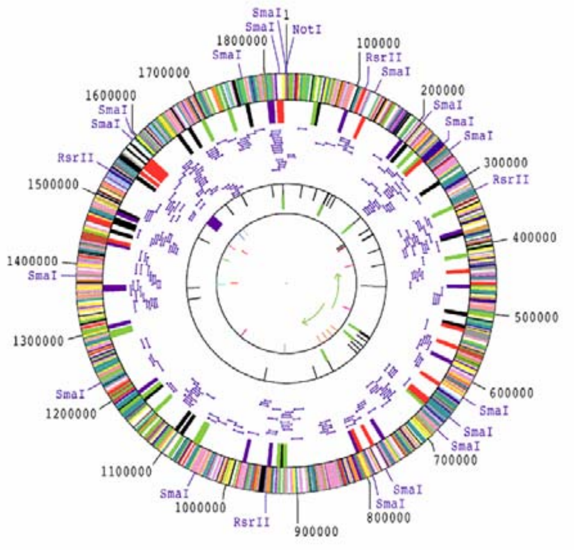
Why Networks? Need for an Edge Ontology

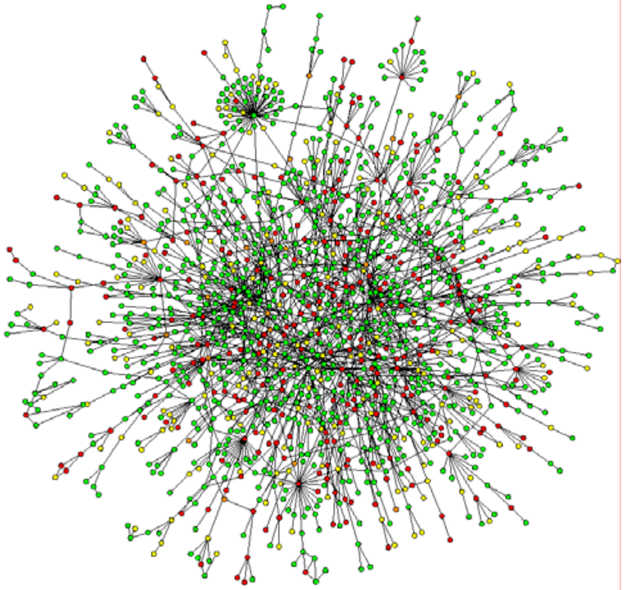
Rapid growth in DBs in science spurring on DB science

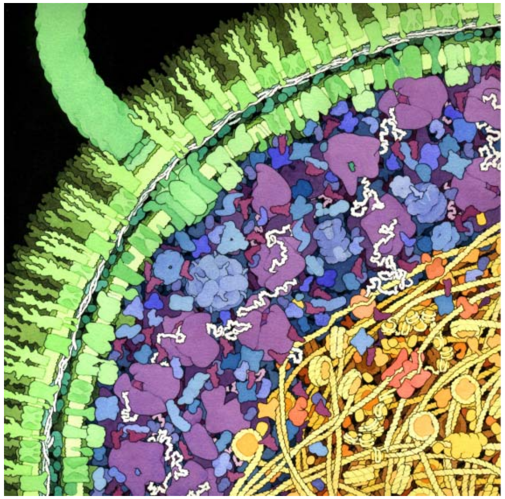




Networks occupy a midway point in terms of level of understanding







1D: Complete Genetic Partslist

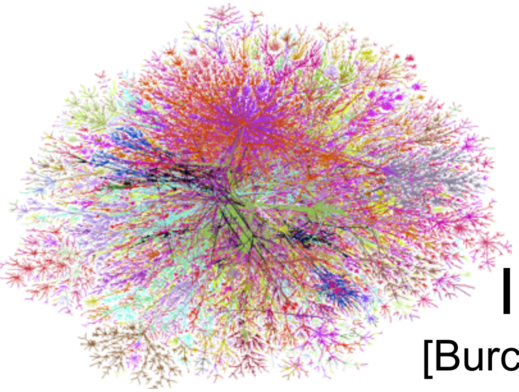
~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

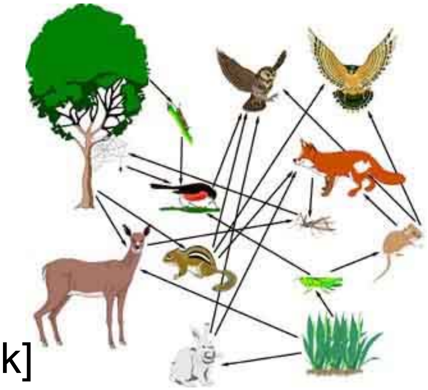
[Fleischmann et al., Science, 269 :496]

[Jeong et al. Nature, 41:411]

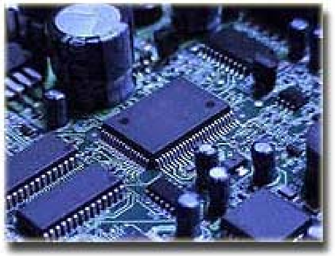
Networks as a universal language



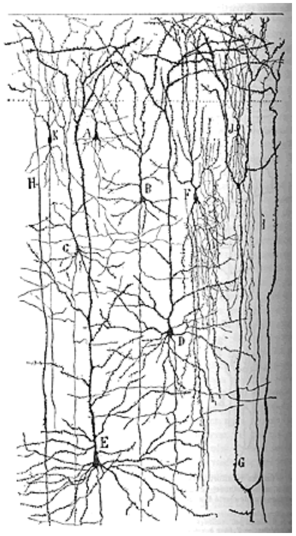
Internet
[Burch & Cheswick]



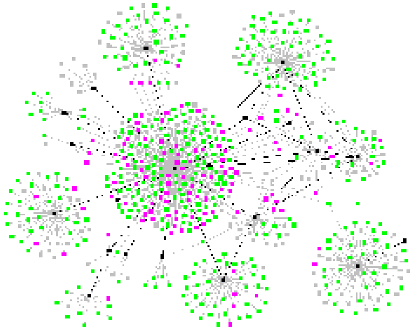
Food Web



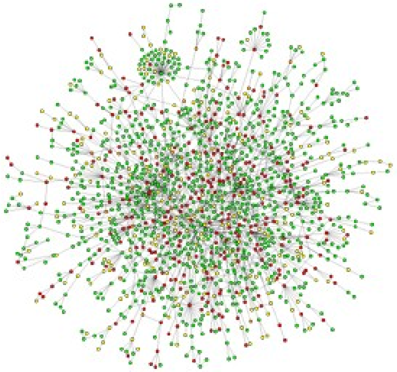
Electronic Circuit



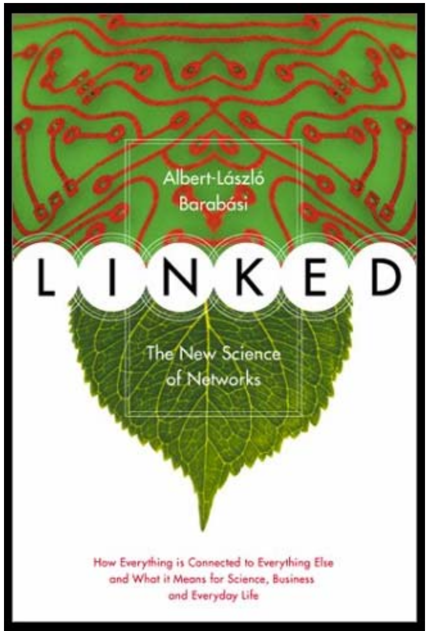
Neural Network
[Cajal]



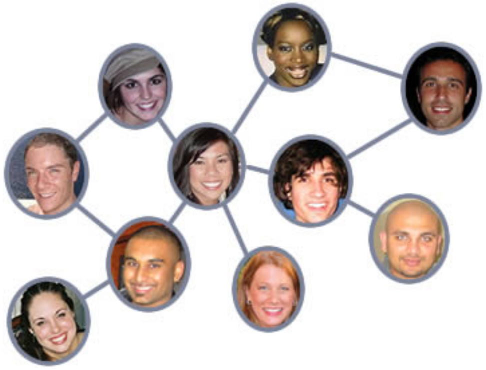
Disease Spread
[Krebs]



Protein Interactions
[Barabasi]

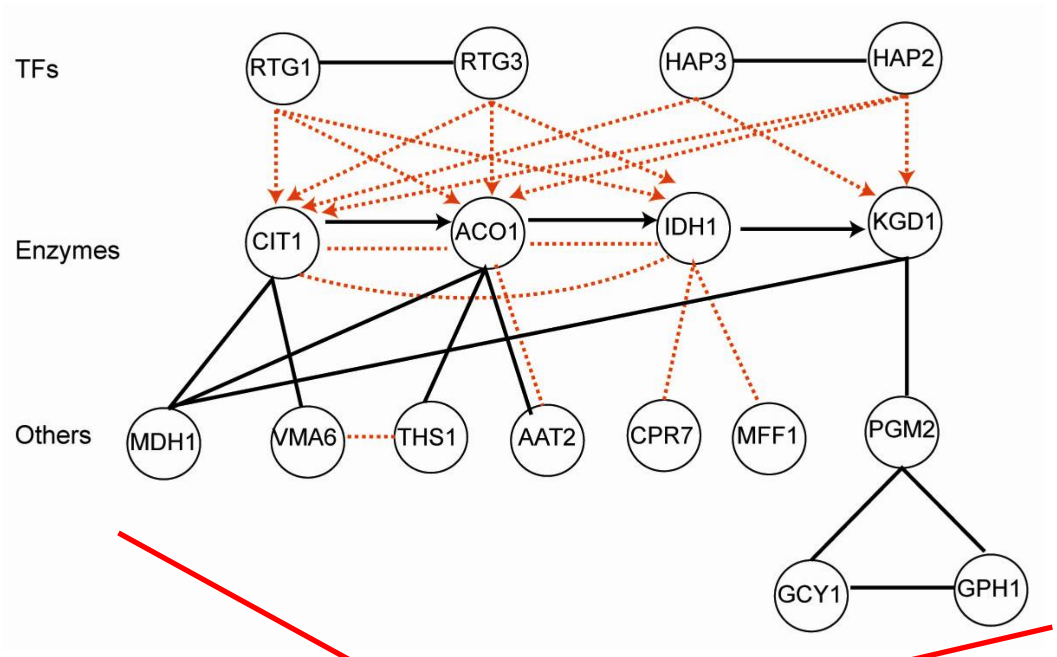


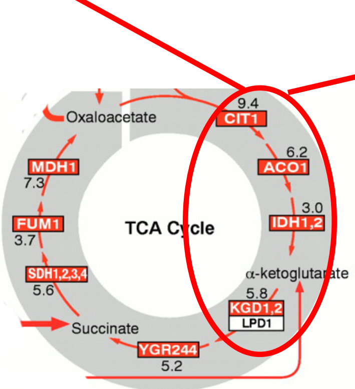
Albert-László Barabási
LINKED
The New Science of Networks
How Everything is Connected to Everything Else and What it Means for Science, Business and Everyday Life



Social Network

Combining networks forms an ideal way of integrating diverse information





Part of the TCA cycle

⟶ **Metabolic pathway**

⋯▶ **Transcriptional regulatory network**

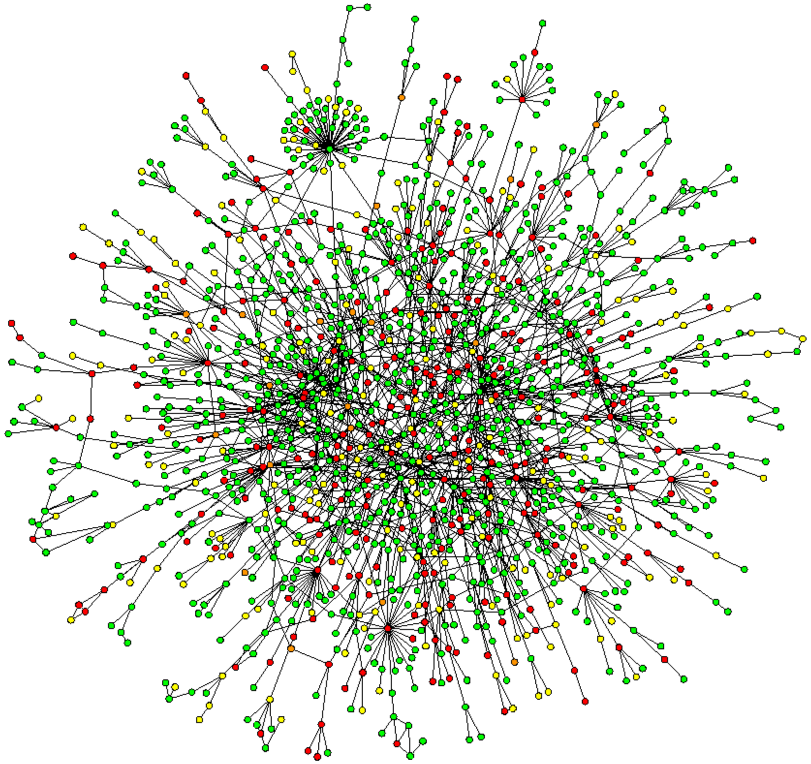
― Physical protein-protein Interaction

⋯⋯ Co-expression Relationship

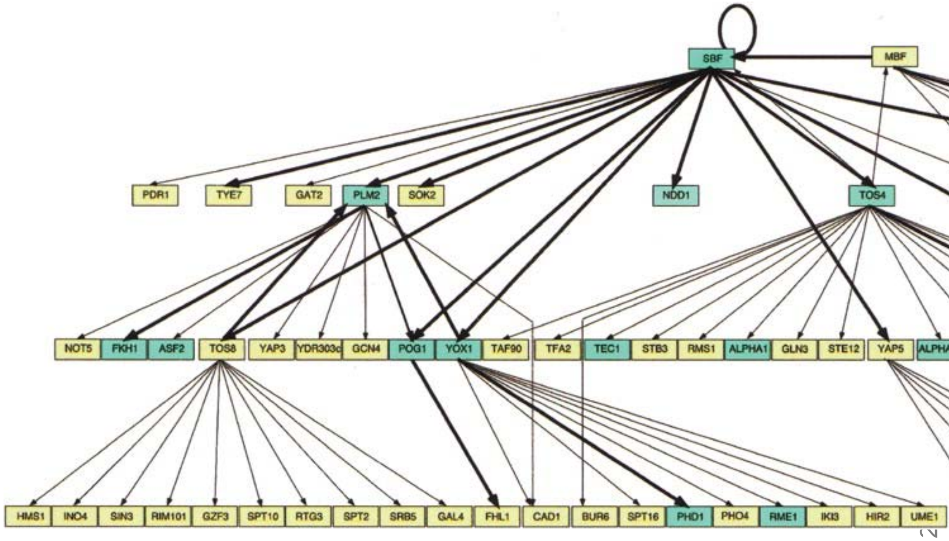
Genetic interaction (synthetic lethal)
Signaling pathways

What do Biological Networks Look Like: High-throughput Networks v Classical Pathways

Highly Standardized, High-throughput Biological networks



Interactions networks

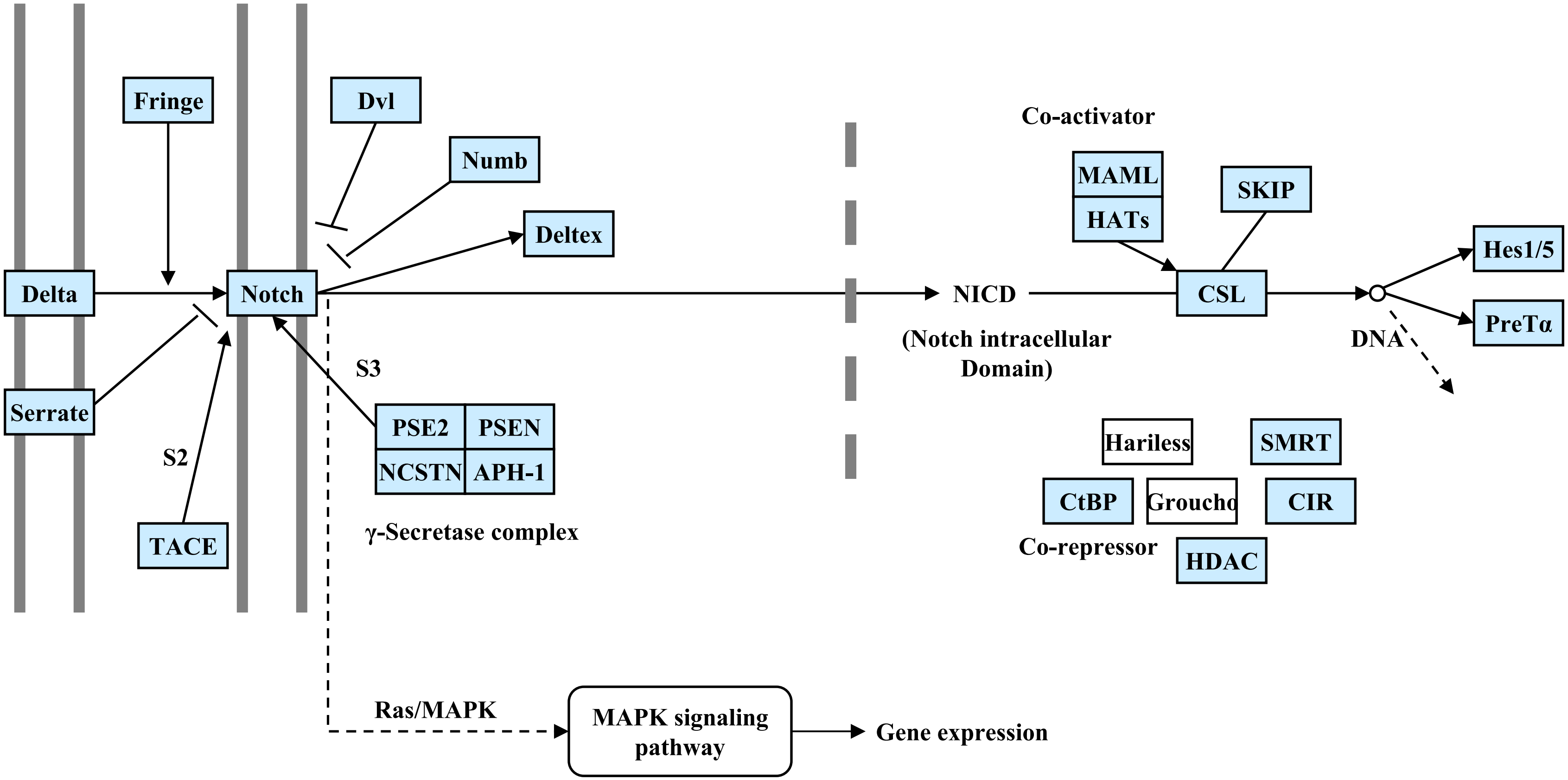


Regulatory networks

[Graphic: Jeong et al, Nature, 41:411]

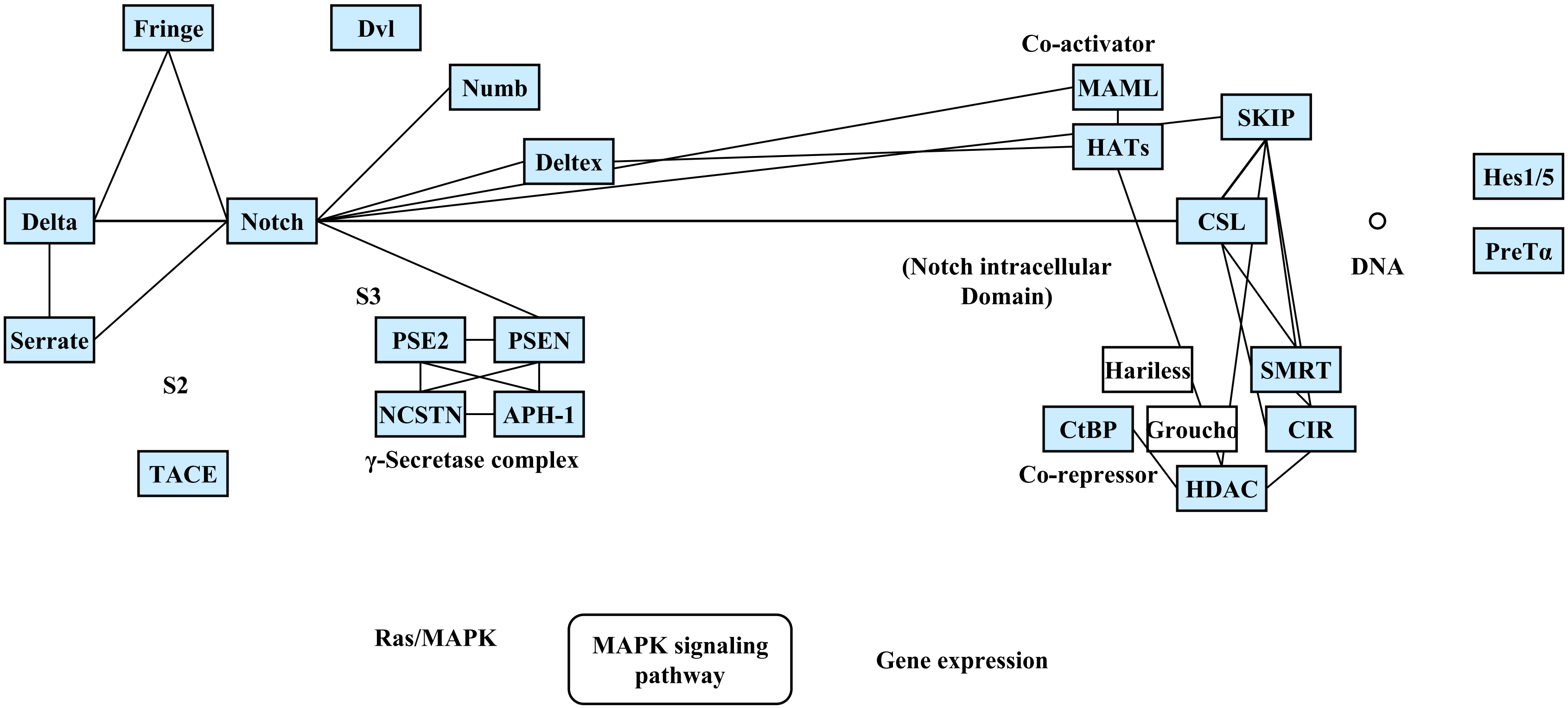
[Horak, et al, Genes & Development, 16:3017-3033]

[Jeong et al, Nature, 41:411]



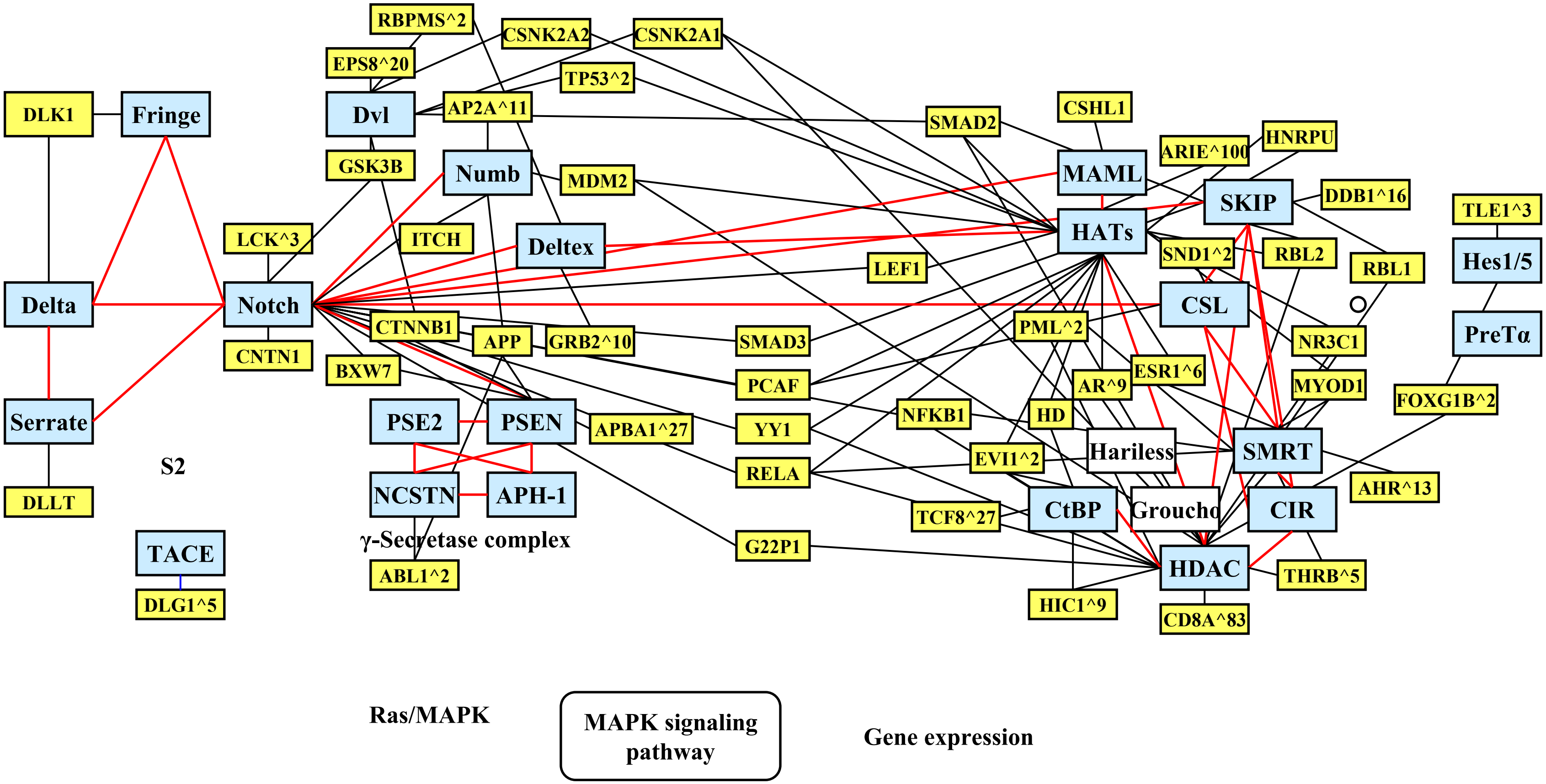
Classical v High-throughput: Classical Notch Pathway

[Lu et al. *TIG* (2007)]



Classical v High-throughput: Notch Embedded in High-throughput Data

[Lu et al. *TIG* (2007)]

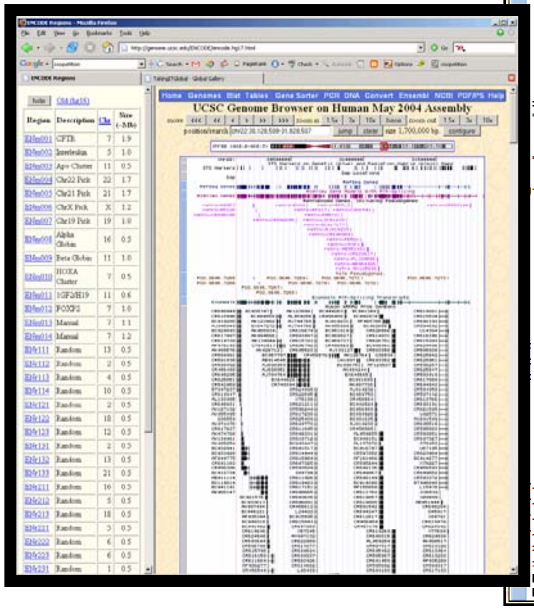


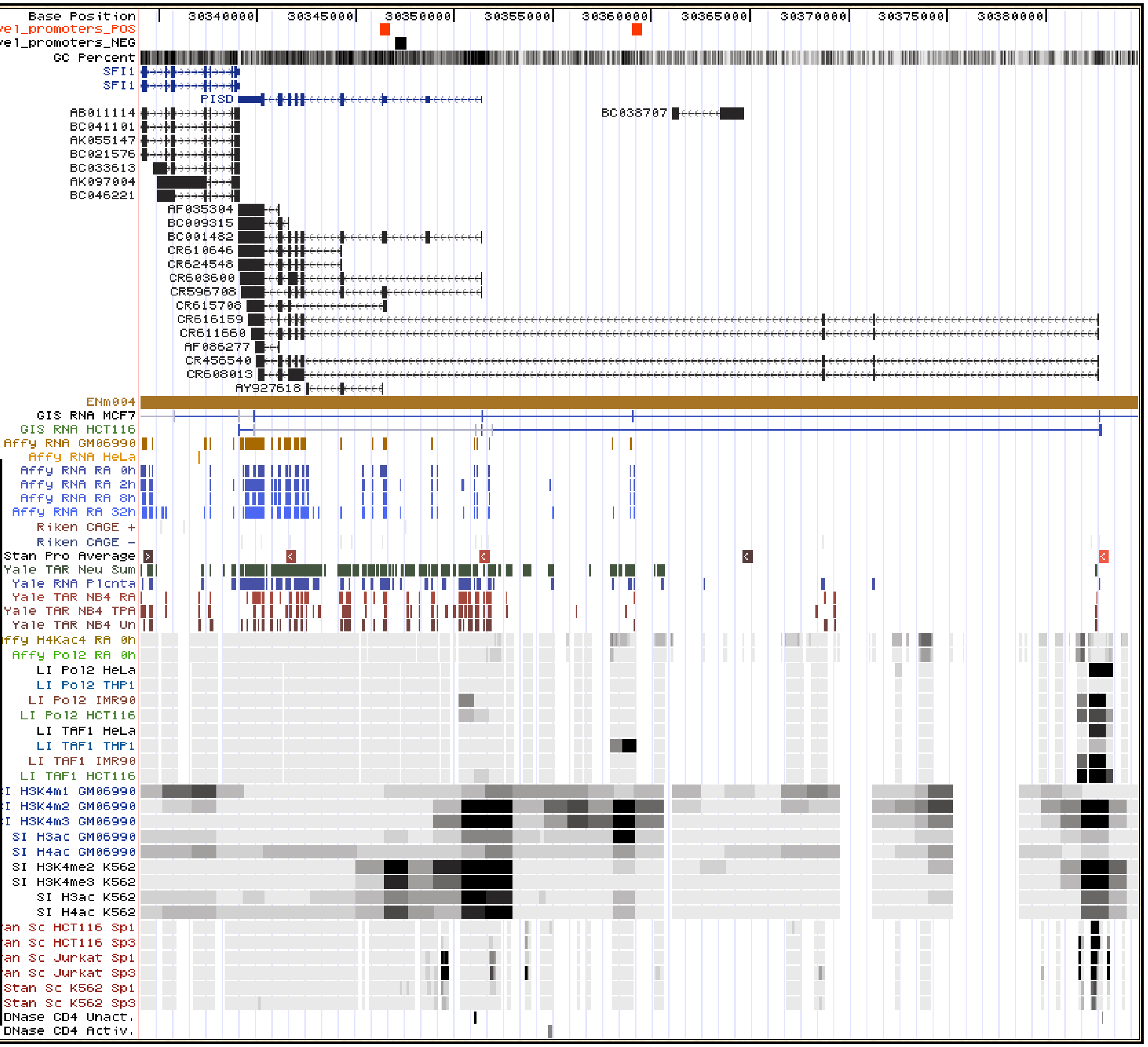
Classical v High-throughput: Core v Extended Interactions

[Lu et al. *TIG* (2007)]

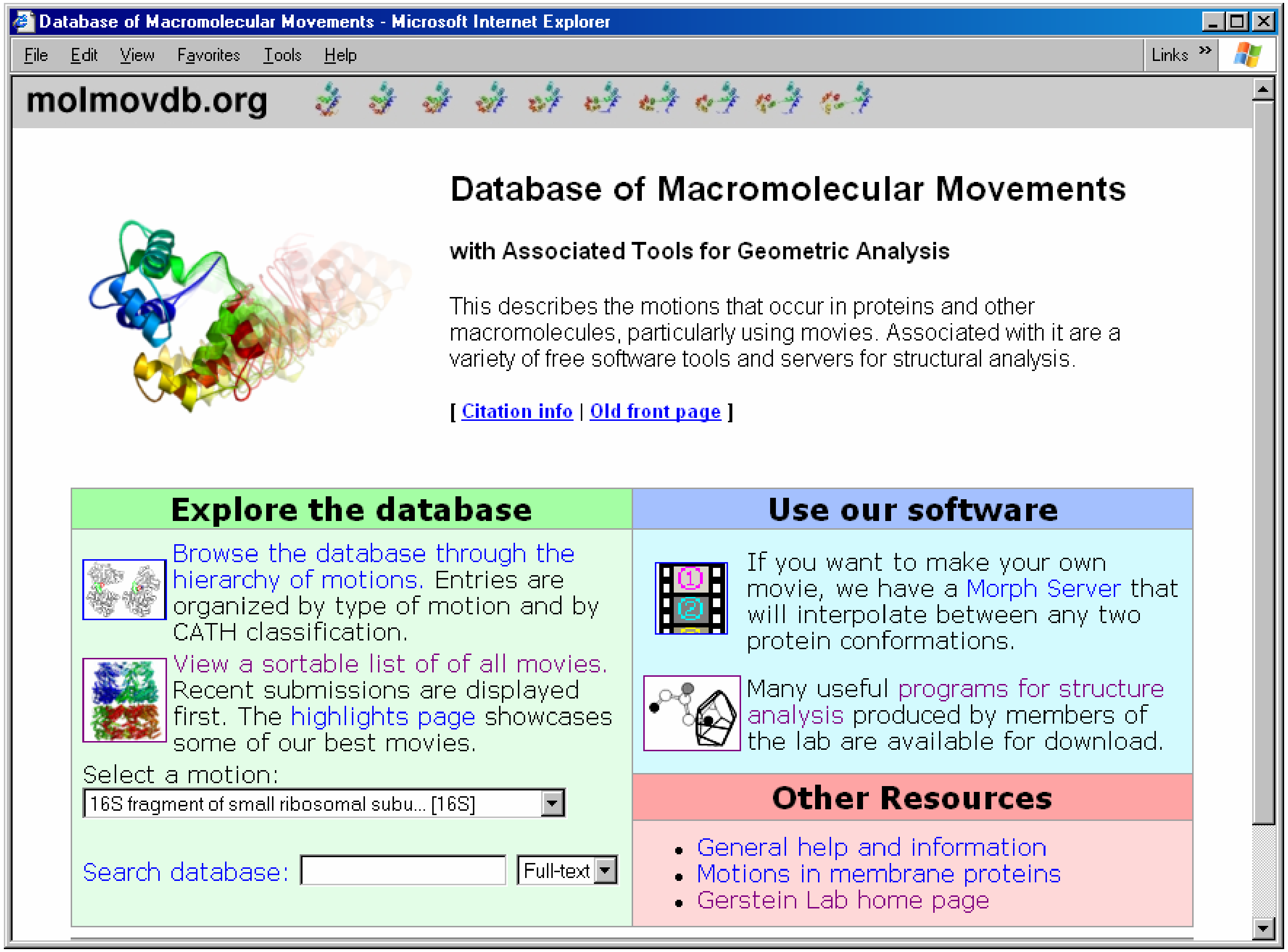
Central Hub DBs

- Genome browser giving overview of whole genome (Google Earth)
- GenBank, UniProt, genome.ucsc.edu, PDB
- Unforeseen "power"





Specialized "Boutique" Databases



Database of Macromolecular Movements - Microsoft Internet Explorer

Database of Macromolecular Movements

with Associated Tools for 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 | Old front page]

Explore the database

Browse the database through the hierarchy of motions. Entries are organized by type of motion and by CATH classification.

View a sortable list of of all movies. Recent submissions are displayed first. The highlights page showcases some of our best movies.

Select a motion: 16S fragment of small ribosomal subu... [16S]

Search database: Full-text

Use our software

If you want to make your own movie, we have a Morph Server that will interpolate between any two protein conformations.

Many useful programs for structure analysis produced by members of the lab are available for download.

Other Resources

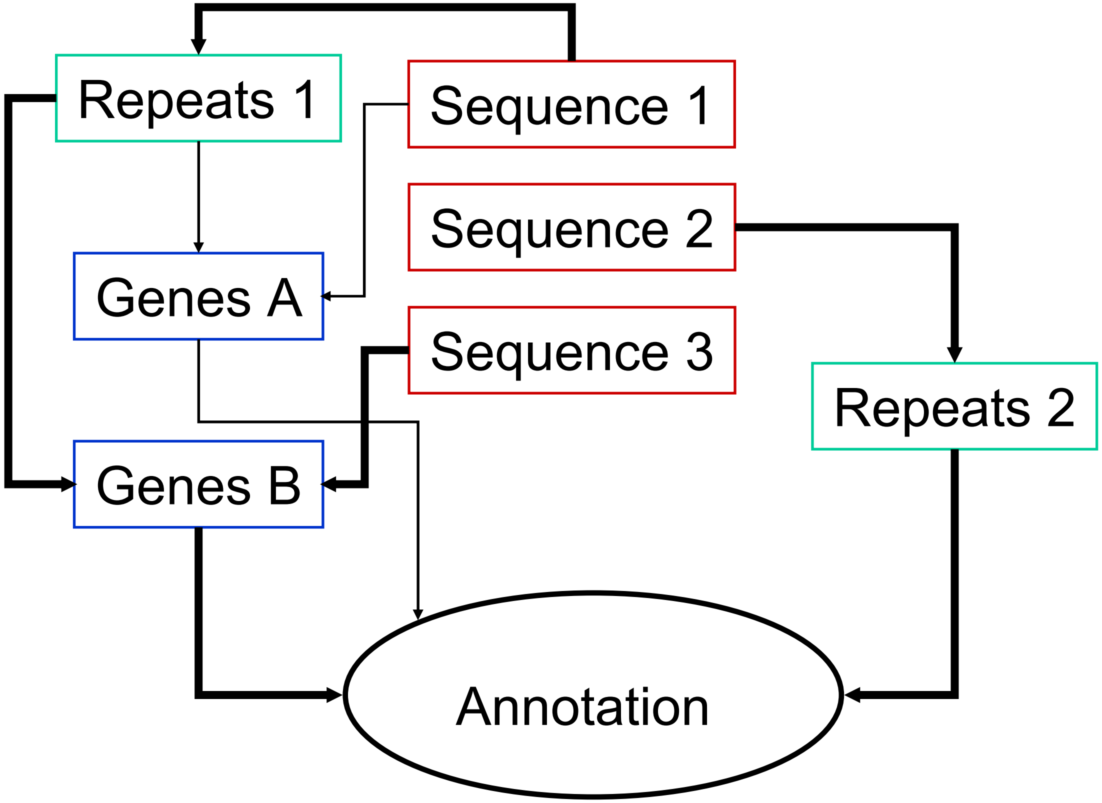
- General help and information
- Motions in membrane proteins
- Gerstein Lab home page

MolMovDB.org - Molecular detail about individual gene

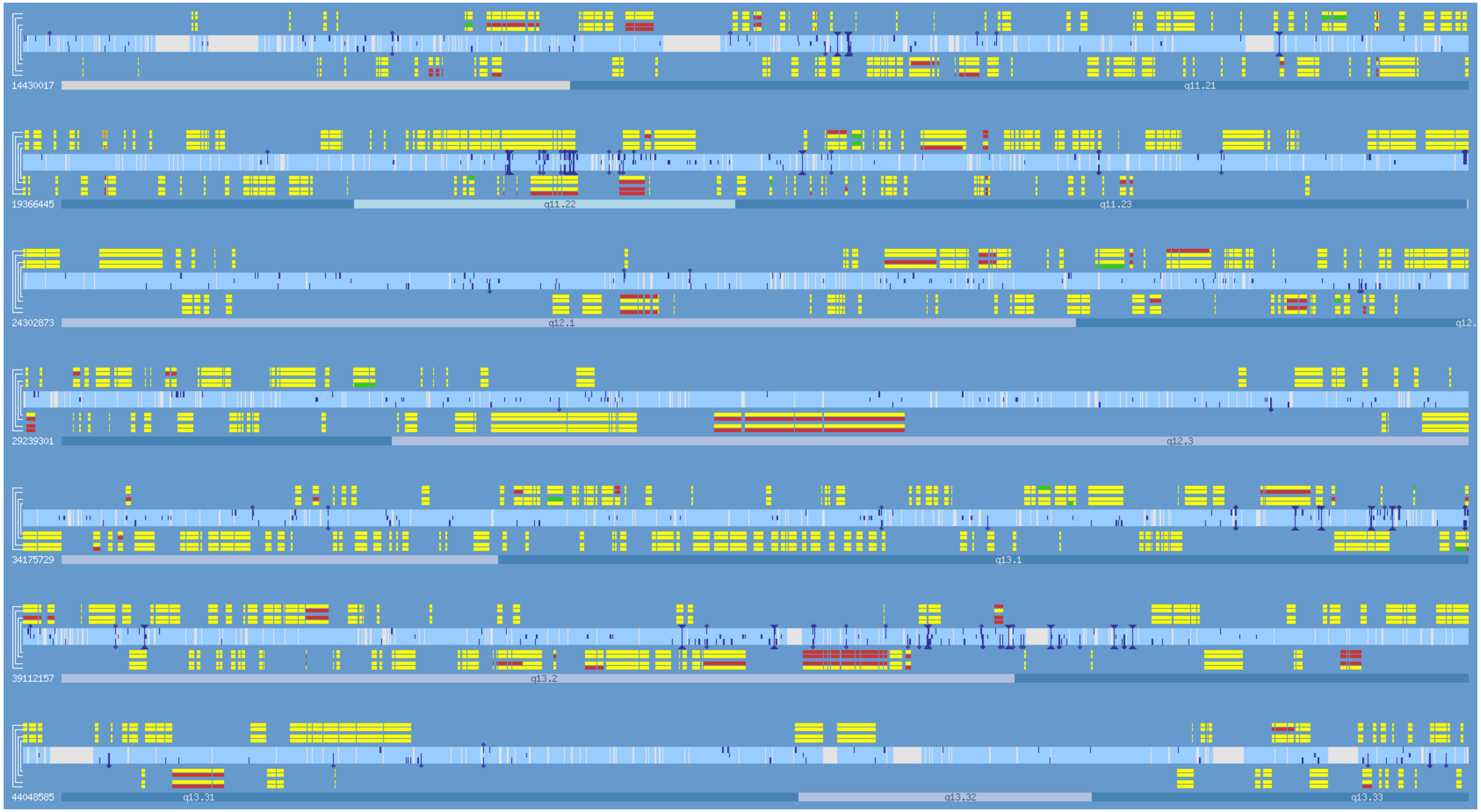
Aspect #1: Intimate Synchronization between Sites, Propagating Dynamic Annotation

- Grappling with changing coordinates & annotation

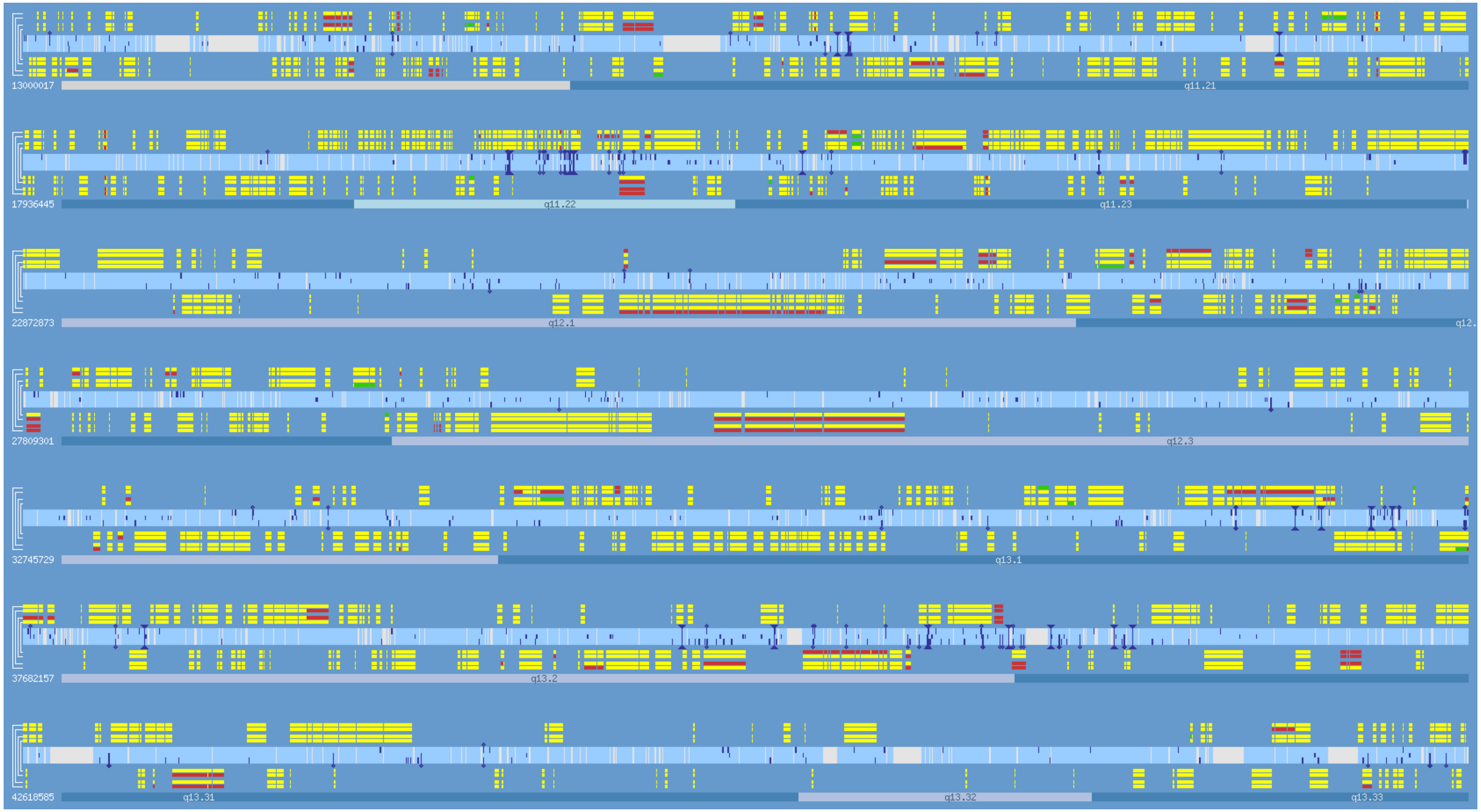
- Complex dependencies between sites



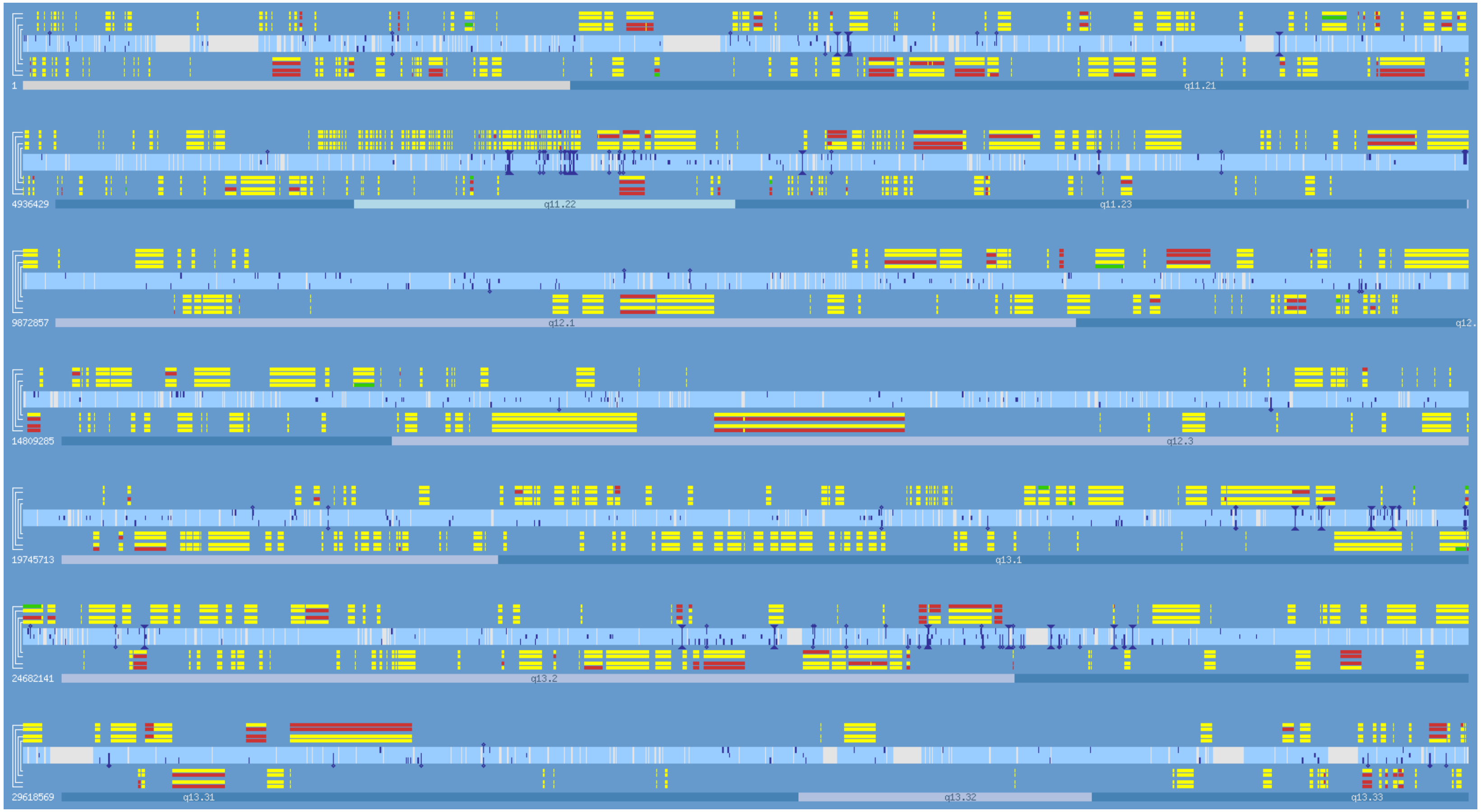
Dynamic Annotation
Ensembl 18.34



Dynamic Annotation
Sanger 2.3



Dynamic Annotation
Sanger 3.1b



law v sci