Structured Digital Literature, a perspective on sharing code and data

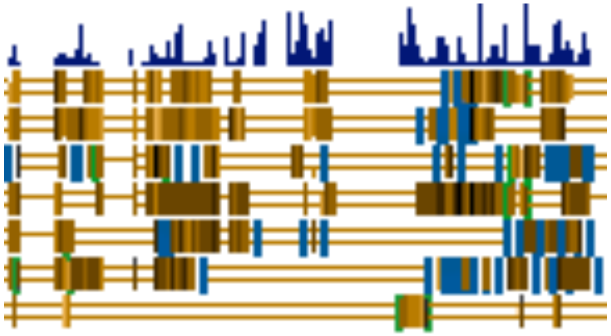
Mark B Gerstein
Yale

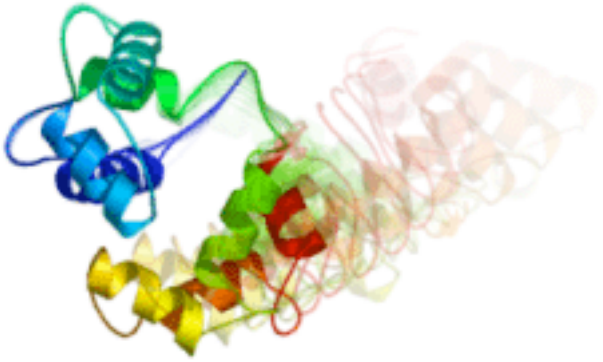
Slides at **Lectures.GersteinLab.org**

(See Last Slide for References & More Info.)

GersteinLab.org Research Overview: Bioinformatics

- **Genome Annotation**
  - ◊ Characterizing the function of non-coding regions of the genome, focusing on protein fossils and novel RNAs (Pseudogene.org + GenomeTech.GersteinLab.org)
- **Molecular Networks**
  - ◊ Using molecular networks to integrate & mine functional genomics information and describe genefunction 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)





In the course of this research....

- Analyze genome-scale experimental datasets
  - ◊ Different scales
    (excel file to >.1 PB next-gen seq. of populations)
- Generate software tools
  - ◊ distributable standalone code packages, webservers, plugins
- Produce large-scale annotation sets
  - ◊ Highly synthetic: reference particular datasets, code versions and "genome builds"
- Work in Consortia
  - ◊ mod/ENCODE, 1KG, PSI, &c

- **Publish results**

How do these connect?

Information Resources & Journals: Two ends of a blurring spectrum

- Distinctions Blurring
  - ◊ Reading Journals via queries
    - Reading DB entries
  - ◊ Towards reading literature with computers
    - Mining text and correlating papers
  - ◊ Distinction between analysis procedure described in article vs. computer code on repository

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

Other Issues with the Current Situtation between DBs & Journals

- Not always a clear linkage between papers & DBs
  - ◊ Keeping entries in DB and paper in sync
    - Numbers of genes in the paper vs on the the webite
- 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

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

The Solution?

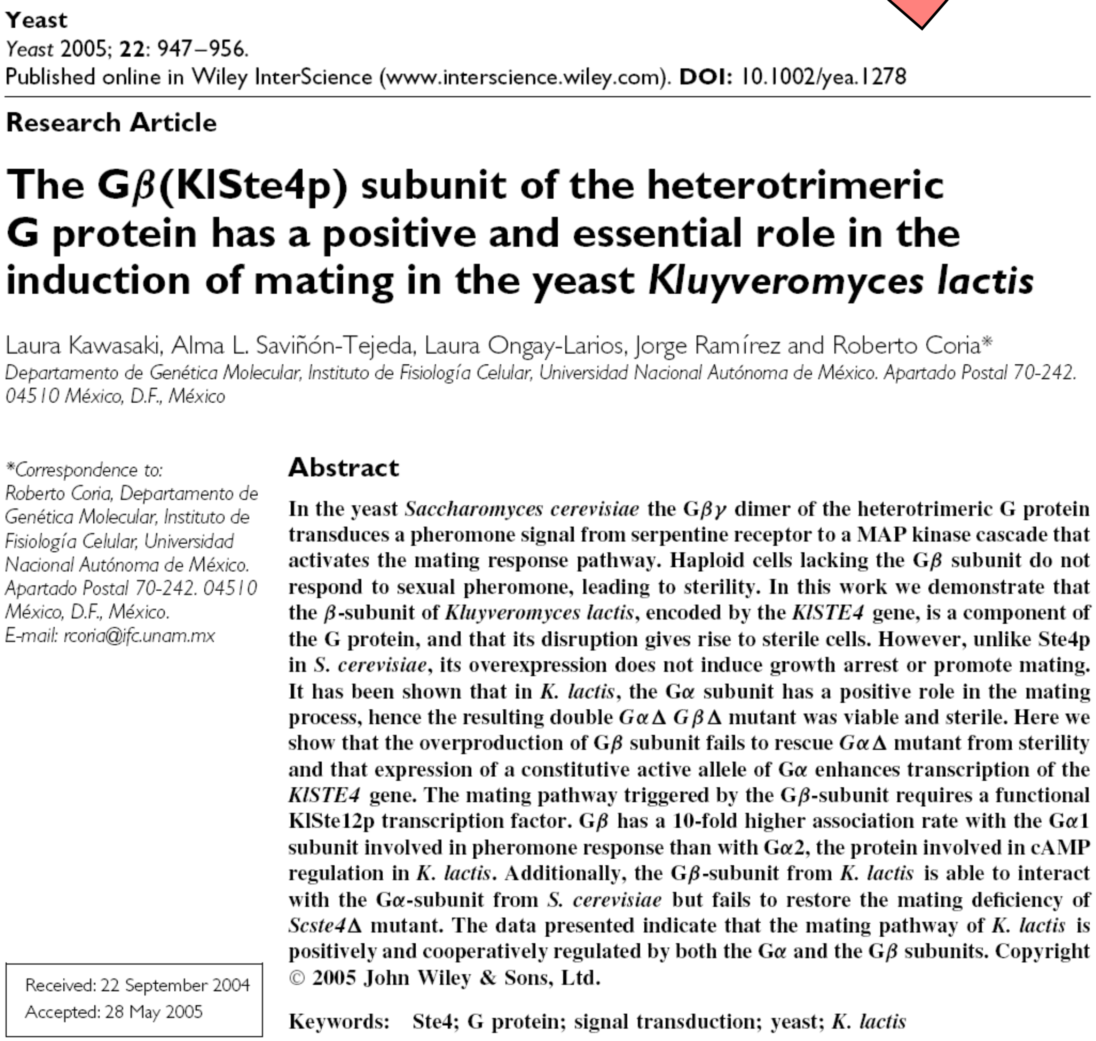
- Ignore papers
  - Just post to blogs, distriubute free software, deposit into datasets, &c
- Structure the scientific literature to make it more compatible with a digital future...
  - Strutured Digital Paper (Structured Abs., Table, Equation...)

Structured Abstract Proposal as a 1st step

- 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

[Seringhaus & Gerstein, FEBS ('08); 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*

Received: 22 September 2004
Accepted: 28 May 2005

Abstract

In the yeast *Saccharomyces cerevisiae* the Gβγ 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*α∆ *G*β∆ mutant was viable and sterile. Here we show that the overproduction of Gβ subunit fails to rescue *G*α∆ 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*∆ 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.

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

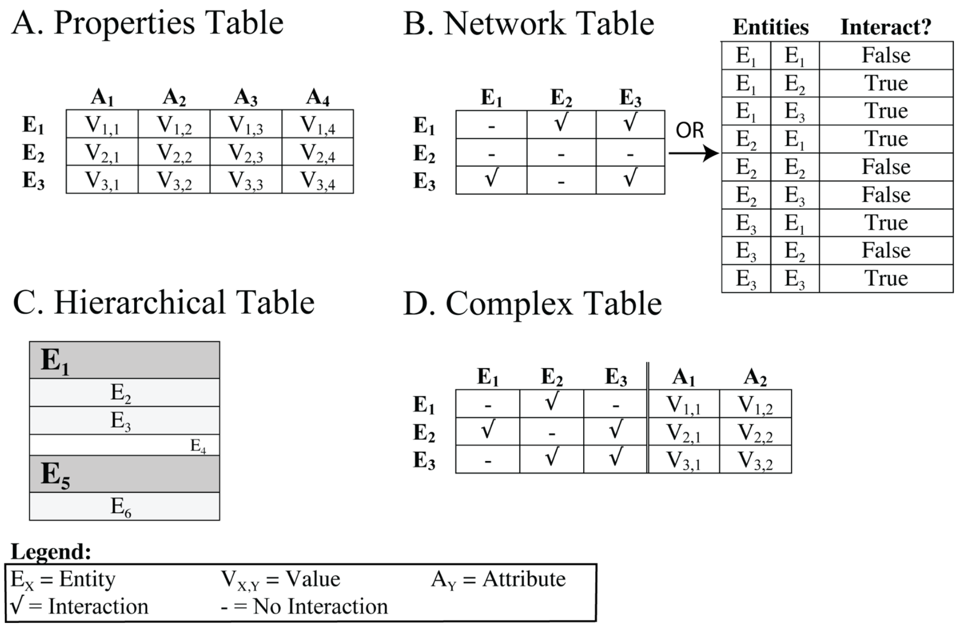
- ***K.lactis*** (species)
  - ◊ **KlSTE4** (gene)
    - **KlSte4p** (protein)
      - – **CLONED**
        - » Available at …
      - – **SEQUENCED**
        - » Sequence ATGTACGCTATAGGC….
      - – **MUTANTS**
        - » **DELETION**
        - » **FUNCTIONAL ASSAYS**
        - » Sterile in both MATa and MAT α
        - » 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)
  - ◊ **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**

[Seringhaus & Gerstein, *BMC Bioinformatics* (2007)]

Structured Digital Table

- Canonical Table Types
- Converting a journal table into these
- Using standardized journal tables as small "stubb" tables for larger datasets

[Cheung et al., MSB, in revision]



A. Properties Table

	A_1	A_2	A_3	A_4
E_1	$V_{1,1}$	$V_{1,2}$	$V_{1,3}$	$V_{1,4}$
E_2	$V_{2,1}$	$V_{2,2}$	$V_{2,3}$	$V_{2,4}$
E_3	$V_{3,1}$	$V_{3,2}$	$V_{3,3}$	$V_{3,4}$

B. Network Table

	E_1	E_2	E_3
E_1	-	√	√
E_2	-	-	-
E_3	√	-	√

OR

Entities		Interact?
E_1	E_1	False
E_1	E_2	True
E_1	E_3	True
E_2	E_1	True
E_2	E_2	False
E_2	E_3	False
E_3	E_1	True
E_3	E_2	False
E_3	E_3	True

C. Hierarchical Table

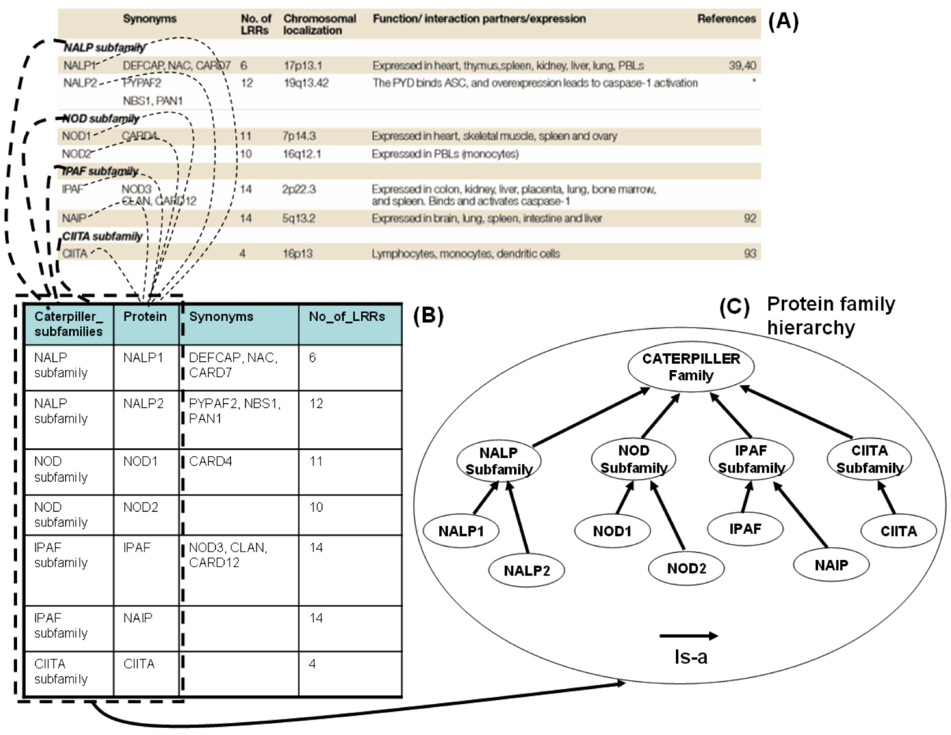
E_1
E_2
E_3
E_4
E_5
E_6

D. Complex Table

	E_1	E_2	E_3	A_1	A_2
E_1	-	√	-	$V_{1,1}$	$V_{1,2}$
E_2	√	-	√	$V_{2,1}$	$V_{2,2}$
E_3	-	√	√	$V_{3,1}$	$V_{3,2}$

Legend:

E_X = Entity $V_{X,Y}$ = Value A_Y = Attribute
√ = Interaction - = No Interaction



(A)

	Synonyms	No. of LRRs	Chromosomal localization	Function/ interaction partners/expression	References
NALP subfamily					
NALP1	DEFCAP, NAC, CARD7	6	17p13.1	Expressed in heart, thymus,spleen, kidney, liver, lung, PBLs	39,40
NALP2	PYPAF2 NBS1, PAN1	12	19q13.42	The PYD binds ASC, and overexpression leads to caspase-1 activation	*
NOD subfamily					
NOD1	CARD4	11	7p14.3	Expressed in heart, skeletal muscle, spleen and ovary	
NOD2		10	16q12.1	Expressed in PBLs (monocytes)	
IPAF subfamily					
IPAF	NOD3 CLAN, CARD12	14	2p22.3	Expressed in colon, kidney, liver, placenta, lung, bone marrow, and spleen. Binds and activates caspase-1	
NAIP		14	5q13.2	Expressed in brain, lung, spleen, intestine and liver	92
CIITA subfamily					
CIITA		4	16p13	Lymphocytes, monocytes, dendritic cells	93

(B)

Caterpiller_ subfamilies	Protein	Synonyms	No_of_LRRs
NALP subfamily	NALP1	DEFCAP, NAC, CARD7	6
NALP subfamily	NALP2	PYPAF2, NBS1, PAN1	12
NOD subfamily	NOD1	CARD4	11
NOD subfamily	NOD2		10
IPAF subfamily	IPAF	NOD3, CLAN, CARD12	14
IPAF subfamily	NAIP		14
CIITA subfamily	CIITA		4

(C) Protein family hierarchy

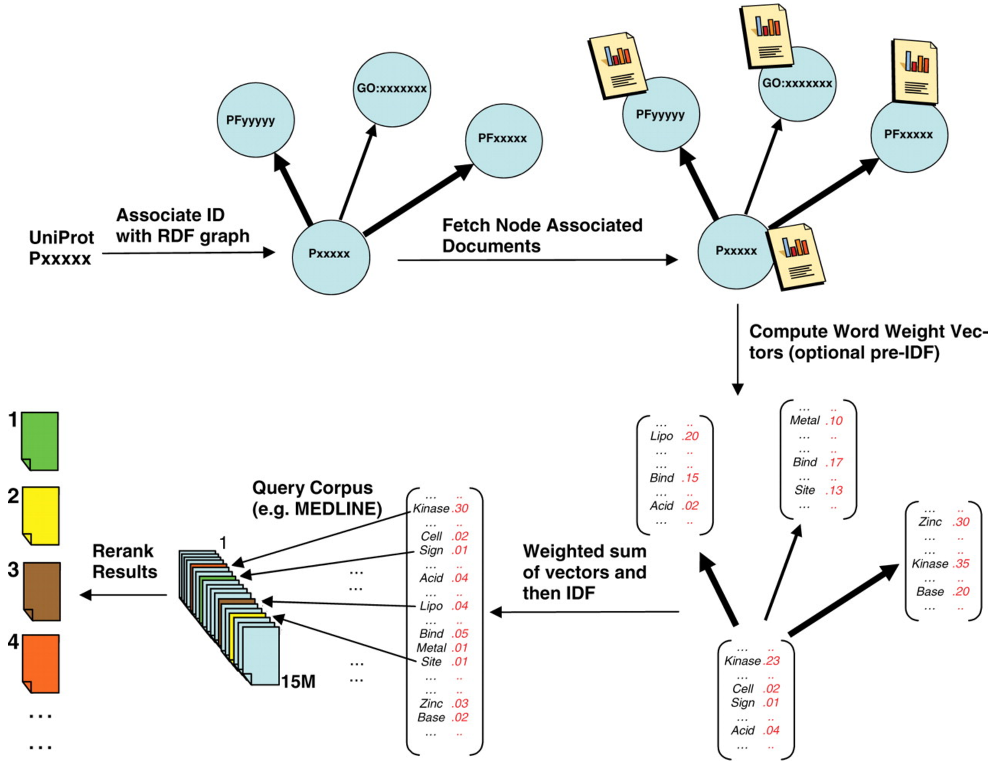
Towards a structured digital literature

- Structured Fig. Captions
  - ◊ MurphyLab @ CMU (A. Ahmed et al. KDD-2009, pp. 39-47)
- Structured equations & pseudocode
  - ◊ Directly convertable into real code

- What are the applications of this...

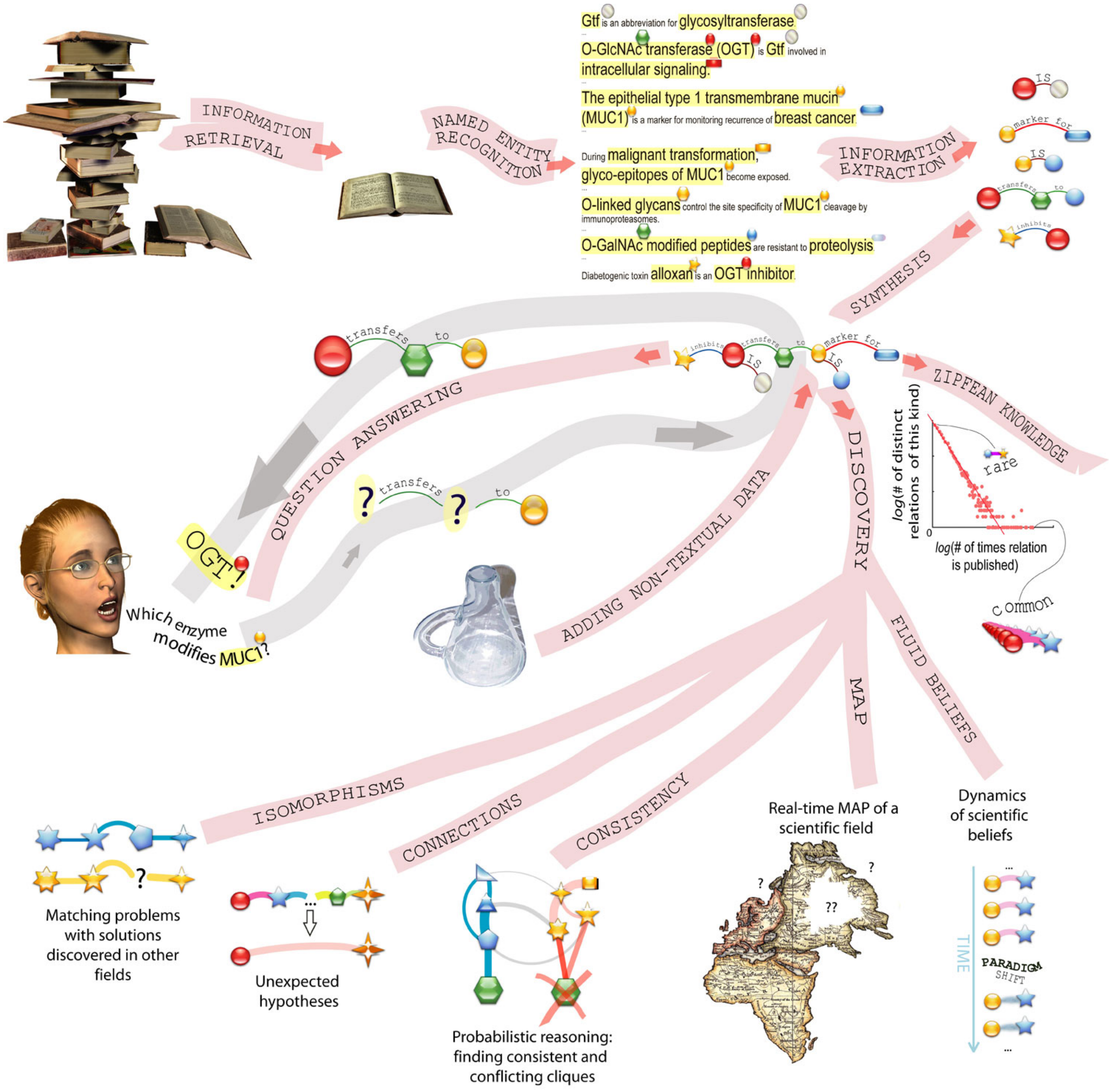
Unsupervised Textmining vs Manually Curated and Structured Documents: Not necessarily a conflict

- Relatively small numb. of structured papers might be good training sets for mining
- Also, gateway to mining (e.g. listing std. names for genes as cast of char., highlighting foreground v. background concepts)



[Smith et al., Bioinformatics ('07)]

Vision for Mining Large-scale Structured Literature



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.

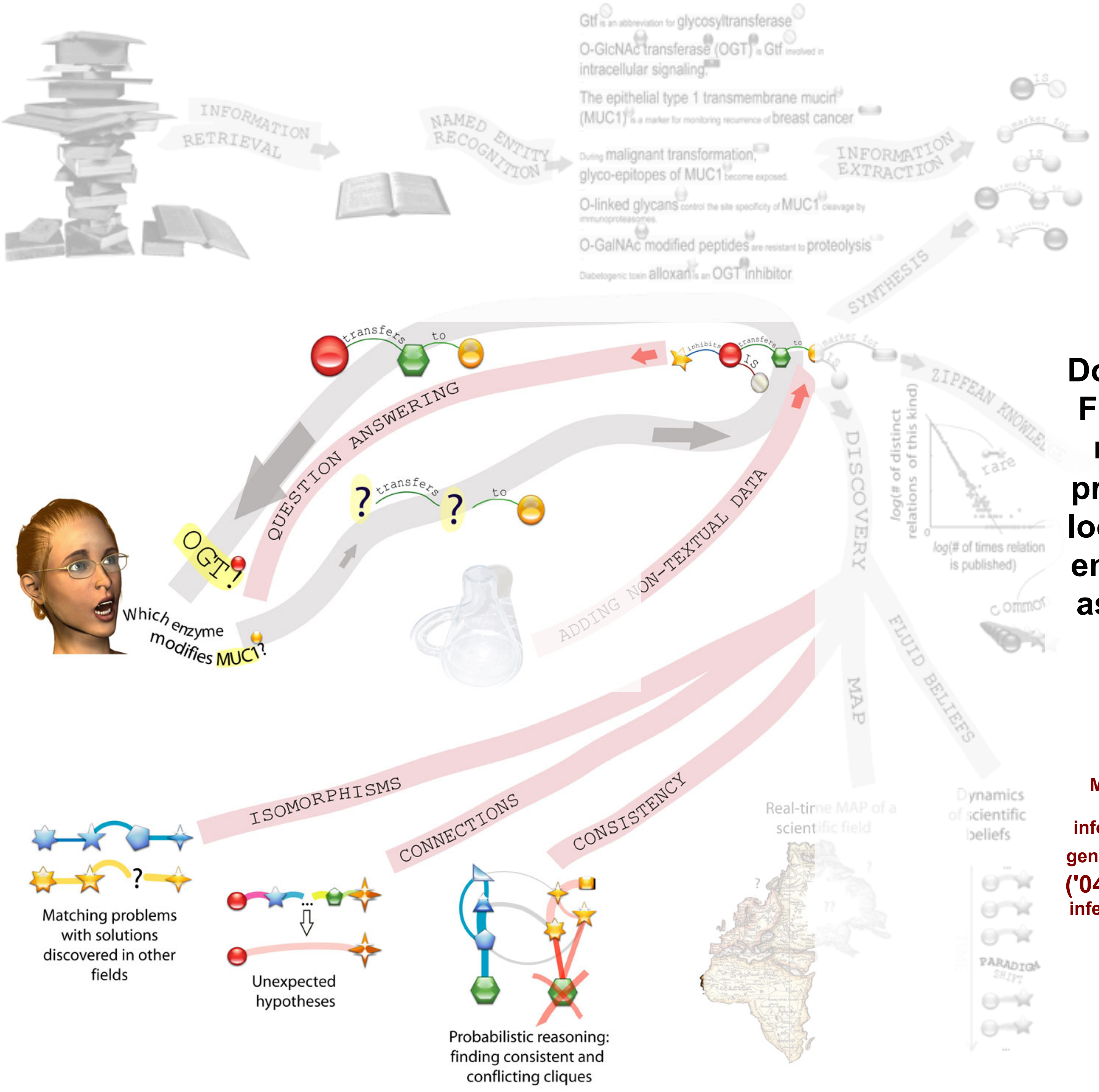
O-GalNAc modified peptides are resistant to proteolysis.

Diabetogenic toxin alloxan is an OGT inhibitor.

Which enzyme modifies MUC1?

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

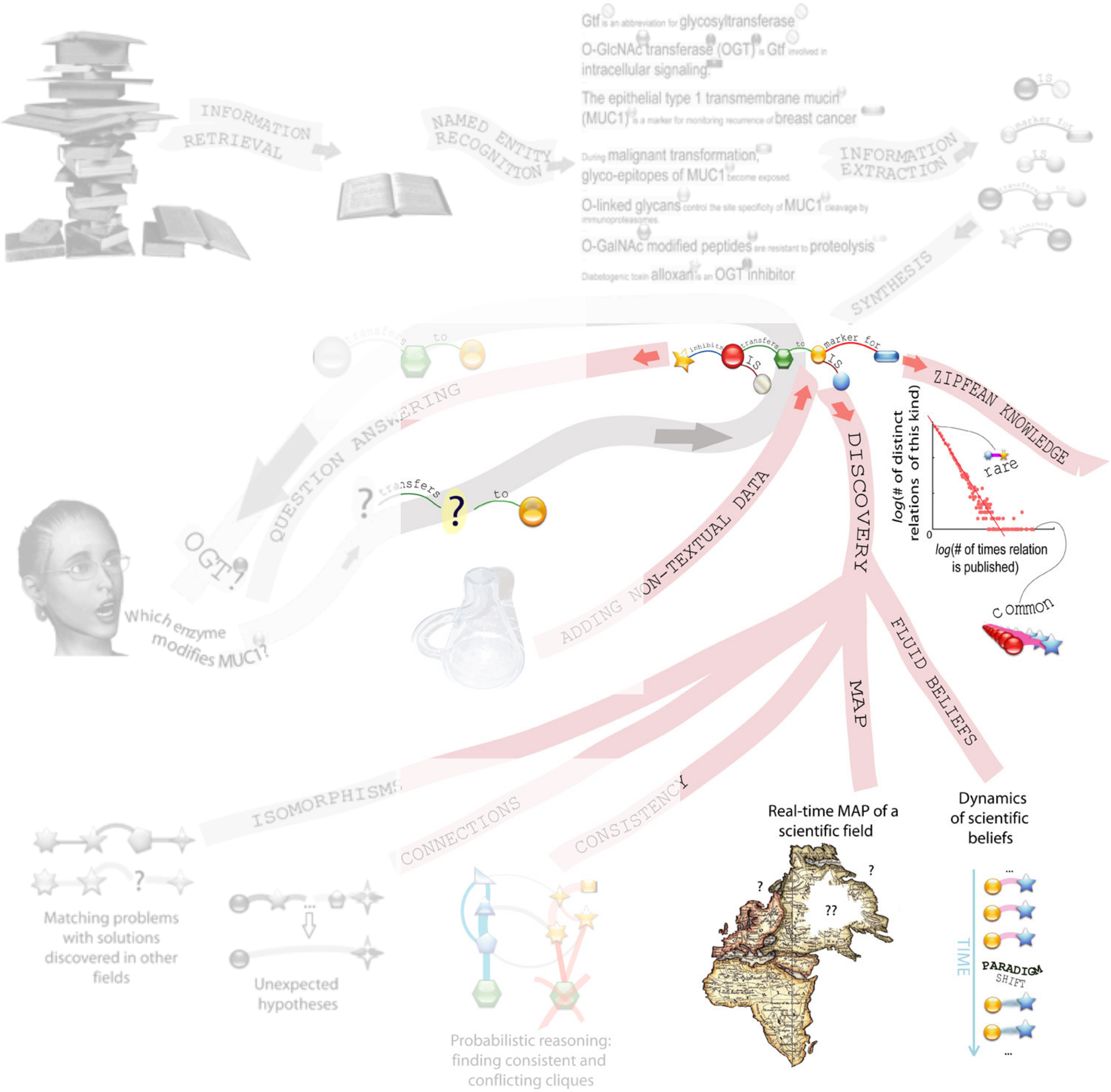
Vision for Mining Large-scale Structured Literature


INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
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
INFORMATION EXTRACTION
IS
marker for
IS
transfers
to
SYNTHESIS
transfers
to
inhibits
transfers
to
IS
marker for
QUESTION ANSWERING
ZIPFEAN KNOWLEDGE
rare
log(# of distinct relations of this kind)
log(# of times relation is published)
common
DISCOVERY
transfers
to
?
?
OGT!
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
MAP
FLUID BELIEFS
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
Real-time MAP of a scientific field
Dynamics of scientific beliefs
PARADIGM SHIFT
Matching problems with solutions discovered in other fields
Unexpected hypotheses
Probabilistic reasoning: finding consistent and conflicting cliques

Doing better science: Finding new protein relationships (e.g. protein interactions), looking for inconsistencies in arguments, assembling consensus definitions automatically

Krauthammer et al. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. PNAS ('04); Iossifov et al. Probabilistic inference of molecular networks from noisy data sources. Bioinformatics ('04)

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]


INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
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
INFORMATION EXTRACTION
IS
marker for
IS
transfers to
conjugate
SYNTHESIS
transfers to
inhibits
transfers
to
marker for
IS
IS
QUESTION ANSWERING
? transfers ? to
OGT?
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
DISCOVERY
ZIPFEAN KNOWLEDGE
log(# of distinct relations of this kind)
log(# of times relation is published)
0
rare
common
FLUID BELIEFS
MAP
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
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
TIME
PARADIGM SHIFT
...
...

Vision for Mining Large-scale Structured Literature

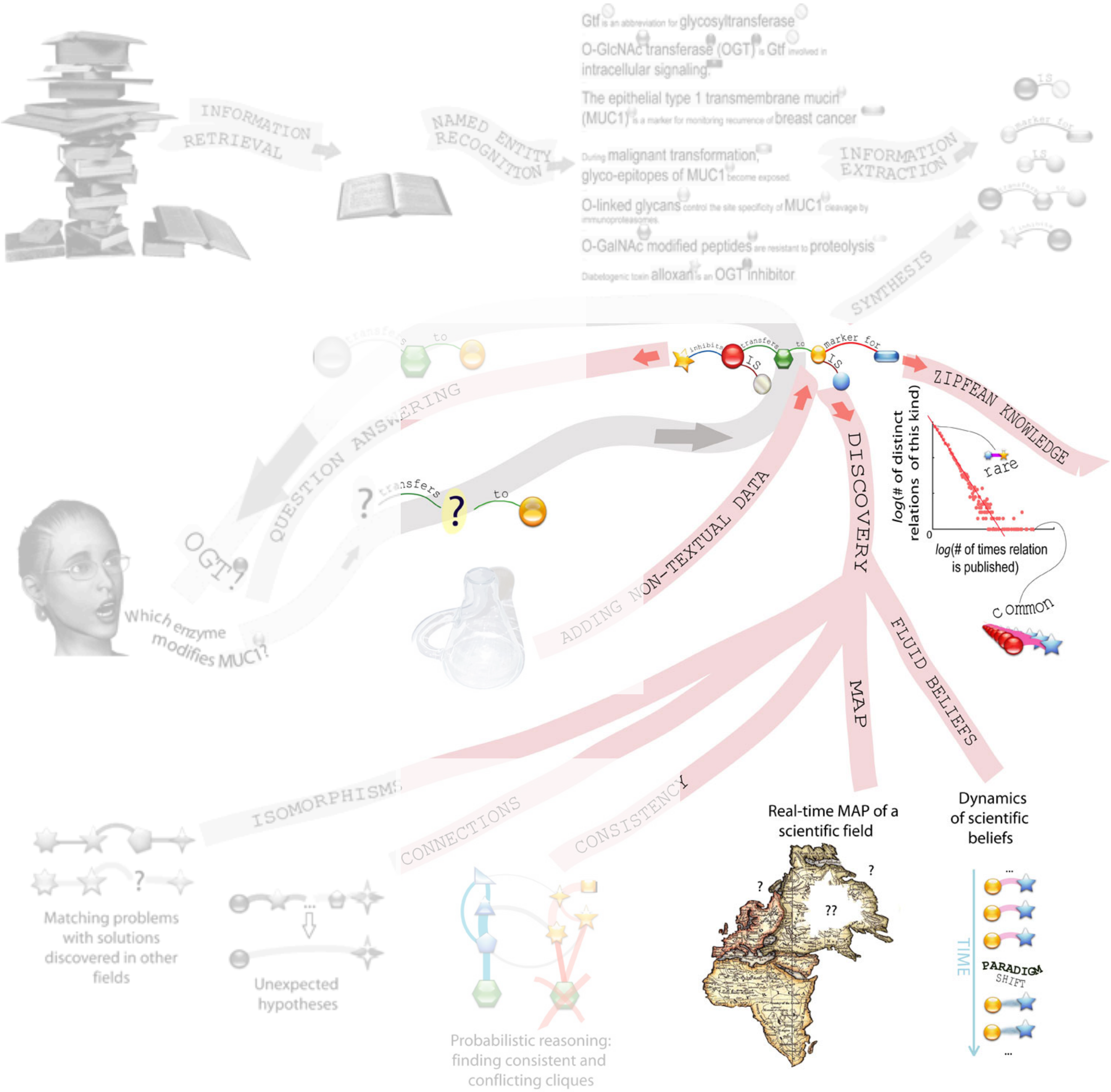
Mapping Science

+

Studying its Dynamics & Evolution

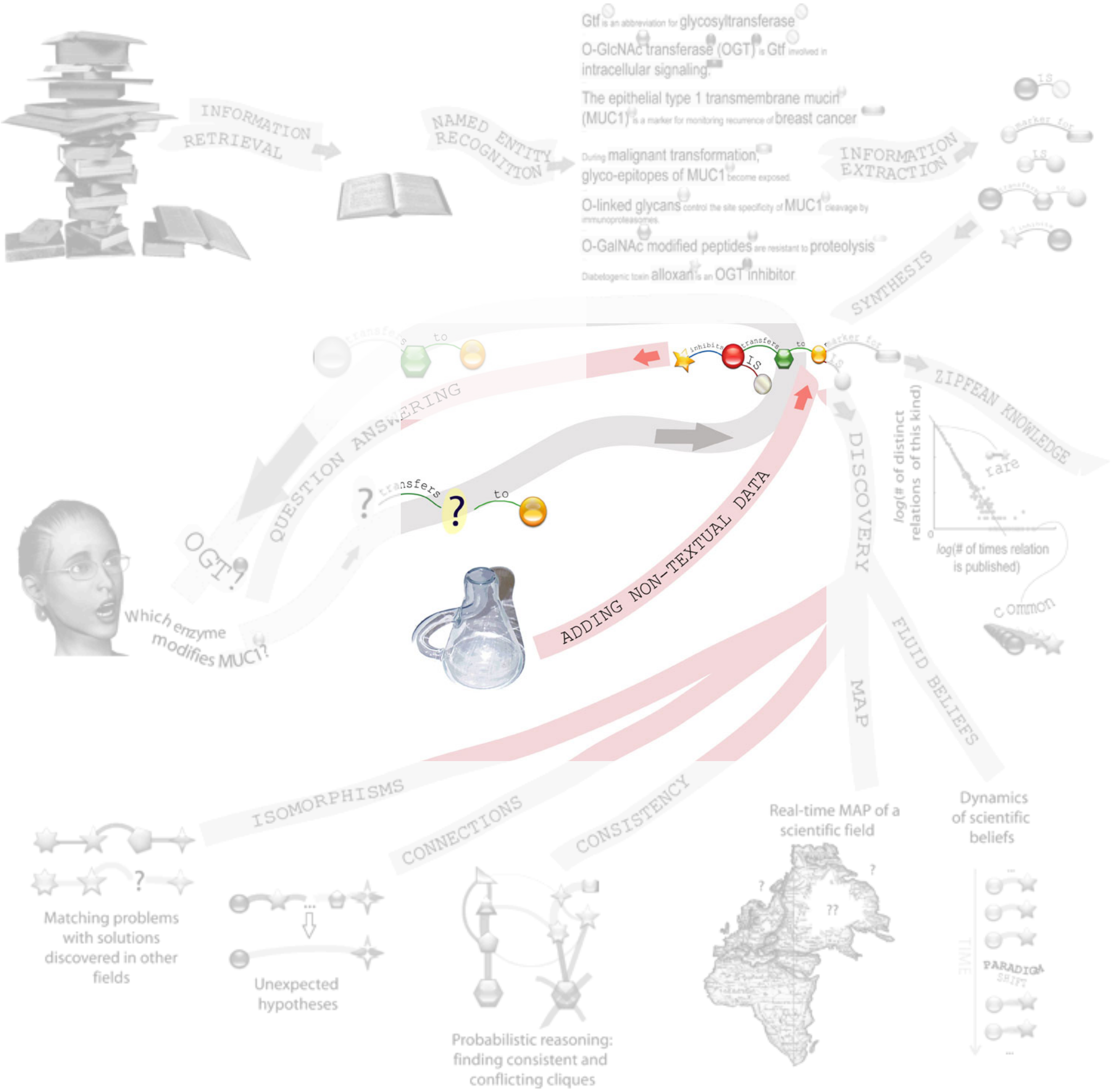
[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

Vision for Mining Large-scale Structured Literature


INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
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
INFORMATION EXTRACTION
IS
marker for
IS
transfers to
SYNTHESIS
inhibits
transfers
to
marker for
IS
IS
transfers to
QUESTION ANSWERING
? transfers ? to
OGT?
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
DISCOVERY
ZIPFEAN KNOWLEDGE
log(# of distinct relations of this kind)
log(# of times relation is published)
0
rare
common
MAP
FLUID BELIEFS
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
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
TIME
...
PARADIGM SHIFT
...

- Revealing patterns of collaboration
- Understanding basis of terms & nomenclature
- Tracking the evolution of ideas
- Models for the evolution of science;
- Helping set policy & research directions

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]


INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
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
INFORMATION EXTRACTION
IS
marker for
IS
transfers
to
inhibits
SYNTHESIS
inhibits
transfers
to
marker for
IS
transfers
to
QUESTION ANSWERING
? transfers ? to
OGT?
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
DISCOVERY
ZIPFEAN KNOWLEDGE
log(# of distinct relations of this kind)
log(# of times relation is published)
rare
common
FLUID BELIEFS
MAP
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
?
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
TIME
PARADIGM SHIFT

Vision for Mining Large-scale Structured Literature

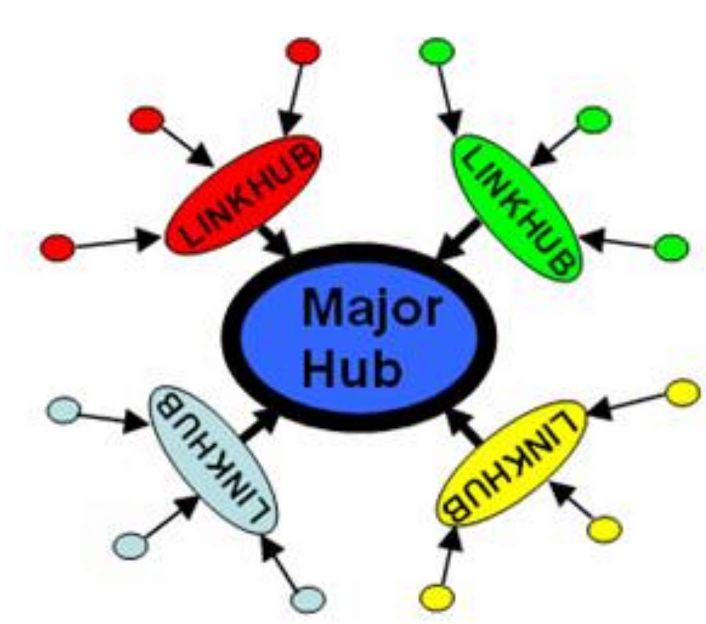
Making it understand-able (through “mashup”)

SciVee, podcasts

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

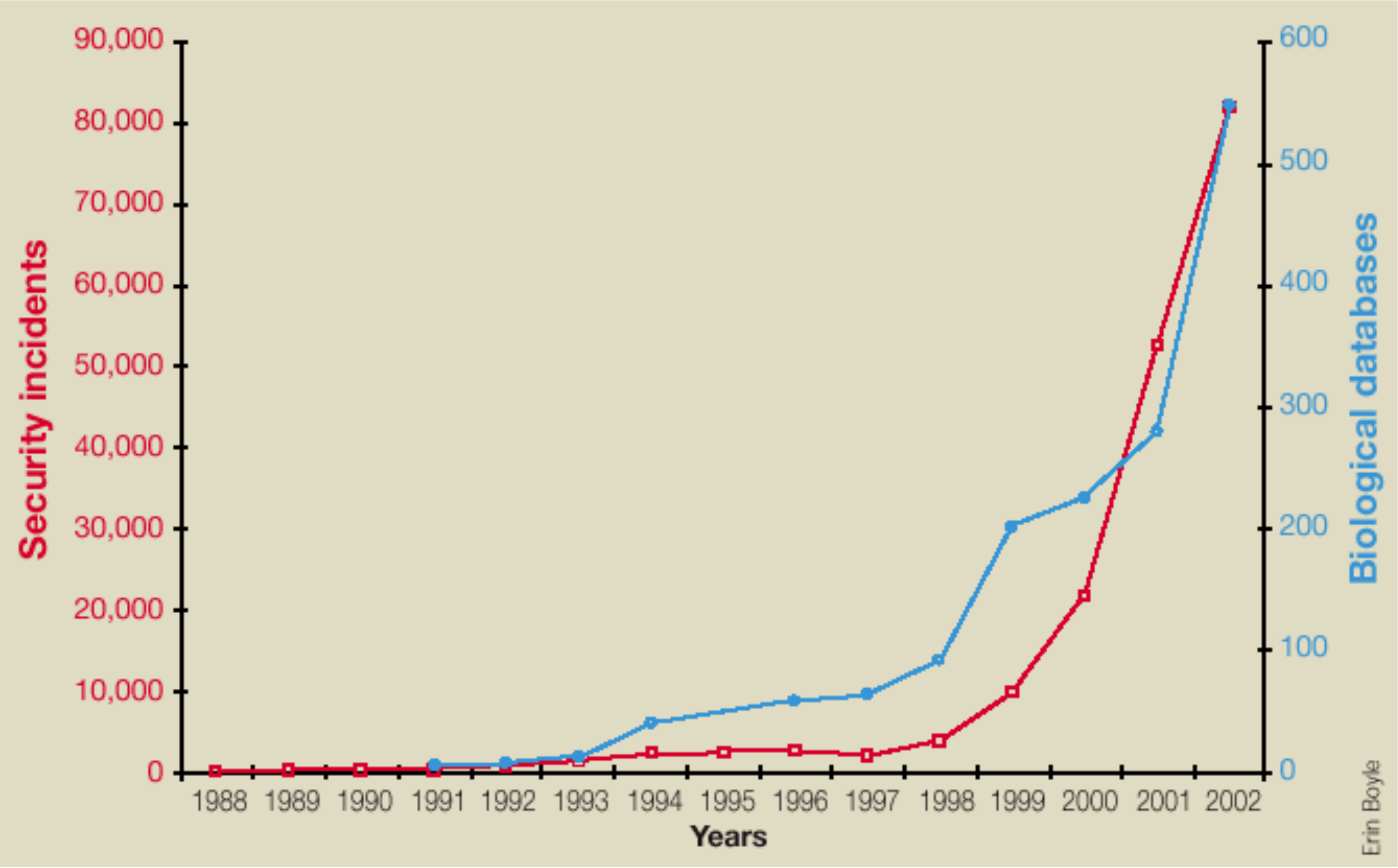
- Need to perform a **"distributed query"** over many information sources
  - ◊ Conventional web links
  - ◊ More complex interfaces
- Genome annotation involves a massive federation of interoperating servers
  - ◊ "Administered" by many disparate people and groups

Federated Information Architecture



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

Vast Computer Security Costs in the "Wild West" Internet



[Greenbaum et al., Nat. Biotech. ('04); Smith et al., GenomeBiol. ('05)]

Summary & Acknowledgements

- Structured Digital Literature
  - ◊ Blurring between digitial information resources & traditional journals
  - ◊ Structured abstracts written by authors, moving through the normal publication process
  - ◊ Structured tables as gateways to large datasets
- Applications
  - ◊ Even a small amount of structured literature is useful as training sets for large scale mining
  - ◊ Using large-scale structured scientific information to look for inconsistencies, see publication trends, and create maps of science

D Greenbaum
M Seringhaus
A Smith
S Douglas
R Auerbach

K Cheung
P Bourne
A Rzhetsky
S Fields

More Information on this Talk

SUBJECT: Textmining

DESCRIPTION:
Data and Code Sharing in Computational Science Meeting, Yale Law 2009.11.21, 9:30-9:40; [I:**ISPSHARING**](Fits into apx. 13 min. with ~10 min of discussion.)

(Works equally well on mac or PC. Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers “ID” on the site. For instance, the topic **pubnet*** can be looked up at **http://papers.gersteinlab.org/papers/pubnet**)

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

.

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: **http://www.flickr.com/photos/mbgmbg/tags/kwpotppt** .

Remember: Setup Show... Advance Slides Manually!