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 gene 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)
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
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 Situation 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 website

• 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, distribute free software, deposit into datasets, &c

• Structure the scientific literature to make it more compatible with a digital future...
  ◊ Structured Digital Paper (Structured Abs., Table, Equation...)

6 Lectures.GersteinLab.org
(c) Mark Gerstein, 2002, bioinfo.mbb.yale.edu
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

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  
E-mail: rcoria@ifc.unam.mx

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 GαΔ GβΔ mutant was viable and sterile. Here we show that the overproduction of Gβ subunit fails to rescue GαΔ 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 KlSte4p transcription factor. Gβ has a 10-fold higher association rate with the Gα 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 SscSte4Δ 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.

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

• K.lactis (species)
  ◊ KISTE4 (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)

◊ KlGPA1 (gene)
  • KlGpa1p (protein)
    – INTERACTIONS
      » TWO-HYBRID
      » KISte4 = XXX
    • KlGpa1p* (protein)
      – INTERACTIONS
        » TWO-HYBRID
        » KISte4 = XXX

◊ KlGPA2 (gene)
  • KlGpa2p (protein)
    – INTERACTIONS
      » TWO-HYBRID
      » KISte4 = XXX

◊ S.cerevisiae (species)
  ◊ SCGPA1 (gene)
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]
Towards a structured digital literature

• Structured Fig. Captions
  ◊ MurphyLab @ CMU
    (A. Ahmed et al. KDD-2009, pp. 39-47)

• Structured equations & pseudocode
  ◊ Directly convertible 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

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]
Vision for Mining Large-scale Structured Literature

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


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

- 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]
Vision for Mining Large-scale Structured Literature

Making it understandable (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

[Smith et al., BMC Bioinfo. ('07)]
Vast Computer Security Costs in the "Wild West" Internet

[Graph showing the increase in security incidents and biological databases over time from 1988 to 2002.]

(Greenbaum et al., Nat. Biotech. (04); Smith et al., GenomeBiol. (05))
Summary & Acknowledgements

• Structured Digital Literature
  ◇ Blurring between digital 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  K Cheung
M Seringhaus  P Bourne
A Smith      A Rzhetsky
S Douglas    S Fields
R Auerbach
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!