

Linking Scientific Literature with public databases & ontologies

Dr. Ian Lewin

European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge

1st November, 2011

Some of the things we need. . .

coordination between the worldviews and vocabularies of diverse medical and scientific specialists

open and well-documented standards and protocols and controlled vocabularies

working practices and attitudes should change

avoid technology fixes that don't fit in with how people [doctors] actually work

A perspective from text mining the biomedical literature

Text Mining: the extraction of structured information from natural language text by automatic means

A perspective from text mining the biomedical literature

Why BioMedicine?

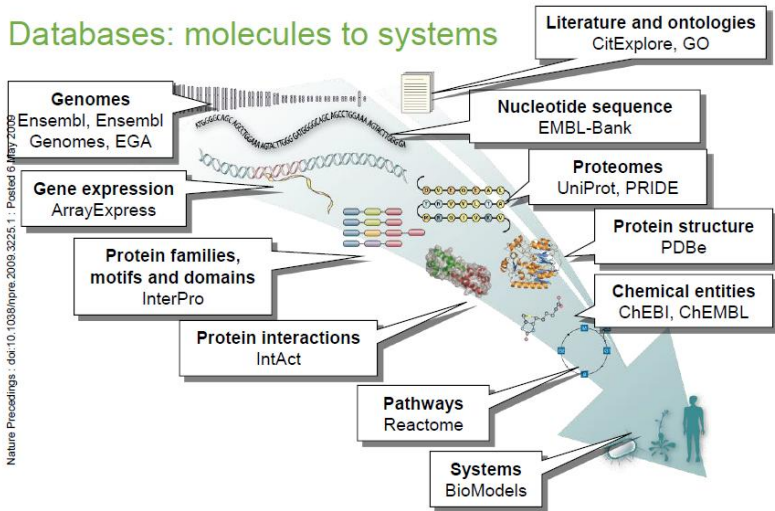
- Biology has been (the?) leading scientific discipline in promoting open access *data*
- and *ontologies*

A perspective from text mining the biomedical literature

Why BioMedicine?

- Biology has been (the?) leading scientific discipline in promoting open access *data*
- and *ontologies*

Databases: molecules to systems



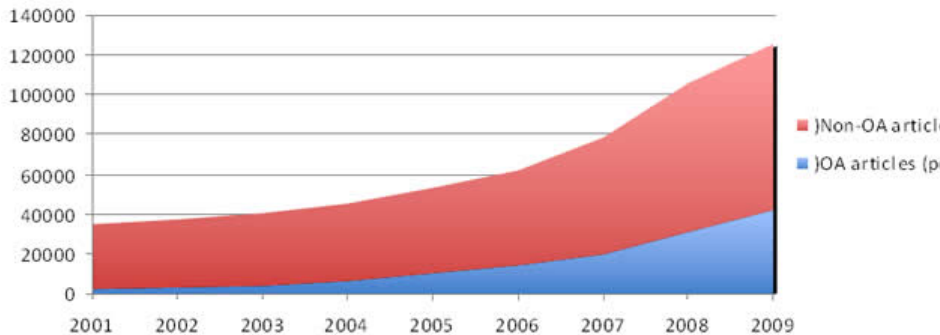
EMBL-EBI



A perspective from Text Mining and Biology

- Biology has been (the?) leading scientific discipline in promoting open access *data*
- and *ontologies*
- and *literature*

Open Access articles in UKPMC 2001 - 2009



Making the literature-to-database roundtrip



Text mining for making the links

*Mortality associated with resistant **cytomegalovirus** among patients with AIDS*

Substring

cytomegalovirus

Associated data

entity-type=species

database-id=10358

classname=scientific-name

rank=genus

Text mining for making the links

*Mortality associated with resistant **cytomegalovirus** among patients with AIDS*

Substring	Associated data
cytomegalovirus	entity-type=species database-id=10358 classname=scientific-name rank=genus

Current UKPMC *objects of interest*

Names

Genes/proteins

Diseases

Organisms

Gene Ontology terms

Chemicals

Accession Numbers

Databases

Uniprot

UMLS

NBCI Taxonomy

GO

Chebi

Embl, Swissprot, Pdb, Genbank, ...

Current UKPMC *objects of interest*

Names

Genes/proteins

Diseases

Organisms

Gene Ontology terms

Chemicals

Accession Numbers

Databases

Uniprot

UMLS

NBCI Taxonomy

GO

Chebi

Embl, Swissprot, Pdb, Genbank, ...

What are the technical issues, in general?

- Term variation
- Ambiguity
 - ▶ overlap with ordinary words
 - ▶ overlap with technical words in other domains
 - ▶ abbreviations
- New names are born (and old ones die)

What is. . . a gene accession number?

*A C. elegans nidogen homologue was identified in the cosmid F54F3 (GenBank **Z79696**).*

*The accession numbers of SLM-1 and SLM-2 are **AF098796** and **AF099092** respectively.*

*Briefly, eggs were dejellied with 2.5% thioglycolic acid (pH 8.2) and activated by treatment with 0.2 g/ml of calcium ionophore **A23187** for 3 min.*

Text mining cues

- The format(s) of accession numbers themselves
- Database names
 - ▶ genbank—gen—swiss-prot—swissprot—sprot—uniprot ...
 - ▶ pdb—protein data bank—protein db— protein databank ...
- Key phrases
 - ▶ Accession— Accession number — Acc number — Acc no — Acc #...
- *Where* these cues appear
 - ▶ in the same sentence
 - ▶ in a footnote
 - ▶ just somewhere in the paper
- Combinations of the above, and their reliability

How good is it?

- processed 39,975 UKPMC XML at random
- using publisher own `ext-link` tags as benchmark
- 2579 accession numbers overall in 533 documents
- automatic system: 95% precision
- automatic system: 92% recall
- and, actually, many 'errors' were really errors *in the benchmark*

We (UKPMC) can do it ... but who should?

How good is it?

- processed 39,975 UKPMC XML at random
- using publisher own `ext-link` tags as benchmark
- 2579 accession numbers overall in 533 documents
- automatic system: 95% precision
- automatic system: 92% recall
- and, actually, many 'errors' were really errors *in the benchmark*

We (UKPMC) can do it ... but who should?

How good is it?

- processed 39,975 UKPMC XML at random
- using publisher own `ext-link` tags as benchmark
- 2579 accession numbers overall in 533 documents
- automatic system: 95% precision
- automatic system: 92% recall
- and, actually, many 'errors' were really errors *in the benchmark*

We (UKPMC) can do it ... but who should?

Who should do it? Publishers? Database curators?

UKPMC Model:

- The papers are there
- The databases are there
- UKPMC adds links between them

Wormbase model

- Include the databases in the publication cycle
- Include the authors in the database cycle

Publishing Interactive Articles: Integrating Journals and Biological Databases

GSA A Publication of The Genetics Society of America

GENETICS

Search

Home Journal Information Subscriptions & Services Collections Previous Issues Current Issue Future Issues

Institution: CALIFORNIA INSTITUTE OF TECHNOLOGY | Sign in as user Name/Password

Originally published as Genetics Published Ahead of Print on June 30, 2010.
Genetics, Vol. 186, 126-145, September 2010, Copyright © 2010
doi:10.1534/genetics.113.117341

A Ubiquitin E2 Variant Protein Acts in Axon Termination and Synaptogenesis in *Caenorhabditis elegans*

Gloriana Trujillo^{1,2}, Katsunori Nakata^{1,2,3}, Dong Yan^{1,2}, Ichi N. Maruyama¹ and Yishi Jin^{1,2,3}

¹Neurobiology Section, Division of Biological Sciences, University of California, San Diego, CA 92093; ²Humanoid Processing Biology Unit, Division of Health, Science and Technology, Chiral Soft, Division 504-02, Japan and ³Howard Hughes Medical Institute, LA Jolla, CA 92093

³ Corresponding author: 8000 Gilman Dr., MC 0986, University of California, San Diego, CA 92093-0986. E-mail: yjin@ucsd.edu

Manuscript received April 6, 2010; Accepted for publication June 19, 2010.

THIS ARTICLE

Abstract ********

Full Text (PDF)

Supporting Information

All Versions of This Article:
genetics.113.117341.v1
106/1/35 **view record**

Alert me when this article is cited

Alert me if a correction is posted

SERVICES

Email this article to a friend

Similar articles in this journal

Similar articles in PubMed

Home Genome Synteny Blast / Blast WormMart Markers Genetic Maps Submit Searches

WormBase Release WS218

WormBase

Find: Any Gene

Exact match Results as XML Literature Search Wormbase Suggest

Web Site Directory

About this release [New/Changed Genes](#), [release notes](#)

General Searches [WormBase Class Browser](#), [Wormbase Query Language Search](#), [AGL Search](#)

Sequences [C. elegans Genome](#), [C. briggsae Genome](#), [Gene](#), [Blast / Blast-e-PCR](#), [Gene Ontology](#), [Synteny Viewer](#), [Co-Elements \(CoOrtho\)](#)

Cells and Gene Expression
[Cell and Pedigree](#), [Neurons](#), [Expression Pattern](#), [Expression profile](#)

Genetics, Strains, and Phenotypes [Genetic Interval](#), [Rearrangements](#), [Balancer](#), [Clone](#), [Allele](#), [SNPs](#), [Markers](#), and [Strains](#), [Strain Report](#), [Phenotypes](#), [RNAi](#)

Batch Queries [WormMart Jobout...](#), [Batch Genes](#), [Batch Sequences](#)

Downloads and Data Mining [Bulk Downloads](#), [Linking to WormBase and Data Mining](#)

Community Worm Meetings [Worm Community Discussion Forum](#), [WormBase Wiki](#), [Mailing Lists](#)

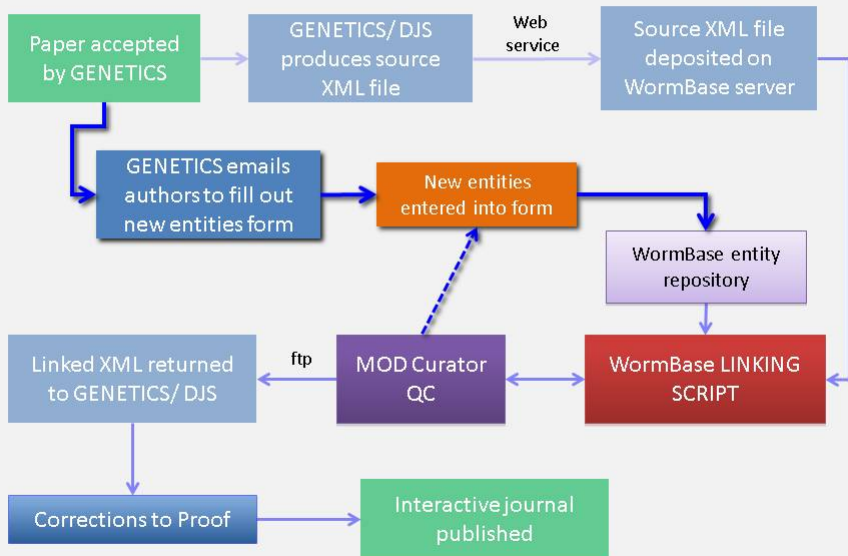
News and Notes

- 02 Oct 2010: Unscheduled service outage; all services restored**
The colocation facility hosting WormBase suffered a catastrophic failure of a critical component today around 8AM ET. System administrators worked throughout the day to restore service. We are happy to report that all services are back online and apologize for any inconvenience.
- 29 Sep 2010: New release of WormBase: WS218**
WormBase has been updated to the WS218 release of the database. We've added a new Genome Browser for *Haemaphysalis contortrix*, bringing the number of species housed at WormBase to 10. Detailed release notes are available on the [WormBase Wiki](#).
- 30 Aug 2010: Possible service interruptions, 31 August 2010**
We're relocating some services to a new hosting facility beginning at 10AM ET (GMT -5), Tuesday August 31, 2010. We plan to maintain systems at the old facility during this transition, but because these upgrades require modifications to the global domain name system records, you may encounter intermittent service interruptions or "server not found" errors. We apologize in advance for any difficulties.

Karen Yook

Genetics Society of America, Dartmouth Journal Services,
Textpresso and WormBase

GSA-WB markup pipeline includes author participation for new entities



Costs and Benefits

- Publishers (& Databases) keep ownership/control of their data
- Databases keep up to date with Literature
- Authors act as QA

- Authors, database owners, publishers must *actively* collaborate
- (Slower?) Interleaved publication & database processes
- It's a bilateral publisher-database agreement. Does it scale?

Textmining for joining up biological data

- Everyone communicates through natural language text
- Textmining can link texts
 - ▶ together
 - ▶ to specialist resources
 - ▶ to acceptable levels of accuracy
 - ▶ especially in the presence of open, shared, *adhered to* standards
- But even when it does, integration *still* depends on ...
- Data ownership and control
- Data responsibility
- Securing participation of interested parties
- Some players may have economic interests to defend

Textmining for joining up biological data

- Everyone communicates through natural language text
- Textmining can link texts
 - ▶ together
 - ▶ to specialist resources
 - ▶ to acceptable levels of accuracy
 - ▶ especially in the presence of open, shared, *adhered to* standards

- But even when it does, integration *still* depends on ...
- Data ownership and control
- Data responsibility
- Securing participation of interested parties
- Some players may have economic interests to defend

Acknowledgments

EMBL-EBI



- www.ebi.ac.uk
- www.ukpmc.ac.uk
- For re-use of slides
 - ▶ Janet Thornton (ebi)
 - ▶ Karen Yook (wormbase)