The Comparative Toxicogenomics Database (CTD): Text Mining Study

Thomas C. Wiegiers

*The Mount Desert Island Biological Laboratory
Salisbury Cove, Maine*

CTD Actor Controlled Vocabularies

Controlled Vocabularies

MeSH

Chemicals

chemical-gene
interactions

chemical-disease
relationships

CTD Action Terms

Entrez-Gene

Genes

gene-disease
relationships

Diseases

MeSH
OMIM

CTD Curation Process Overview

Nature Precedings : doi:10.1038/npre.2009.3142.1 : Posted 22 Apr 2009

PubMed Search
Chemical + " gene"

1: **Nature**. 1995 Jun 29;375(6534):791-5.

nature

Absence of radius and ulna in mice lacking hoxa-11 and hoxd-11.

Davis AP, Witte DP, Hsieh-Li HM, Potter SS, Capecchi MB.

Howard Hughes Medical Institute, Department of Human Genetics, University of Utah School of Medicine, Salt Lake City 84112, USA.

Mice with targeted disruptions in Hox genes have been generated to evaluate the role of the Hox complex in determining the mammalian body plan. This complex of 38 genes encodes transcription factors that specify regional information along the embryonic axes. Early in vertebrate evolution an ancestral complex shared with invertebrates was duplicated twice to give rise to the four linkage groups (Hox A, B, C and D). As a consequence, corresponding genes on the separate linkage groups, called paralogues, are most closely related to each other. Based on sequence similarities, the Hox genes have been subdivided into 13 paralogous groups. The five most 5' groups (Hox 9-13) pattern the posterior region of the vertebrate embryo and the appendicular skeleton. Mice with individual mutations in the paralogous genes hoxa-11 and hoxd-11 have been described. By breeding these two strains together we have generated double mutants which have dramatic phenotypes not apparent in mice homozygous for the individual mutations. The radius and the ulna of the forelimb are almost entirely eliminated, the axial skeleton shows homeotic transformations, and there are severe kidney defects not present in either single mutant. The limb and axial phenotypes are quantitative: as more mutant alleles are added to the genotype, the phenotype becomes progressively more severe. The appendicular skeleton defects suggest that paralogous Hox genes function together to specify limb outgrowth and patterning along the proximodistal axis.

PMID: 7596412 [PubMed - indexed for MEDLINE]

Interaction Entries

Chemical/Gene

Chemical/Disease

Gene/Disease

Excel Spreadsheet

Microsoft Excel - ALL ANNOTATIONS COMBINED_2.xls

	E	F	G	H
	PMID	Interaction	Evidence	Organism
1				
2806	17033922	C1 b G1/p	IDA	Homo sapiens
2807	17033922	C1/n b G1/p	IDA	Homo sapiens
2808	17033922	C1 b -act G1/p	IDA	Homo sapiens
2809	17033922	C1/n b -act G1/p	IDA	Homo sapiens
2810	17033922	C2 -rxn [C1 +act G1/p]	IDA	Homo sapiens
2811	17033922	C2/n -rxn [C1 +act G1/p]	IDA	Homo sapiens
2812	16904708	C1/n -act G1/p	IDA	Homo sapiens
2813	16904708	G2/p +rxn [C1/n -act G1/p]	IDA	Homo sapiens
2814	16904708	C1/n act G1/p	IDA	Homo sapiens
2815	16904708	C1/n act [G1/p b G2/p]	IDA	Homo sapiens
2816	16904708	C1/n +act G1/p	IDA	Homo sapiens
2817	16904708	C1/n +act G1/p	IDA	Homo sapiens
2818	16904708	C1 +act G1/p	IDA	Homo sapiens
2819	16901503	G1/p b C1	IDA	Rattus norvegicus
2820	16901503	G1/p b C1	IDA	Rattus norvegicus
2821	16876394	C1 -rxn [G1/p b G2/p]	IDA	Bos taurus
2822	16876394	C1 +act G1/p	IDA	Bos taurus
2823	16876394	C1 -exp G1/p	IDA	Bos taurus
2824	16876394	C1 -exp G1/p	IDA	Bos taurus
2825	16876394	C1 -exp G1/m	IDA	Bos taurus
2826	16876394	C1 -exp G1/m	IDA	Bos taurus
2827	16787663	C1 b G1/p	IDA	Homo sapiens

CTD Web Interface

Comparative Toxicogenomics Database

Search by Keyword

Chemicals: CAS RN, ID

Connect, Compare

The Comparative Toxicogenomics Database (CTD) elucidates molecular mechanisms by which environmental chemicals affect human disease. Chemical-gene/protein interactions and chemical- and gene-disease relationships are curated from the published literature, and integrated with other data (chemicals, gene/proteins, human diseases, references, sequences, vertebrate and invertebrate organisms, and the Gene Ontology) to facilitate environmental health research.

Discover:

- Which human diseases are associated with a **gene/protein**?
- Which gene/proteins interact with a **chemical**?
- Which chemicals interact with a **gene/protein**?
- Which references report a **chemical-gene/protein interaction**?
- Which cellular functions (GO terms) are affected by a **chemical**?

ctd.mdibl.org

Major Deficiencies in Existing Curation Process

- PubMed used for chemical search
- PubMed search capability is limited
- Hits sorted in PubMed ID order
- Many false-positives
- Actors are not identified
- Actor mapping to CTD very difficult
- Documents are arbitrarily eliminated

Major Text Mining Goals for CTD

- Identify relevant references more effectively
 - Reduce number of false positives
- Prioritize and rank references for curation
 - Frequency of occurrence of biological actors
 - Location of occurrence of biological actors
- Enhance curation efficiency
 - Highlighting/extracting vocabulary terms
- No arbitrary elimination of documents from curation
 - All documents reviewed, either computationally or manually

Text Mining Process Overview

Major Text Mining Tools Used

- [Oscar3](#) - Chemical recognition
- [Abner](#) - Gene/protein recognition
- [MetaMap/MMTx](#) - Disease, chemical and gene recognition
- [Lucene](#) - Document relevancy ranking tool

Study Overview

- 1,593 documents tested
 - 10 target chemicals
 - 6,367 curated actors
 - 27,495 hyperlinked actors
- Each document manually triaged/curated by CTD curators
- Expert curated results are used as Gold Standard
- 3 Major areas of Study:
 - Curated Actor Identification
 - False Positives
 - Document Ranking

Actor Identification Methodology

- Actor synonyms used:
 - Abstract Text: “ ...breast cancer...”
 - Text Mining Tools: “ Breast Cancer”
 - MeSH and Curator: “ Breast Neoplasm”

Any synonym match to curated term counts as a *match*
- Curator vs. Tool Specificity
 - Text: “ ...uterine malignant mixed mesodermal tumor...”
 - Tools: uterine neoplasms, malignant mixed tumor, uterine cancer
 - Curator: *Mullerian mixed tumor*

If no synonyms match, counts as a *miss*
- **Problem:** Curation sometimes done in full-text

Actor Identification Results

Gross Actor Identification Ratio	80%*
Adjusted Actor Identification Ratio	93%*

* Calculated using *macro-averaging* to account for microarray data

False Positive Results

66% False Positive Ratio
(hyperlinked actors were NOT involved in an interaction)

Ranking Schemes Compared

- Control Case Ranking
 - Default PubMed ranking based on PubMed ID
- Lucene Ranking
 - Ranking based on the Lucene tool
- Customized Rule-Based Formula
 - Ranking using a formula developed as part of the project

Ranking Criteria and Weighting

- Term in title
- Frequency of terms in abstract
- Position of terms in abstract
- High priority journal
- Microarray content
- Co-occurrence of action terms and actors

Ranking Results

Ranking Mechanism	Mean Average Precision	Rank-to-Rank Correlation Coefficient	Score-to-Ixn Rank Correlation Coefficient
Control Case (PubMed)	63%	14%	-
Lucene	72%	30%	27%
Custom Formula	73%*	38%**	43%

*16% improvement in MAP over the control case

**171% increase in the correlation to the number of interactions over the control case

The same 354 abstracts

Conclusions

- Identified curated actors
- False positive rate was high
- Significantly increased already-high *MAP* rates
- Significantly improved ranking of interaction-rich documents
- Rule-based formula ranked docs better than Lucene
- Great processing times for Abner and Oscar3 (1sec/doc) vs. MMTx (1min 17s/doc)

Some Next Steps

- Implement a curatorial tool using text mining:
 - Rank documents
 - Tag and color code actors
 - Facilitate data entry
- Study the effect of the curation tool and text mining on productivity
- Study the effect of document ranking on curation:
 - Does the ranking improve curation efficiency?
 - Are we curating the right documents?

Acknowledgements

CTD Team

Carolyn J. Mattingly, PhD

Scientific Curators

Allan Peter Davis, PhD

Cindy Murphy, PhD

Cynthia Richards, PhD

Scientific Software Engineers

Michael Rosenstein, JD

Thomas Wieggers

System Administrator

Roy McMorran

MITRE Team:

Lynette Hirschman, PhD

Kevin Bretonnel Cohen

<http://ctd.mdibl.org>

Funding

NIEHS (ES014065 and ES003828)

NCRR (RR016463)