M Library
UNIVERSITY OF MICHIGAN

IMPROVING GENE NAME DICTIONARIES

Marci Brandenburg, University of Michigan Health Sciences Libraries, Ann Arbor, Michigan

NCIBI
National Center for Integrative Biomedical Informatics

BACKGROUND:

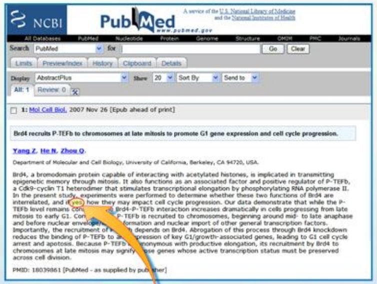
The National Center for Integrative Biomedical Informatics (NCIBI) is one of seven National Centers for Biomedical Computing (NCBC) in the NIH Roadmap. The Center is developing the framework of conceptual models, computational infrastructure and an integrated knowledge repository, which modern scientists need to make effective use of the wealth of data flowing from molecular biology and translational research. NCIBI used a process for detecting and selecting gene information from sentences within scientific articles. These sentences were then tagged as referring to a specific gene. However, this gene name curation process failed to tag about 30% of appropriate sentences. NCIBI asked the Health Sciences Libraries for assistance in improving recall.

OBJECTIVE:

To demonstrate that librarian interaction increases gene name curation efficiency by improving the gene name dictionary currently maintained by the medical center's informatics unit.

METHODS:

This project began with a manual review of sentences associated with the genes commonly not tagged and case-by-case identification of synonyms to improve recall. A review of the literature was conducted to determine current practices, such as the creation of rules, mining existing dictionaries, and using natural language processing. Using professional health sciences library experience and subject expertise in biology, I identified potential methods for improving the gene name curation process.



□ 1: Mol Cell Biol. 2007 Nov 26 [Epub ahead of print]

Brd4 recruits P-TEFb to chromosomes at late mitosis to promote G1 gene expression and cell cycle progression.

Yang Z, He N, Zhou Q.

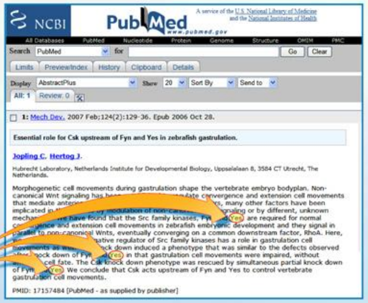
Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.

Brd4, a bromodomain protein capable of interacting with acetylated histones, is implicated in transmitting epigenetic memory through mitosis. It also functions as an associated factor and positive regulator of P-TEFb, a Cdk9-cyclin T1 heterodimer that stimulates transcriptional elongation by phosphorylating RNA polymerase II. In the present study, experiments were performed to determine whether these two functions of Brd4 are interrelated, and if yes, how they may impact cell cycle progression. Our data demonstrate that while the P-TEFb level remains con[illegible] Brd4-P-TEFb interaction increases dramatically in cells progressing from late mitosis to early G1. Con[illegible] P-TEFb is recruited to chromosomes, beginning around mid- to late anaphase and before nuclear envelo[illegible] formation and nuclear import of other general transcription factors. Importantly, the recruitment of [illegible] depends on Brd4. Abrogation of this process through Brd4 knockdown reduces the binding of P-TEFb to a[illegible] expression of key G1/growth-associated genes, leading to G1 cell cycle arrest and apoptosis. Because P-TEFb [illegible] synonymous with productive elongation, its recruitment by Brd4 to chromosomes at late mitosis may signify [illegible] genes whose active transcription status must be preserved across cell division.

PMID: 18039861 [PubMed - as supplied by pub[illegible]sher]

Yes = common word in the English language

Yes = Proto-oncogene tyrosine-protein kinase



□ 1: Mech Dev. 2007 Feb;124(2):129-36. Epub 2006 Oct 28.

Essential role for Csk upstream of Fyn and Yes in zebrafish gastrulation.

Jopling C, Hertog J.

Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.

Morphogenetic cell movements during gastrulation shape the vertebrate embryo bodyplan. Non-canonical Wnt signaling has [illegible] convergence and extension cell movements that mediate anter[illegible]rs, many other factors have been implicated in [illegible] by modulation of non-cano[illegible]gnaling or by different, unknown mechan[illegible] we have found that the Src family kinases, Fy[illegible] Yes, are required for normal co[illegible]gence and extension cell movements in zebrafish embryonic development and they signal in parallel to non-canonical Wnts, eventually converging on a common downstream factor, RhoA. Here, [illegible]ative regulator of Src family kinases has a role in gastrulation cell movements as w[illegible]ck down induced a phenotype that was similar to the defects observed after knock down of Fy[illegible] Yes in that gastrulation cell movements were impaired, without [illegible] cell fate. The Csk knock down phenotype was rescued by simultaneous partial knock down of Fyn [illegible] Yes. We conclude that Csk acts upstream of Fyn and Yes to control vertebrate gastrulation cell movements.

PMID: 17157484 [PubMed - as supplied by publisher]

RESULTS:

Developing a comprehensive gene dictionary is challenging, as each gene name has many synonyms. In addition, it can be difficult to distinguish between gene names and protein names during the curation process. Another challenge is distinguishing between gene name symbols and common words in text. For example, "yes" is a gene symbol for "proto-oncogene tyrosine-protein kinase YES." Therefore, even with an extremely comprehensive gene name dictionary, there are still problems.

There have been a number of methods used to create gene name dictionaries, but none have resolved all of the above problems. A variety of heuristics have been used to eliminate gene or protein name variations and ambiguities. Examples of such heuristics are: using both numeric and Roman numerals for gene names containing numbers (Koike and Takagi), removing Greek letters from names (Zhou, et al.), and converting all letters to lower-case (Mika and Rost). A combination of methods seems to have yielded the best results. However, a comprehensive dictionary for gene name curation is yet to be developed.

Table 1. Examples of gene names, showing the gene name symbol before and after applying a specific rule.

Gene Name (Full)	Gene Name Symbol (Original)	Rule	Gene Name Symbol After Applying Rule
Nitric oxide synthase 2, inducible	Nos2	Use both numeric and Roman numerals	Nos2, NosII
Interleukin 1β	Il1β	Remove Greek letters from name	Il1, Il1B, Il1 Beta
Tumor necrosis factor	TNF	Convert all letters to lower-case	tnf

REFERENCES:

Koike A. and Takagi T. 2004. Gene/protein/family name recognition in biomedical literature. BioLINK2004: Linking Biological Literature , Ontologies, and Databases. 9-16.

Mika S. and Rost B. 2004. Protein names precisely peeled off free text. Bioinformatics 20(Suppl. 1): i241-i247.

Zhou GD, et al. 2005. Recognition of protein/gene names from text using an ensemble of classifiers. BMC Bioinformatics 6(Suppl. 1): S7-S13.

FUTURE WORK:

To automate the process of improving gene name dictionaries by creating additional heuristics and using natural language processing.

ACKNOWLEDGEMENTS:

I would like to thank Jean Song, Health Sciences Libraries Bioinformationist, and David States, Director of Bioinformatics Training Program, for their help and support on this project. This work was supported by the Health Sciences Libraries and the National Center for Integrative Biomedical Informatics, NIH Grant#U54-DA021519.