MANCHESTER 1824

The University of Manchester

Linking gene expression mentions to anatomical locations

Martin Gerner *
Goran Nenadic **
Casey Bergman *

*: Faculty of Life Sciences, University of Manchester
**: School of Computer Science, University of Manchester

MANCHESTER 1824

The University of Manchester

Talk overview

- Project goals, motivation
- System description
- Evaluation method, results
- Data interface demonstration
- Conclusion

MANCHESTER 1824

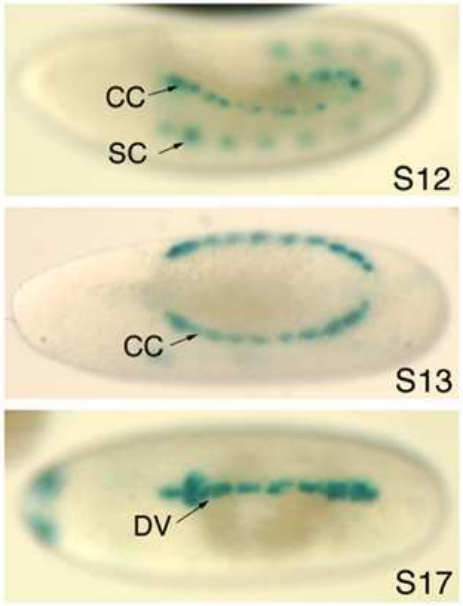
The University of Manchester

Introduction

- The expression of genes vary between cell types

- This knowledge is critical:
  - Understanding of a gene
  - Understanding of a cell type

- Some information available in databases



Source: Fly Embryo RNAi project

MANCHESTER 1824

The University of Manchester

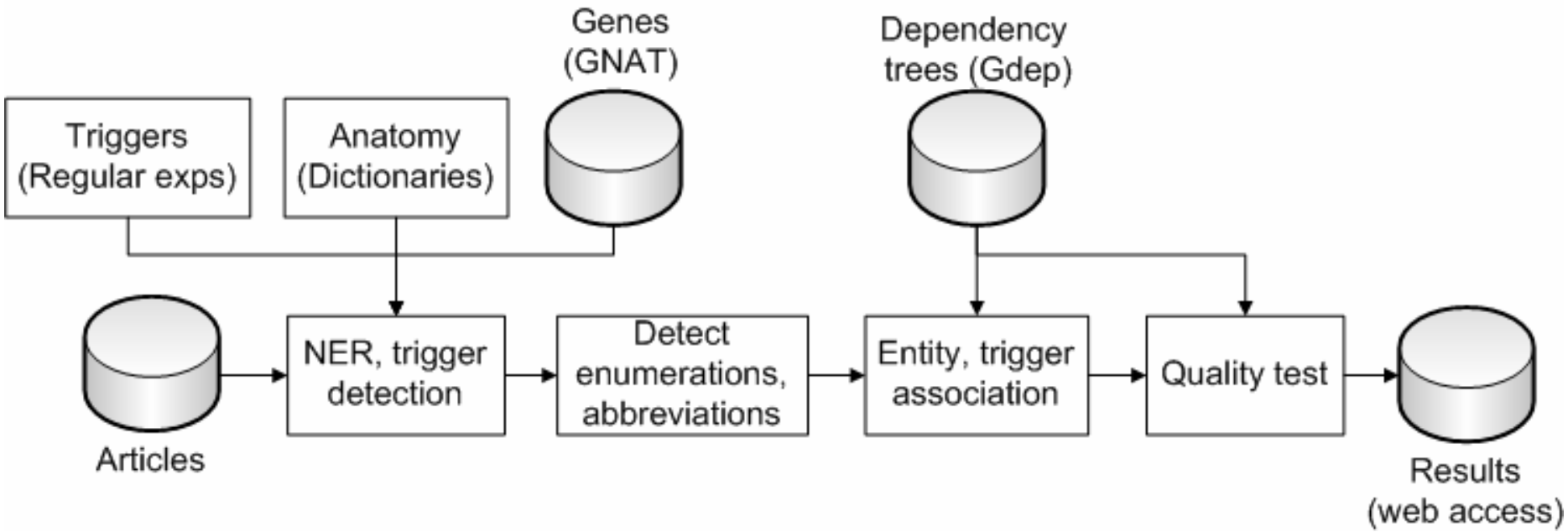
Goals

- Identify where authors discuss gene expression in the context of an anatomical location
- Link these mentions to normalized gene and anatomical location identifiers
- Apply to MEDLINE/PMC
- Example: “Regulation of interleukin-2 induced interleukin-5 and interleukin-13 production in human peripheral blood mononuclear cells”

MANCHESTER 1824

The University of Manchester

GETM system overview

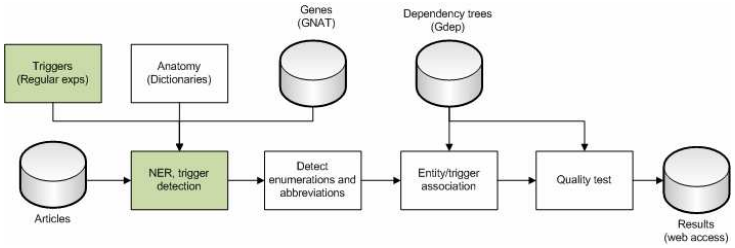


MANCHESTER 1824

The University of Manchester

Trigger detection

- Recogniton of trigger keywords, e.g. “expression”and “production”
- Regular expressions are constructed from these trigger keywords
- Simple negation detection (“not expressed”, “negative expression”, etc.)



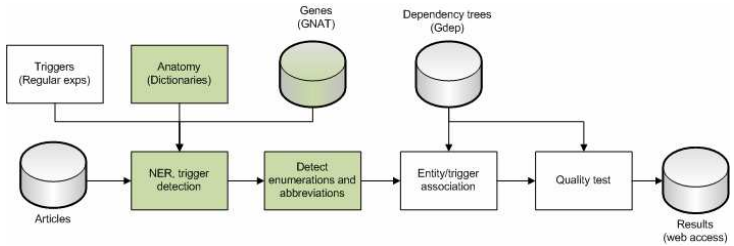
MANCHESTER 1824

The University of Manchester

Named entity recognition (NER)

- Gene NER:
  - GNAT (dictionary-based; aided by LINNAEUS) (Hakenberg *et al*., 2008)

- Anatomical NER (expanded dictionaries):
  - OBO Foundry ontologies (160,000 entities)
  - Cell-line ontology (9,500 entities)

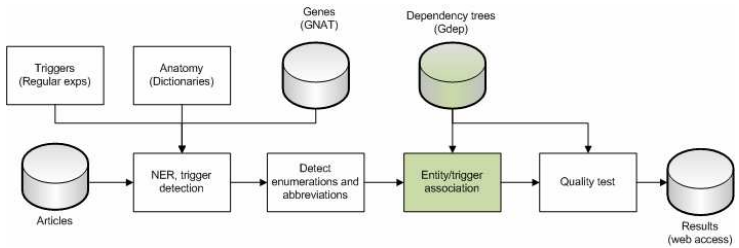


MANCHESTER 1824

The University of Manchester

Associating entities to triggers

1. Sentence splitting
2. If there is only one gene and one cell type, associate them with the trigger
3. If the genes, cell types and trigger conform to certain patterns, associate them with the trigger
4. Associate trigger to the gene and cell type mentions with shortest tree distance to the trigger



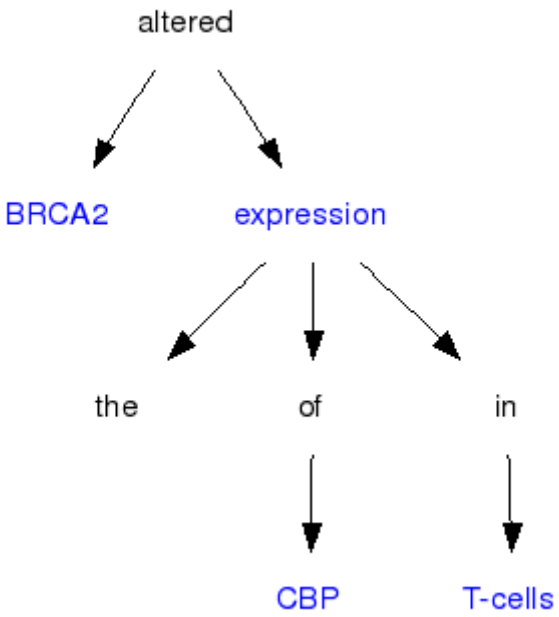
MANCHESTER 1824

The University of Manchester

Determining reliability

- Test if the gene mention ”depends” on the trigger (-> class A, B)

- Example:
  - BRCA2 altered the expression of CBP in T-cells
  - CBP -> class A



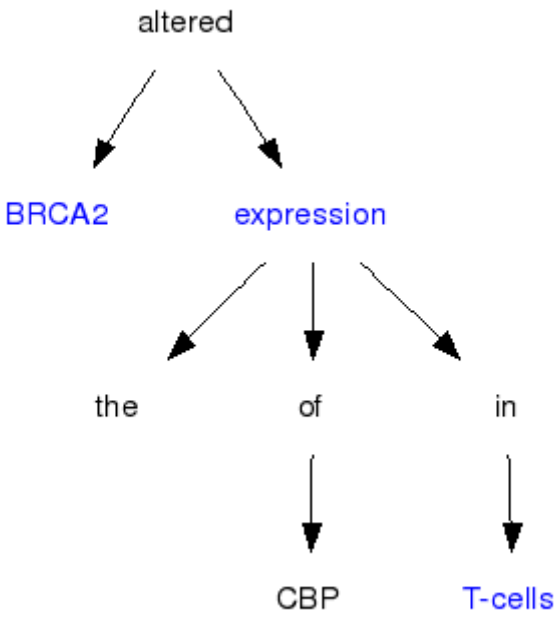
MANCHESTER 1824

The University of Manchester

Determining reliability

- Test if the gene mention ”depends” on the trigger (-> class A, B)

- Example:
  - BRCA2 altered the expression of CBP in T-cells
  - BRCA2 -> class B



MANCHESTER 1824

The University of Manchester

Evaluation

- BioNLP 2009 shared task corpus:
  - Contains mention-level gene expression annotation (without tissues)

- Extended annotations for 150 abstracts (267 of 377 entries)

MANCHESTER 1824

The University of Manchester

Evaluation results

	Precision	**Recall**
Category A + B	0.60	0.24
Category A only	0.74	0.19

- Categorization improved precision, at a relatively small cost of recall

MANCHESTER 1824

The University of Manchester

False negative (FN) analysis

Problem type	% of corpus
Gene NER FNs	50 %
Cell-type NER FNs	26 %
Trigger detection	21 %
Multi-sentence entries	23 %
Overall	74 %

- NER results main reason for low recall

MANCHESTER 1824

The University of Manchester

Large-scale application results

- 747,000 extracted mentions (58% A)
  - MEDLINE: 578,000
  - PubMed Central (PMC): 144,000
- Results range across:
  - 28,000 different genes (top: TNF)
  - 3,900 different anatomical locations (top: T cells).
  - 240,000 different gene/location combinations (60% mentioned once)

MANCHESTER 1824

The University of Manchester

Interface demonstration

- Goal: make results available

- Interface and data live on

http://getm-project.sourceforge.net

MANCHESTER 1824

Gene Expression Text Miner (GETM)

GETM is a tool which is capable of extracting information about the expression of genes from biomedical literature. Using the data extracted by GETM, it is possible to get an overview of the cell types that are discussed in context with gene expression of a particular gene (example), and vice versa (example).

For questions, suggestions or bug reports, please contact me.
To navigate back: Martin Gerner's personal page, the Bergman lab or the Nenadic group.
The files on this webpage can also accessed from this project's SourceForge project page.

GETM dataset query interface

The dataset extracted by GETM can be queried using the form below. Enter either an anatomical location or a gene name and click submit.

Location:

Gene:

Submit

GETM tool and evaluation corpus availability

- getm.tar.gz (4.5 MB): Compiled GETM .jar file, suitable for running directly. Also contains documentation, the GETM source code and javadoc.
- getm-manual-corpus.tar.gz (86 kB): A set of 150 MEDLINE abstracts, containing annotations of gene expression events, extended to also contain information about the anatomical locations of those events. The original gene expression annotation was performed by members of the Tsujii lab.

MANCHESTER 1824

The University of Manchester

Home > Gene search

Gene search results for 'BRCA1':

The following genes have been identified as matching your search query. If you would like to search for your specific search term only, choose the first alternative ("only search the term..."). If you would like to retrieve data for a specific gene (including potential synonyms), click its name among the lower alternatives. For human genes, you can choose to retrieve results from not only the human gene but also orthologs in any other species.

- Only search the term 'BRCA1' (?)

- BRCA1: breast cancer 1, early onset (Homo sapiens) (include orthologs)
- Brca1: breast cancer 1 (Mus musculus)
- Brca1: breast cancer 1 (Rattus norvegicus)
- BRCA1: breast cancer 1, early onset (Gallus gallus)
- BRCA1: breast cancer 1, early onset (Bos taurus)
- brca1: breast and ovarian cancer susceptibility protein (Xenopus laevis)
- BRCA1: breast cancer 1, early onset (Equus caballus)
- BRCA1: breast cancer 1, early onset (Canis lupus familiaris)
- BRCA1: breast cancer 1, early onset (Pan troglodytes)
- BRCA1: breast cancer 1, early onset (Felis catus)
- BRCA1: breast cancer 1, early onset (Sus scrofa)
- BRCA1: breast cancer 1, early onset (Macaca mulatta)

MANCHESTER 1824

The University of Manchester

Home > Gene search > Choose location

GETM results for BRCA1: breast cancer 1, early onset (Homo sapiens) (and orthologs in other species)

Shown below are the anatomical locations that most often are discussed in relation to the gene above. Click the number in the 'frequency' column to view the mentions for this particular combination of gene and location, or 'view for other genes' to see the other genes that are associated with that particular anatomical location.

View results for:

Location	Frequency	Other genes
All locations	491	
MCF-7	56	View for other genes
epithelial cells	49	View for other genes
mammary gland	33	View for other genes
HCC1937	27	View for other genes
germline	23	View for other genes
E2	21	View for other genes
ovary	19	View for other genes
lung	18	View for other genes
MCF7	15	View for other genes
granulosa cells	11	View for other genes
MDA-MB-468	11	View for other genes
testis	10	View for other genes
colon	10	View for other genes
stem cells	8	View for other genes
surface epithelium	6	View for other genes
fibroblasts	6	View for other genes

MANCHESTER 1824

The University of Manchester

Document	Gene	Location	Sentence
Rosell et al. (2009)	BRCA1	lung	Customized treatment in non-small-cell *lung* cancer based on EGFR mutations and *BRCA1* mRNA *expression*
⊞ Rosell et al. (2009) (3)	BRCA1	lung	In addition to the potential predictive role of BRCA1, *BRCA1* *overexpression* confers aggressive behavior in transgenic models of small cell and squamous cell *lung* carcinomas, as well as in a subset of lung adenocarcinomas harboring the intrinsic T/t-antigen cancer signature.[30] Poor prognosis has also been associated with BRCA1 overexpression in early NSCLC.[31] In the present study, two-year survival was 41% in patients with the lowest levels of BRCA1, 16% in those with intermediate levels and 0% in those with the highest levels
Boukovinas et al. (2008)	BRCA1	lung	We have retrospectively examined the effect of *RRM1*, *RRM2* and *BRCA1* *expression* on outcome to gemcitabine plus docetaxel in advanced non-small-cell *lung* cancer (NSCLC) patients
⊟ Lee et al. (2007) (4)	BRCA1	lung	Anchorage-dependent growth after reexpression of these genes was examined in a *lung* cancer cell line that originally lacked *BRCA1* and *BRCA2* *expression*
	BRCA1	lung	RESULTS: The data indicated that low protein *expression* of *BRCA1* and *BRCA2* was frequent in *lung* adenocarcinomas (42-44%), whereas low XRCC5 protein expression was more prevalent among squamous cell carcinoma (32%)
	BRCA1	lung	In addition, low *BRCA1* *expression* was significantly associated with low RB expression, especially in *lung* adenocarcinoma
	BRCA1	lung	CONCLUSIONS: Our retrospective study provides compelling evidence that low mRNA and protein *expression* in the *BRCA1*/*BRCA2* and *XRCC5* genes occur in *lung* adenocarcinoma and squamous cell carcinoma, respectively, and that promoter hypermethylation is the predominant mechanism in deregulation of these genes

MANCHESTER 1824

The University of Manchester

GETM results for lung

Shown below are the genes that most often are discussed in relation to the anatomical location above. Click the number in the 'frequency' column to view the mentions for this particular combination of gene and location, or 'view for other genes' to see the other anatomical locations that are associated with that particular gene.

View results for:

Gene	Frequency	Other locations
All genes	17068	
TP53: tumor protein p53 (Homo sapiens)	522	View for other locations
VEGFA: vascular endothelial growth factor A (Homo sapiens)	351	View for other locations
EGFR: epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian) (Homo sapiens)	270	View for other locations
ERBB2: v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian) (Homo sapiens)	215	View for other locations
BCL2: B-cell CLL/lymphoma 2 (Homo sapiens)	214	View for other locations
ABCB1: ATP-binding cassette, sub-family B (MDR/TAP), member 1 (Homo sapiens)	178	View for other locations
TNF: tumor necrosis factor (TNF superfamily, member 2) (Homo sapiens)	148	View for other locations
Tnf: tumor necrosis factor (Mus musculus)	129	View for other locations

MANCHESTER 1824

The University of Manchester

Conclusion

- GETM: links gene expression mentions to normalized genes and anatomical locations
- Extracted 747,000 entries (60% precision), of which 58% have higher precision (74%)

- First time anatomical NER has been done on this scale

MANCHESTER 1824

The University of Manchester

Future work

- Improve recall (NER)
- Negation detection
  - Differentiate entries on whether x *is* or *is not* expressed

- Combine with figure mining from PMC

MANCHESTER 1824

The University of Manchester

Acknowledgements

- Jörg Hakenberg (GNAT)
- The members of the Bergman and Nenadic groups
- Supported by: University of Manchester, BBSRC, and BioMed Central

http://getm-project.sourceforge.net

(source code, extracted data, evaluation corpus)