

Concept-based text mining for Systems Biology

Kevin Bretonnel Cohen

Center for Computational Pharmacology

U. Colorado School of Medicine

Interfaces 2007: Systems Biology

Model organisms

- You can experiment on them
- They are like us, or are otherwise interesting



Model organism databases



Search for

in these sections

- All sections
- Gene symbols/names**
- Accession IDs
- Phenotype/Human Disease
- Gene Expression
- Gene Ontology
- Anatomical Dictionary
- Phenotype Ontology (MP)

Advanced search for...

Search Categories

- [All Search Tools](#)
- [Genes/Markers](#)
- [Phenotypes/Alleles](#)
- [Strains/Polymorphisms](#)
- [Expression Sequences](#)
- [Comparative Maps/Data](#)
- [Mouse Maps/Data](#)

?
Gene Detail
Your Input Wel

Symbol Name ID	Brca1 breast cancer 1 MGI:104537	Nomenclature												
Genetic Map	Chromosome 11 60.5 cM, cytoband D Detailed Genetic Map ± 1 cM Mapping data(25)													
Sequence Map	Chr11:101304854-101368045 bp, - strand (From VEGA annotation of NCBI Build 36) VEGA ContigView Ensembl ContigView UCSC Browser NCBI Map Viewer	<p style="text-align: center;">MGI Mouse GBrowse</p>												
Mammalian homology	human; dog, domestic; rat (Mammalian Orthology) Comparative Map (Mouse/Human Brca1 ± 2 cM) Protein SuperFamily: transcriptional regulator, BRCA1 type													
Sequences	<table border="0" style="width: 100%;"> <thead> <tr> <th style="text-align: left;">Representative Sequences</th> <th style="text-align: right;">Length</th> <th style="text-align: right;">Strain/Species</th> </tr> </thead> <tbody> <tr> <td><input type="checkbox"/> genomic OTTMUSG00000002870 VEGA Gene Model MGI Sequence Detail</td> <td style="text-align: right;">63192</td> <td style="text-align: right;">C57BL/6J</td> </tr> <tr> <td><input type="checkbox"/> transcript NM_009764 RefSeq MGI Sequence Detail</td> <td style="text-align: right;">6469</td> <td style="text-align: right;">-</td> </tr> <tr> <td><input type="checkbox"/> polypeptide P48754 UniProt EBI MGI Sequence Detail</td> <td style="text-align: right;">1812</td> <td style="text-align: right;">Not Applicable</td> </tr> </tbody> </table>	Representative Sequences	Length	Strain/Species	<input type="checkbox"/> genomic OTTMUSG00000002870 VEGA Gene Model MGI Sequence Detail	63192	C57BL/6J	<input type="checkbox"/> transcript NM_009764 RefSeq MGI Sequence Detail	6469	-	<input type="checkbox"/> polypeptide P48754 UniProt EBI MGI Sequence Detail	1812	Not Applicable	
Representative Sequences	Length	Strain/Species												
<input type="checkbox"/> genomic OTTMUSG00000002870 VEGA Gene Model MGI Sequence Detail	63192	C57BL/6J												
<input type="checkbox"/> transcript NM_009764 RefSeq MGI Sequence Detail	6469	-												
<input type="checkbox"/> polypeptide P48754 UniProt EBI MGI Sequence Detail	1812	Not Applicable												

122 references...

References

(Earliest) [J:31493](#) Hall JM *et al.*, "Linkage of early-onset familial breast cancer to chromosome 17q21." *Science* 1990 Dec 21;250(4988):1684-9

(Latest) [J:117113](#) Clark-Knowles KV *et al.*, "Conditional inactivation of Brca1 in the mouse ovarian surface epithelium results in an increase in preneoplastic changes." *Exp Cell Res* 2007 Jan 1;313(1):133-45

All references([122](#))

How long does it take?

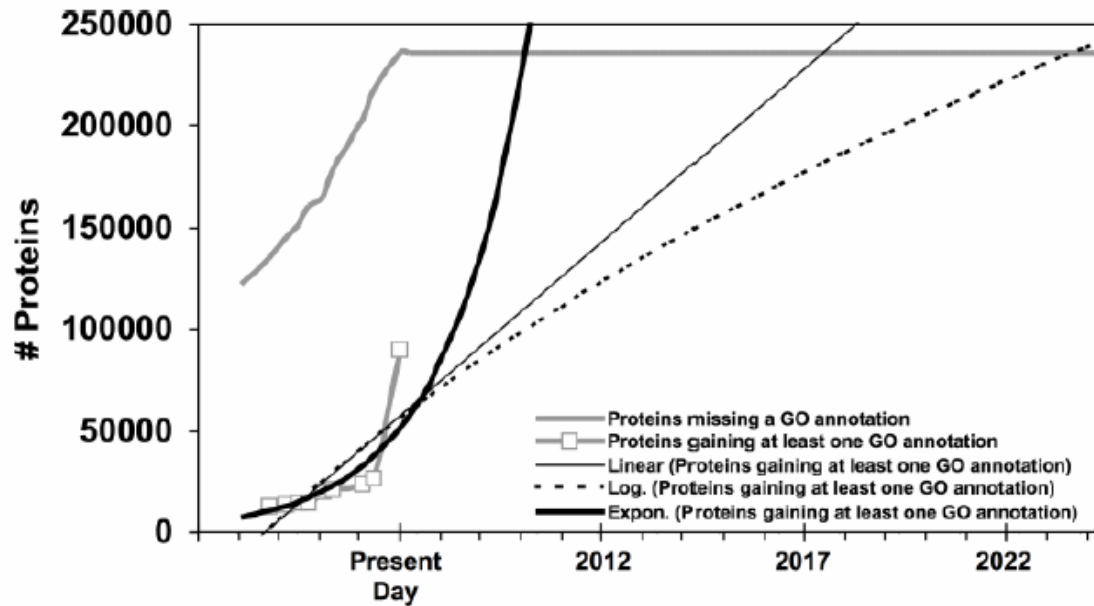


Fig. 11. GO annotation of all proteins in Swiss-Prot, with functions fitted to the gained-annotations line. For this plot, proteins were required to only have at least one annotation to be considered *fixed*.

How long does it take?

Data type	linear	R^2	exponential	R^2	logarithmic	R^2
Swiss-Prot Drosophila GO annotations	1.16	0.9570	0.55	0.9506	1.38	0.9572
Swiss-Prot Mouse GO annotations	3.06	0.8778	0.90	0.8436	3.75	0.8845
Swiss-Prot all species GO annotations	10.5	0.5746	3.05	0.7852	16.68	0.5530
Swiss-Prot all species <i>function</i> annotations	99.0	0.9807	9.12	0.8870	1.07×10^9	0.8207
Entrez Gene Human GeneRIFs	13.0	0.9788	0.003	0.7132	24.83	0.9784
Entrez Gene Mouse GeneRIFs	38.3	0.9777	0.40	0.7227	629,396	0.9221

So, how do you do
text mining?

Two approaches that are not
coexisting peacefully

Two approaches to NLP

Knowledge-based

Statistical/machine
learning

First approach to NLP

- Rule-based
- AI, linguistics
 - Ontologies
 - Knowledge bases
- Patterns (regular, context-free...)
- Procedures

K-based: procedural

- Patterns (regular, context-free, ...)
- Procedures

```
if (currentWordEndsWith-ing) {  
    if (previousWordIsThe) {  
        if (nextWordIsOf) {
```

K-based: regex

- Patterns (regular, context-free, ...)
- Procedures

```
$geneName = "[A-Za-z]+--[0-9]";
```

```
$input =~ /interaction of ($geneName) with ($geneName)/;
```

```
$interactionAssertion->setGene1($1);
```

```
$interactionAssertion->setGene2($2);
```

K-based: CFGs

- Patterns (regular, context-free, ...)
- Procedures

`NounPhrase -> NounPhrase+ Conjunction NounPhrase`

`NounPhrase -> Predeterminer Determiner+ Adjective+ Noun`

Knowledge-based approaches

Why they work

- Patterns are real
 - Psychologically
 - Formally adequate (mostly)
- Intuition works
- No need for training data

Knowledge-based approaches

Why they're hard

- Knowledge takes time to get
- Process of developing large rule sets can be slow
 - Consider English syntax...

Second approach to NLP

- Mosteller & Wallace
- Bayesian
- Other machine learning techniques

Statistical/ML approaches

- Frame the NLP task as a series of classification problems
 - Which POS is this?
 - Which word meaning?
 - Which phrasal grouping?

Statistical approaches

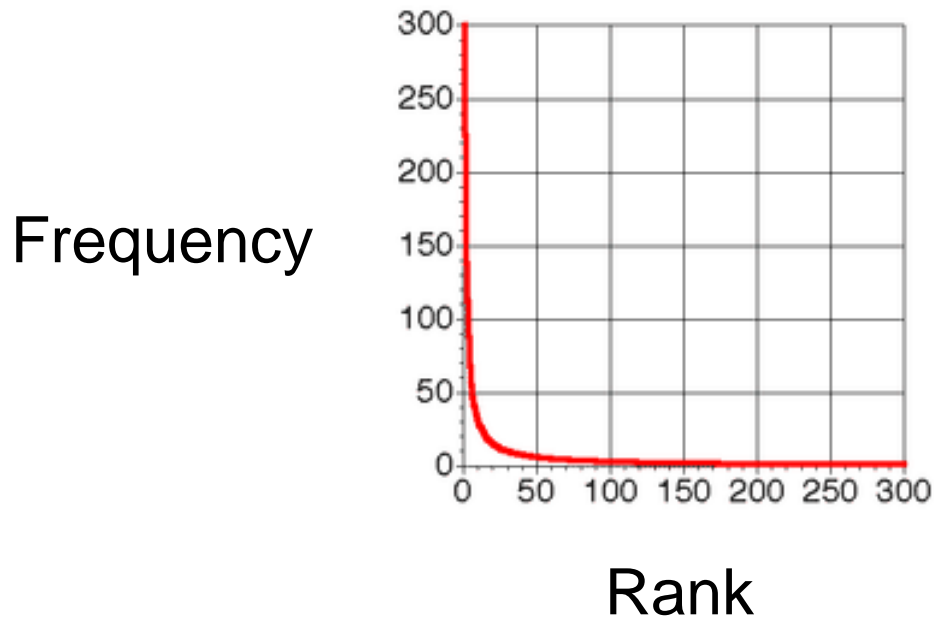
Why they work

- Statistics can be proxy for knowledge
- Some interesting stuff is frequent enough to be tractable

Statistical approaches

Why they're hard

- Problem: sparse data



Statistical approaches

Why they're hard

- Solutions: smoothing, back-off

Statistical approaches

Why they're hard

- Problem: labelled training data is expensive

Statistical approaches

Why they're hard

- Solutions:
 - spend money
 - figure out how to use other people's
 - "weakly labelled" data

**Knowledge-based or
statistical: what to do??**

Knowledge-based vs. statistical approaches

- Pragmatic answer #1: if you must pick one...
 - Is it cheaper to label more training data, or to put time into developing patterns?

Knowledge-based vs. statistical approaches

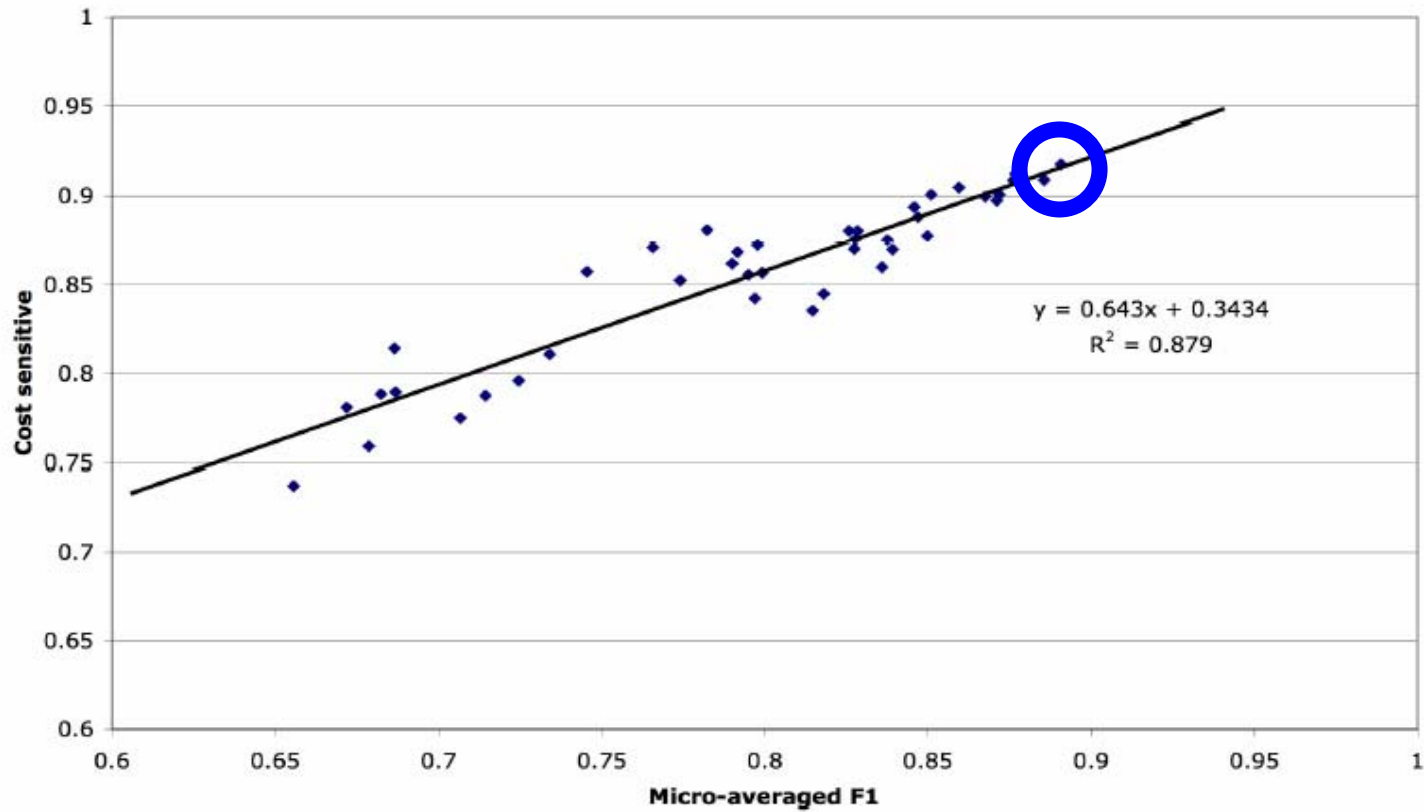
- Researcher's answer:
 - Use one as the baseline for the other

Knowledge-based vs. statistical approaches

- Pragmatic **the 2.5th** **ne them**
 - Do
 - Statistical **approach** **rule-**
based

"Natural language processing is never pure and rarely simple."

Which works better?



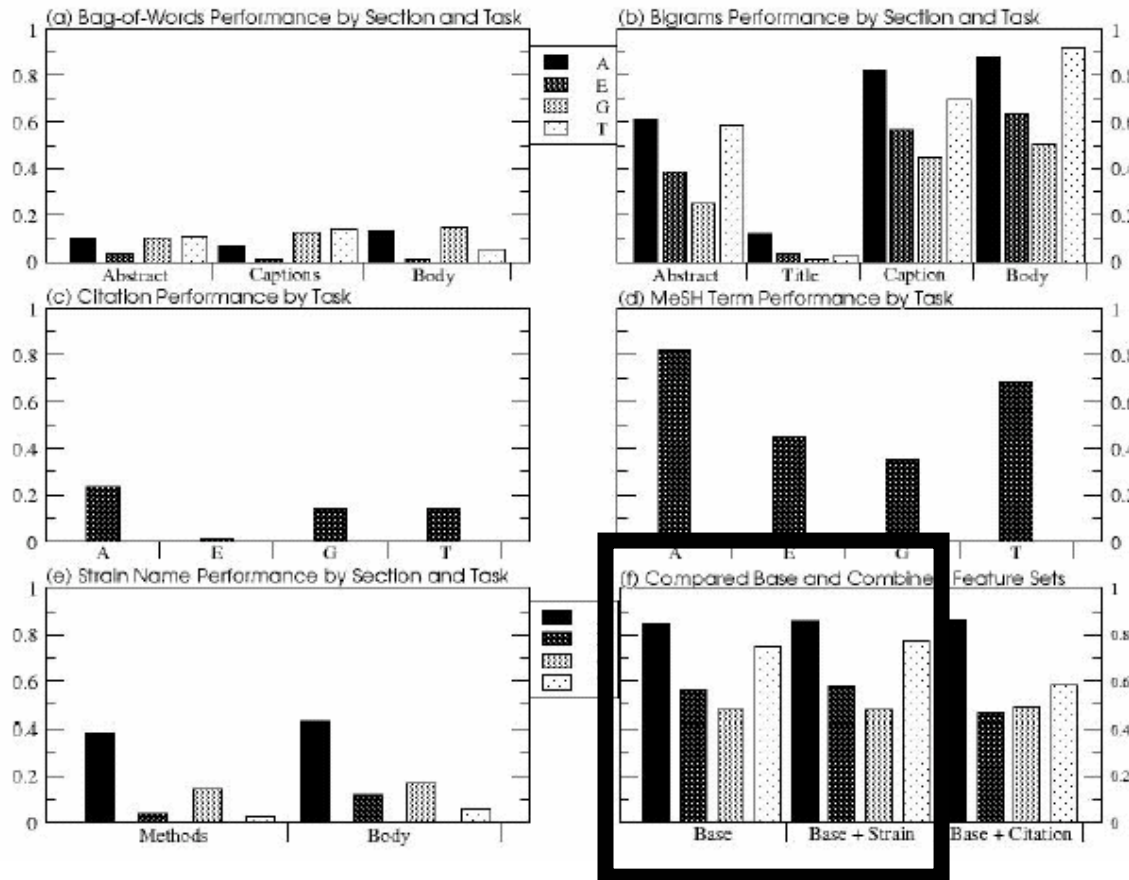
Pestian et al. (2007)

A rapprochement

Conceptual features for information retrieval

- Task: retrieve sentences that contain mentions of mutations.
- Keyword approach: 1,092
- Recognize mutation mentions: additional 2,171

Conceptual features in document classification



Conceptual features in document classification

Table 2: Feature values with highest information gain by task. Values beginning with ‘MGI:’ are mouse strain names, normalized to their MGI identifiers. Italicised values are MeSH terms, and the remaining values are bigrams.

A	E	G	T
<i>Mice, Knockout</i>	mous embryo	<i>Mice</i>	<i>Mice</i>
<i>Mice</i>	<i>Mice</i>	<i>Mice, Knockout</i>	<i>Mice Knockout</i>
MGI:2160041	situ hybrid	MGI:2160041	MGI:2160041
MGI:2160085	neural tube	<i>Animals</i>	MGI:2160085
southern blot	<i>Gene Expression Regulation, Developmental</i>	MGI:2160085	southern blot
target vector	volk sac	mous tissu	tumor incid
<i>Animals</i>	embryo b	southern blot	histolog analys
mutant mice	sagitt section	MGI:2159769	apc mice
<i>Mice, Inbred C57BL</i>	transvers section	MGI:2161069	<i>Mice, Transgenic</i>
wild-typ mice	branchial arcl	b southern	<i>Intestinal Neoplasms</i>

Untapped conceptual types

Malignancies (F = 0.84)

PMID: 15316311

Morphologic and molecular characterization of renal cell carcinoma in children and young adults. A new WHO classification of renal cell carcinoma has been introduced in 2004. This classification includes the recently described renal cell carcinomas with the ASPL-*TFE3* gene fusion and carcinomas with a PRCC-*TFE3* gene fusion. Collectively, these tumors have been termed Xp11.2 or *TFE3* translocation carcinomas, which primarily occur in children and young adults. To further study the characteristics of renal cell carcinoma in young patients and to determine their genetic background, 41 renal cell carcinomas of patients younger than 22 years were morphologically and genetically characterized. Loss of heterozygosity analysis of the von Hippel-Lindau gene region and screening for VHL gene mutations by direct sequencing were performed in 20 tumors. *TFE3* protein overexpression, which correlates with the presence of a *TFE3* gene fusion, was assessed by immunohistochemistry. Applying the new WHO classification for renal cell carcinoma, there were 6 clear cell (15%), 9 papillary (22%), 2 chromophobe, and 2 collecting duct carcinomas. Eight carcinomas showed translocation carcinoma morphology (20%). One carcinoma occurred 4 years after a neuroblastoma. Thirteen tumors could not be assigned to types specified by the new WHO classification: 10 were grouped as unclassified (24%), including a unique renal cell carcinoma with prominently vacuolated cytoplasm and WT1 expression. Three carcinomas occurred in combination with nephroblastoma. Molecular analysis revealed deletions at 3p25-26 in one translocation carcinoma, one chromophobe renal cell carcinoma, and one papillary renal cell carcinoma.

Jin et al. (2006)

Mouse strains

- CAST/EiJ
- C57BL
- SJL/J
- SEG
- C3H/He
- RIII
- DBA/1

Mutations

- *Ala64→Gly*
- *Ala64Gly*
- *A376G*

Point/Counterpoint

Contradictory findings

- TREC 2003: "...searching in the MeSH and substance name fields, along with filtering for species, accounted for the best performance" (Hersh and Bhupatiraju 2003, Caporaso et al. 2005)
- TREC 2004: "Approaches that attempted to map to controlled vocabulary terms did not fare as well" (Hersh et al. 2004)

Understanding the TREC 2004 results

- Poor choice of concepts
 - MeSH terms only, which is known to have problems even if manually indexed
- “Conceptual” systems weren't very good (or didn't try very hard) at concept recognition
 - Even synonymy not detected well (1 case)
 - Methods not described, so presumably not a focus of the work (2 cases)
- Hersh et al. (2004) overstate role of concepts in these systems
 - Synonym source only (1 case)
 - Only one of several features (1 case)

I'm convinced in theory, but will it scale?

- Jin et al. (2006): for malignancy mentions, relatively small amount of training data sufficed
- Caporaso et al. (2007): mutation patterns were learnable with small person-hour investment

Conclusion

- Statistical and conceptual approaches to text mining can coëxist peacefully
 - Statistical and rule-based concept recognizers can work well
 - Concepts are good features for statistical systems

Acknowledgements

- Bill Baumgartner
- J. Gregory Caporaso
- Larry Hunter
- Pete White