Using GO to improve text information access: A case study using human disease genes in yeast

Colleen Crangle, Ph.D.[1,2] (with Lynne Sopchak, Ph.D.[1] and Alex Zbyslaw, M.S.[1])

[1]ConverSpeech LLC, Palo Alto, CA & [2]Faculty of Informatics, Univ. of Ulster, N. Ireland

In the context of the human genome project … How do we make all that interesting natural-language information out there easily accessible?

Wouldn't it be nice ...

1. … "to have a database which combines all the information ever gathered and published about a single gene and its expression in response to whatever treatment…"
2. … to be able to get for any gene "[t]he putative function, ... biochemical pathways, ... chromosomal loci, ... functional categories ... gene families."
3. … "to link the gene to historical or archived data [on]... phenotypic variations resulting from mutations like translocations, inversions, and deletions…." "... and allelic variations."
4. … to compare "the chromosomal loci of genes of interest between species ..."

From work by Timothy Patrick and Lillian Folk at the University of Missouri-Columbia

Wouldn't it be nice ...

… to actually *find* all the MEDLINE articles that are relevant to a gene of interest

and then, for citations or full-text articles,

- Identify protein functions
- Detect relations between genes
- Extract functionally coherent gene groups
- Identify gene-disease connections
- Automatically associate genes with GO terms
- Identify relations among genes and proteins that form functional networks
- …

The Problem

Different names for the same entity

IFNG, Interferon gamma, Ifg IFN-g, IFN-gamma, IFNgamma, IFNG2, gamma interferon, interferon type ii

PubMed Searches July 12, 2002

Search term	# Citations returned		Details of generated PubMed search	PubMed Translation
IFNG (also Ifng)	**79**		IFNG[All Fields] (also Ifng[All Fields])	
interferon, gamma (also Interferon gamma)	**32,845**		interferon, gamma[All Fields] (also Interferon gamma[All Fields])	("interferon type ii"[MeSH Terms] OR "interferon gamma"[Text Word] (also …"Interferon gamma"[Text Word])
Ifg	**432**		Ifg[All Fields]	
IFN-g	**71**		IFN-g[All Fields]	
IFN-gamma	**24,245**		IFN-gamma[All Fields]	
IFNG2	"No items found. One of your terms is not found in the database."		IFNG2[All Fields]	
IFNgamma	**1026**		IFNgamma[All Fields]	
gamma-interferon	**28,493**		gamma-interferon[All Fields]	("interferon type ii"[MeSH Terms] OR gamma-interferon[Text Word])
interferon type ii	**26,561**		interferon type ii[All Fields]	("interferon type ii"[MeSH Terms] OR interferon type ii[Text Word])

The Problem

Different names for the same entity

IFNG, Interferon gamma, Ifg IFN-g, IFN-gamma, IFNgamma, IFNG2, gamma interferon, interferon type ii

Some names are really descriptions

ATP6V1H: ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H *"... vacuolar H(+)-ATPase ..."*

The Problem

Different names for the same entity

IFNG, Interferon gamma, Ifg IFN-g, IFN-gamma, IFNgamma, IFNG2, gamma interferon, interferon type ii

Some names are really descriptions

ATP6V1H: ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H

"... vacuolar H(+)-ATPase ..."

The same name can refer to distinct entities

Both these genes have the alias *SFD*, which is used in the literature

TIMP3: tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)

“SFD”

Sorsby’s fundus dystrophy
spray freeze drying
somatoform disorders
standard fat diet (also “standard chow”; “standard food”; “high-salt/high-fat diet”; “solid food diet”)
source to field distance (also “source to film distance”)
Société Française de Distilleries (1)
sulfur hexafluoride
sulfonamide derivatives
small-for-date
symptom-free days
sulfadiazine
schizophreniform disorder
Shenfu Decoction
sap flow density
seasonal facial dermatitis
semantic processing deficit
single fiber density
solitary focal demyelination
sfdA and sfdB for suppressors of flbD
sequential fractal dimension
spectra of spinach ferredoxin spectrofluorometric detection
survival free period (of the disease)
Silo filler’s disease
Special Filter Drabkin
severely folate-deficient
soluble fibrinogen derivatives
summary focal dose (also “single focal dose”)
misspelling of “SDF” for “superficial digital flexor”
in address code “CB2 SFD”
in “SFD HIV _ test”
cell line (NB4-R1(SFD))
in “the silicon detector SFD from …”
in “the first author (SFD)” [S. F. Dye]
fractional D shortening
superficial femoral dissection
staphylococcal food-borne disease
standardized forced diuresis
stuffy (sfy) and stuffed (sfd) zebra fish
suppressors of fatty acid (stearoyl) desaturase deficiency (sfd) mutants
silicon detector SFD from Scanditronix
scrofuloderma
sense frequency deltas

and

sub fifty-eight-kDa doublet or sub-fifty-eight-kDa dimmer

The Problem

Different names for the same entity

IFNG, Interferon gamma, Ifg IFN-g, IFN-gamma, IFNgamma, IFNG2, gamma interferon, interferon type ii

Some names are really descriptions

ATP6V1H: ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H

"... vacuolar H(+)-ATPases ..."

The same name can refer to distinct entities

Both these genes have the alias *SFD*, which is used in the literature

TIMP3: tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)

An entity can be referred to by a more general term

"The ***TIMPs*** tango with MMPs and more in the central nervous system"

"Different levels of ***TIMPs*** and MMPs in human lateral and medial rectus muscle tissue …"

What if we want to find all articles about CGI-11?

Why cgi-11? Search on "cgi-11 and pde7a" for spastic paraplegia 5A

LocusID	51606
Org	*Hs*
Symbol	**ATP6V1H**
Description	ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H
Position	8p22-q22.3

PubMed returns 2 citations!

We created BioMedPlus, a federated ontology for *term expansion* and *results filtering*

Drawn from LocusLink, SGD, GO, …

cgi-11; ATP6V1H; ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H; SFD; SFDalpha; SFDbeta; VMA13

Found 216 citations …

TIMP-3, collagen, and elastin immunohistochemistry and histopathology of Sorsby's fundus dystrophy.
PMID: 10711711 [PubMed - indexed for MEDLINE]

Ultrasonographic tissue characterization of equine superficial digital flexor tendons by means of gray level statistics.
PMID: 10685695 [PubMed - indexed for MEDLINE]…

WHICH ARE REALLY ABOUT CGI-11?

We created BioMedPlus, a federated ontology for *term expansion* and *results filtering*

cgi-11; ATP6V1H; ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H; SFD; SFDalpha; SFDbeta; VMA13

What do we use for the filter?

Found 216 citations …

TIMP-3, collagen, and elastin immunohistochemistry and histopathology of Sorsby's fundus dystrophy.
PMID: 10711711 [PubMed - indexed for MEDLINE]

Ultrasonographic tissue characterization of equine superficial digital flexor tendons by means of gray level statistics.
PMID: 10685695 [PubMed - indexed for MEDLINE]…

WHICH ARE REALLY ABOUT CGI-11?

We created BioMedPlus, a federated ontology for *term expansion* and *results filtering*

cgi-11; ATP6V1H; ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H; SFD; SFDalpha; SFDbeta; VMA13

What do we use for the filter?

GO annotations for the expanded list of terms.

Found 216 citations …

TIMP-3, collagen, and elastin immunohistochemistry and histopathology of Sorsby's fundus dystrophy.
PMID: 10711711 [PubMed - indexed for MEDLINE]

Ultrasonographic tissue characterization of equine superficial digital flexor tendons by means of gray level statistics.
PMID: 10685695 [PubMed - indexed for MEDLINE]…

WHICH ARE REALLY ABOUT CGI-11?

More Problems …

The language of GO, like all natural language, is complex in structure and morphology (derivational and inflectional)

- **"hydrolase activity"** (a GO term) can appear in the literature as
  - "…**activity** in sucrose **hydrolysis**…"
  - "…**hydrolase activities**…"
  - "…inhibitory **activity** against recombinant Plasmodium falciparum SAH **hydrolase**…"
  - "…three major 20S proteasome **activities** (chymotrypsin-like, trypsin-like, and peptidylglutamyl-peptide **hydrolase)**…"
  - "…S-adenosyl-L-homocysteine-**hydrolase** (SAHH) **activity**…**"**
  - "…SAH**H activity**…**"**
  - "…ATP-**hydrolase reaction**…**"**
  - … and so on

Article 67J Biol Chem 1999 May;274(22):15913-9

Recombinant SFD isoforms activate vacuolar proton pumps.

…… Although SFD is essential to the **activation of ATPase and proton** pumping activities catalyzed by holoenzyme, its constituent polypeptides have not been separated to determine their respective roles in ATPase functions. Recent molecular characterization of these subunits revealed that they are isoforms that arise through an alternative splicing mechanism …. We determined that purified recombinant proteins, rSFDalpha and rSFDbeta, when reassembled with SFD-depleted holoenzyme, are functionally interchangeable in restoration of **ATPase and proton** pumping **activities**. In addition, we determined that the V-pump of chromaffin granules has only the SFDalpha isoform in its native state and that rSFDalpha and rSFDbeta are equally effective in restoring **ATPase and proton** pumping **activities** to SFD-depleted enzyme. Finally, we found that SFDalpha and SFDbeta structurally interact not only with V1, but also withV0, indicating that these activator subunits may play both structural and functional roles in coupling **ATP** hydrolysis to **proton** flow.

Full, Partial and Overlapping Full and Partial Match

Full Matches

- **ATP binding** (2): From: GO Annotation
- **ATP biosynthesis** (2): From: GO Annotation
- **proton transport** (11): From: GO Annotation

Partial Matches

- **hydrogen-exporting ATPase activity, phosphorylative mechanism** (4): From: GO Annotation
- **proton-exporting ATPase** (12): From: GO Annotation

Normalize the GO term and stem the words

Some words are “optional”

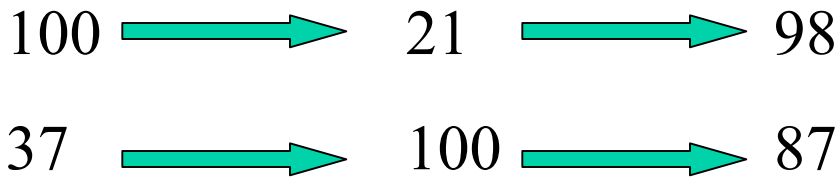
Some words are “insufficient”

Some words are ignored

RESULTS

(for “cgi-11 and pde7a”)_

PubMed Precision 100% (19/19) Recall *at most* 37% (19/52) There are *at least* 52 MEDLINE citations that are relevant.	**Distiller with Term Expansion** Precision 21% (52/252) Recall *possibly as high as* 100% (52/52) **Distiller with Term Expansion and Results Filtering** **Precision 98% (45/46)** **Recall *at most* 87% (45/52)**



CONCLUSION

The BioMedPlus ontology improves access to gene-related text information if it is used not only for term expansion but also results filtering.

GO terms play an essential role in this process.

METHOD

Case Study. We chose a set of genes identified as new candidate genes for the putative mitochondrial-related disorder of spastic paraplegia 5A [*].

We then:

Searched the MEDLINE citation database for all articles relevant to this set of genes, using the query "CGI-11 OR LOC85479 OR PDE7A" submitted to PubMed.	Submitted this same query to the ConverSpeech Distiller, a front-end to the PubMed database that has integrated into it the BioMedPlus ontology for **term expansion** and results **filtering**.

Compared the results.

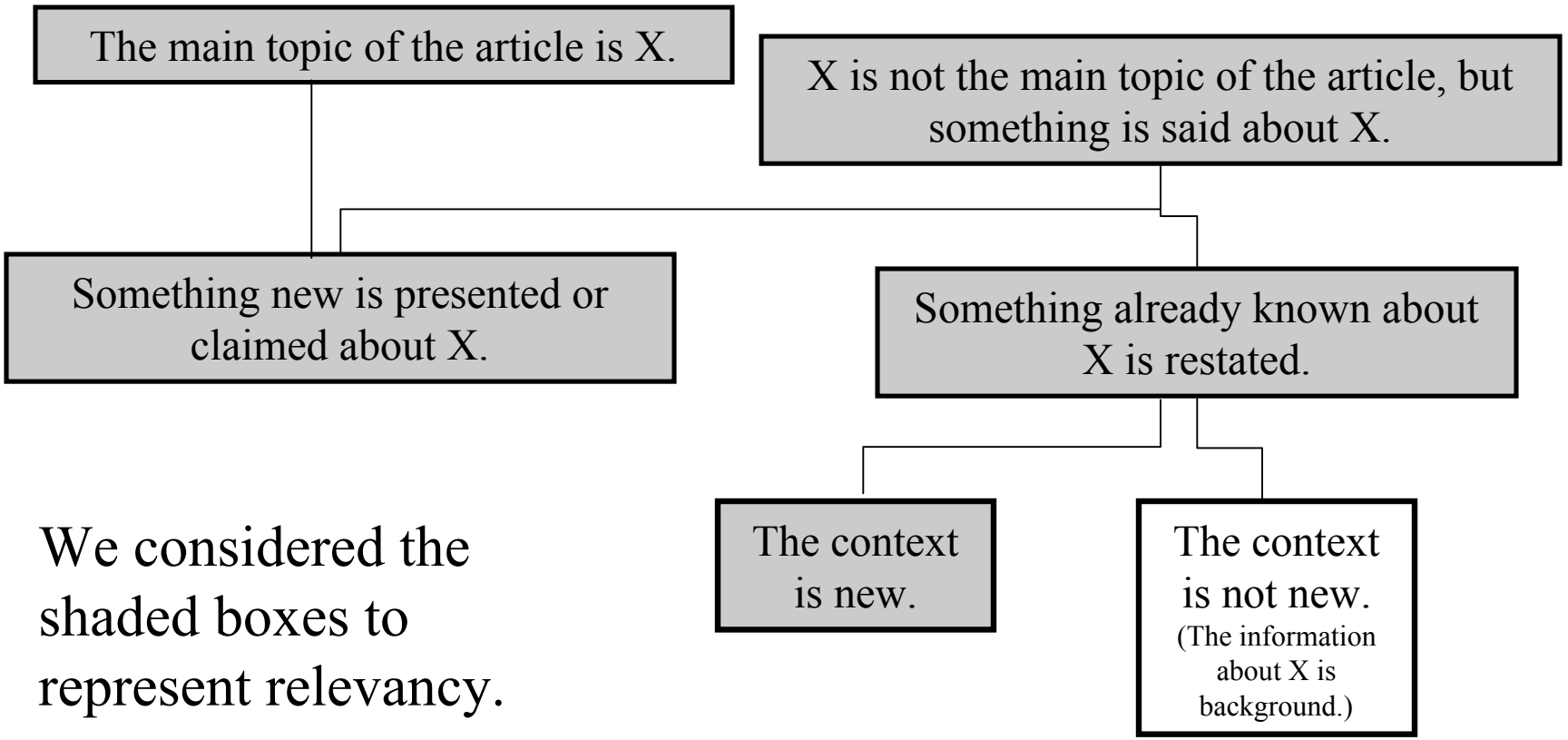
Term expansion adds aliases and other synonyms to a search.

Filtering examines the results of a search and makes a judgment as to which are truly relevant.

[*] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. Nat Genet. 2002 Aug;31(4):400-4. Systematic screen for human disease genes in yeast.

How to Judge Relevancy?

When is an article (MEDLINE citation) *about* some specific thing X?



We considered the shaded boxes to represent relevancy.

METHOD (cont.)

Compared Three Results Using the Measures
Recall and **Precision**

the PubMed search	the Distiller search with *term expansion* only the Distiller search with *term expansion* and *results filtering*

Let RETURNED be the number of citations returned by a search.
Let RELEVANT be the number of citations that are actually relevant to a search.
Let RETREL be the number of citations returned by a search that are relevant.

Recall = RETREL / RELEVANT
Precision = RETREL / RETURNED

RESULTS

- The PubMed search returned 19 citations.
- The Distiller search with term expansion returned 252 citations.
- Of the 252 citations returned by the Distiller search with term expansion, 52 were judged relevant by domain expert (2nd author).
- The Distiller search with term expansion and results filtering returned 46 citations.
- Of the 46 citations returned, 45 were judged relevant by domain expert (2nd author).

DISCUSSION

To improve the Distiller's precision without jeopardizing its recall …

…we had conducted a failure analysis. We analyzed the 200 non-relevant results from the Distiller search with term expansion only to understand what each article was about and why the citation may have been returned by the Distiller.

Failure analysis showed that term expansion introduced references to over two dozen different biomedical entities.

> For example, "SFD," one of the aliases for "cgi-11," is widely used as an acronym in the biomedical literature. See **Abbreviation "SFD"** page. But "sfd" as an acronym for "sub fifty-eight-kDa Doublet" or "sub-fifty-eight-kDa dimmer" was needed to retrieve 4 citations.

Results filtering eliminated all citations containing these extraneous referring phrases, but also excluded seven relevant citations.

DISCUSSION

We designed the results filtering of the Distiller to test the following hypothesis:

Of the 252 citations returned for the Distiller search on "cgi-11 OR LOC85479 OR pde7a" with term expansion, the relevant ones will contain in their titles reference to one or more "core concepts" related to cgi-11, LOC85479, and pde7a.

How to get the "core concepts"?

We used the BioMedPlus ontology to find related biomedical terms – e.g., synonyms, more specific (e.g., gene products), terms from definitions or descriptions, GO annotations – and determined significant collocations and high-frequency biomedical terms. These were the "core concepts."

What is an ontology?

- An ontology represents "what there is" in a domain. An ontology includes a vocabulary (which promotes a standard way of naming the concepts of the domain) and a system of hierarchical and other relations between and among the concepts and the vocabulary items.
- The ConverSpeech ontology – BioMedPlus – is a language-oriented ontology. Concepts in BioMedPlus are represented as synonym sets. A synonym set is a set of words or phrases that express the same meaning or refer to the same biomedical entity in at least one context. Any particular word or phrase will have more than one sense if it means something different in different contexts. The most obvious example of a term having more than one sense is when it can be used to refer to more than one distinct biomedical entity.
- For example, {*ATP6V1H; ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H; CGI-11; SFD; SFDalpha; SFDbeta; VMA13*} is the synonym set for the *Hs* gene with official symbol *ATP6V1H*. The term *SFD,* however, has many distinct senses (see later).
- BioMedPlus is presently built up from several source ontologies, including the following FIVE. It contains over 785,000 senses, 696,000 distinct words or phrases, and 500,000 synonym sets.
  - The three ontologies of GO, the Gene Ontology (http://www.geneontology.org/index.shtml). GO aims to provide controlled vocabularies for the description of the molecular function, biological process and cellular components of gene products. In our adaptation of GO, we provide synonyms and pointers to hypernyms (more general) and holonyms, which are also reversed to provide pointers to hyponyms (more specific) and meronyms (part-whole).
  - The National Center for Biotechnology Information (NCBI)'s LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/). LocusLink has descriptive information about genetic loci, including information on official nomenclature and aliases.
  - Linguistic data from SGD™ , a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae* (http://www.yeastgenome.org/).

OBJECTIVE

To understand how a biomedical ontology can improve access to natural-language, free-text information about gene-related discoveries.