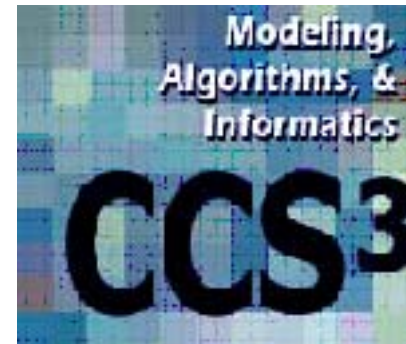


The LANL BioCreAtIvE submission

Karin Verspoor, verspoor@lanl.gov

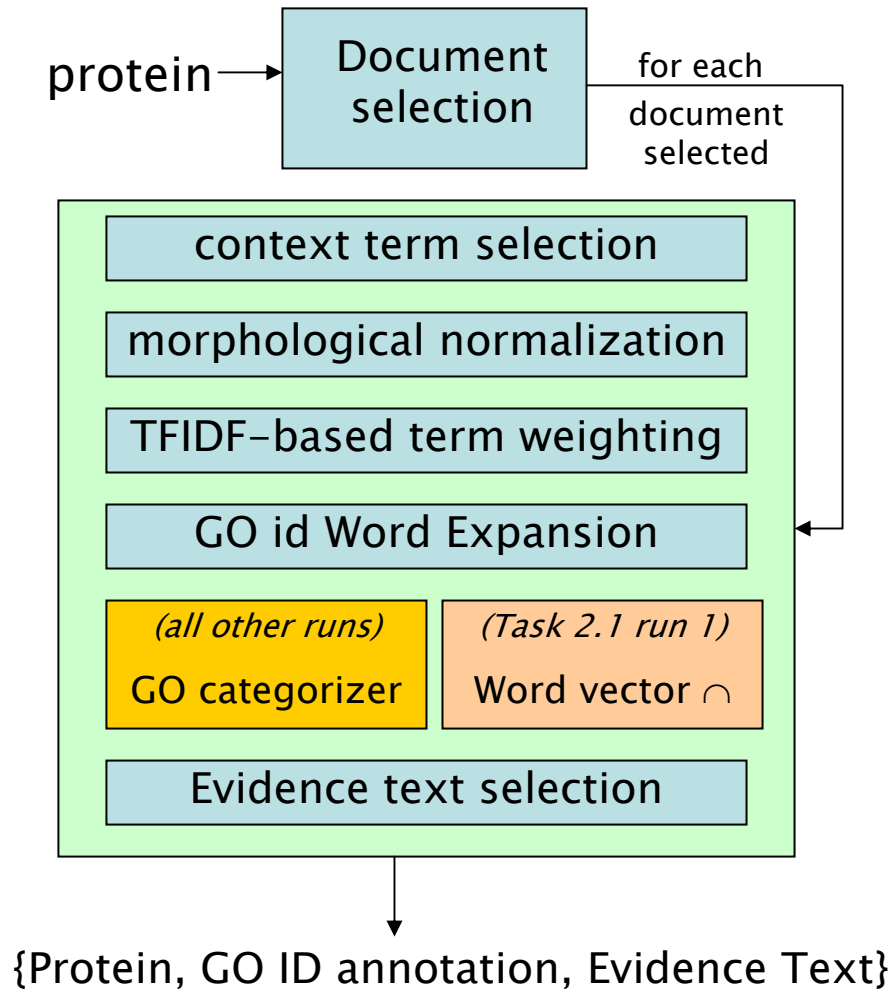
Judith Cohn, Cliff Joslyn, Sue Mniszewski,
Andreas Rechtsteiner, Luis M. Rocha, Tiago Simas
March 29, 2004



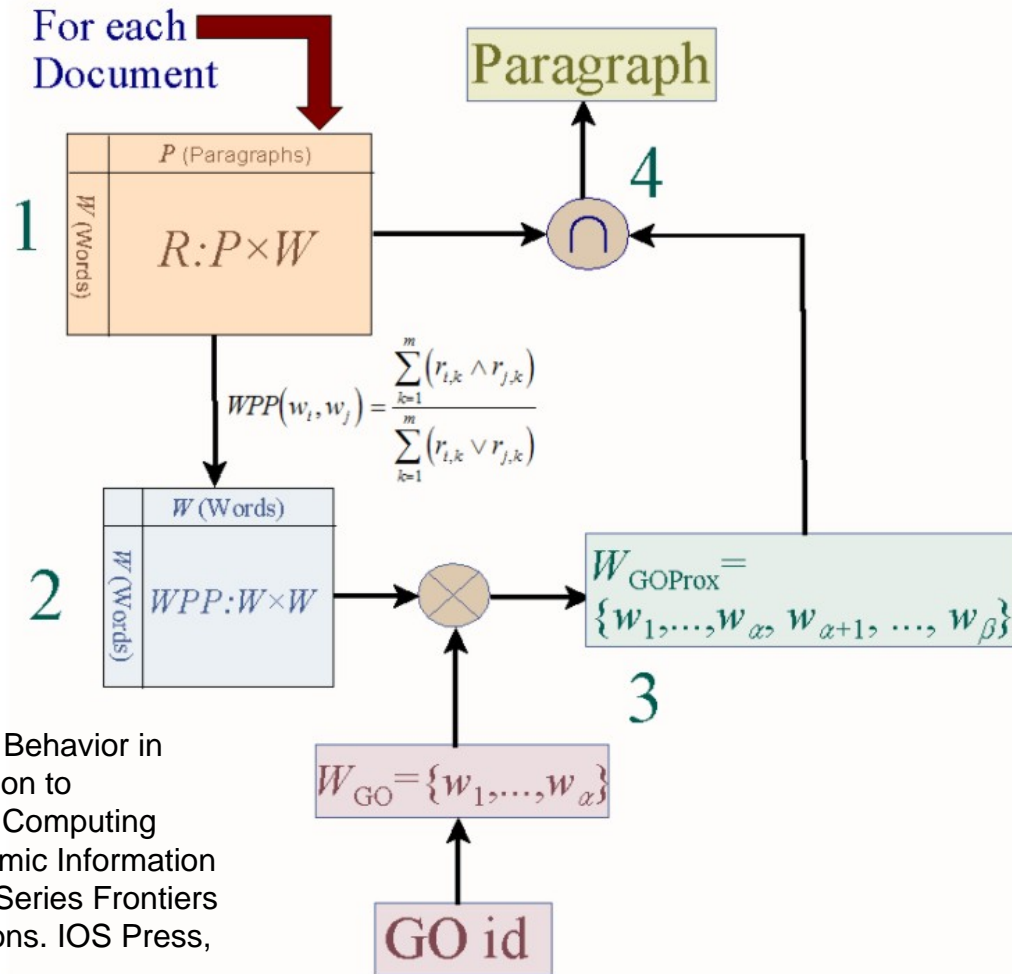
Strategy: Two Technologies

- Annotation as Categorization
 - Application of categorization methodology utilizing the Gene Ontology structure to find the best covering nodes given a set of node “hits”.
 - “Hits” in this case are based on overlaps between input terms and GO node labels
- Word Proximities
 - Based on all available documents, establish an association between words based on in-paragraph co-occurrence
 - Used to associate additional words with GO node labels
 - Used to select paragraphs close to a set of words

LANL BioCreative System Architecture



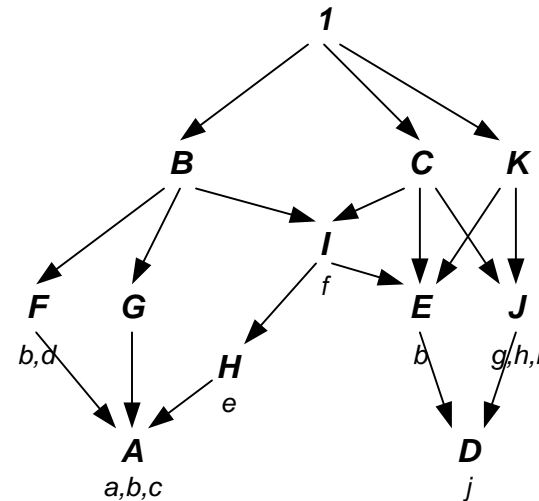
Proximity (Rocha and Simas): Task 2.1, Run 1 architecture



Rocha, Luis M. (2002). "Semi-metric Behavior in Document Networks and its Application to Recommendation Systems". In: Soft Computing Agents: A New Perspective for Dynamic Information Systems. V. Loia (Ed.) International Series Frontiers in Artificial Intelligence and Applications. IOS Press, pp.137-163.

GO Categorizer (Joslyn, Mniszewski, et al)

- Given inputs (c, e, i, \dots), what nodes (e.g. C, I, H) are best to pay attention to? Answer is based on pseudo-distances between comparable nodes, measured according to the structure of the ontology, with rank ordering of nodes balancing *coverage* – covering as many inputs as possible – and *specificity* – covering the inputs at the lowest level possible.
- Inputs are clustered based on comparable high-score nodes.



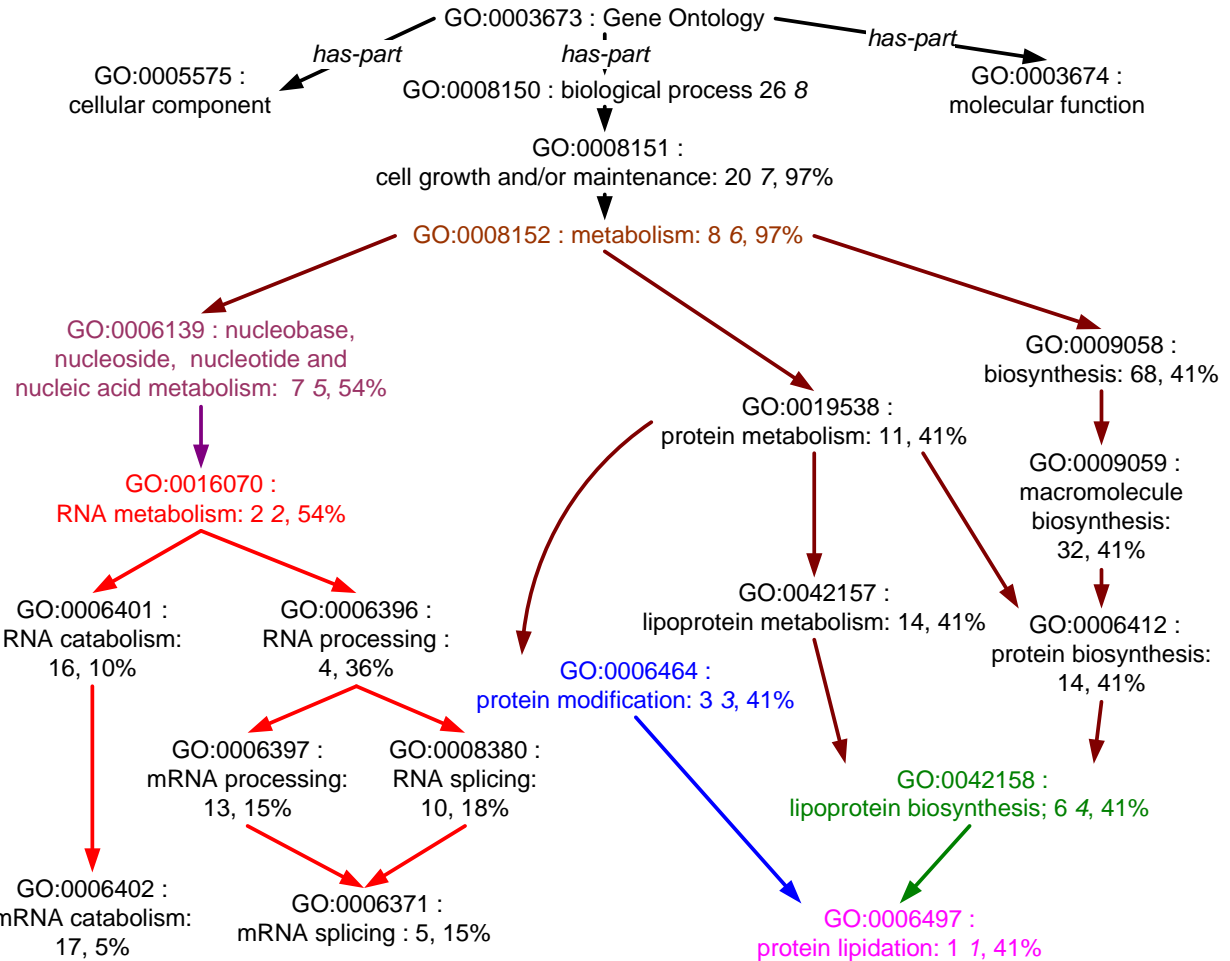
Joslyn, C., S. Mniszewski, A. Fulmer, G. Heaton (2003). "Structural Classification in the Gene Ontology". In Proceedings of the Sixth Annual Bio-Ontologies Meeting (Bio-Ontologies 2003), Brisbane, Australia, June 28, 2003.

Joslyn, C., S. Mniszewski, A. Fulmer, G. Heaton (2004) "The Gene Ontology Categorizer", to be presented at ISMB 04, in press for Bioinformatics.

Clustering of Gene products

Query result for a set of inputs consisting of genes annotated to GO nodes.

The inputs cluster into roughly two groups: under **protein lipidation** and **RNA metabolism**.



Input nodes hit:

GO:0018268, GO:0018270, GO:0018281
 GO:0018009, GO:0042050, GO:0018226
 GO:0042079, GO:0042050, GO:0018227
 GO:0018342, GO:0006507, GO:0018220
 GO:0018008, GO:0006497, GO:0019939
 GO:0018348, GO:0042082, GO:0016070
 GO:0000291, GO:0006396, GO:0000245
 GO:0006371, GO:0006399, GO:0006389
 GO:0045291, GO:0016555, GO:0006374
 GO:0016547, GO:0006365, GO:0000244
 GO:0016550, GO:0006377, GO:0016076
 GO:0000291, GO:0006401, GO:0009452
 GO:0001510, GO:0042245

Annotation Strategy

- Application of a categorization methodology which utilizes the structure of the Gene Ontology to find the best covering nodes given a set of node “hits”.
- The node hits are determined through term overlaps between node labels in the GO and selected text in the selected publication.
 - Terms are collected through analysis of the sentential context of the given protein.
 - The terms are processed to remove morphological endings such as verb endings or plurals.
 - These terms are weighted using a normalized TFIDF (term frequency inverse document frequency) value generated based on statistical analysis of the training documents. The weights represent the “contentfulness” of each term.

Protein References

- Different ways of referring to a given protein, even within a single document
- Must recognize all occurrences of the protein in the document
- Strategy:
 - utilize a list of synonym/acronym mappings constructed by our Procter & Gamble partners from sources such as HUGO, OMIM, etc.
 - expand the set of synonyms associated with individual Swiss-Prot IDs by looking for overlaps between the Swiss-Prot gene/protein names and names in the P&G list
 - fall back to EBI TrEMBL human database; generate variants of these names

Term mismatch: BioMorpher

- Morphological variants
 - GO annotations are fully-inflected noun phrases; text is also fully-inflected
 - So “proteins” in text and “protein” in a GO annotation won’t match using direct string comparison
 - Need to normalize the words to remove inflections
 - Solution: BioMorpher
 - Tool for normalizing words to their base forms
 - A tool for generic English tailored to biological context
 - contains large exception list of words which should not be normalized according to the standard rules
 - recognizes some unusual biological naming conventions and does not attempt normalization of any part of those names
e.g. *5'-GCRTGNCCAT-3'*, *naip-rs2*, *bits1*

GO Categorizer (GOC) as applied to terms

- Transform an input query into a set of node hits:
 - Morphologically normalize GO node labels
 - Look for any overlaps between input terms and terms in the normalized node labels
 - An overlap = a node hit, with strength based on the input weight of the term (from TFIDF)
 - Multiple overlaps on a given node count as multiple hits
- GOC traverses the structure of the GO, percolating hits upwards, and calculating scores for GO nodes.
- GOC returns a set of GO nodes representing cluster heads for weighted term input set, and data on which input terms contributed to the selection of each cluster head: *Annotation predictions*

GO Node matching: Definitions and Proximity

- Need to expand the set of terms that “fire” a GO node
- Use the “definitions” associated with GO nodes
 - Morphologically normalize
 - Remove stop words (including some terms extremely frequent in the definitions, such as “activity”, “cell”, “differentiate”)
- **GO id Word Expansion** [*Luis Rocha and Tiago Simas*]:
 - Expand terms associated with GO nodes based on the proximity of GO node terms to other terms in the document set
- Any hit on a term from a definition or proximity association to a GO node will count as a hit for that node
 - A parameter controls dampening of the hit strength, so that a hit on a node label term can be counted as more important than one of these indirect hits

Evidence Text Selection

- Paragraph selection: using vector intersection on matrix R
 - Measure the overlap of terms contributing to the selected GO annotation with individual paragraphs in the document, using the R ($P \times W$) matrix.
 - Use a vector intersection operation. We choose as evidence text the paragraphs associated with the columns of R (representing words occurring in a specific paragraph) that yield the largest intersection with the set of terms contributing to an annotation.
 - **For Task 2.1, Run 1:** We choose as evidence text for the GO id the paragraphs associated with the columns of R that yield the largest intersection with W_{GOProx} (the expanded set of terms associated with the GO node based on the given document) – *ignoring the protein entirely.*
- Sentence selection: simple overlap algorithm
 - Pick sentence with most overlap with terms contributing to the selected GO annotation

Example: Swiss-Prot protein ID P40337

- Synonyms extracted from the database

vhl, vhlh, von hippel-landau tumor suppressor,
von hippel-lindau, von hippel-lindau protein,
von hippel-lindau tumor suppressor protein

- Weighted term neighborhood

From “The von Hippel-Lindau tumor suppressor stabilizes novel plant homeodomain protein Jade-1” (PMID 12169691)

- jade-1|0.15975

- vhl|0.09216

- renal|0.01643

- 293t17|0.01341

- phd|0.01036

- 786-o|0.00835

- speckle|0.00724

- tubule|0.00562

- immunoprecipitation|0.0051

- cancer|0.00505

- stabilize|0.00496

- del1|0.00488

- aa|0.00446

- cotransfect|0.00441

- antiserum|0.00441

- 67 GO nodes hit directly; 272 hit through associated terms

P40337, PMID 12169691 results

	GO	Protein		
1.	GO:0005333	low	high	norepinephrine transporter activity
2.	GO:0005515	high	high	protein binding
3.	GO:0042382	high	high	paraspeckles
4.	GO:0042583	low	high	chromaffin granule

0005515,
Contributing
key terms:

- tubule
- aa
- jade-1
- cotransfect
- vhl

Sentence selected, Task 2.2, Run 1

E, expression levels of [cotransfected Jade-1](#) (J) (upper panels) or [VHL](#) protein (V) (lower panels) as measured by Western blotting of the same whole cell lysates used for immunoprecipitations in F and G. In 293T17 cells, VHL or empty pFLAG-CMV2 (ev) was [cotransfected](#) with FLAG- or HA-tagged [Jade-1](#) or truncations (see Fig. B for construct schematics), or with empty pCR3.1 uni HA (ev).

Paragraph selected, Task 2.2, Run 2

<P>The von Hippel-Lindau disease gene (<GLOSREF RID="G1"><IT>VHL</IT> </GLOSREF>) is the causative gene for most adult renal cancers. However, the mechanism by which <GLOSREF RID="G1">VHL</GLOSREF> protein functions as a renal tumor suppressor remains largely unknown. To identify low occupancy <GLOSREF RID="G1">VHL</GLOSREF> protein partners with potential relevance to renal cancer, we screened a human kidney library against human <GLOSREF RID="G1">VHL</GLOSREF> p30 using a yeast two-hybrid approach. <IT>Jade-1</IT> (<UNL>g</UNL>ene for <UNL>A</UNL>poptosis and <UNL>D</UNL>ifferentiation in <UNL>E</UNL>pithelia) encodes a previously uncharacterized 64-kDa protein that interacts strongly with <GLOSREF RID="G1">VHL</GLOSREF> protein and is most highly expressed in kidney. Jade-1 protein is short-lived and contains a candidate destabilizing (PEST) motif and plant homeodomains that are not required for the <GLOSREF RID="G1">VHL</GLOSREF> interaction. Jade-1 is abundant in proximal tubule cells, which are clear-cell renal cancer precursors, and expression increases with differentiation. Jade-1 is expressed in cytoplasm and the nucleus diffusely and in speckles, where it partly colocalizes with <GLOSREF RID="G1">VHL</GLOSREF>. <GLOSREF RID="G1">VHL</GLOSREF> reintroduction into renal cancer cells increases endogenous Jade-1 protein abundance up to 10-fold. Furthermore, <GLOSREF RID="G1">VHL</GLOSREF> increases Jade-1 protein half-life up to 3-fold. Thus, direct protein stabilization is identified as a new <GLOSREF RID="G1">VHL</GLOSREF> function. Moreover, Jade-1 protein represents a novel candidate regulatory factor in VHL-mediated renal tumor suppression.

Results Summary: Task 2.1

User, Run	# results	“perfect”	“generally”
4, 1	1048	268 (25.57%)	74 (7.06%)
5, 1	1053	166 (15.76%)	77 (7.31%)
5, 2	1050	166 (15.81%)	90 (8.57%)
5, 3	1050	154 (14.67%)	86 (8.19%)
7, 1	1057	272 (25.73%)	154 (14.57%)
7, 2	1864	43 (2.31%)	40 (2.15%)
7, 3	1703	66 (3.88%)	40 (2.35%)
9, 1	251	125 (49.80%)	13 (5.18%)
9, 2	70	33 (47.14%)	5 (7.14%)
9, 3	89	41 (46.07%)	7 (7.87%)
10, 1	45	36 (80.00%)	3 (6.67%)
10, 2	59	45 (76.27%)	2 (3.39%)
10, 3	64	50 (78.12%)	4 (6.25%)
14, 1	1050	303 (28.86%)	69 (6.57%)
15, 1	524	59 (11.26%)	28 (5.34%)
15, 2	998	125 (12.53%)	69 (6.91%)
17, 1	413	83 (20.10%)	19 (4.60%)
17, 2	458	7 (1.53%)	0 (0.00%)
20, 1	1048	301 (28.72%)	57 (5.44%)
20, 2	1050	280 (26.72%)	60 (5.73%)
20, 3	1050	239 (22.76%)	59 (5.62%)

Our submissions: [User 7](#)

Run 1: Direct from GO ID to paragraph using proximity only (10% no evidence text returned)

Run 2: Using full system with sentence selection (52% no evidence text returned)

Run 3: Using full system with paragraph selection (65% no evidence text returned)

Note that we submitted a higher number of results than other teams – in absolute terms, even our Run 2/3 results are comparable to the [middle band](#) of results. Our Run 1 result is arguably the [best](#).

Results Summary: Task 2.2

User, Run	# results	“perfect”	“generally”
4, 1	661	78 (11.80%)	49 (7.41%)
7, 1	156	1 (0.64%)	1 (0.64%)
7, 2	384	19 (4.95%) [1]	9 (2.34%) [1]
7, 3	263	2 (0.76%)	10 (3.80%)
9, 1	28	9 (32.14%)	3 (10.71%)
9, 2	41	14 (34.15%)	1 (2.44%)
9, 3	41	14 (34.15%)	1 (2.44%)
10, 1	120	35 (29.17%)	8 (6.67%)
10, 2	86	24 (27.91%)	6 (6.98%)
10, 3	116	37 (31.90%)	11 (9.48%)
15, 1	502	3 (0.60%)	8 (1.59%)
15, 2	485	16 (3.30%)	26 (5.36%)
17, 1	247	52 (21.05%) [1]	23 (9.31%) [0]
17, 2	55	1 (1.82%)	0 (0.00%)
17, 3	99	1 (1.01%)	1 (1.01%)
20, 1	673	20 (2.97%)	30 (4.46%)
20, 2	672	38 (5.65%)	26 (3.87%)
20, 3	673	58 (8.62%)	27 (4.01%)

Our submissions: [User 7](#)

Run 1: Using full system with sentence selection (50% no evidence text returned)

Run 2 [Rescored]: Using full system with paragraph selection (16% no evidence text returned, down from 60% in the original scoring)

Run 3: Using full system without context neighborhood selection, just fall-back scenario and paragraph selection (60% no evidence text returned)

Results discussion

- The evaluators conflated the two main variables of the test: GO annotation and evidence text selection were evaluated simultaneously in one score
- This method of evaluation proved problematic for us since our main focus was on the annotation component
 - Possible to get annotation correct and still get it “wrong” if the evidence text selected was poor
 - The sentence and paragraph selection depended on identification of terms relevant for a given predicted annotation

Annotation task results:

Run	Precision, Direct	Precision, Indirect	Precision, Total	Recall, Direct	Recall, Indirect	Recall, Total	F-score, Total
Run 1	0.061	0.185	0.246	0.059	0.181	0.241	0.243112
Run 2	0.061	0.185	0.246	0.059	0.181	0.241	0.243112
Run 3	0.057	0.228	0.285	0.059	0.238	0.298	0.291323

Results discussion continued

- The vast majority of the proteins in the test set did not have associated names in the database
 - 198/256 (77%) did not give an entry in our db
 - We were unable to hone in on the context window and “fell back” to using the top-ranked terms in the full document
 - This was problematic given the strategy for selecting sentences/paragraphs based on overlap with the terms used for annotation – the selected text was unlikely to reference the protein
 - (in retrospect, there were other strategies we could have used for selecting sentences/paragraphs which may have been more effective under these conditions)

Future Work

- Parameter testing
 - GOC parameters
 - Size of context window
 - Thresholds and functions used in Proximity
- Other ideas
 - Global word proximity matrix
 - Incorporation of “smarter” protein detection
 - IE techniques for honing in on protein relations
 - Investigation of needed adjustment of TFIDF to GO context

Thank you!

- Some websites with more information:

Karin Verspoor

<http://public.lanl.gov/verspoor>

Cliff Joslyn

<http://www.c3.lanl.gov/~joslyn>

Luis M. Rocha

<http://www.c3.lanl.gov/~rocha>