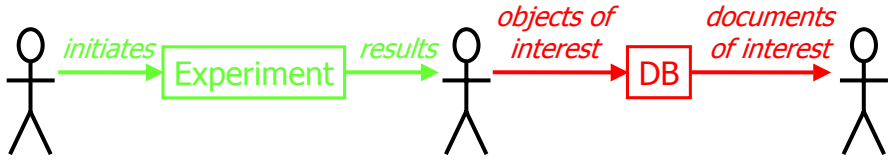
Extracting Keyphrases to Annotate Biological Objects of Interest

Rich Maclin
Mark Craven

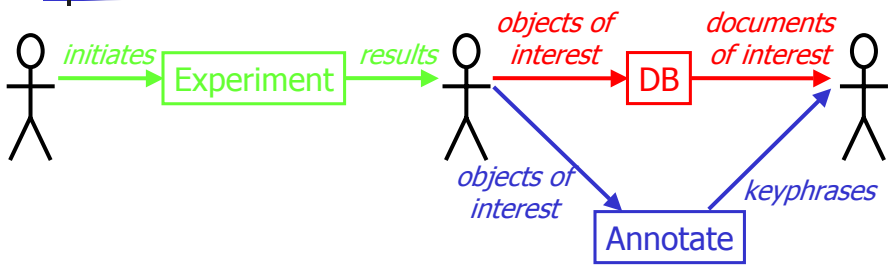
Steps in an Experimental Process



Objects of interest might be a gene, genes, protein, etc.

What happens next?
Someone sits down,
reads the documents of interest, and
tries to figure out what they have in common

Annotate: Extracting Keyphrases



Idea: automatically examine the documents of interest and summarize them with *keyphrases*

Keyphrases

- Could be as simple as keywords (single words, unigrams)
- As complex as whole sentences
- What makes a good keyphrase?
  - Highly indicative of biological objects of interest
    - Used (often) when describing those objects
    - Not used when describing other objects in the universe of similar objects

Annotation Notes

- Note, annotation not meant to replace other analysis, but to add to it
- Annotation makes use of the biological databases
- Leads to another task – evaluating the results of an annotation:

  What makes a good annotation?

Key Annotation Processes

- Defining universe of keyphrases
  - Which documents to use to define the universe?
  - How do we extract keyphrases from that universe?
- Ordering keyphrases
  - How do we count occurrences of the keyphrase?
  - What indicates a significant difference?

Andrade & Valencia, 1998

- *Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families*
- Used to annotate protein families
  - Protein families based on PDBSELECT
  - Family proteins based on similarity
- Extracts keywords and key sentences
  - Significance tests based on frequency
  - Sentences chosen based on average significance of words in sentence

Documents for A&V

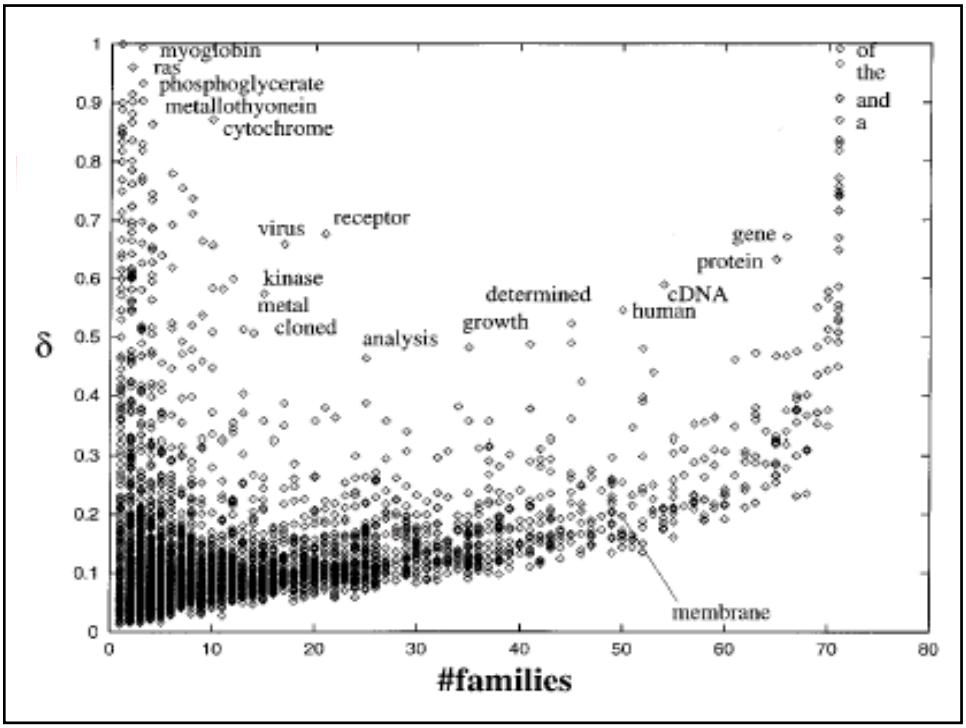
- Select proteins from PDBSELECT (<25% sequence similarity)
- Protein families from HSSP
  - Family members selected based on at least 40% sequence similarity
  - Small protein families excluded
  - Abstracts from SwissProt connections for family members

Identifying Keywords for A&V

- Hyphenated words at end of sentences connected
- Non letter/digit characters replaced with spaces
- Words consisting only of digits (numbers) removed
- Stemming – words that are the same except for one or two letters at the end considered the same (non standard stemming)

Statistics for A&V

- $Freq_w^f = Count_w^f / |f|$
- $\delta_w = \Sigma_{f \in families} Freq_w^f$ /#families word occurs in
- $MeanFreq_w = \Sigma_{f \in families} Freq_w^f / |families|$
- $\sigma_w = \sqrt{(1/(n-1)\ \Sigma_{f \in families}(Freq_w^f - MeanFreq_w)^2)}$
- $z_w^f = (Freq_w^f - MeanFreq_w) / \sigma_w$
- $z_s^f = (\Sigma_{w \in sentence\ s}\ z_w^f) / |sentence\ s|$



Keywords extracted by AbXtract for query ataxia

Text Sentences

keyword	freq.	z	found at
disorder	0.37	12.30	[95312868] [96255945] [96018070] [89070677] [96391788] [97288735] [96105020] [97262209] [96154672] [94141360] [91171851] [90174198] [92097021] [96254972] [90235178] [92058549] [95150036] [93233707] [93206979] [95357487] [83107389] [91169545] [95048379] [95381456] [95364870] [82029766] [89151839] [89139395] [96390593] [87164160] [89250669] [97041722] [83259005] [96404417] [94221101] [97294602] [95138623] [90316537] [97123513]
autosomal	0.33	9.02	[95312868] [96255945] [89070677] [97288735] [97262209] [96154672] [95372371] [90321562] [91171851] [96038263] [92097021] [96254972] [90235178] [92058549] [93233707] [86061765] [93206979] [95357487] [91169545] [96338579] [95048379] [88314587] [95213012] [82029766] [89003946] [92035738] [89139395] [83259005] [96404417] [94221101] [92072632] [90316537] [90259001] [97123513]
disease	0.30	5.29	[95312868] [96255945] [93104606] [97288735] [95372371] [96038263] [92194830] [96254972] [91288572] [92298322] [92058549] [86061765] [93144253] [95357487] [83107389] [91169545] [96338579] [95048379] [88314587] [96105008] [95213012] [95187960] [82029766] [92035738] [87239245] [96390593] [96081682] [86061784] [96377133] [94221101] [92072632]

generated by **AbXtract** (Fri Nov 21 19:36:47 GMT 1997)

Sentences extracted by AbXtract for query ataxia

Text Keywords

medline	sentence	score
90235178	cancer predisposition of ataxia telangiectasia heterozygotes	14.70
85254461	workshop on ataxia telangiectasia heterozygotes and cancer.	12.31
96154672	a t heterozygotes are moderately cancer prone	9.47
92298322	enhanced levels of radiation induced g2 phase delay in ataxia telangiectasia heterozygotes	8.36
86061784	cellular hypersensitivity to chronic gamma radiation in cultured fibroblasts from ataxia telangiectasia heterozygotes	7.72
96255945	ataxia telangiectasia is a genetic disorder with an autosomic recessive transmission	6.77
96154672	ataxia telangiectasia is an autosomal recessive disorder involving cerebellar degeneration, immunodeficiency radiation sensitivity, and cancer predisposition	6.69
92072632	patients with ataxia telangiectasia and cells derived from homozygotes and heterozygotes are unusually sensitive to ionizing radiation	6.62

Shatkay, Edwards, Wilbur, Boguski, 2000

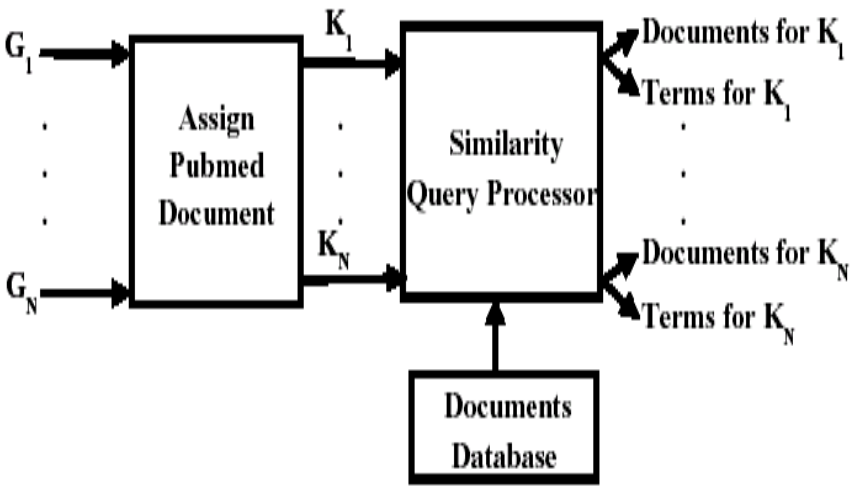
- *Genes, Themes and Microarrays: Using Information Retrieval for Large-Scale Gene Analysis*
- Used to annotate genes
  - Set of kernel documents is selected for the genes of interest
  - Kernel documents used to select other similar documents (similarity queries)
- Extracts keyphrases (unigrams, bigrams)
  - Model based on assumption of a Bernoulli generation of documents
- Finds functional relationships among genes
  - Relationships among genes based on

Finding Documents and Terms

- Document: $<d_1, d_2, \ldots, d_M>$ for *M* possible terms t_i, 1 if term appears in document and 0 otherwise
- For *theme* T (documents associated with gene), presence of terms in document *d* based on one of three Bernoulli distributions:
  - p_i^T – $Pr(term_i \in d \mid d \in T)$
  - q_i^T – $Pr(term_i \in d \mid d \notin T)$
  - DB_i – $Pr(term_i \in d \mid d \in DB)$
    - probabilities estimated from the entire collection
- Other key parameters
  - P_d – a priori probability document is in theme (set to 0.01)
  - λ_i – probability that DB_i is used to generate t_i

Similarity Queries

- Parameters DB_i and P_d are set initially, and p_i^T, q_i^T, λ_i found with EM:
  - Parameters are initialized using the kernel document and its comparison to the rest of the dataset
  - E step – determine likelihood for each document to be part of theme based on current settings
  - M step – find new model parameters to maximize likelihood of partitioning into on-theme/off-theme documents
- Top documents are selected based on their likelihood
- Best keyphrases have high values of the ratio p_i^T/q_i^T

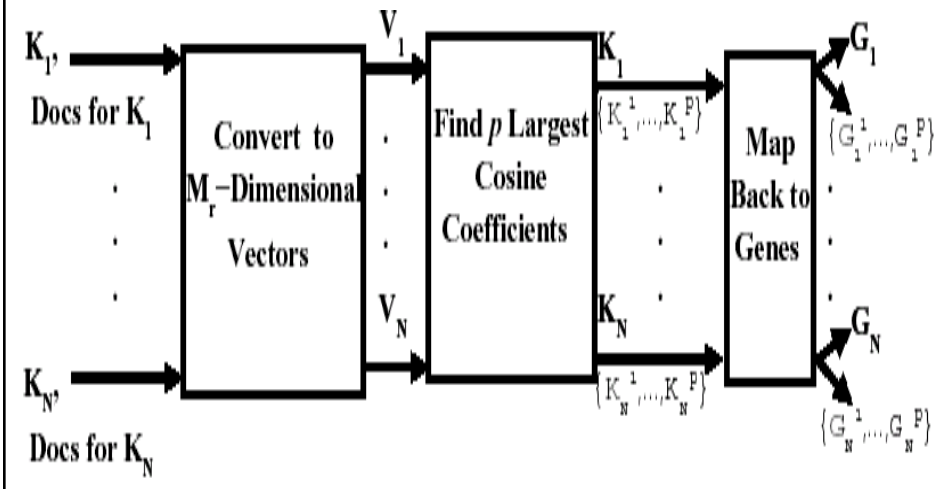


Finding Relationships Among Genes

- Top 50 documents selected for each gene
- PubMed identifiers R are found for each document
- Identifiers that occur for only one gene are dropped from R
- For each gene *g* construct vector V_g, where entries are 0 when identifier is not used in *g* and *1/#identifiers for g* if identifier is used for *g*:

  $<v_1^g, v_2^g, \ldots, v_{|R|}^g>$
- Calculate cosine coefficient between each pair of genes:

  $\cos(g1,g2) = V_{g1} \bullet V_{g2}$



Our Current Research

- Used to annotate clusters of genes
  - What keyphrases characterize all (or some) of the members of the cluster
  - Meant to help annotate results of large scale mechanisms such as microarrays
  - Builds on previous work by Andy Pohl and Mark Craven
- Extracts keyphrases (unigrams, bigrams, trigrams)
  - Currently testing for yeast genes – documents are abstracts collected from Medline
  - Keyphrases extracted using standard text processing methods
- Simply looks for good (characteristic) keyphrases
  - Initial results are promising

Collecting the Documents

- Abstracts are collected using queries on Medline
- For each gene, query looks for
  - (1) the occurrence of the locus name of that gene (or any other accepted aliases), and
  - (2) the occurrence of the phrase "saccharomyces cerevisiae"
  - If no abstracts are found, the second condition is dropped
- For a list of 6,267 possible yeast genes, 15,885 abstracts are found for 3,193 of the genes
  - Median number of abstracts found for the 3,193 = 3
  - Average number of abstracts = 5.029
  - Highest number of abstracts is 59 (LYS1) followed by SPO11 and SLY2 at 39

Selecting the Keyphrases

- Tokens may contain letters, digits and some internal punctuation ,()'- (some unique to this domain)
  - Note that numbers may be tokens
  - Hyphenated words at the end of sentences are joined
- Stemming is done using the *Porter* stemmer
  - List of stemming rules (e.g., ends in "ational", change to ends in "ate")
  - We have tried three approaches to stemming: no stemming, stemming everything, and "dictionary" stemming (our current approach)
    - "dictionary" stemming – only stem words in /usr/dict/words (leave biological terms alone)

Selecting the Keyphrases (cont)

- Stop words (standard list) are eliminated
  - Stop words: a, about, above, across, after, afterwards, etc.
- Bigrams and trigrams are kept only when there is no punctuation or stop words separating the tokens
- Keyphrases that only occur once are eliminated
- Keyphrase stats:
  - 120,808 possible keyphrases (most frequent, "gene" occurs 30,691 times in 10,407 abstracts for 2,833 of the genes)
  - Total of 2,344,183 occurrences of keyphrases
  - Median frequency is 3, average frequency is 19.404
  - 32,259 unigrams, ave. freq 53.462, 67,462 bigrams, ave. freq. 7.787, 21,087 trigrams, ave. freq. 4.469
  - 44,417 keyphrases occur 2 times, 20,197 occur 3 times, etc.

Ordering the Keyphrases

- General approach: compare a *count* for those genes in the cluster to *count* for genes outside cluster
- Counts previously investigated:
  - Frequency: how often the keyphrase occurs
  - Abstracts: how many abstracts the keyphrase occurs in
  - Genes: how many genes the keyphrase occurs for
  - Frequency and Abstracts tend to favor those genes with large numbers of abstracts
  - Genes counts would equate a term occurring once for five abstracts with a term that occurs 17 times for 20 abstracts
    - Empirically, lots of low frequency terms occur for one or two genes and end up scoring high
- Note, unigrams of gene names in cluster eliminated

Normalizing Abstract Counts

- Counting abstracts would work if there were the same number of abstracts per gene
  - Combines aspects of Abstracts and Genes counts
- Idea: estimate the percentage occurrence of abstracts per gene and project this number to some number to a count (of say 20)
- Key consideration: projecting percentage for keyphrase that occurs once for a gene's one abstract fairly when comparing it to keyphrase that occurs 18 times for another gene's 23 abstracts?
  - Idea: use M estimate to normalize percentages based on small numbers towards the population percentage

Counting Normalized Abstracts

if ($\#abstracts_{gene} < A$)

    $actual\_adjusted = \#abstracts_{k,g}$

else

    $actual\_adjusted = \#abstracts_{k,g} * (A / \#abstracts_{gene})$

fi

$$m\_estimate = \frac{\#abstracts_{k,g} + M * (\#abstracts_{keyphrase} / \#abstracts)}{\#abstracts_{gene} + M}$$

$estimated\_abstracts_{k,g} = \max(actual\_adjusted, m\_estimate)$

M is 10, A is 20

Computing Significance Statistics

- Can use t-tests, compare mean occurrence in and out of cluster
- Also Chi Square (which we prefer)

Count of keyphrase k in cluster	Count of keyphrase k outside of cluster
Count of other keyphrases in cluster	Count of all other keyphrases (neither keyphrase k, nor in cluster)

A Cluster Query

- Cluster from Ahlquist lab:

1: TEF1 - YPR080W : 19 abs
2: SPP81 - YOR204W : 17 abs
3: DCP1 - YOL149W : 17 abs
4: MDM2 - YGL055W : 14 abs
5: PIP1 - YER032W : 13 abs
6: TYS1 - YGR185C : 9 abs
7: TEF2 - YBR118W : 9 abs
8: SDH2 - YLL041C : 8 abs
9: GCD10 - YNL062C : 7 abs
10: SGS1 - YMR190C : 7 abs
11: SKI1 - YGL173C : 6 abs
12: LHP1 - YDL051W : 6 abs
13: JIP1 - YNL078W : 5 abs
14: MRT1 - YCR077C : 5 abs
15: CDC33 - YOL139C : 5 abs
16: LSM4 - YER112W : 4 abs
17: CCA1 - YER168C : 4 abs
18: TIF4632 - YGL049C : 3 abs
19: RNC1 - YKR056W : 3 abs
20: LOS1 - YKL205W : 3 abs
21: SIZ2 - YOR156C : 2 abs
22: PUS4 - YNL292W : 2 abs
23: ARC1 - YGL105W : 2 abs
24: SNU56 - YDR240C : 1 abs
25: TIF4631 - YGR162W : 1 abs
26: LSM7 - YNL147W : 1 abs
27: SPB8 - YJL124C : 1 abs
28: GCD14 - YJL125C : 1 abs
29: LSM6 - YDR378C : 0 abs
30: SMX4 - YLR438C-A : 0 abs
31: SNP3 - YBL026W : 0 abs
32: LSM5 - YER146W : 0 abs
33: YJL109C : 0 abs
34: YKL082C : 0 abs

Some Results

1. 561.1 **decapping** (17/4 : 10/9)
2. 503.1 **decapping enzym** (12/2 : 2/2)
3. 393.5 ***mrna decapping*** (9/3 : 2/2)
4. 360.6 **tyrosyl-trna** (8/2 : 1/1)
5. 309.4 ***tyrosyl-trna synthetase*** (7/2 : 1/1)
6. 299.3 **decapping activ** (8/1 : 1/1)
7. 298.8 pat1p (6/3 : 2/2)
8. 282.9 lsm protein (5/3 : 1/1)
9. 257.9 gcd14p (4/2 : 0/0)
10. 219.3 tef2 gene (6/2 : 1/1)
11. 213.1 lsm (5/3 : 3/3)
12. 208.8 pat1 gene (4/1 : 1/1)
13. 198.3 eif-4e (5/2 : 4/4)
14. 193.9 pl10 (5/1 : 0/0)
15. 184.3 intrins protein (7/1 : 4/1)
16. 178.2 ded1 gene (6/1 : 2/1)
17. 175.5 ef-1 alpha (7/2 : 5/3)
18. 174 exonucleolytic degrad
19. 170.7 **cap structur** (10/5 : 23/15)
20. 165.3 eif-4f (2/2 : 0/0)
21. 163 e3-like factor (2/1 : 0/0)
22. 163 e3-like (2/1 : 0/0)
23. 161.7 **membran intrins protein** (5/1 : 1/1)
24. 161.7 **plasma membran intrins** (5/1 : 1/1)
25. 160 **cap-binding** (7/5 : 18/15)
26. 159.6 ***aminoacylation*** (6/2 : 7/7)
27. 159.3 delta9 (5/1 : 1/1)
28. 158 **nucleotidyltransferase** (3/1 : 1/1)
29. 158 ***trna nucleotidyltransferase*** (3/1 : 1/1)
30. 151 **cap-binding protein** (5/4 : 10/10)

Advantages/Disadvantages of Our Work

- Advantages:
  - Very little preprocessing required
  - Easy to adapt to any universe of documents
  - Seems to do well (in initial tests)
  - (Working on) providing significance values
- Disadvantages:
  - Simple document retrieval may miss related documents
  - Infrequent terms abound, may score high and be shown by random chance
  - Results somewhat dependent on hand-set parameters (M,A)

Future Work

- Calculating significance values
  - Permutation tests to deal with FEW
- Building online query mechanism
- More effectively eliminating terms associated with gene names
- Annotating results
  - Showing genes used in decisions
  - Recognizing when results overlap (membran intrins protein and plasma membran intrins)
  - Combining results that cluster based on the same genes
- Incorporating other online data
  - Protein to protein mapping information

Conclusions

- As the set of online documents pertaining to biological objects grows, techniques for automatically annotating become critical
- One simple technique to annotate is to search for keyphrases
- Keyphrases are based on statistics concerning the distribution of the keyphrases for the objects of interest versus the entire population
- Techniques have been introduced to annotate protein families, annotate genes, compare gene similarity, and annotate gene clusters