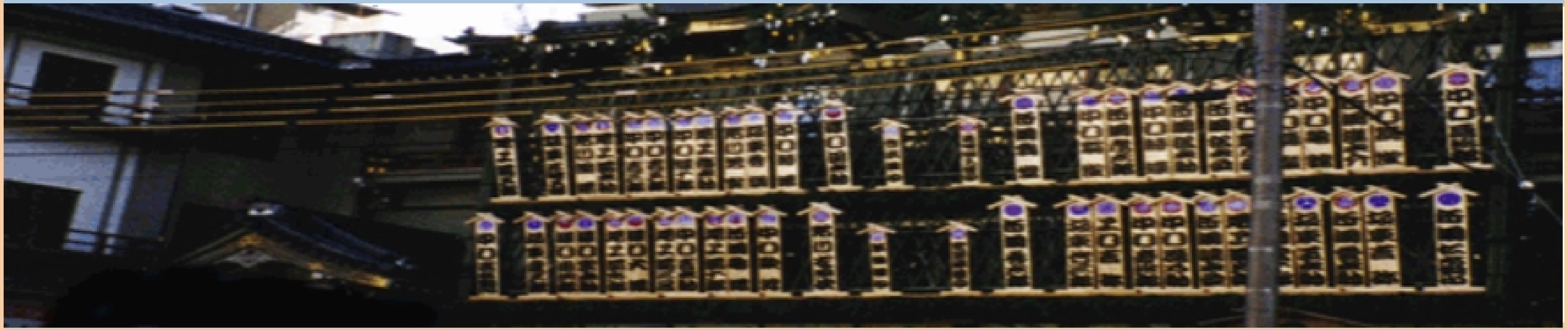
informatics
luis rocha 2007

bibliome informatics



text-mining for computational biology

luis m. rocha
Indiana university
school of informatics and cognitive science program
901 East Tenth Street, Bloomington IN 47408
and
Instituto Gulbenkian de Ciencia
Computational Biology
Oeiras, Portugal

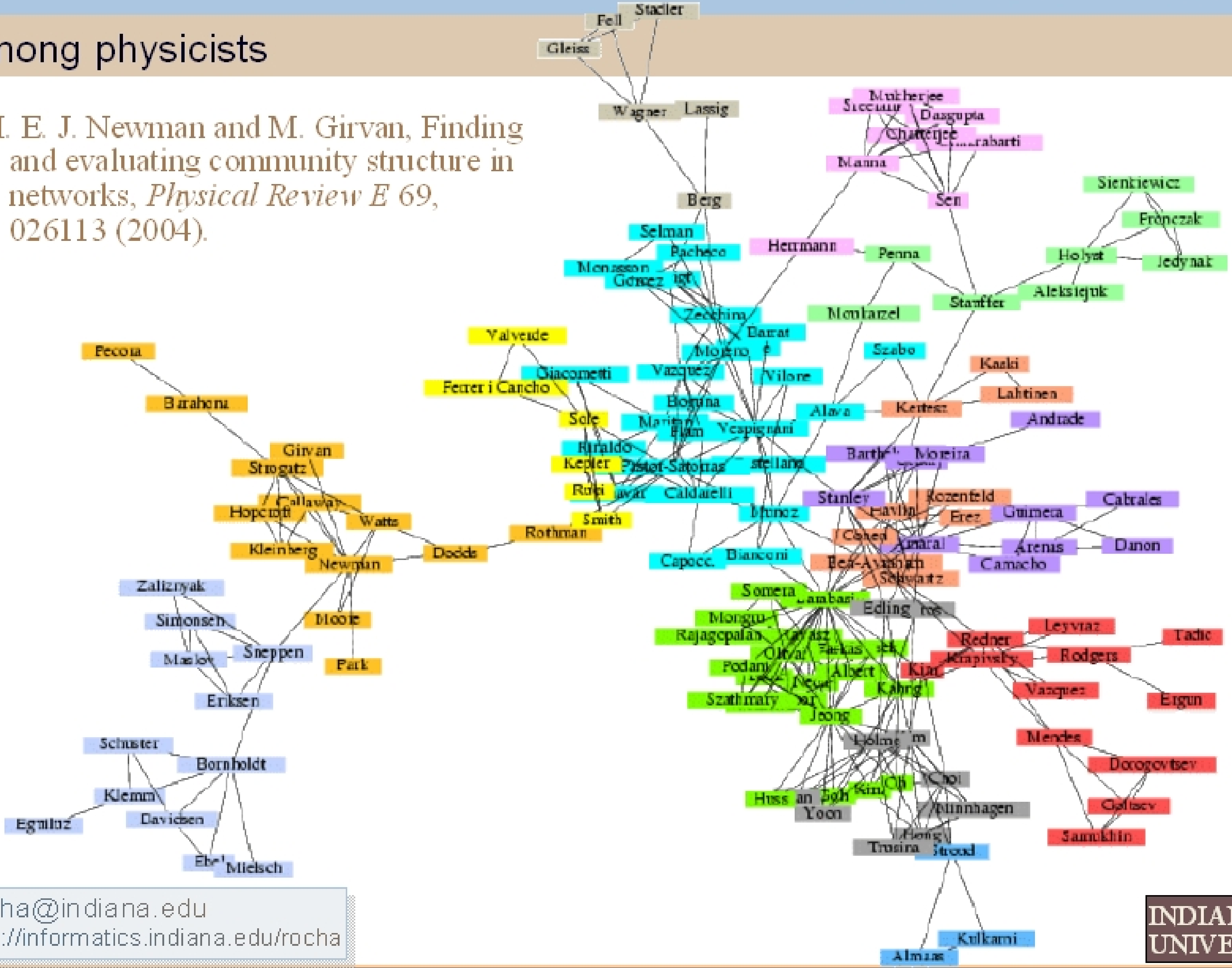
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Cognitive Science Program
the biocomplexity institute
INDIANA UNIVERSITY

collaboration in science

among physicists

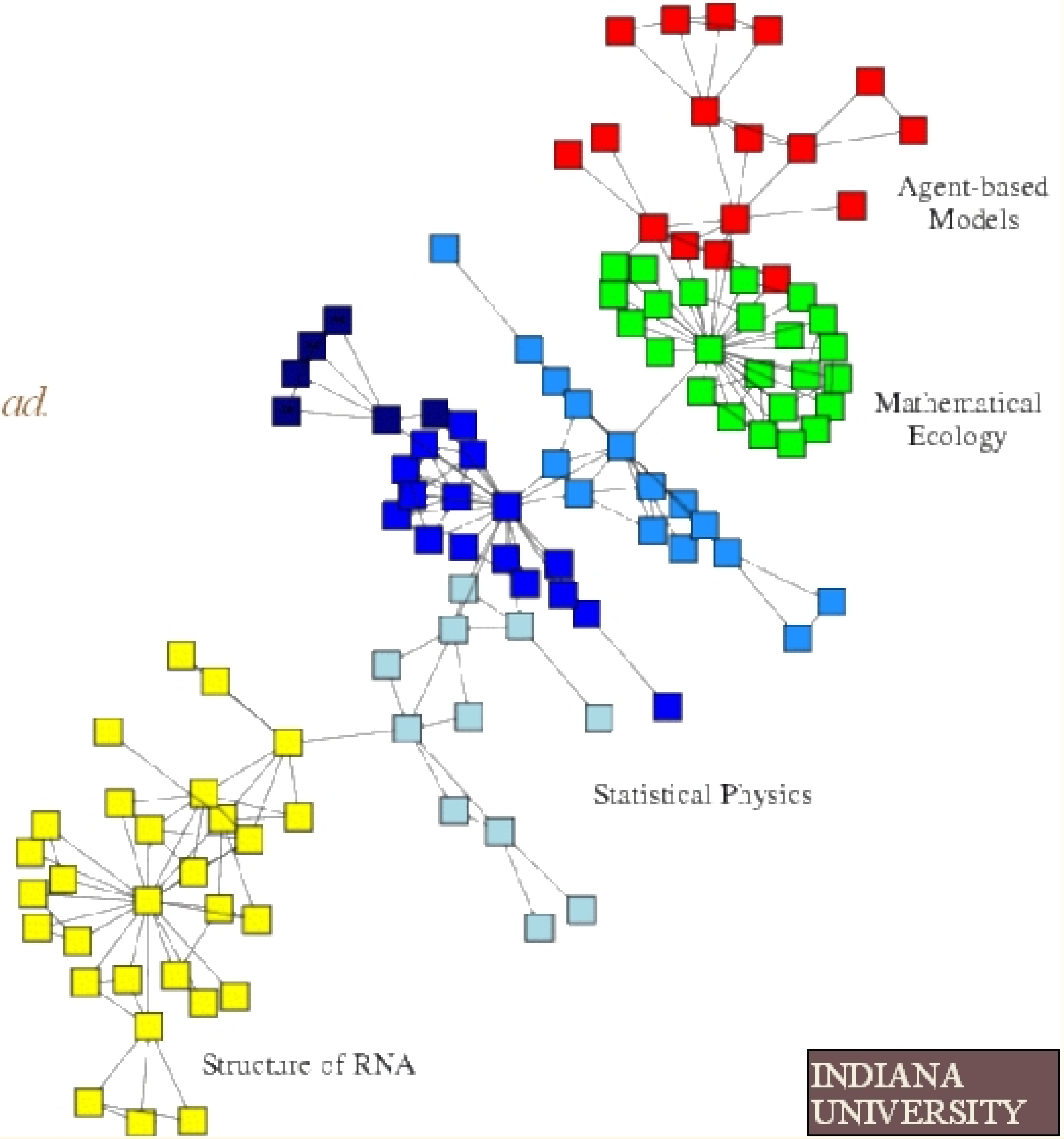
M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical Review E* 69, 026113 (2004).



collaboration in science and technology

typical pattern

M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", *Proc. Natl. Acad. Sci.* USA 99, 8271-8276 (2002).



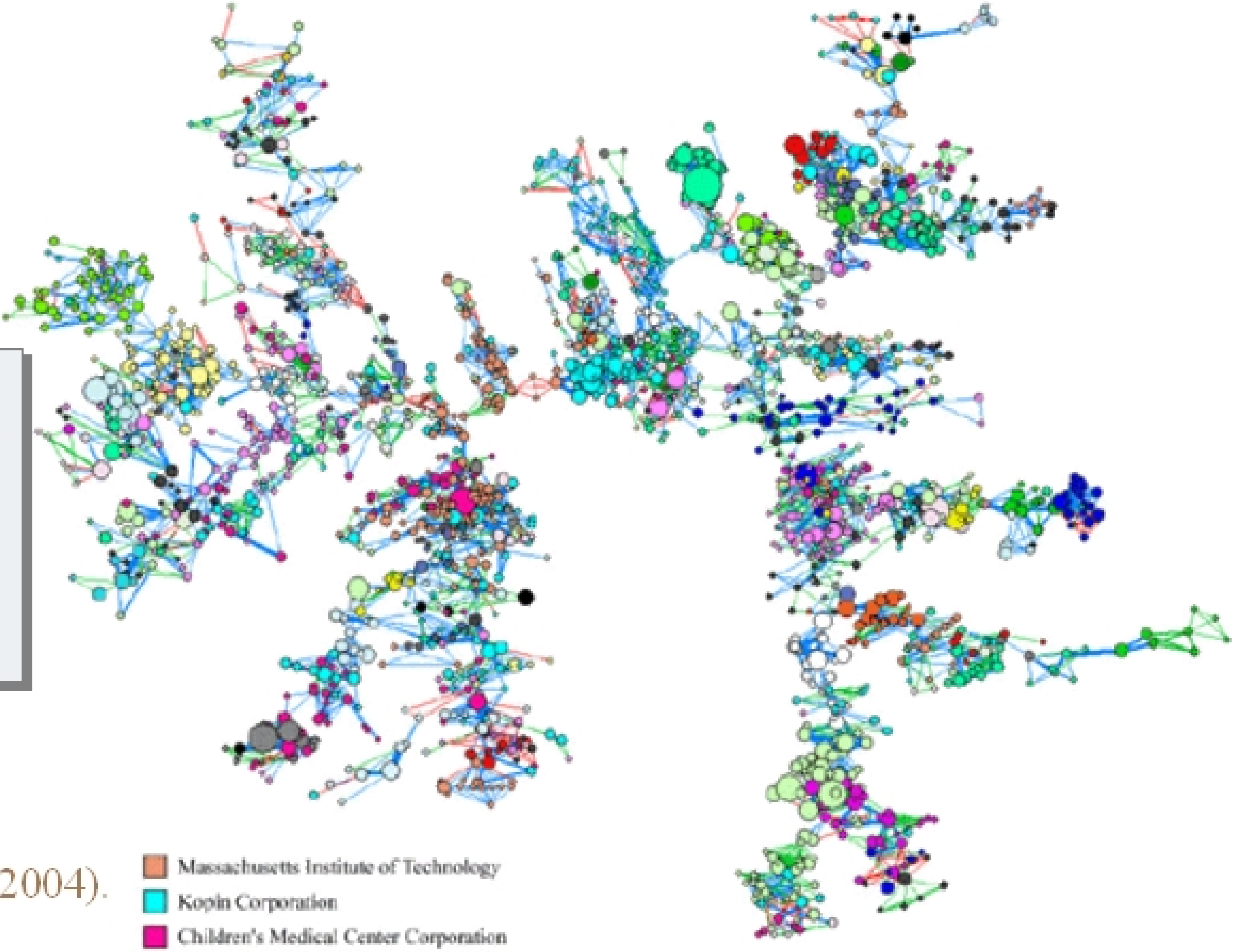
rocha@indiana.edu
http://informatics.indiana.edu/rocha

patent network

Boston (90's): the biotechnology story

Fleming, Lee. "Perfecting Cross-Pollination." *Harvard Business Review* 82, no. 9 (September 2004): 22-24.

- Subgraph
- 30 years of patents
  - Nodes are inventors
  - 3 milion patents
  - 2 milion inventors


🟧 Massachusetts Institute of Technology
🟦 Kopin Corporation
🟪 Children's Medical Center Corporation

Fleming, Lee, and Adam Juda. "A Network of Invention." *Harvard Business Review* 82, no. 4 (April 2004).

patent network

a few facts

- Highest impact inventions (economic measures)
  - Distant connections are more important
  - Negative influence of regional clusters
    - Except in very diverse clusters
- At the end of the 90's, half of the inventors are connected via some path in the network
  - Knowledge keeps flowing via such paths, years after the connection origin
  - Inventors network is a "small-world"
  - *Know-how* depends highly on ***know-whom***
    - Companies seek people with expertise and capacity for collaboration

Fleming, Lee. "Perfecting Cross-Pollination." *Harvard Business Review* 82, no. 9 (September 2004): 22-24.

biotecnology growth patterns

Collaboration in the life sciences

- R&D in biotecnology requires collaboration among diverse types of organozations
  - Walter Powell, Jason Owen-Smith, Douglas White, Kenneth Kopout
    - “Interorganizational collaboration and the locus of innovation: networks of learning in Biotechnology”. *Administrative Science Quarterly* 41(1):116-45.
    - “Practicing polygamy with good taste: the evolution of interorganizational collaboration in the Life sciences”.
    - “A comparison of U.S. and European University-Industry relations in the Life Sciences”
- Studied the biotech network evolution
  - Collaboration is the norm in the US
  - In Europe very little cross-city and even less cross-national collaboration

FLAD Computational Biology Colaboratorium

strengthen international and inter-organization collaboration and re-integration

FLAD
Computational Biology Collaboratorium

- Open organization to enhance productive collaboration among national and international organizations
  - A central designed to enable a network of collaboration
  - Dovetailing with Phd on Computational Biology
  - Objectives:
    - Facilities for hosting scientists and research
    - Add value to the visiting professor schedule of the PhD Program
    - Increase the value of the program and its visitors to the Portuguese network
    - Develop and host relevant informatics technology
    - Attract high-quality students for program, high-quality supervisors for them, and facilitate integration into portuguese scientific community

http://bc.igc.gulbenkian.pt /collaboratorium
/PDBC

Collaboration host

FLAD

Computational Biology Collaboratorium

- short-term research partnerships to tackle a specific pro ject
  - Papers, grants, projects
- Host workgroup meetings towards long-term projects
- Proposals
  - A scientific advisory committee evaluates collaboration grant proposals
    - Short-term proposals can be submitted at any time and evaluated within 2 months
    - Proposals for short-term courses or participation in program modules
    - Workgroup proposals: hosting support proposals can be submitted at any time

http://bc.igc.gulbenkian.pt

uncovering global patterns of functional behavior in biology

via knowledge integration

- Microarray (gene expression) analysis discovers patterns of expression behavior in groups of genes:
  - numerical expression values without functional or semantic characterization
- The biological reasons of gene groupings must be ascertained by biologists
  - Need to be able to integrate knowledge about a large number of possible underlying biological mechanisms for a large number of genes in microarrays
- Uncover “implicit” gene-gene, protein-protein, TF-gene relations
- Methods
  - Integration of available sources of functional knowledge
    - databases with biomedical publications and data
  - Validation
    - Relevant associations

Aims to assist biologists with automated annotation
reducing the number and proposing new functional explanations

acknowledgements

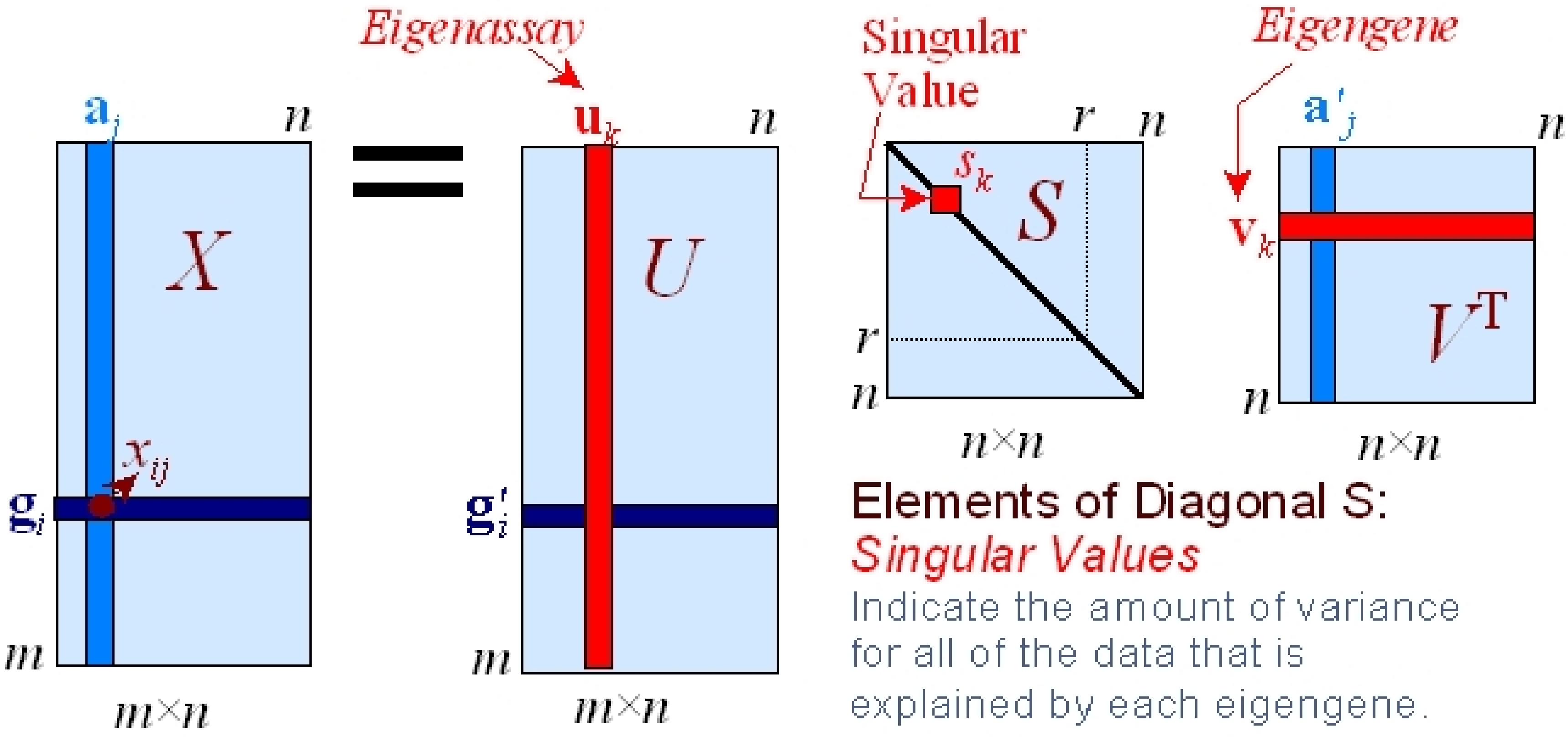
present work

- **Main collaborators**
  - Andreas Retchsteiner (IU)
  - Ana Maguitman (Bahia Blanca)
  - Alaa Abi Haidar (IU)
  - Jasleen Kaur (IU)
  - Predrag Radivojac (IU)
  - Zhiping Wang (IU)
- **Other Researchers Involved**
  - Tiago Simas (IU)
  - Karin Verspoor (LANL)
  - Jean Challacombe (LANL)
  - Charlie Strauss (LANL)
  - Michael Wall (LANL)

http://casci.informatics.indiana.edu

singular value decomposition

for microarray analysis



Rows of V$^{\mathrm{T}}$: *eigengenes* (colums are time steps)

Each gene's expression pattern is a linear combination of the eigengene patterns.

$$\mathbf{g}_i = \sum_{k=1}^{r} u_{ik} s_k \mathbf{v}_k, \quad i{:}1,\ldots,m$$

Elements of Diagonal S: *Singular Values*

Indicate the amount of variance for all of the data that is explained by each eigengene.

$$X = USV^{\mathrm{T}}$$

Gene Expression Matrix: Columns are assays (time steps) and rows are genes

Columns of U: *eigenassays* (rows are genes) **describe how each component contributes to a single gene's expresssion pattern**

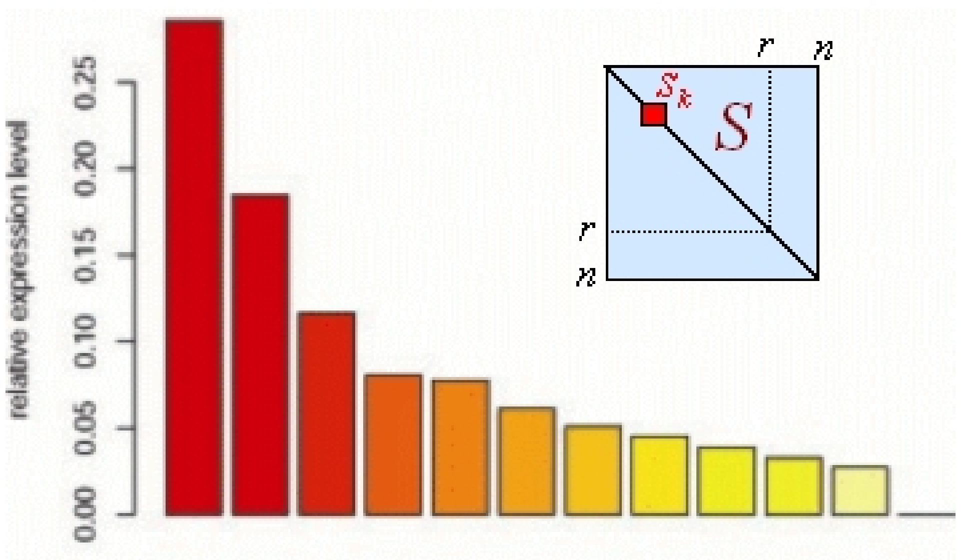
$$\mathbf{a}_i = \sum_{k=1}^{r} v_{jk} s_k \mathbf{u}_k, \quad j{:}1,\ldots,n$$

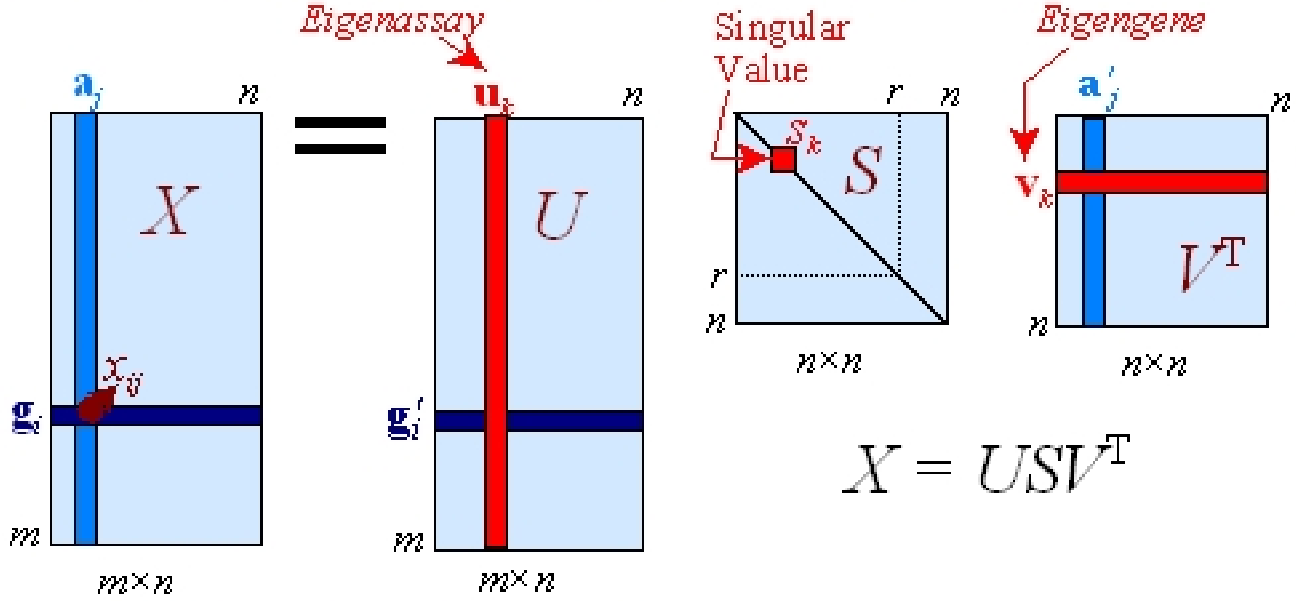
Wall, Rechtsteiner and Rocha [2002]. "Singular value decomposition and principal component analysis". In *Understanding and Using Microarray Analysis Techniques: A Practical Guide*. D.P. Berrar, W. Dubitzky, M. Granzow, eds.

informatics
luis rocha 2007

singular value decomposition

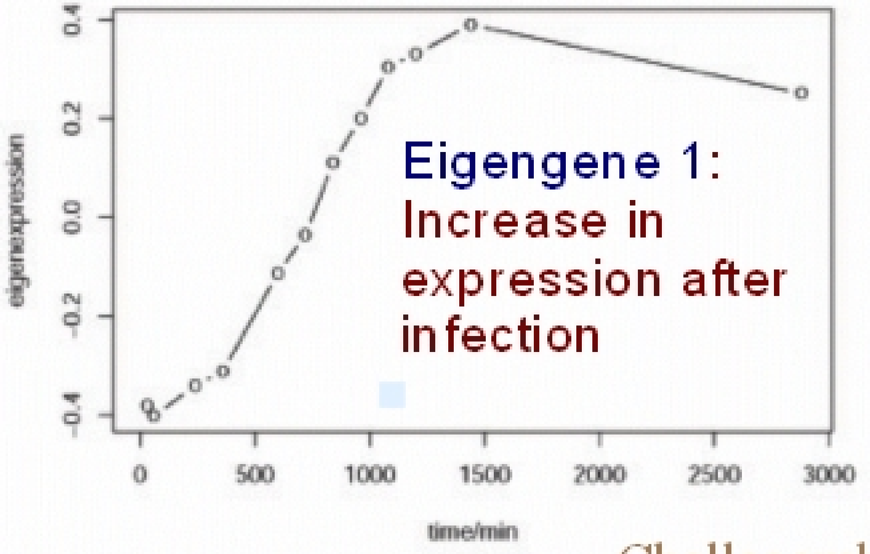
gene expression (13000 genes) after infection with herpes virus



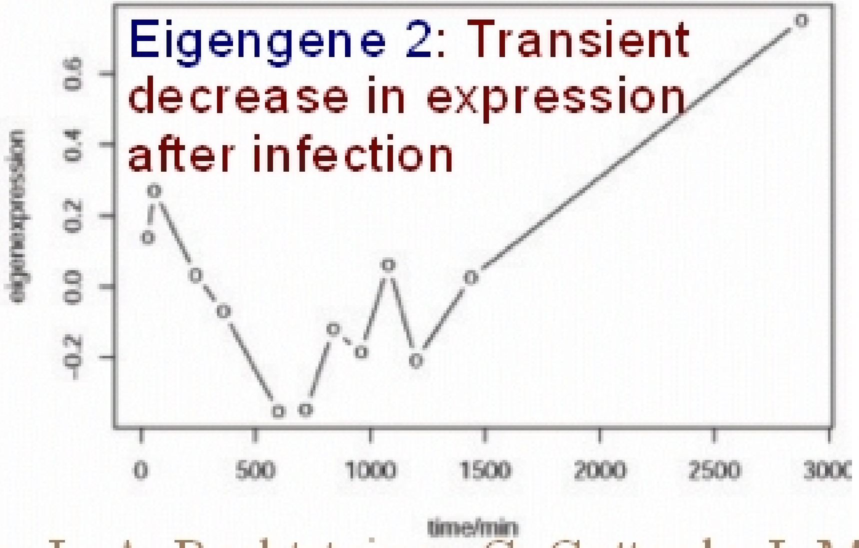


$$X = USV^T$$

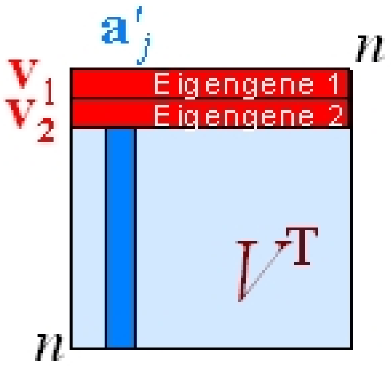
12 point time series (30min - 48hrs)



Eigengene 1: Increase in expression after infection



Eigengene 2: Transient decrease in expression after infection



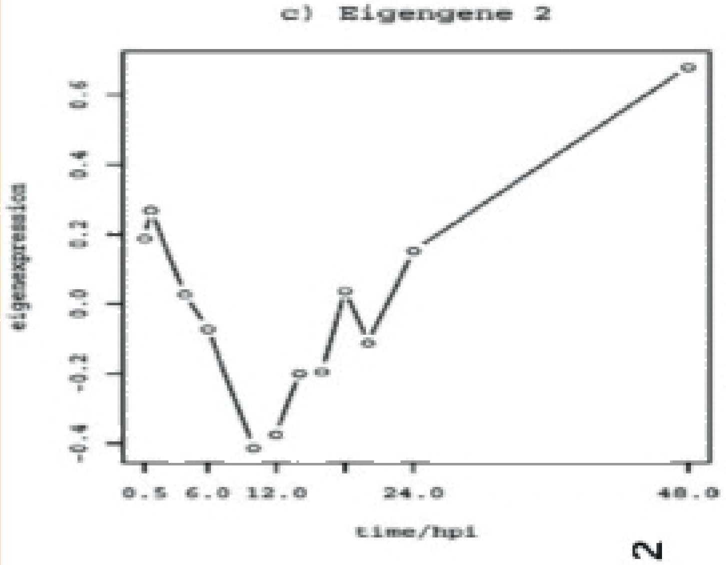
Challacombe, J., A. Rechtsteiner, G. Gottardo, L.M. Rocha, E.P. Brown, T. Shenk, M. Altherr, T. Brettin [2004]. "Evaluation of the host transcriptional response to human cytomegalovirus infection". *Physiol. Genomics*. **10**.1152

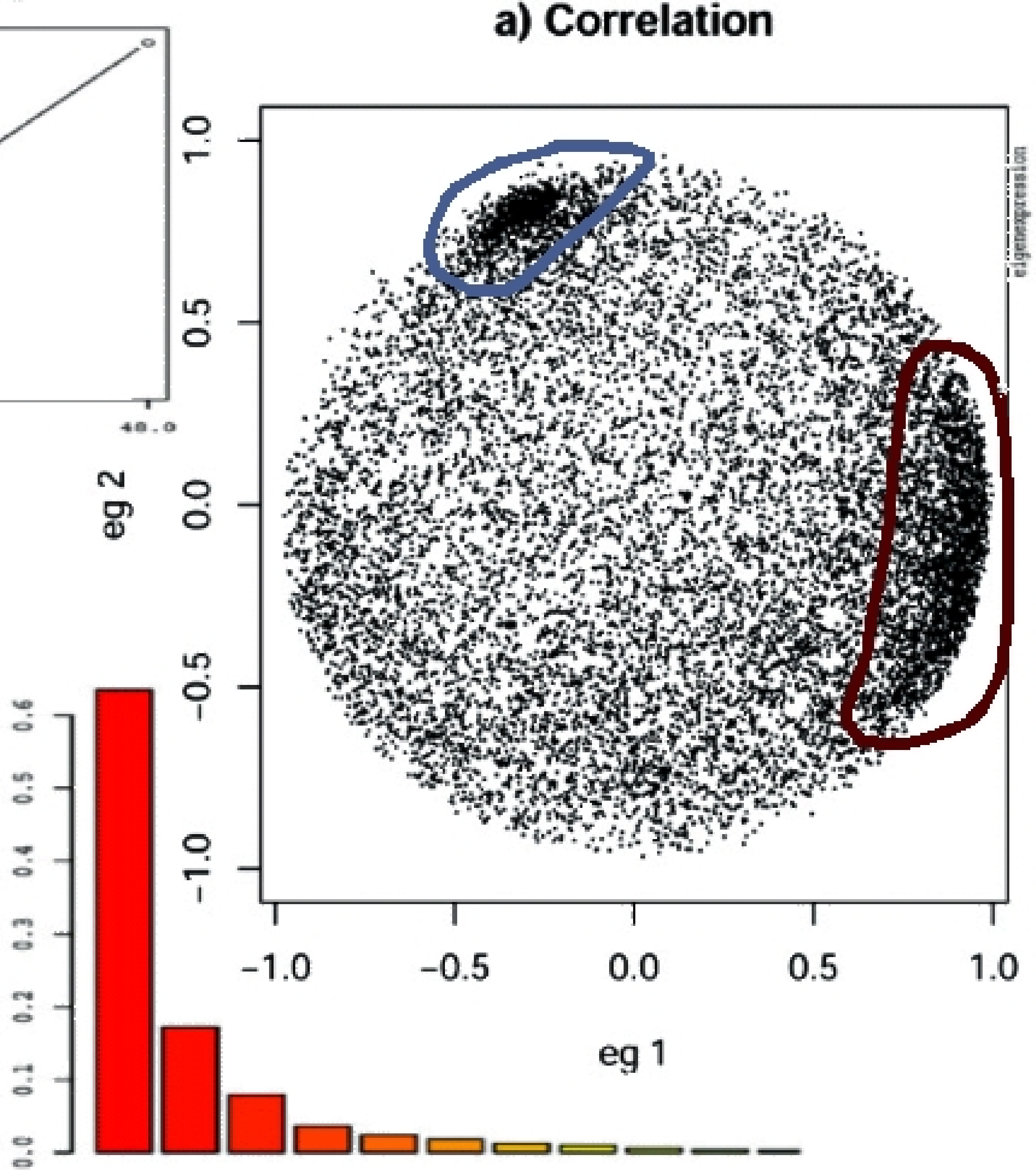
rocha@indiana.edu
http://informatics.indiana.edu/rocha

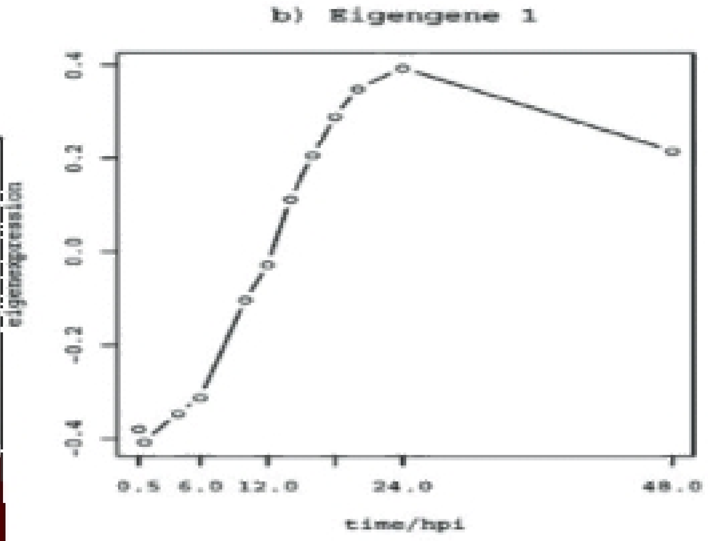
INDIANA UNIVERSITY

biological discovery via SVD

eigenassay coefficient plot: human cytomegalovirus infection







Cluster 2:
Genes involved in immune system regulation, signal transduction and cell adhesion. Also mainly in cluster 2, genes targeted by HCMV's immune evasion strategies.

Cluster 1:
Genes involved in ***transcriptional regulation***, oncogenesis and cell cycle regulation. Also mainly in cluster 1, genes involved in the host response to HCMV infection.

Challacombe et al. *Physiol. Genomics*. **10**.1152. 2004.

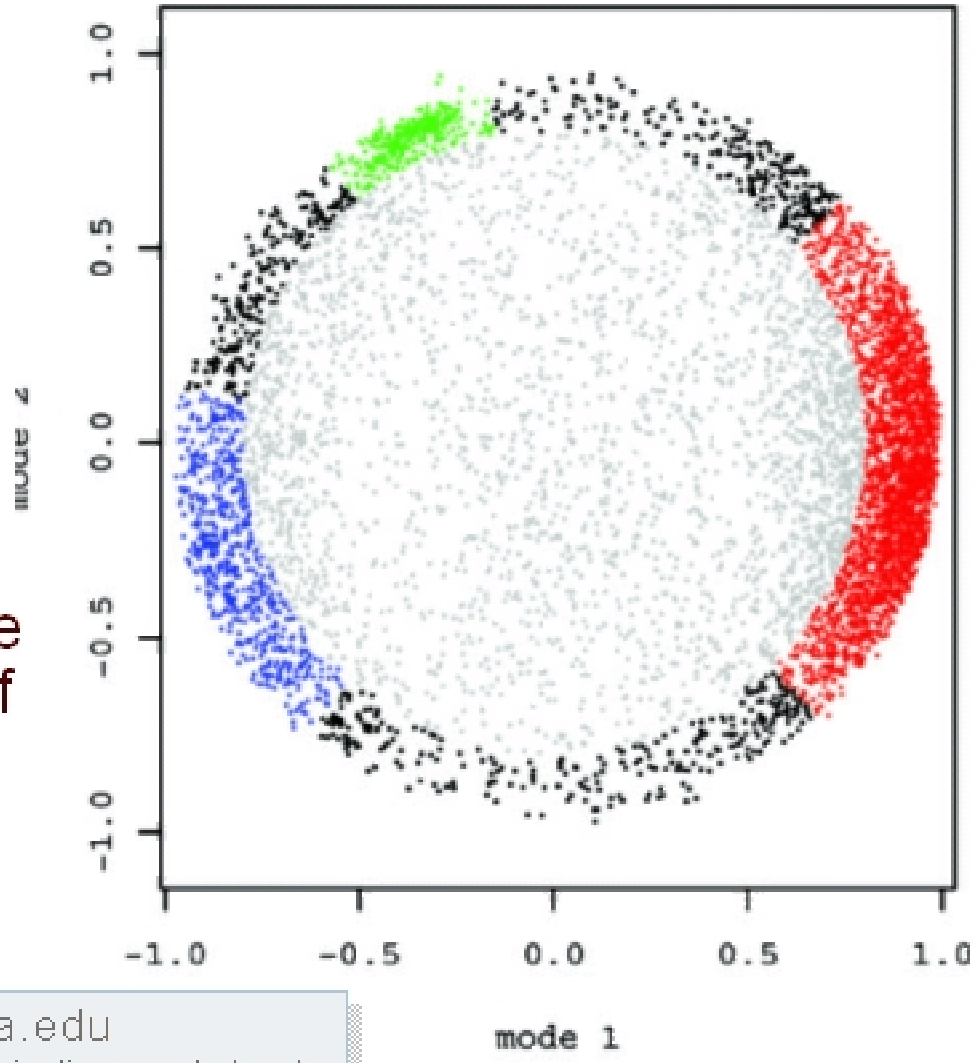
density estimation of polar angles

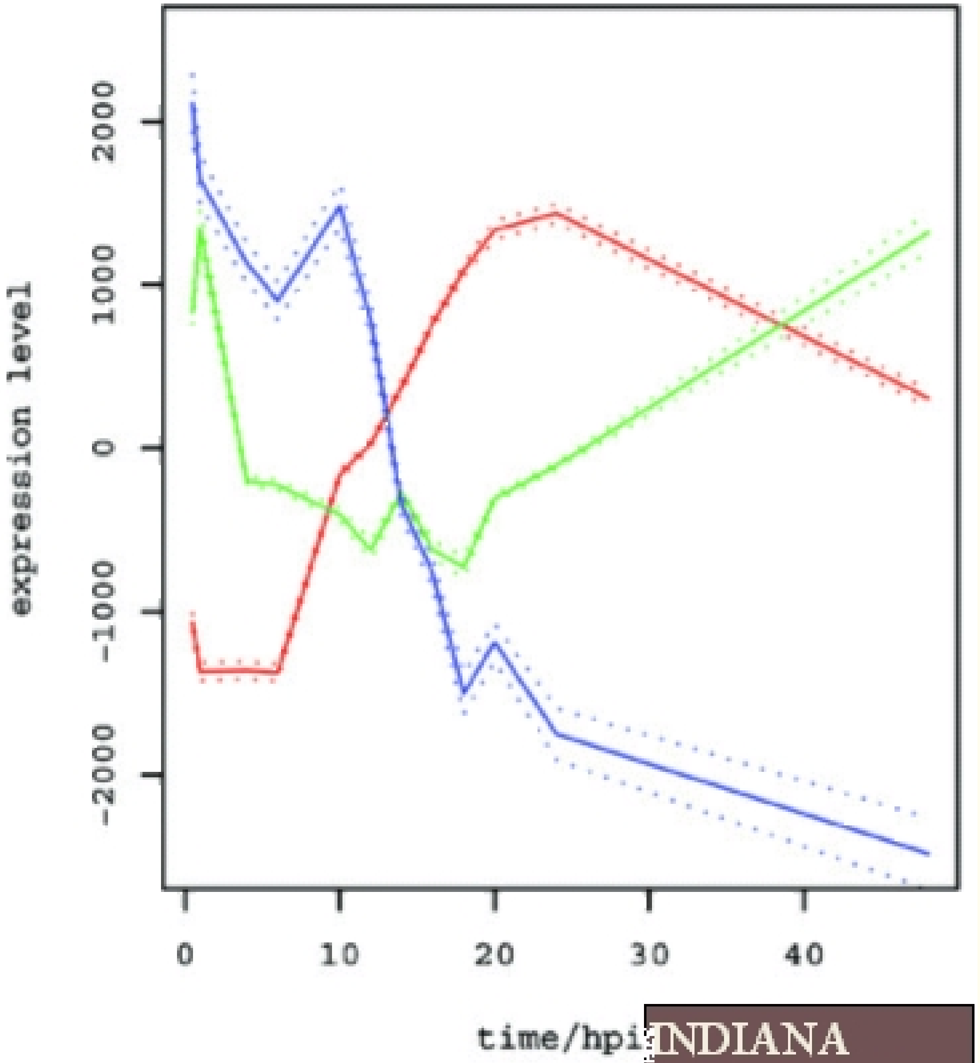
in SVD subspace (after serial correlation filtering)

- Boundary in space
  - largest rate of change of polar angle density from uniform
- Choose regions of higher density
  - By density of polar angles

Rechtsteiner, A. and L.M. Rocha [2004]. "MeSH Key Terms for Validation and Annotation of Gene Expression Clusters". *RECOMB 2004*, pp. 212-213.

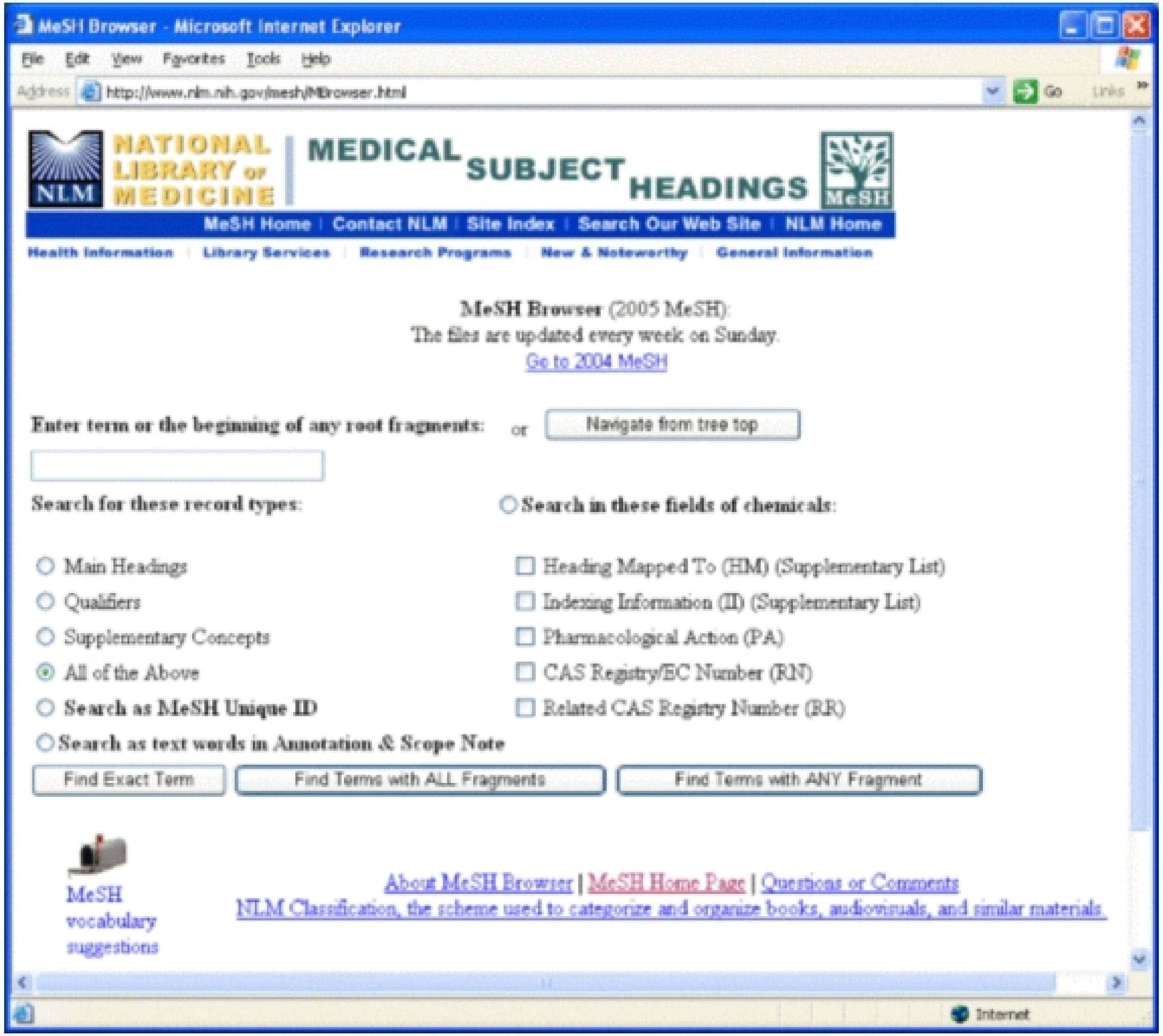
What is the Function of genes in clusters?





Medical Subject Headings

- Well designed, controlled, hierarchically organized vocabulary (22,568 descriptors).
- Used by the National Library of Medicine to index all publications in MEDLINE/PubMED
    - average of 10 headings per paper.
    - Updated continuously by its staff of 10.



MeSH Browser - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.nlm.nih.gov/mesh/MBrowser.html
Go
Links
NLM NATIONAL LIBRARY of MEDICINE
MEDICAL SUBJECT HEADINGS
MeSH
MeSH Home | Contact NLM | Site Index | Search Our Web Site | NLM Home
Health Information | Library Services | Research Programs | New & Noteworthy | General Information

MeSH Browser (2005 MeSH):
The files are updated every week on Sunday.
Go to 2004 MeSH

Enter term or the beginning of any root fragments: or Navigate from tree top

Search for these record types:

○ Main Headings
○ Qualifiers
○ Supplementary Concepts
● All of the Above
○ **Search as MeSH Unique ID**
○ **Search as text words in Annotation & Scope Note**

○ **Search in these fields of chemicals:**

□ Heading Mapped To (HM) (Supplementary List)
□ Indexing Information (II) (Supplementary List)
□ Pharmacological Action (PA)
□ CAS Registry/EC Number (RN)
□ Related CAS Registry Number (RR)

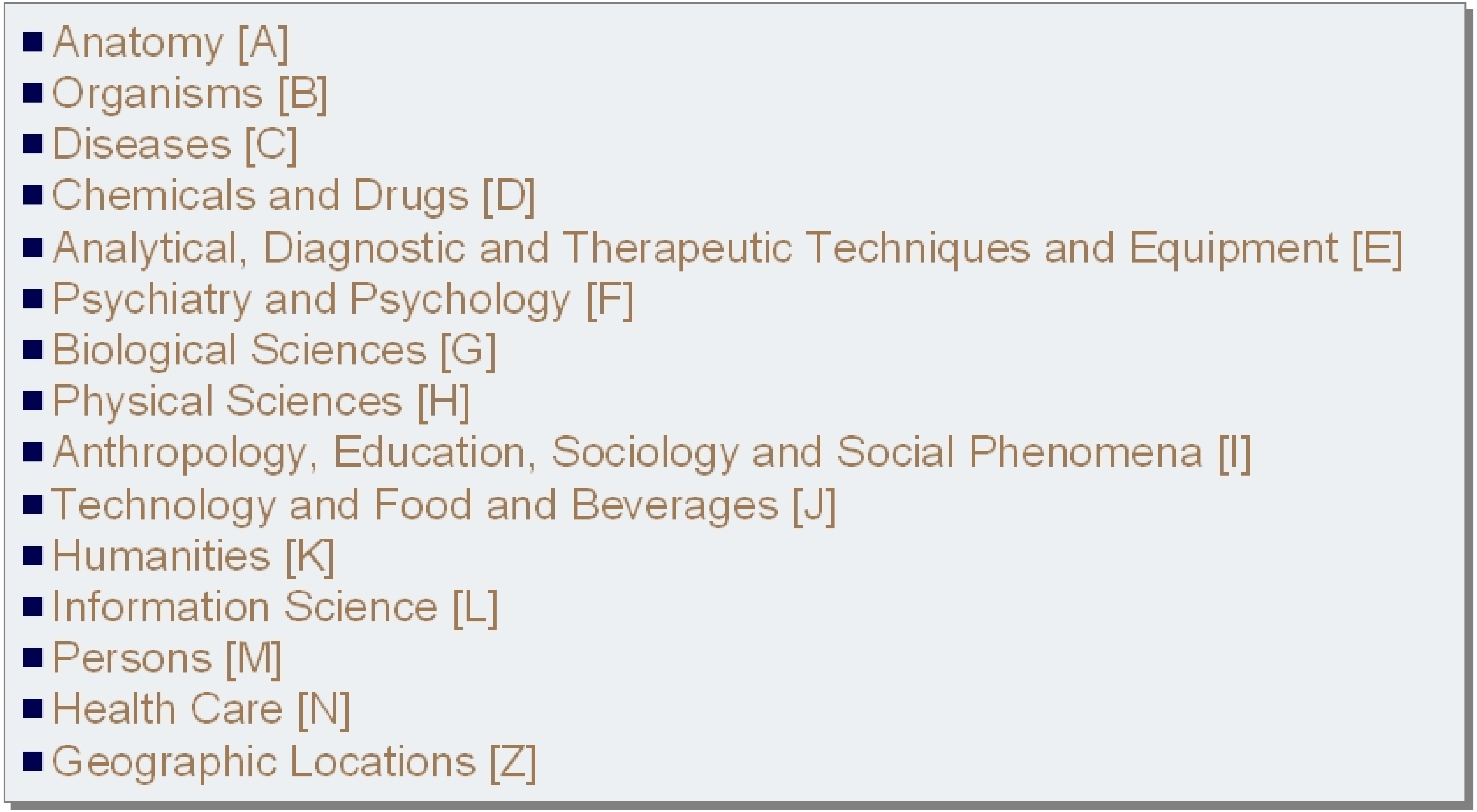
Find Exact Term | Find Terms with ALL Fragments | Find Terms with ANY Fragment

MeSH vocabulary suggestions

About MeSH Browser | MeSH Home Page | Questions or Comments
NLM Classification, the scheme used to categorize and organize books, audiovisuals, and similar materials.

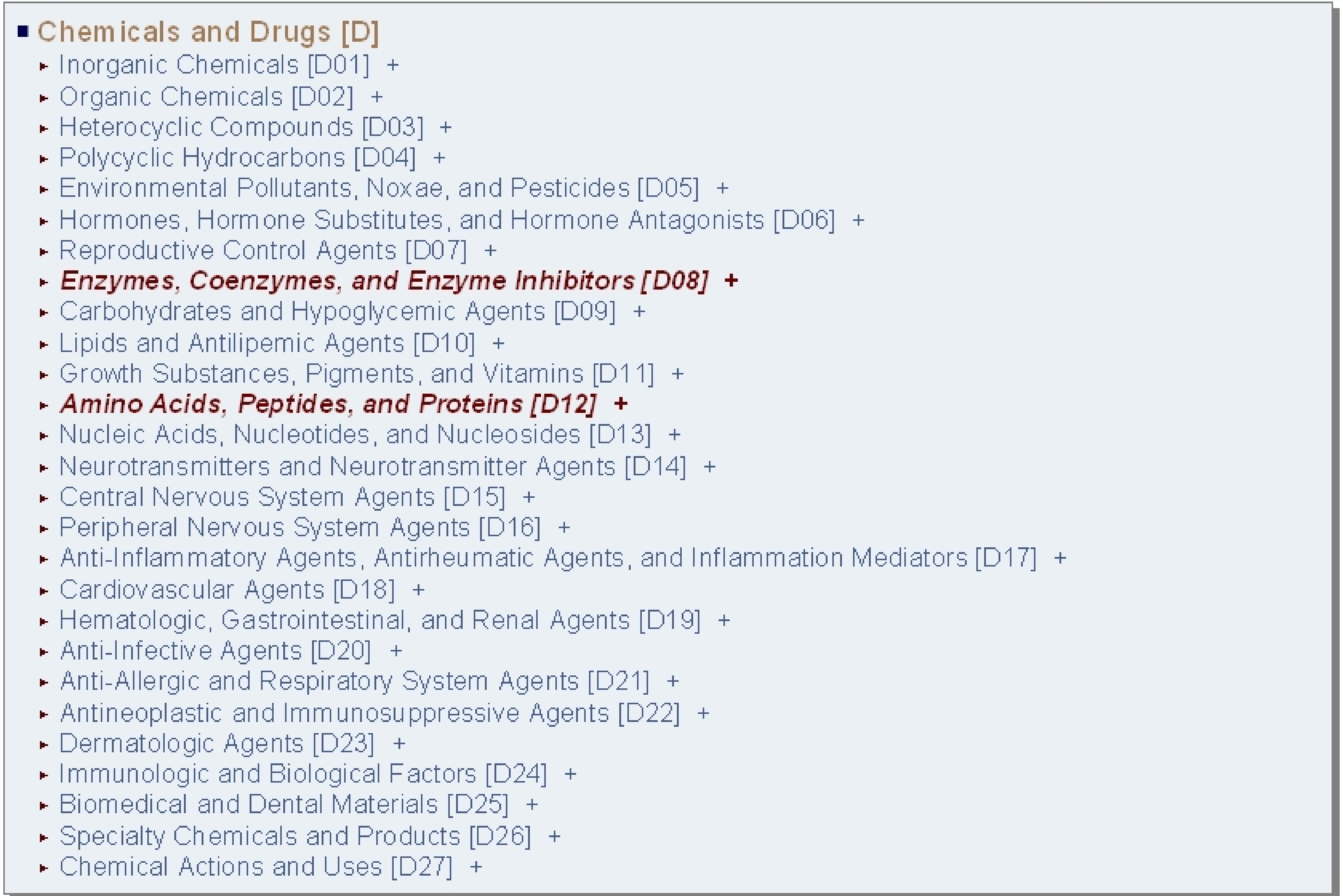
MeSH

Browse from Tree Top



- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
- Psychiatry and Psychology [F]
- Biological Sciences [G]
- Physical Sciences [H]
- Anthropology, Education, Sociology and Social Phenomena [I]
- Technology and Food and Beverages [J]
- Humanities [K]
- Information Science [L]
- Persons [M]
- Health Care [N]
- Geographic Locations [Z]

MeSH: Browse from Tree Top



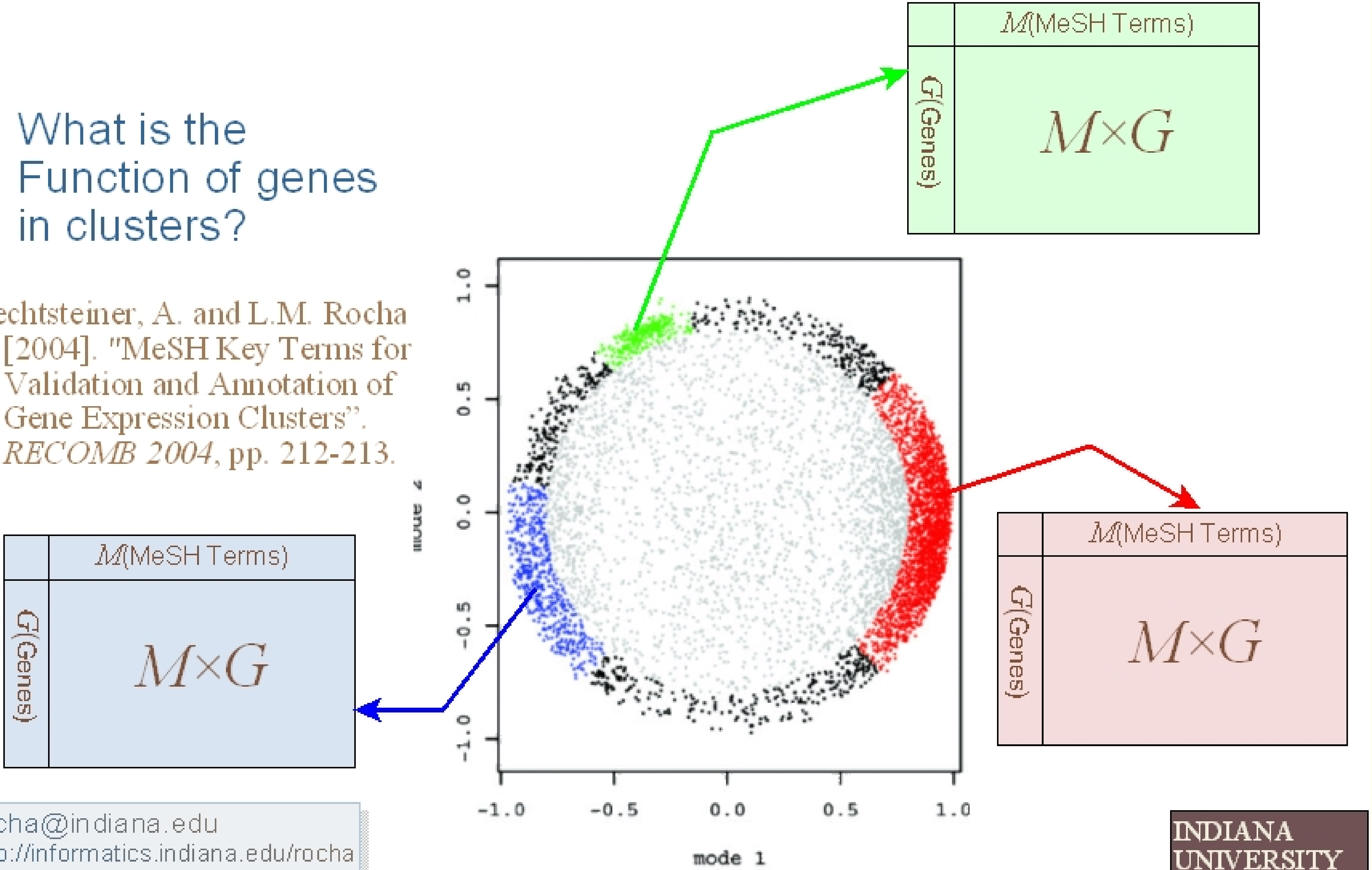
- Chemicals and Drugs [D]
  - Inorganic Chemicals [D01] +
  - Organic Chemicals [D02] +
  - Heterocyclic Compounds [D03] +
  - Polycyclic Hydrocarbons [D04] +
  - Environmental Pollutants, Noxae, and Pesticides [D05] +
  - Hormones, Hormone Substitutes, and Hormone Antagonists [D06] +
  - Reproductive Control Agents [D07] +
  - ***Enzymes, Coenzymes, and Enzyme Inhibitors [D08] +***
  - Carbohydrates and Hypoglycemic Agents [D09] +
  - Lipids and Antilipemic Agents [D10] +
  - Growth Substances, Pigments, and Vitamins [D11] +
  - ***Amino Acids, Peptides, and Proteins [D12] +***
  - Nucleic Acids, Nucleotides, and Nucleosides [D13] +
  - Neurotransmitters and Neurotransmitter Agents [D14] +
  - Central Nervous System Agents [D15] +
  - Peripheral Nervous System Agents [D16] +
  - Anti-Inflammatory Agents, Antirheumatic Agents, and Inflammation Mediators [D17] +
  - Cardiovascular Agents [D18] +
  - Hematologic, Gastrointestinal, and Renal Agents [D19] +
  - Anti-Infective Agents [D20] +
  - Anti-Allergic and Respiratory System Agents [D21] +
  - Antineoplastic and Immunosuppressive Agents [D22] +
  - Dermatologic Agents [D23] +
  - Immunologic and Biological Factors [D24] +
  - Biomedical and Dental Materials [D25] +
  - Specialty Chemicals and Products [D26] +
  - Chemical Actions and Uses [D27] +

automatic functional gene annotation

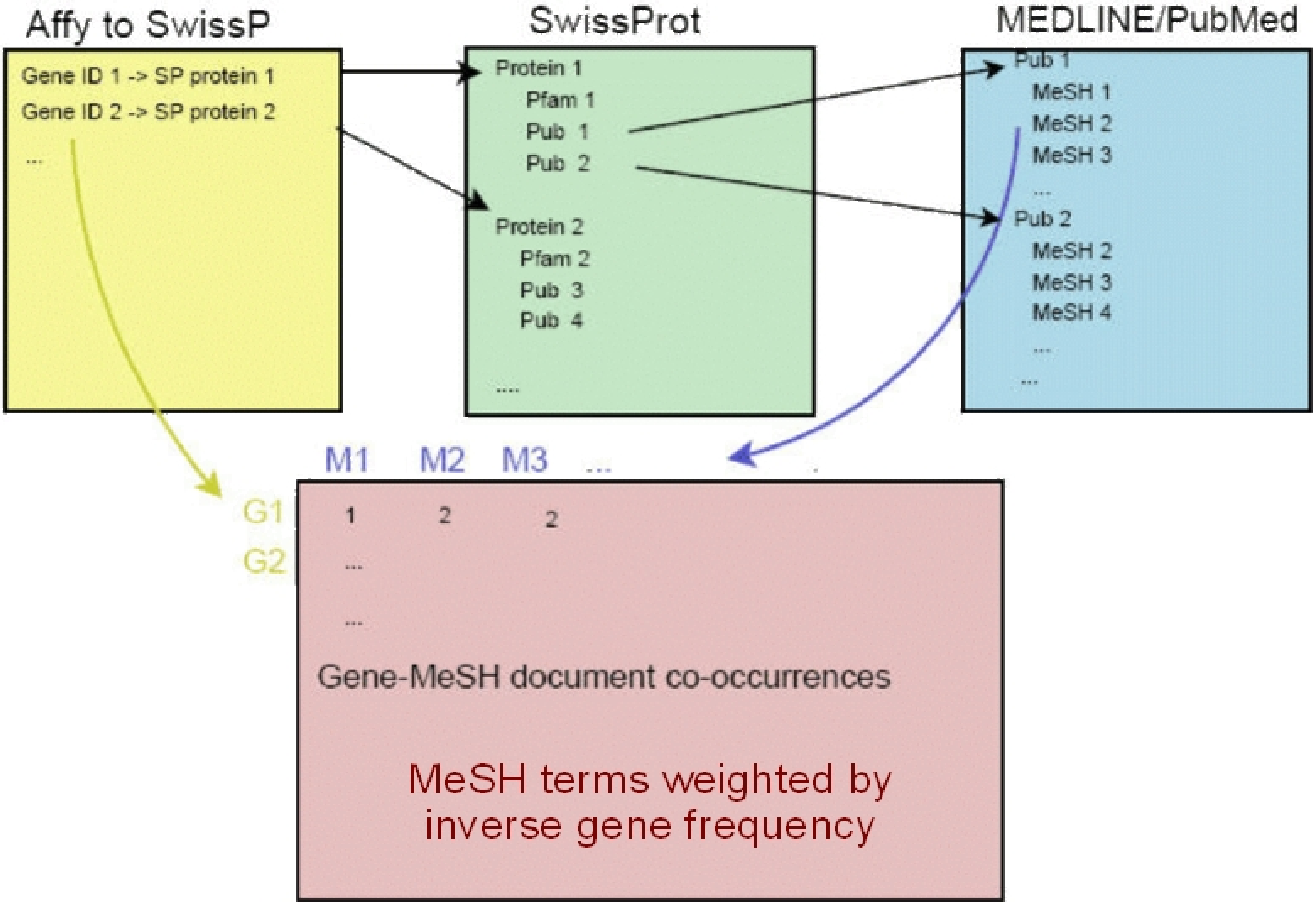
using the biomedical literature

What is the Function of genes in clusters?

Rechtsteiner, A. and L.M. Rocha [2004]. "MeSH Key Terms for Validation and Annotation of Gene Expression Clusters". *RECOMB 2004*, pp. 212-213.



gene to MeSH term linkage



Affy to SwissP

Gene ID 1 -> SP protein 1
Gene ID 2 -> SP protein 2
...

SwissProt

Protein 1
Pfam 1
Pub 1
Pub 2

Protein 2
Pfam 2
Pub 3
Pub 4

....

MEDLINE/PubMed

Pub 1
MeSH 1
MeSH 2
MeSH 3
...
Pub 2
MeSH 2
MeSH 3
MeSH 4
...
...

	M1	M2	M3	...
G1	1	2	2	
G2	...			
	...			

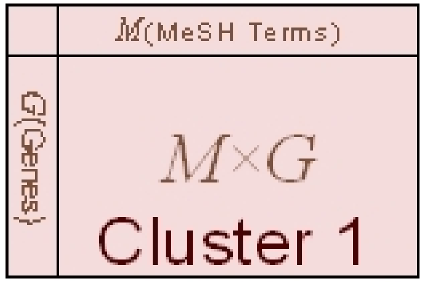
Gene-MeSH document co-occurrences

MeSH terms weighted by inverse gene frequency

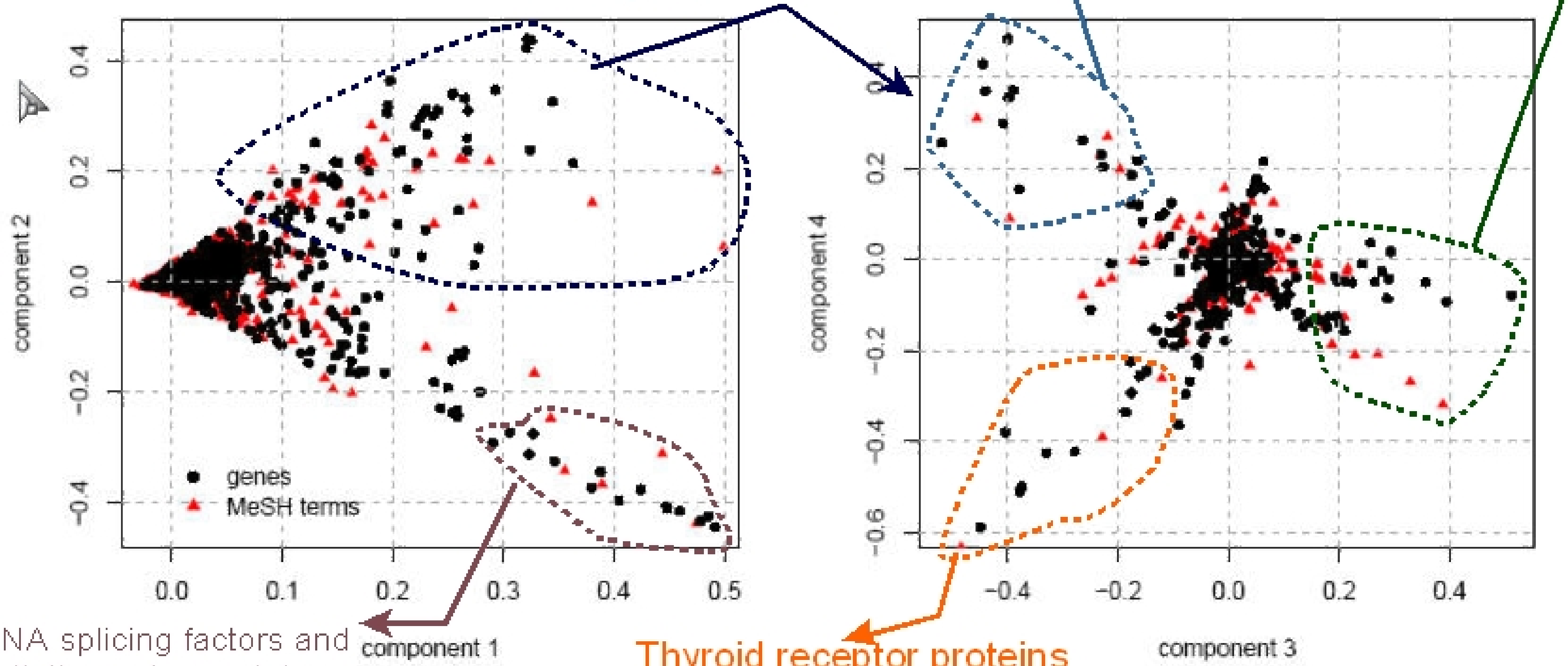
Rechtsteiner, A. [2005]. PhD Dissertation.

SVD of Gene/MeSH co-ocurrence

for each gene co-expression cluster: uncovering "functional themes"



Unvovered in: Challacombe et al. *Physiol. Genomics*. **10**.1152. 2004.



Transcription Factors (many associated with the Cell Cycle)

Transcription Factors and Retinoblastoma proteins

mRNA splicing factors and small ribonucleo proteins

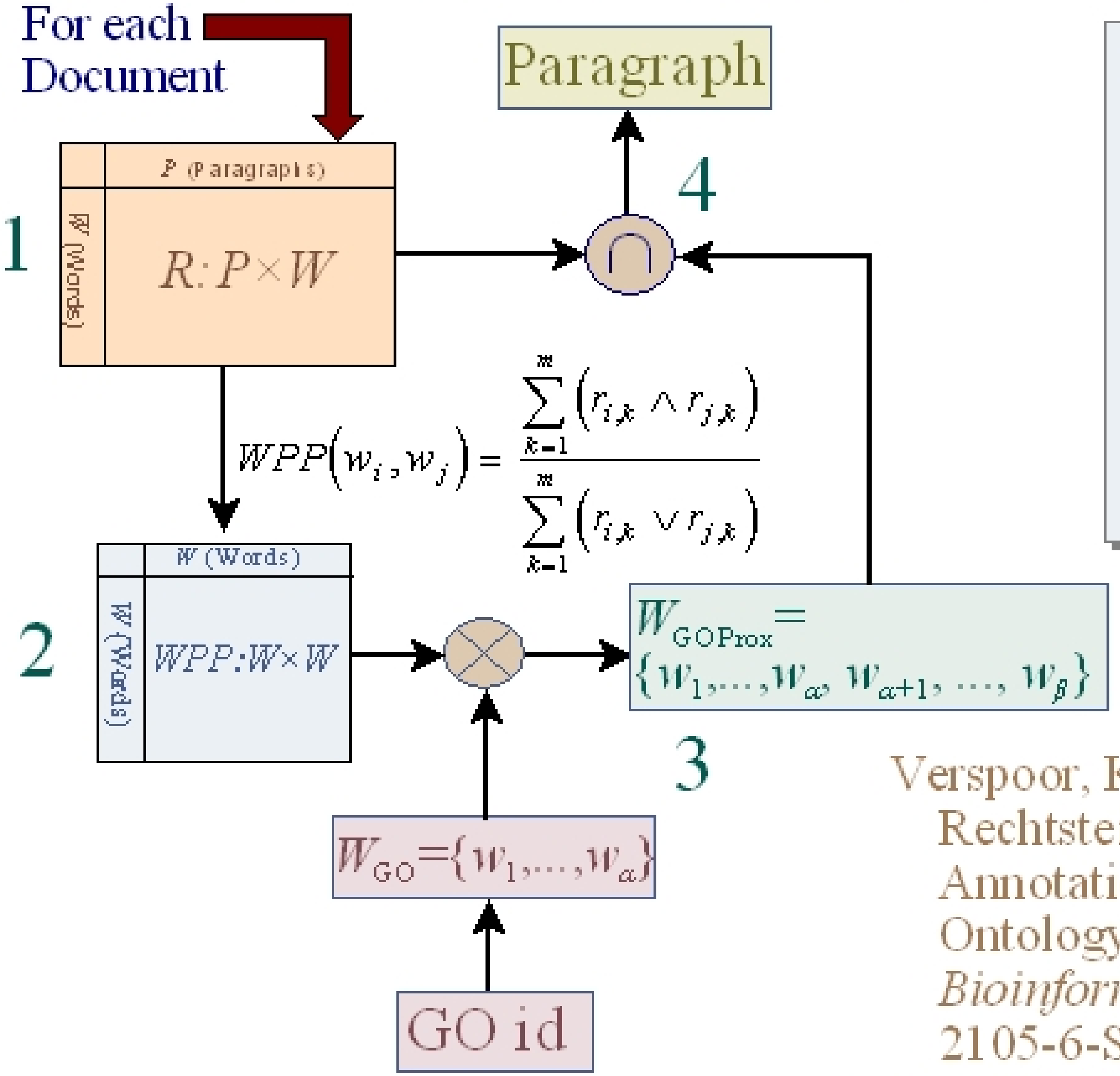
Thyroid receptor proteins or proteins binding to this receptor

Rechtsteiner, A. [2005]. PhD Dissertation.

INDIANA UNIVERSITY

Biocreative competition (EMBO Workshop)

a critical assessment of text mining methods in molecular biology

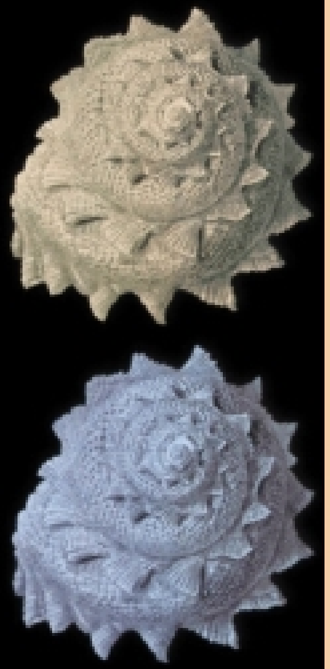


$$WPP(w_i, w_j) = \frac{\sum_{k=1}^{m} (r_{i,k} \wedge r_{j,k})}{\sum_{k=1}^{m} (r_{i,k} \vee r_{j,k})}$$

- Task 2: Given a document, discover the portion of text most appropriate to annotate the protein's function, and produce appropriate Gene Ontology node for annotation
  - Learning set: triplets (protein, document, GO id)
  - Test set: documents

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2005]. "Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks". *BMC Bioinformatics*, **6**(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20

Rocha, Luis M. [2002]. In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) IOS Press, pp. 137-163.

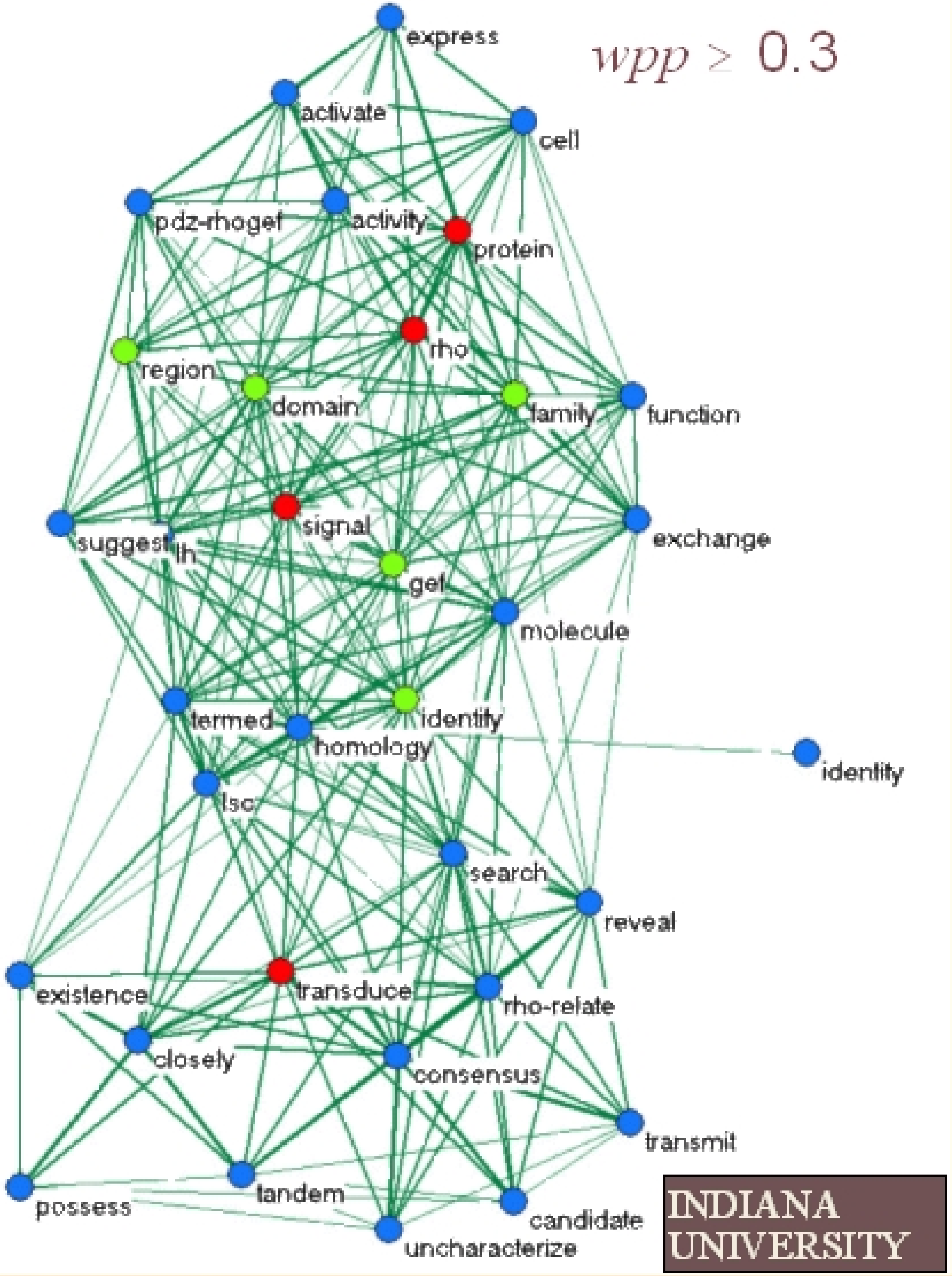


word in paragraph proximity

example document

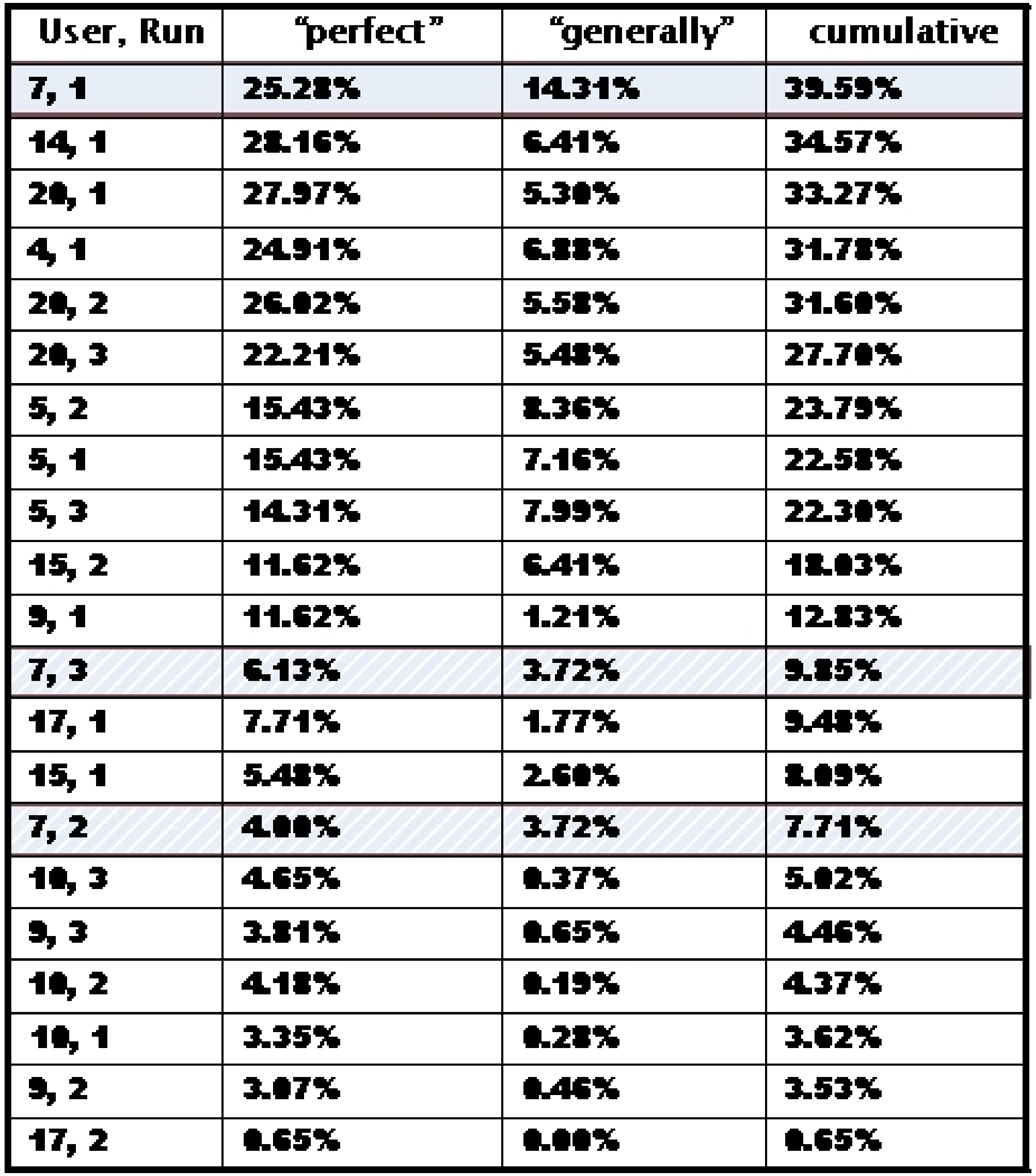
- document bc005868
  - *WPP* contains 1102 words
  - Subgraph of 34 words
    - Red nodes: words removed from the respective GO annotation (0007266): Rho, protein, signal, transducer).
    - Blue nodes: words that co-occur very frequently (*wpp* > 0.5) with at least one of the red nodes
    - Green nodes: additional words recommended with largest average proximity to all input words (red nodes)

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2005]. "Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks". *BMC Bioinformatics*, **6**(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20



Task 2.1 Results

Proximity-based run



User, Run	"perfect"	"generally"	cumulative
7, 1	25.28%	14.31%	39.59%
14, 1	28.16%	6.41%	34.57%
20, 1	27.97%	5.30%	33.27%
4, 1	24.91%	6.88%	31.78%
20, 2	26.02%	5.58%	31.60%
20, 3	22.21%	5.48%	27.70%
5, 2	15.43%	8.36%	23.79%
5, 1	15.43%	7.16%	22.58%
5, 3	14.31%	7.99%	22.30%
15, 2	11.62%	6.41%	18.03%
9, 1	11.62%	1.21%	12.83%
7, 3	6.13%	3.72%	9.85%
17, 1	7.71%	1.77%	9.48%
15, 1	5.48%	2.60%	8.09%
7, 2	4.00%	3.72%	7.71%
10, 3	4.65%	0.37%	5.02%
9, 3	3.81%	0.65%	4.46%
10, 2	4.18%	0.19%	4.37%
10, 1	3.35%	0.28%	3.62%
9, 2	3.07%	0.46%	3.53%
17, 2	0.65%	0.00%	0.65%

Verspoor, K., et al [2005]. *BMC Bioinformatics*, **6**(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20

in search of a gold standard

need for validation tools

- Bibliome tools for data-driven experiments are typically tested by sampling some of their output and presenting it to experts, but
  - experts typically disagree
  - cannot be an expert on all topics involved,
  - get tired of manually testing the output of mechanic algorithms, leading to potentially unreliable answers
- Tools for automatic validation are needed!

Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C.E., Rocha, L.M. [2006]. "Large-Scale Testing Of Bibliome Informatics Using Pfam Protein Families". In: *Pacific Symposium on Bioinformatics 2006*: **11**:76-87.

protein family prediction

studying the quality of links in biomedical resources

- Large scale study to explore how well publications about proteins can predict the Pfam families of proteins.
- Pfam families do cluster and are largely separable in publication space
  - Pfam families for 15,217 proteins from 1611 Pfam families
  - For 76% of the proteins the correct Pfam family was the first predicted
  - For 89% of proteins the correct Pfam family was found within the first 5 predicted families.
  - Many of the mispredictions occur between closely related families.
  - Prediction success depends on family size and the number of publications referenced

Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C.E., Rocha, L.M. [2006]. "Large-Scale Testing Of Bibliome Informatics Using Pfam Protein Families". In: *Pacific Symposium on Bioinformatics 2006*: **11**:76-87.

informatics
luis rocha 2007

extracting data

from SwissProt/UniProt

Publications referenced in SwissProt

PubMed ID	MeSH term	MeSH ID	Protein ID
7532594	CHO Cells	A11.251.210.200	62913
7532594	Hamsters	B02.649.865.635 .325	62913
7532594	Rats	B02.649.865.635 .560	62913
.....	.....	.....	.....
8125992	Molecular Sequence Data	L01.453.245.667	3200

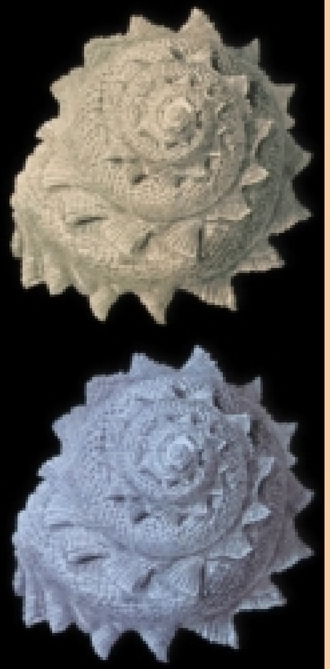
Pfam protein families

Protein ID	PFAM ID
62913	PF00001
62913	PF00001
62913	PF00001
.....	.....
3200	PF04988

75,649 publications

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY



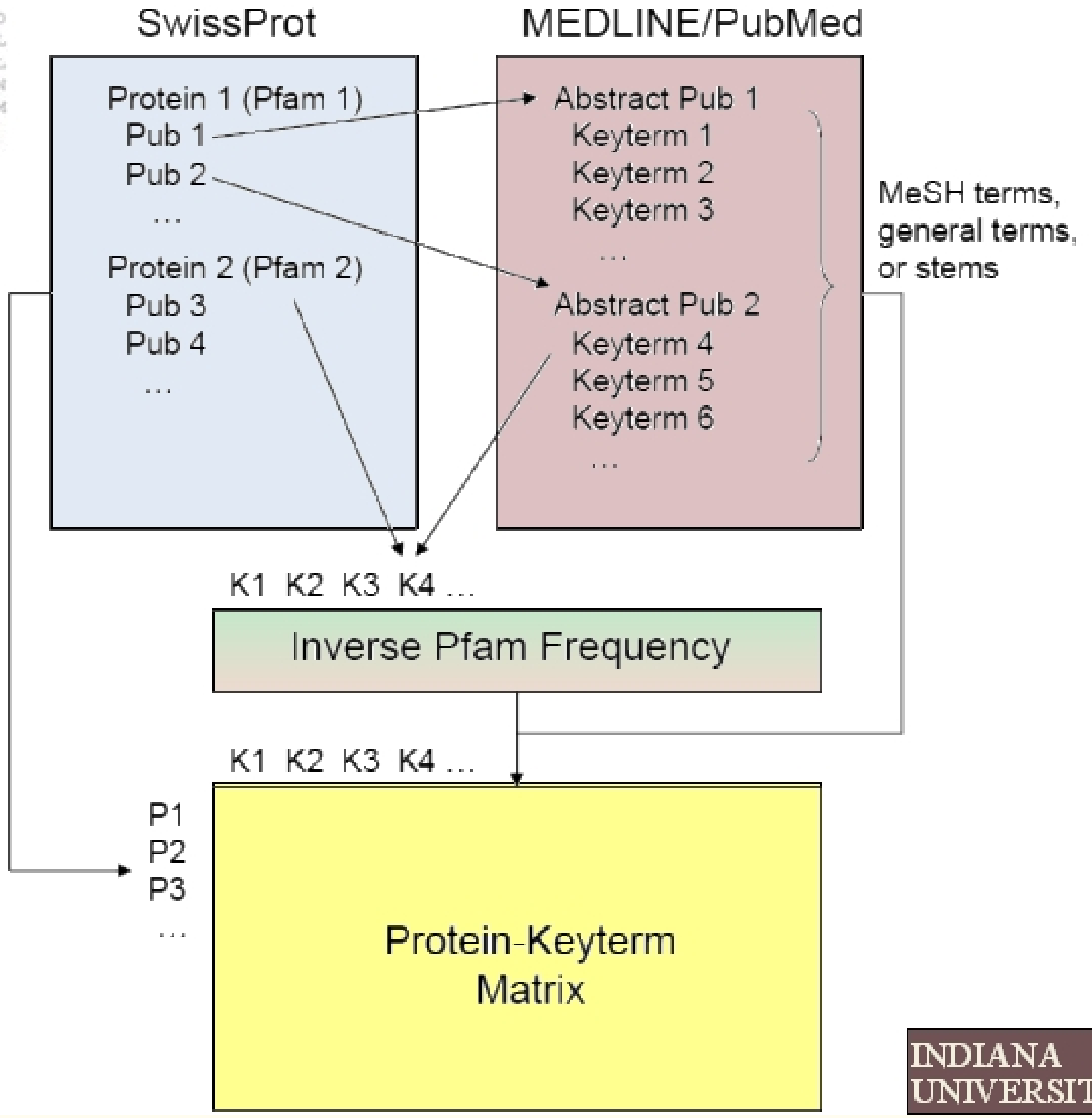
building a protein-keyterm matrix

from Medline/PubMed

UniProt
the universal protein resource

SIB swissprot EBI

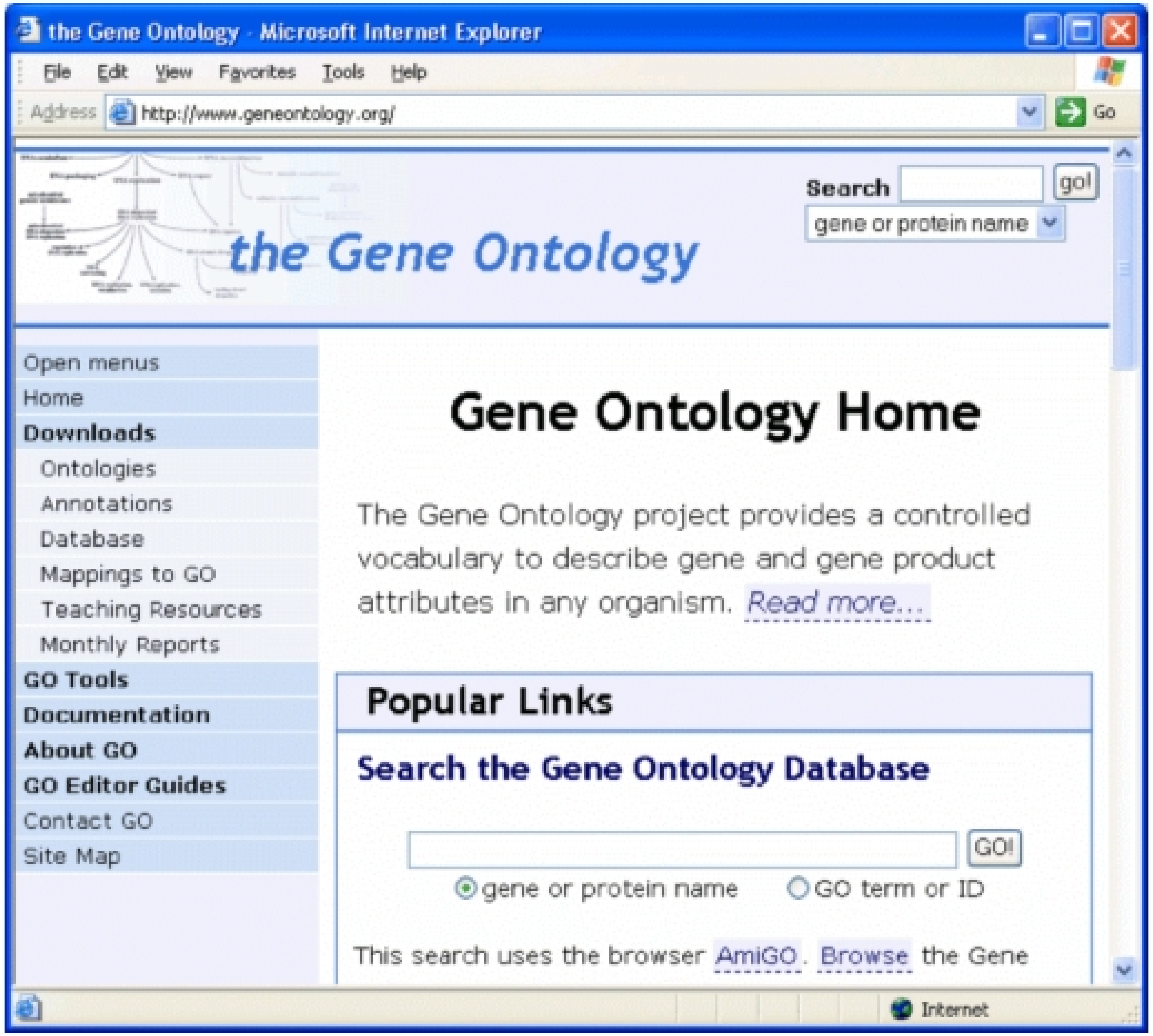
15,217 Proteins
1611 Pfam Families



the Gene Ontology (GO)

another source of keyterms

Controlled vocabulary to describe gene and gene product attributes in any organism.



Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Read more...

Popular Links

Search the Gene Ontology Database

GO!

● gene or protein name ○ GO term or ID

This search uses the browser AmiGO. Browse the Gene

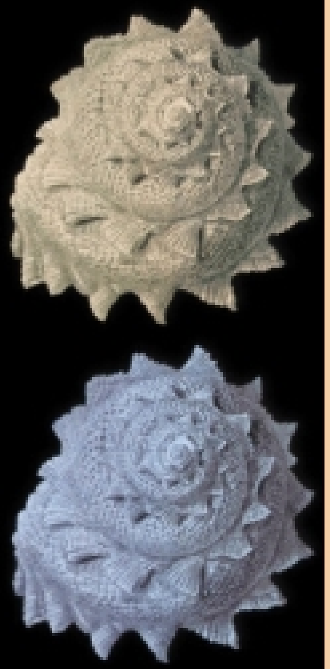
from the Gene Ontology

Gene Ontology Annotations

GO term	GO ID	Protein ID
reproduction	GO:0000003	62913
cell cycle checkpoint	GO:0000075	62913
mitotic metaphase	GO:000008	62913
.....	.....	.....
DNA replication checkpoint	GO:000007	3200

Pfam protein families

Protein ID	PFAM ID
62913	PF00001
62913	PF00001
62913	PF00001
...	...
3200	PF04988



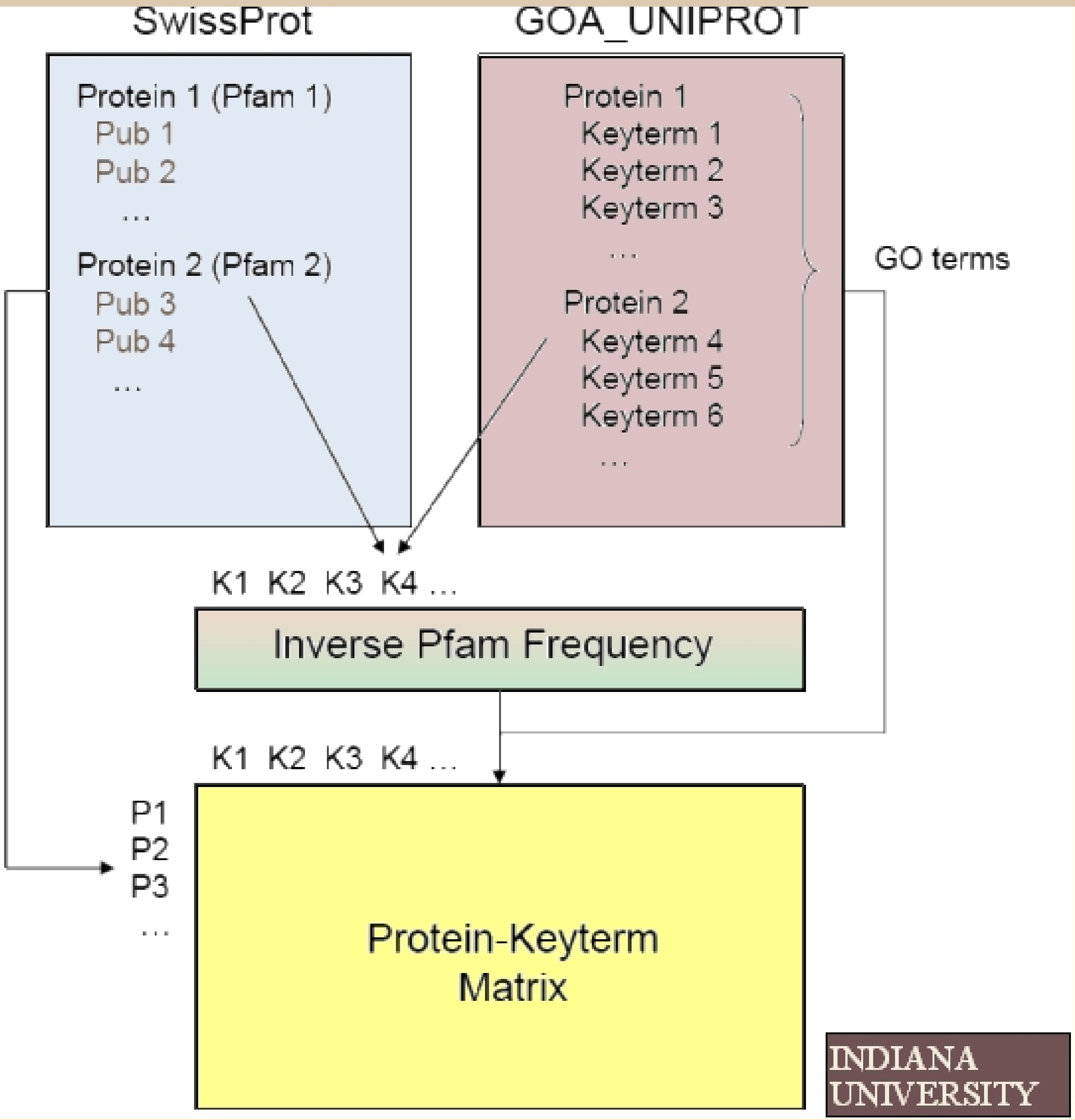
building a protein-keyterm matrix

from GO/GOA/Uniprot

the Gene Ontology

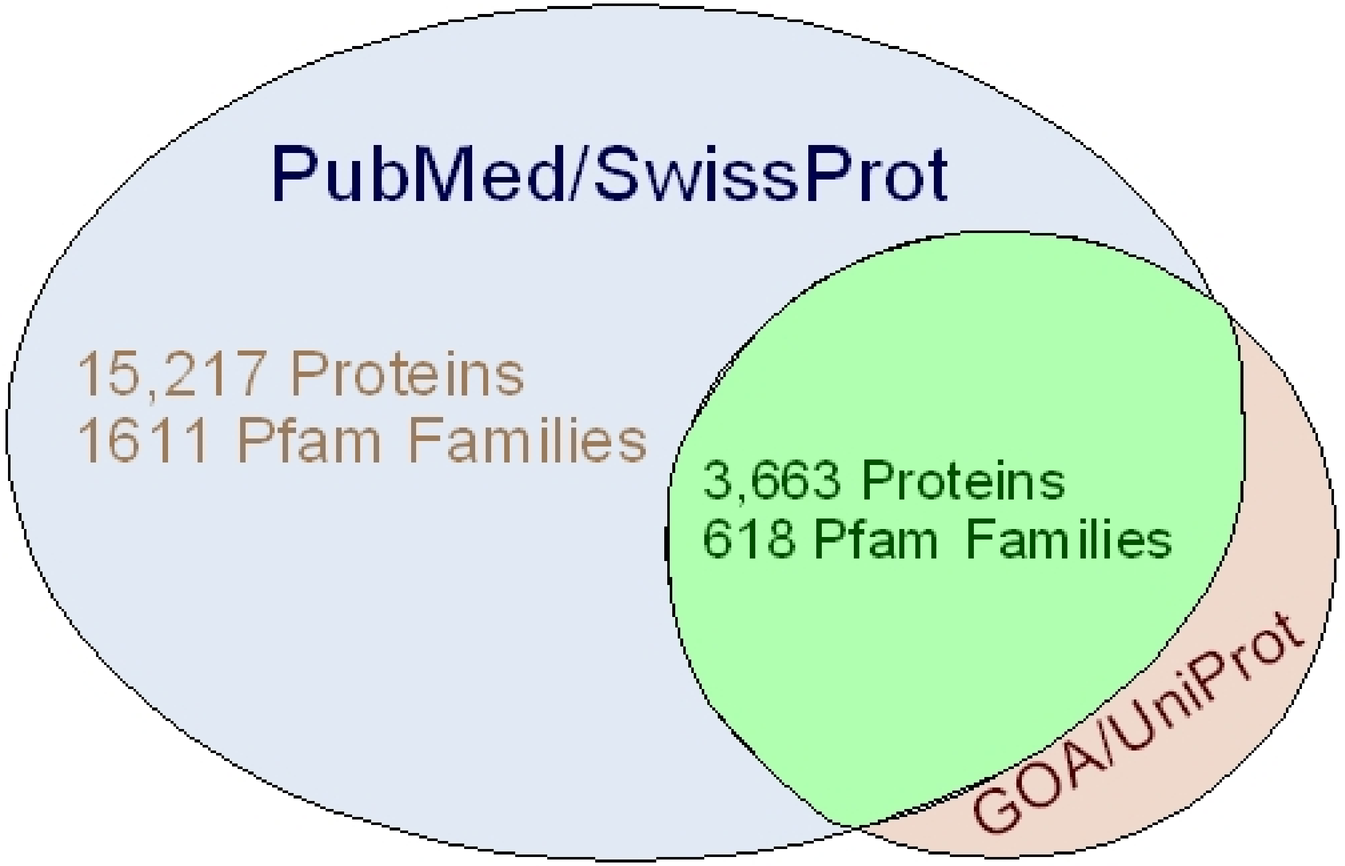
Annotation@EBI

UniProt
the universal protein resource



the reduced set of proteins

a common set between PubMed and GO



PubMed/SwissProt

15,217 Proteins
1611 Pfam Families

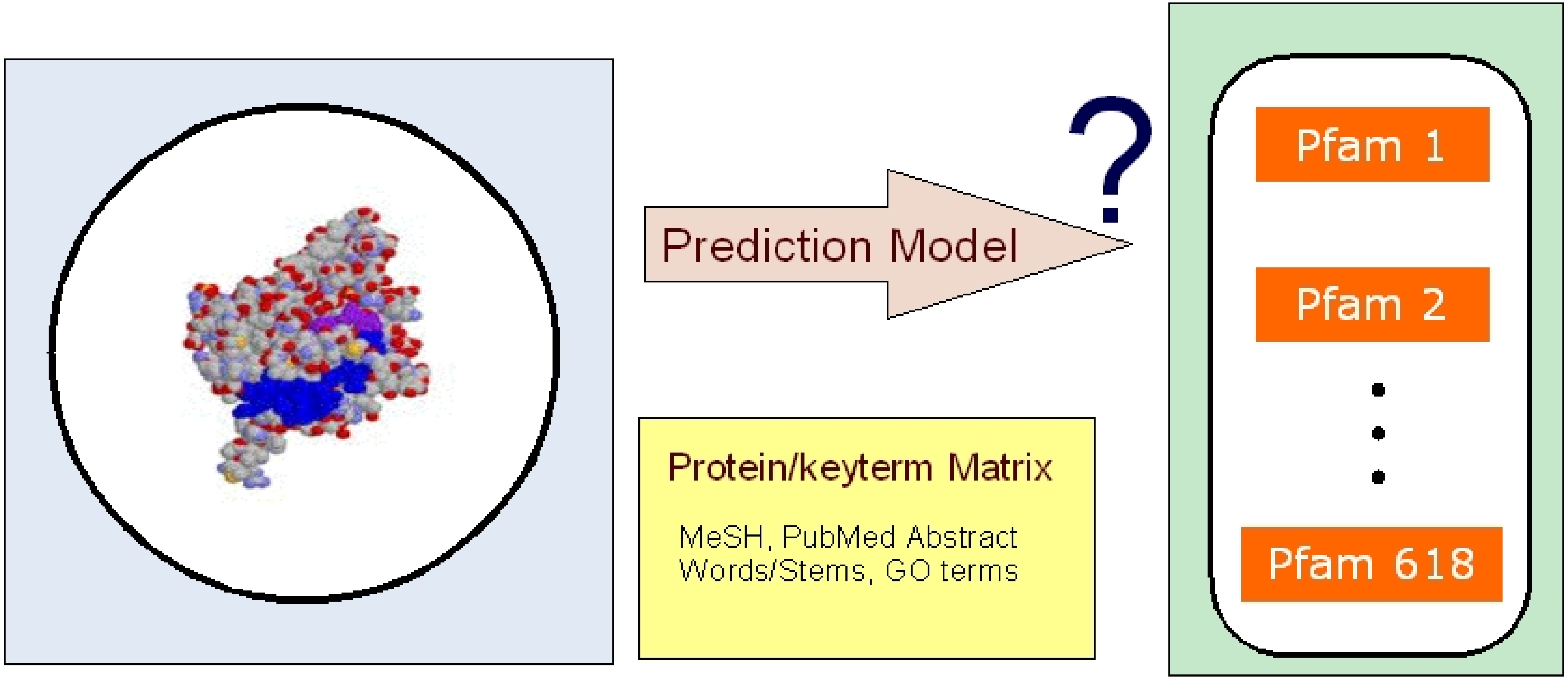
3,663 Proteins
618 Pfam Families

GOA/UniProt

- Families with at least 3 proteins
  - Mean=5.9
  - Median=5
  - Standard Deviation=3.3
  - 179 families with only 3 proteins
  - Largest 3 families contain 17 proteins
  - Average keyterms per protein
    - MeSh: 27
    - PubMed Words: 153
    - PubMed Stems: 132
    - GO: 4

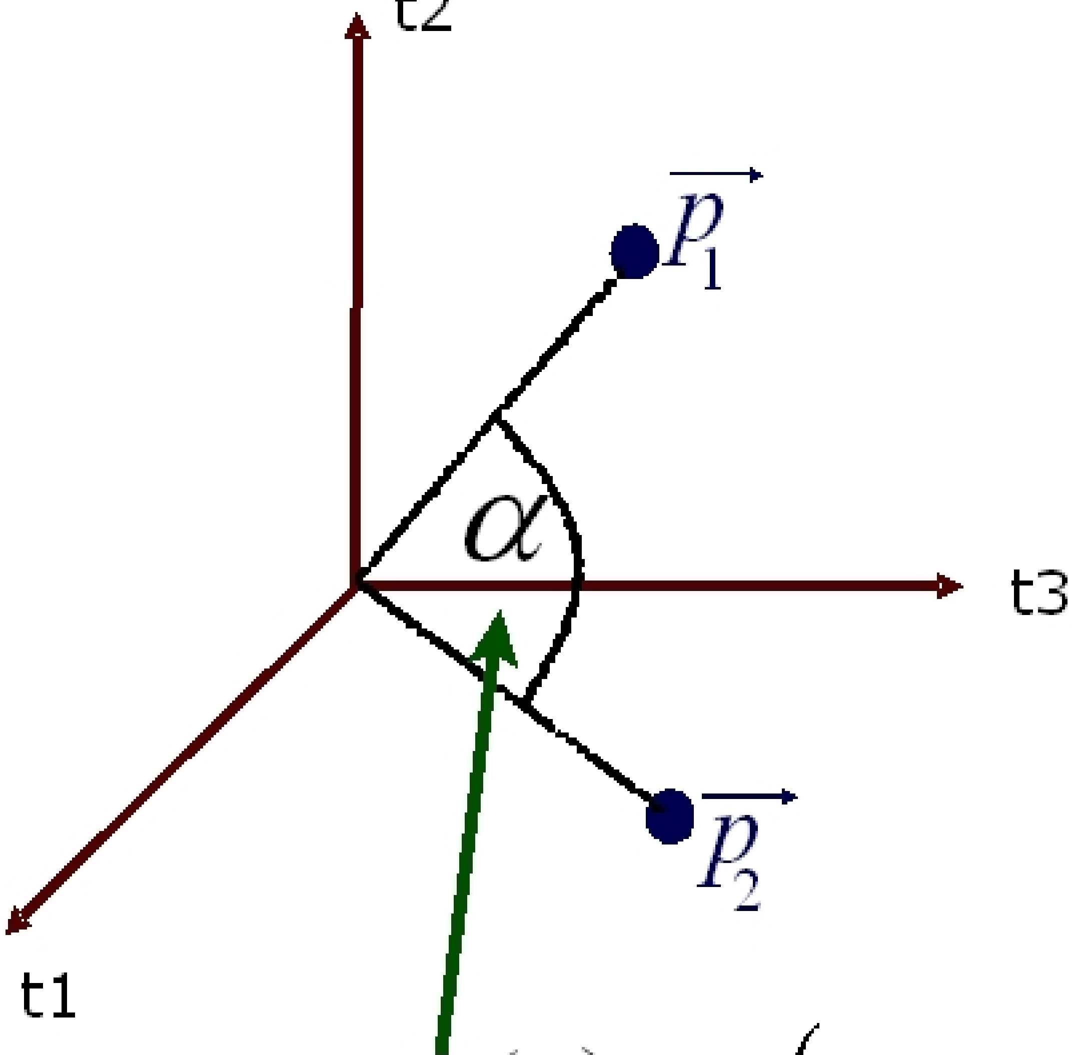
Pfam family prediction

from keyterm associations



vector space model

cosine similarity



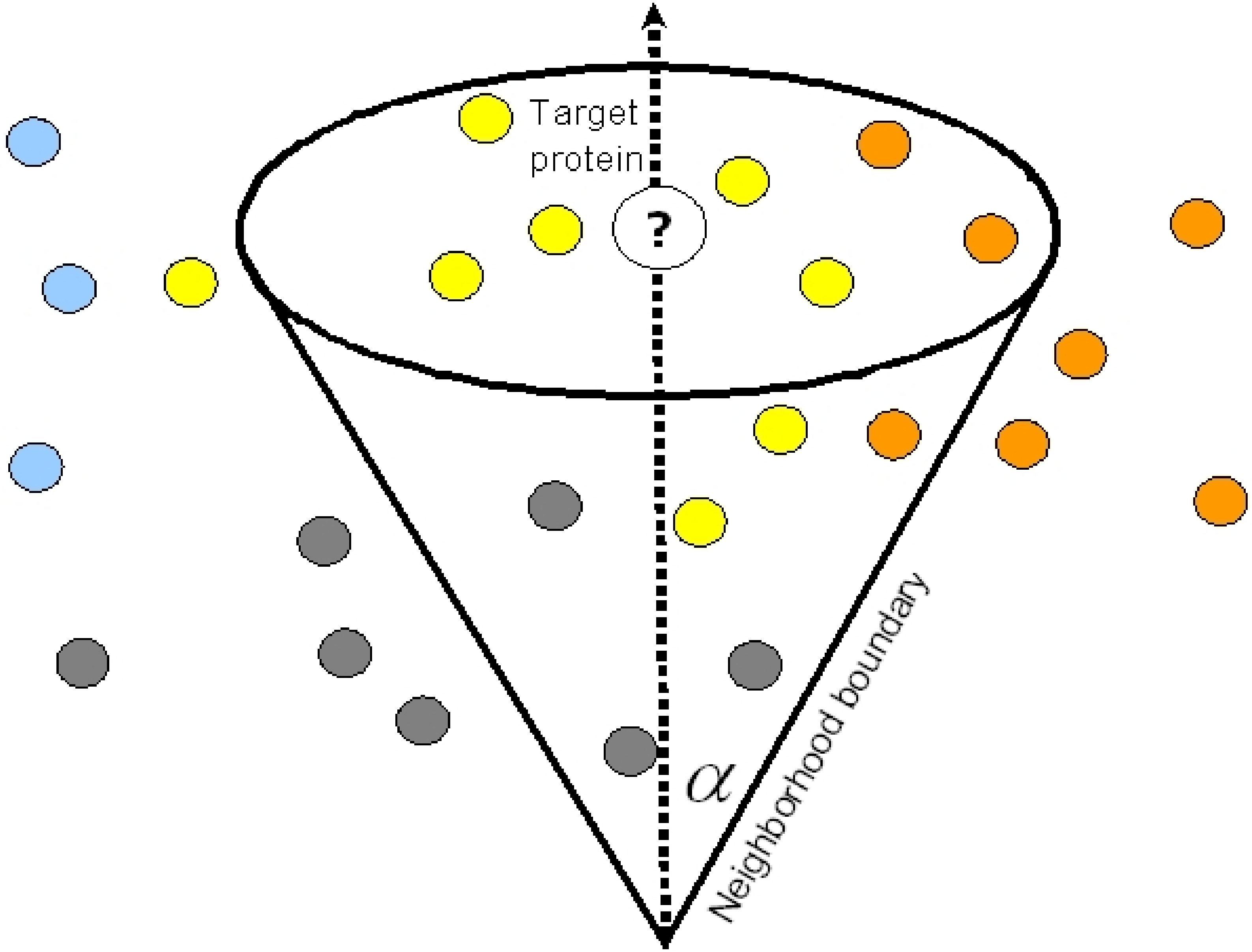
$$\cos(\alpha) = \sigma(p_1, p_2) = \frac{\vec{p}_1 \cdot \vec{p}_2}{\|\vec{p}_1\| \cdot \|\vec{p}_2\|}$$

Publications referenced in SwissProt

PubMed ID	MeSH term	MeSH ID	Protein ID
7532594	CHO Cells	A11.251.210.200	62913
7532594	Hamsters	B02.649.865.635 .325	62913
7532594	Rats	B02.649.865.635 .560	62913
…..	…..	…..	…..
8125992	Molecular Sequence Data	L01.453.245.667	3200

Pfam family prediction model

neighborhood angle

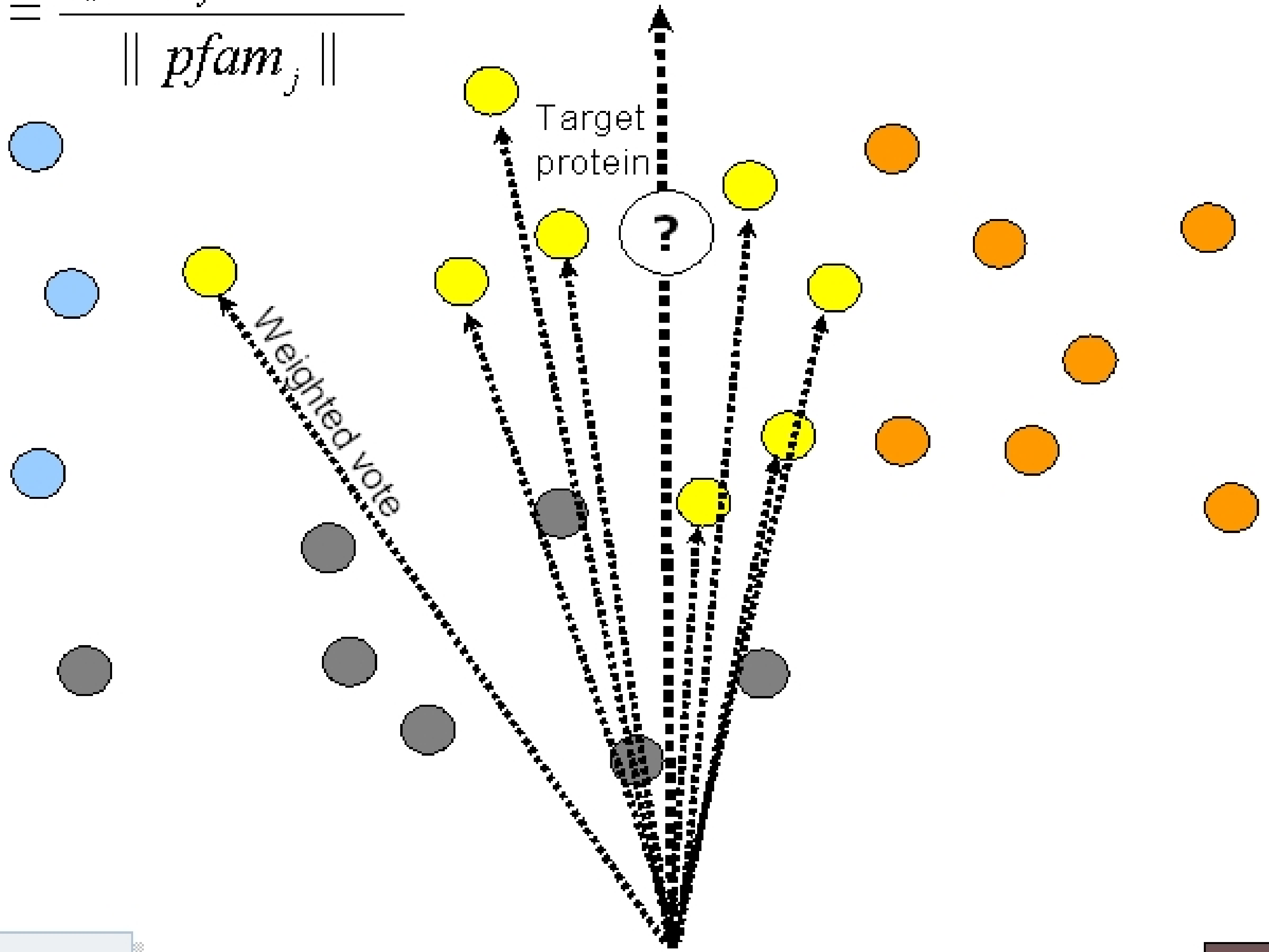


Rechtsteiner, A. [2005]. PhD Dissertation.

Pfam family prediction model

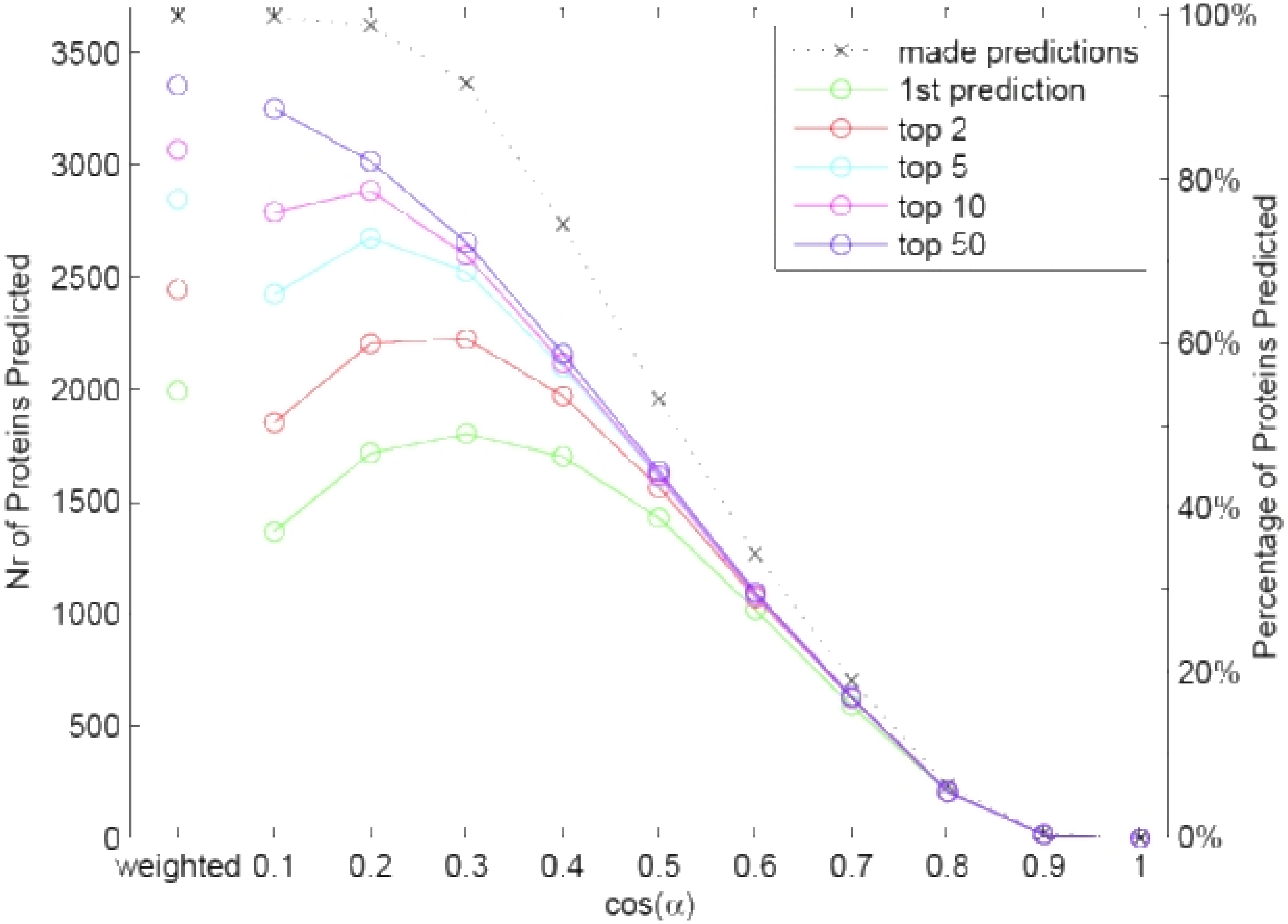
voting model

$$A_{WV} : Pfam_j(p_i) = \frac{\sum_{p_k \in pfam_j} \sigma(p_k, p_i)}{\| pfam_j \|}$$



Protein family prediction

MesH

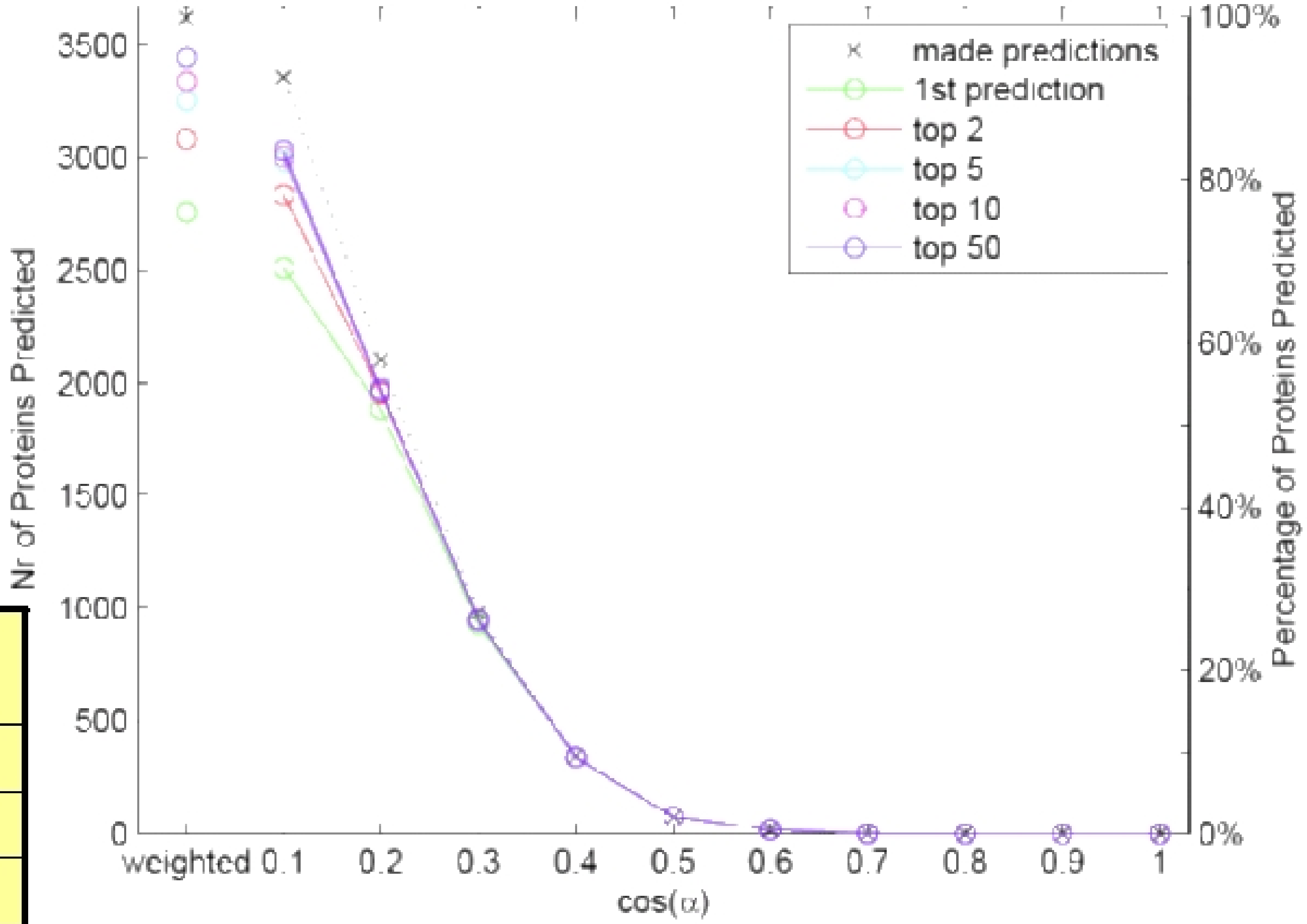


	MeSH terms
1st prediction	54.35%
top 2	66.72%
top 5	77.70%
top 10	83.76%
top 50	91.54%

informatics
luis rocha 2007

protein family prediction

PubMed abstract words



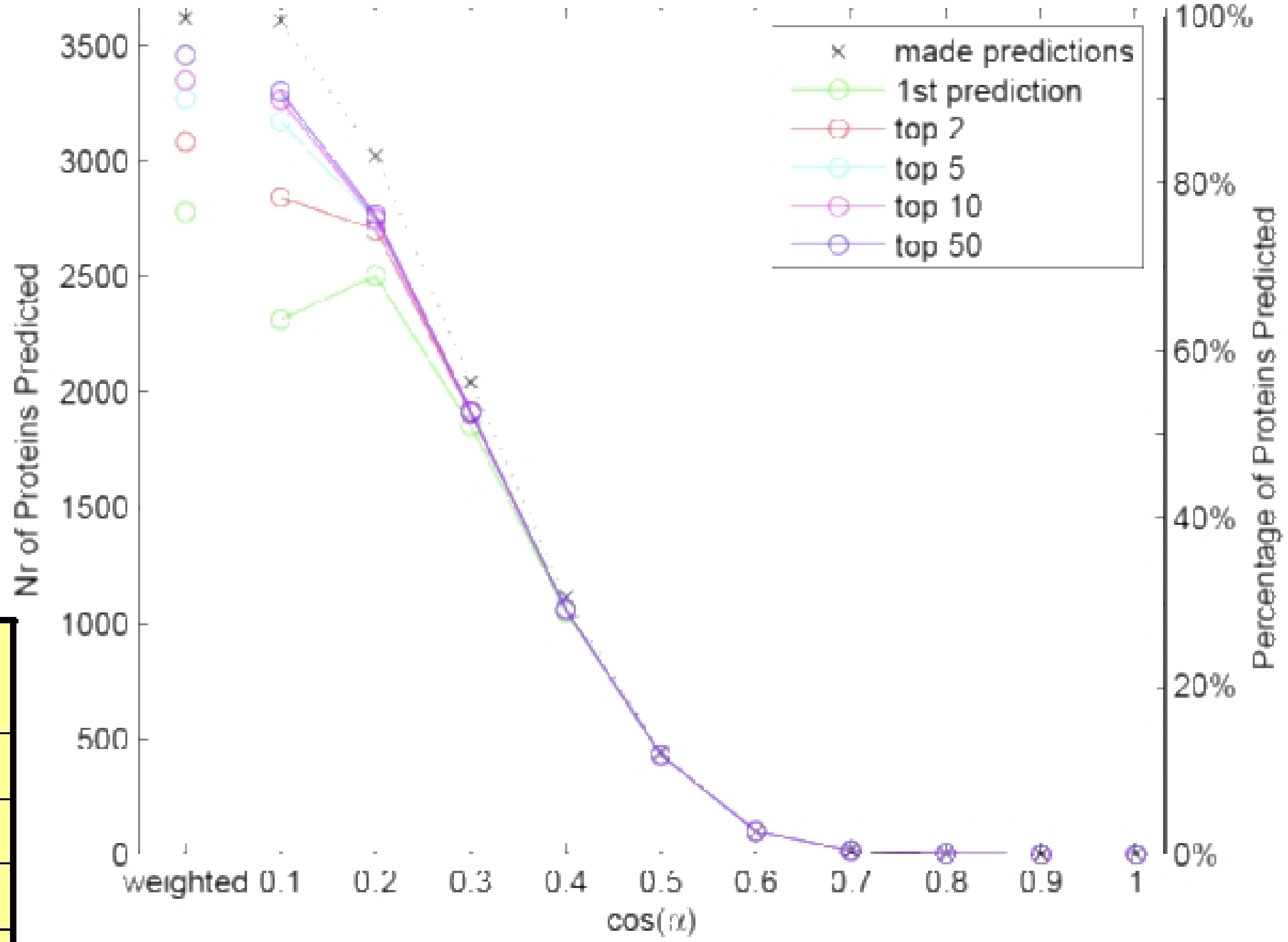
	PubMed Words
1st prediction	75.27%
top 2	84.17%
top 5	88.83%
top 10	91.13%
top 50	94.02%

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

protein family prediction

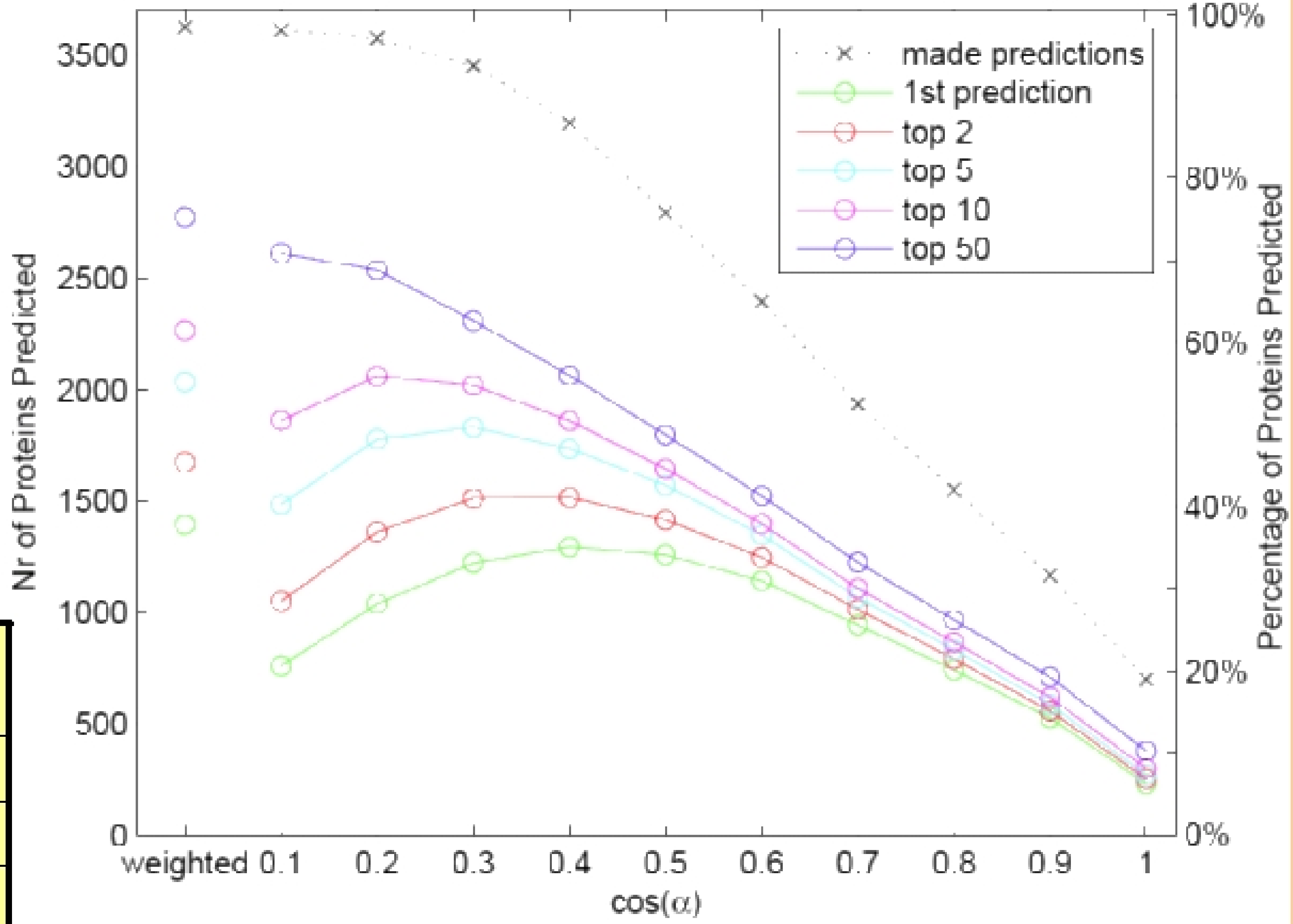
PubMed abstract stems



	PubMed Stems
1st prediction	75.89%
top 2	84.22%
top 5	89.30%
top 10	91.48%
top 50	94.40%

protein family prediction

GO terms



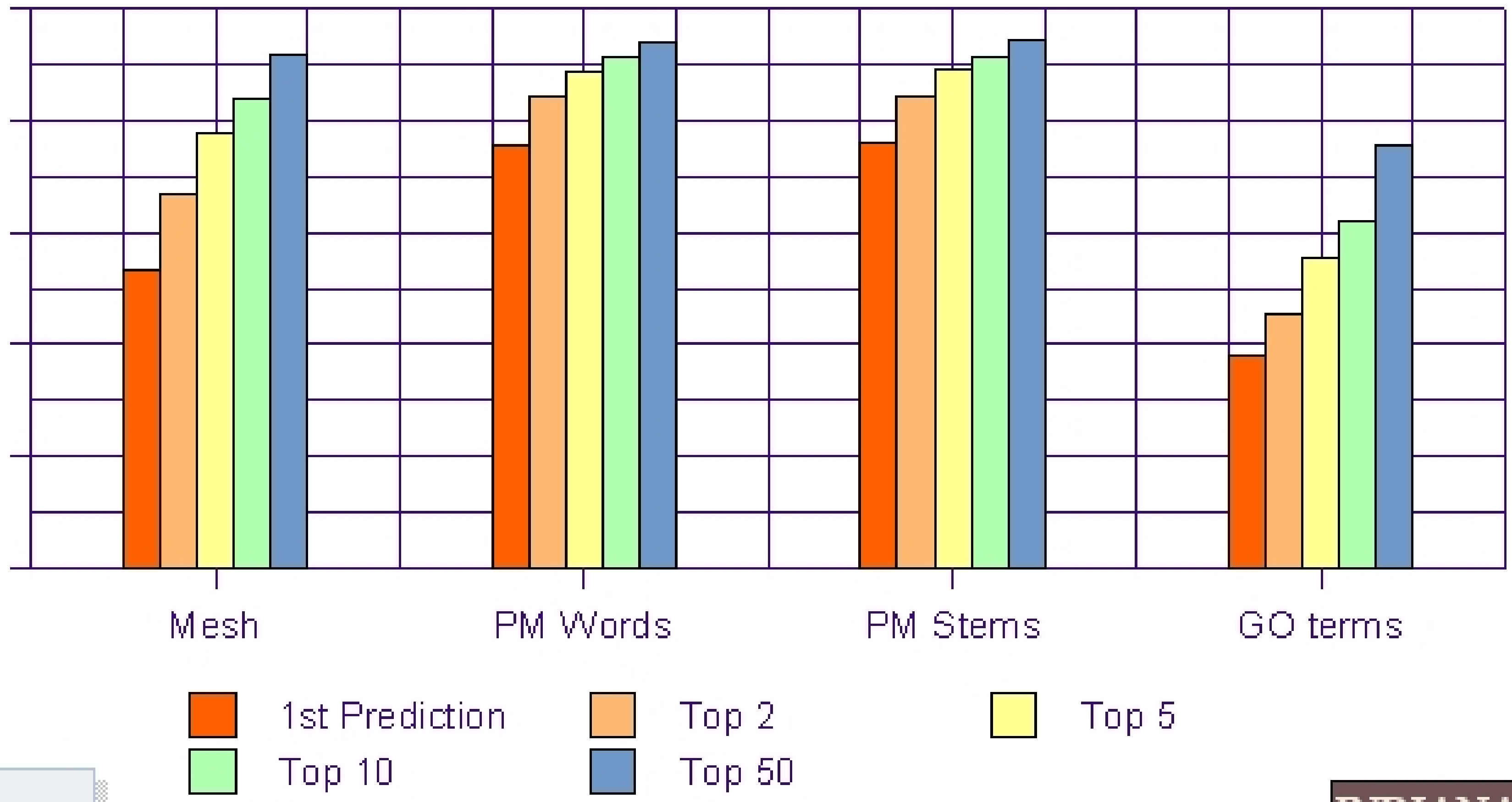
	GO terms
1st prediction	38.08%
top 2	45.65%
top 5	55.53%
top 10	61.86%
top 50	75.59%

informatics
luis rocha 2007

Pfam family prediction

comparison of different keyterm sets

	Mesh	PM Words	PM Stems	GO terms
1st Prediction	53.35	75.37	75.89	38.08
Top 2	66.72	84.17	84.22	45.65
Top 5	77.7	88.83	89.3	55.53
Top 10	83.76	91.13	91.48	61.86
Top 50	91.54	94.02	94.4	75.59



rocha@indiana.edu
http://informatics.indiana.edu/rocha

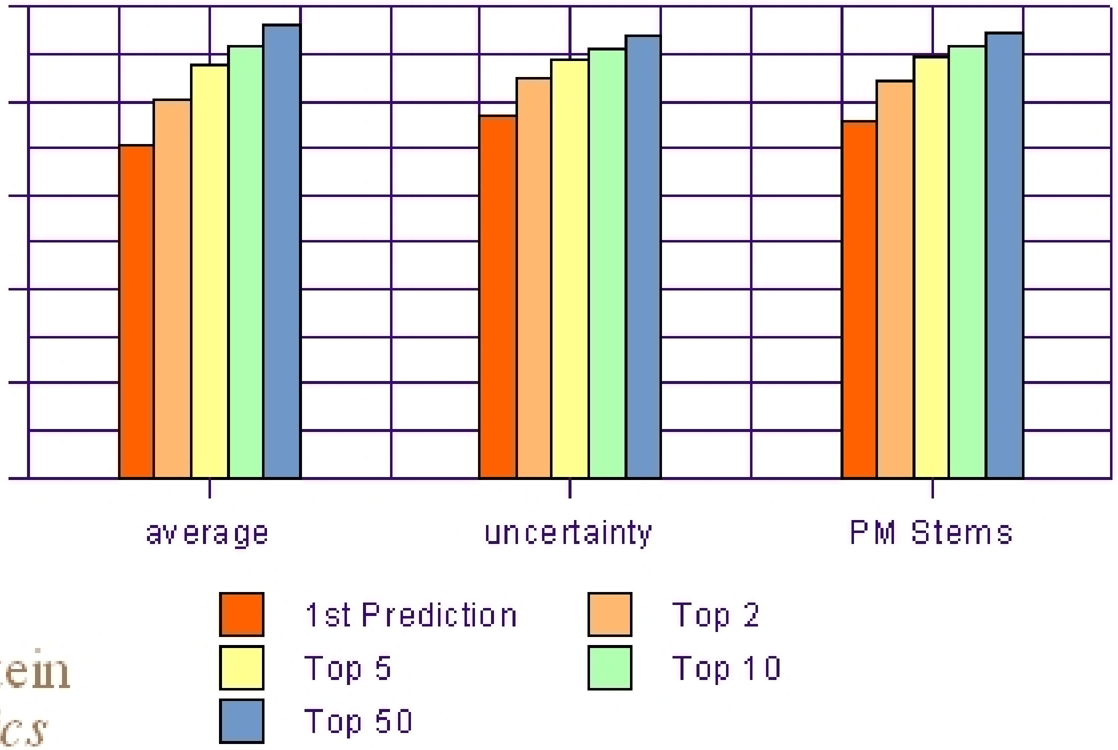
INDIANA
UNIVERSITY

Pfam family prediction

Integration of different sources

	average	Uncertainty	PM Stems
1st Predicti>	70.84	77.15	75.89
Top 2	80.02	84.77	84.22
Top 5	87.5	88.86	89.3
Top 10	91.35	90.88	91.48
Top 50	95.93	93.8	94.4

- Uncertainty Method
  - Choose prediction from least uncertain source
    - Measured by Shannon's entropy measure
    - On probability of selecting a given family distribution



Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C.E., Rocha, L.M. [2006]. "Large-Scale Testing Of Bibliome Informatics Using Pfam Protein Families". In: *Pacific Symposium on Bioinformatics 2006:* **11**:76-87.

informatics
luis rocha 2007

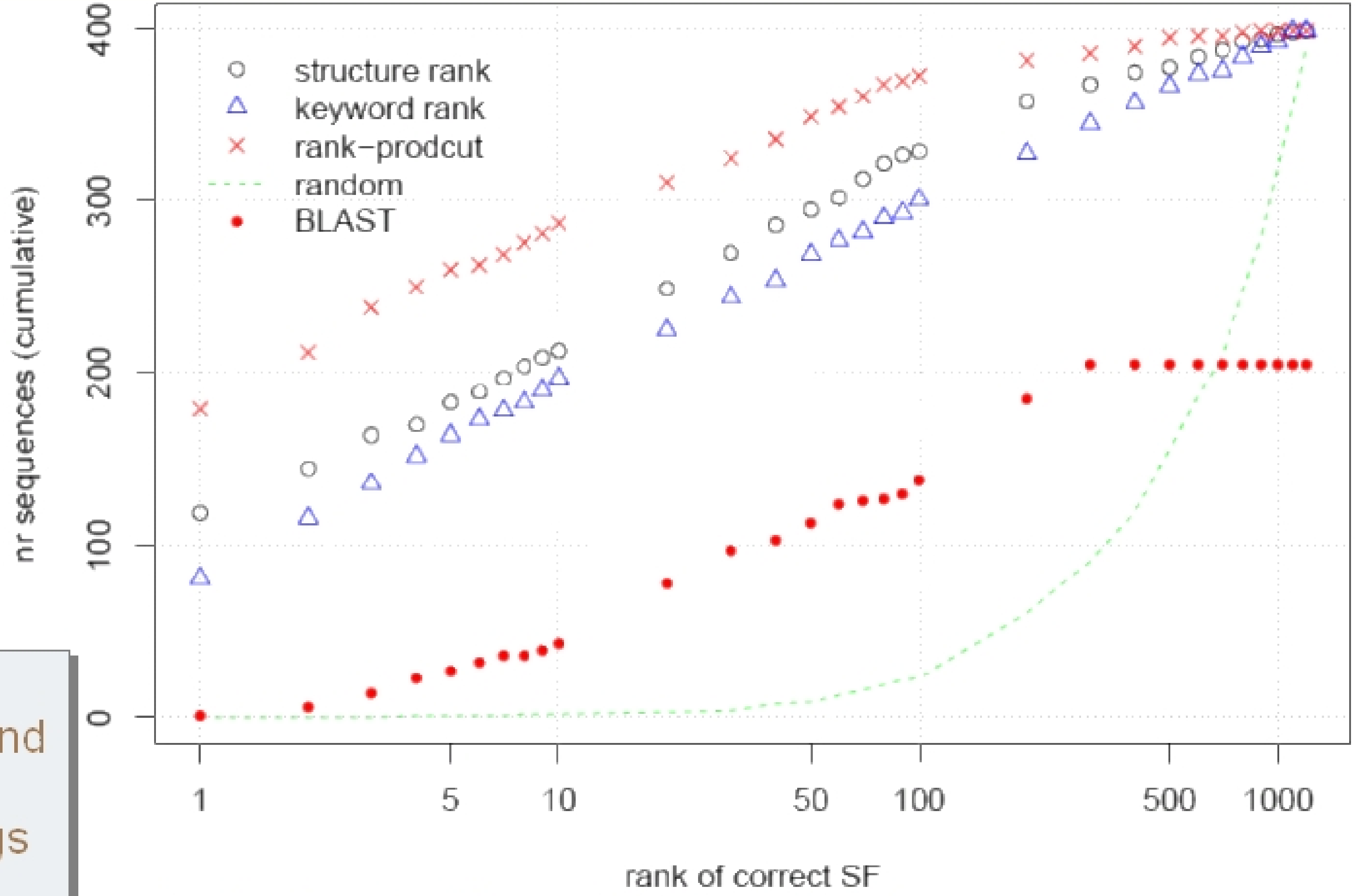
text mining for protein structure prediction

in lack of sequence homology

- Structure Prediction
  - 40-60% of proteins in a new genome are reliably predicted by sequence comparison with previously annotated genomes
    - Typically the genes we care least about….
  - Ab-inition structure prediction (Rosetta and Mammoth)
    - Predicts proteins' approximate structure and compares it to the structure of proteins of known function.
    - Does not require homologs
- Large set of sequences of known structure
  - 400 test sequences (with known structures)
    - MeSH keyword information
  - SCOP super-families
    - Representative MeSH keyword frequency vectors obtained
    - Using BLAST, homologs of test sequences were removed
  - Cosine vector similarity
    - between each SCOP family vector and all the keyword vectors of test sequences
    - Rank SCOP super-families by decreasing similarity for each test sequence.

SCOP family prediction

comparison



- Rank product to combine ab-initio and keyword ranks
- Sequence homologs removed

Rechtsteiner, A., Luinstra, J., Rocha, L.M., Strauss, C.E., [2006]. "Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology". In: *ISMB/BioLink 2006: In Press*

uncovering protein-protein interactions in the bibliome

BioCreative II --- Group T11

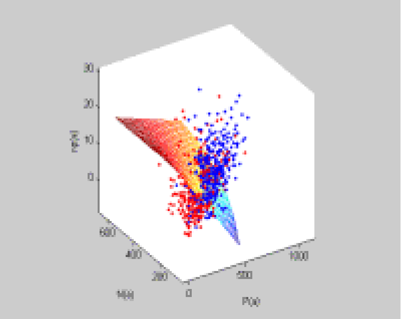
Alaa Abi-Haidar, Jasleen Kaur, Ana Maguitman,
Predrag Radivojac, Andreas Retchsteiner,
Karin Verspoor, Zhiping Wang, **Luis M. Rocha**

Indiana University, USA

Instituto Gulbenkian de Ciencia, Portugal

Universidad Nacional del Sur, Argentina

Los Alamos National Laboratory, USA



Proc. of the Second BioCreative Challenge Evaluation Workshop (ISBN 84-933255-6-2), pp. 247-255

Genome Informatics. In Press.

rocha@indiana.edu
http://informatics.indiana.edu/rocha

IAS (IPS and ISS)

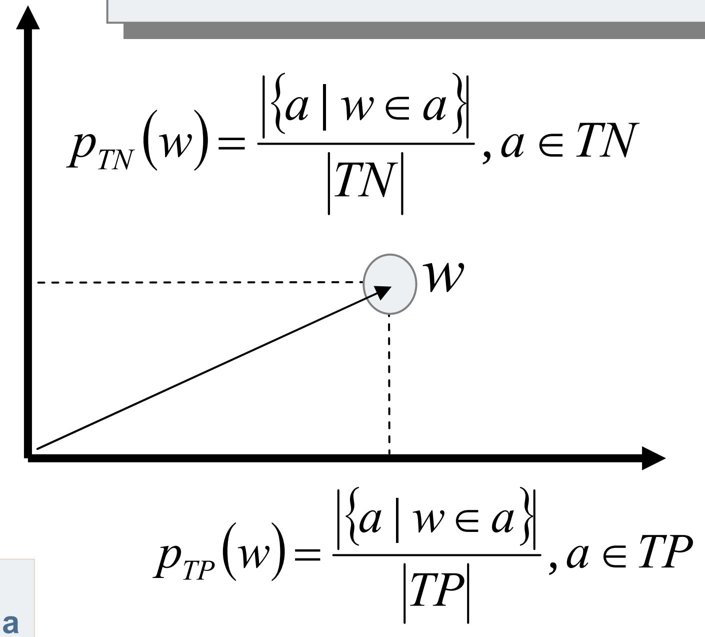
TP **3536** | **TN** **1959** | **Official Training Data**

TP* **13K** — **Noisy Positive Distributed by Biocreative**

TP^{M} **367** — **From MIPS database**

TN^{S} **427** — **Likely negatives from Santiago Schnell**

- Single Words
  - Top 650 w_i
    - with $S(w_i)=|p_{TP}(w_i)-p_{TN}(w_i)|$.
- “word bigrams”
  - $S^{bi}(w_i w_j)$
- “Window-10 Word Pairs”
  - $S^{10}(w_i, w_j)$.
- Number of protein Mentions
  - $np(a)$
    - Using Settles’ ABNER (*A Biomedical Named Entity Recognizer*)



$$p_{TN}(w)=\frac{|\{a \mid w \in a\}|}{|TN|}, a \in TN$$

$$p_{TP}(w)=\frac{|\{a \mid w \in a\}|}{|TP|}, a \in TP$$

IAS:Run 1: Support Vector Machine (SVM)

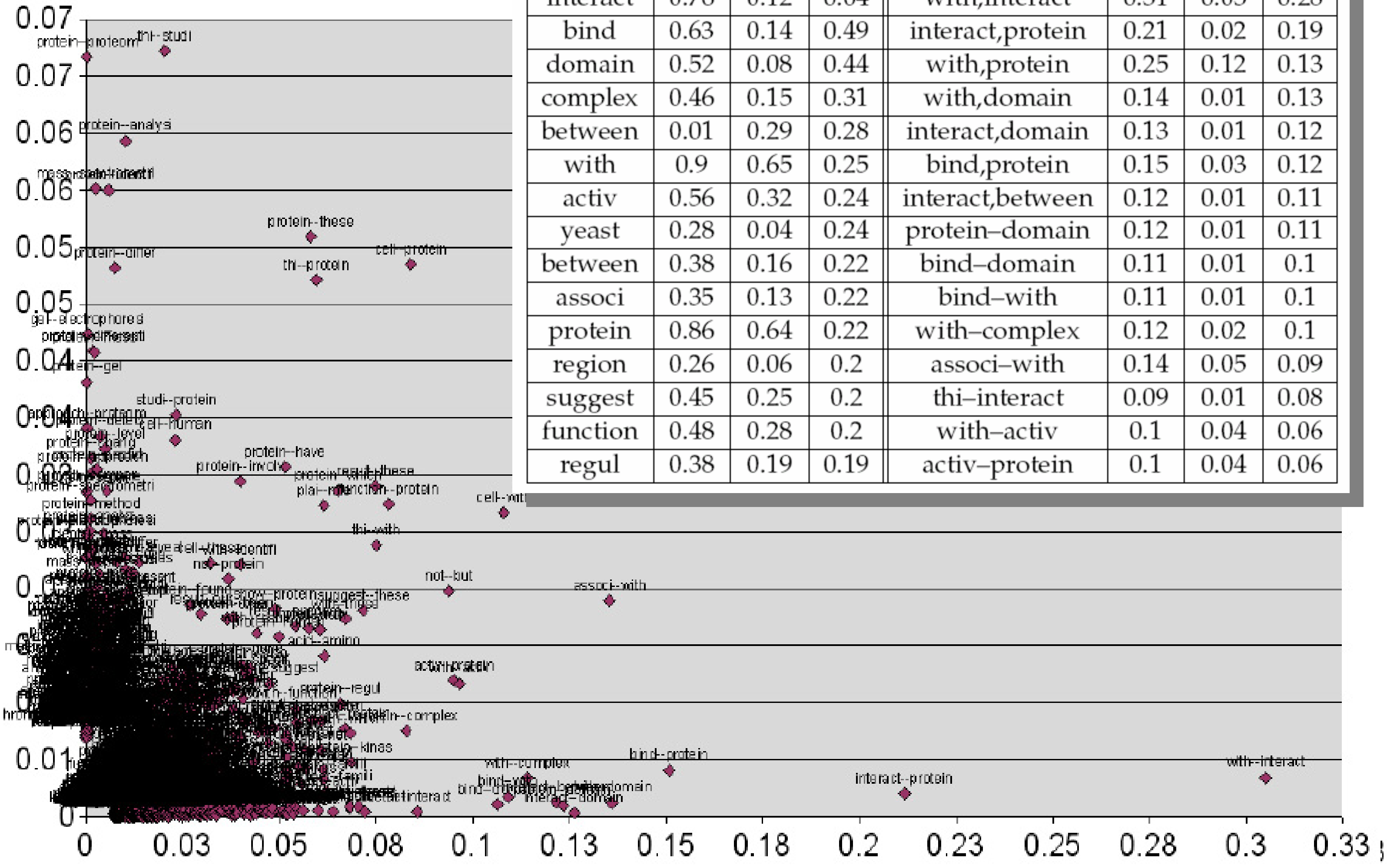
- Feature Selection
  - Top 650 Words plus number of protein mentions
  - Filtered via t-test
  - Dimensionality reduction via PCA
- Final configuration
  - linear support vector machine.
- Results
  - Our best AUC: **0.7995**
- Post-results
  - Selecting features differently leads to same results
  - Training and test set very different
    - An SVM predictor for labeled vs. unlabeled data
      - AUC = 69%, F-score = 92%
  - Bootstrapping from unlabeled data
    - Making training data more similar to test data
      - AUC = 81.5% (on 650 word features(

Informatics
luis rocha 2007

feature example

“window-10 word-pairs”

Top 15 words for S				Top 15 pairs for S^{10}			
w	P_{TP}	P_{TN}	S	w_i, w_j	P_{TP}	P_{TN}	S^{10}
interact	0.76	0.12	0.64	with,interact	0.31	0.03	0.28
bind	0.63	0.14	0.49	interact,protein	0.21	0.02	0.19
domain	0.52	0.08	0.44	with,protein	0.25	0.12	0.13
complex	0.46	0.15	0.31	with,domain	0.14	0.01	0.13
between	0.01	0.29	0.28	interact,domain	0.13	0.01	0.12
with	0.9	0.65	0.25	bind,protein	0.15	0.03	0.12
activ	0.56	0.32	0.24	interact,between	0.12	0.01	0.11
yeast	0.28	0.04	0.24	protein–domain	0.12	0.01	0.11
between	0.38	0.16	0.22	bind–domain	0.11	0.01	0.1
associ	0.35	0.13	0.22	bind–with	0.11	0.01	0.1
protein	0.86	0.64	0.22	with–complex	0.12	0.02	0.1
region	0.26	0.06	0.2	associ–with	0.14	0.05	0.09
suggest	0.45	0.25	0.2	thi–interact	0.09	0.01	0.08
function	0.48	0.28	0.2	with–activ	0.1	0.04	0.06
regul	0.38	0.19	0.19	activ–protein	0.1	0.04	0.06



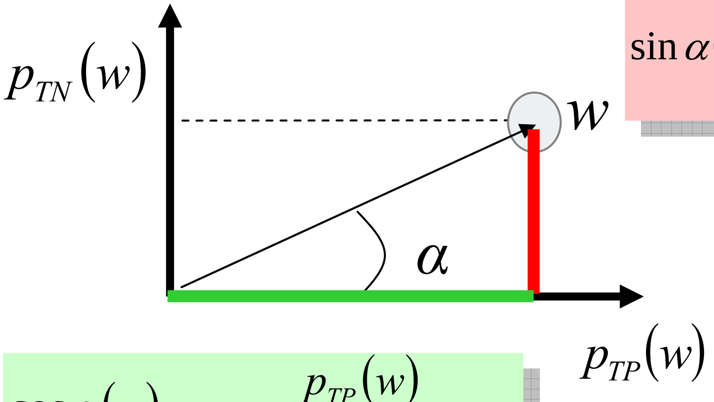
IAS: Run 2: Variable Trigonometric Threshold (VTT)

- Feature Selection
  - “Window-10 word pairs” plus number of protein mentions
    - Also “bigrams” for Run 3
- Linear Decision Model
  - λ: relative cost of features
  - β: number of protein mentions
- Results
  - Our most balanced run
    - **F1: 0.745, AUC: 0.7567, accuracy: 0.7371**

$$P(a) = \sum_{w \in a} \cos(\alpha(w))$$

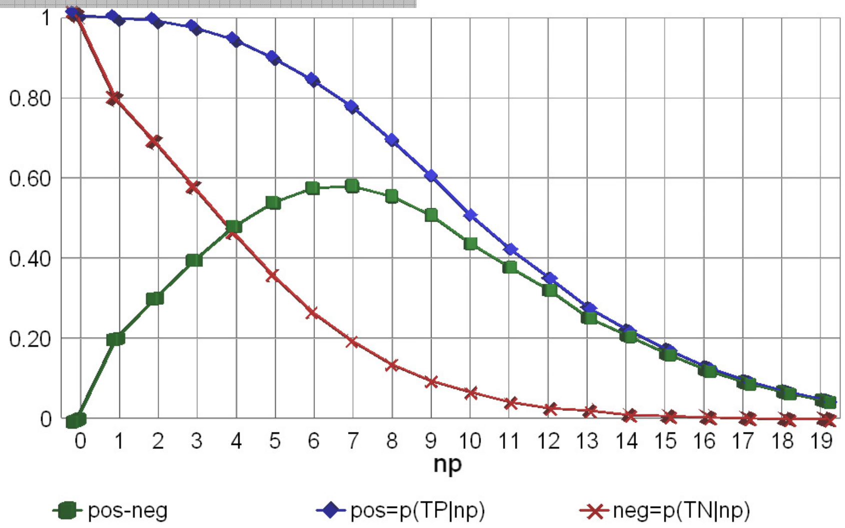
$$N(a) = \sum_{w \in a} \sin(\alpha(w))$$

$$\begin{cases} a \in TP & if \quad \dfrac{P(a)}{N(a)} \geq \lambda_0 + \dfrac{\beta - np(a)}{\beta} \\ a \in TN & otherwise \end{cases}$$



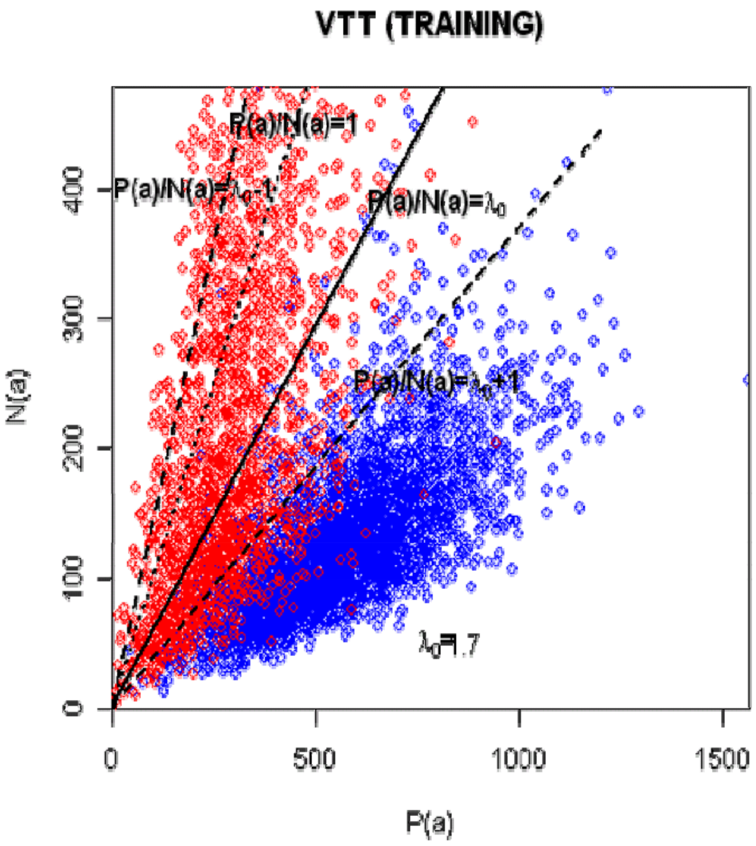
$$\sin\alpha(w) = \frac{p_{TN}(w)}{\sqrt{p_{TP}^2(w) + p_{TN}^2(w)}}$$

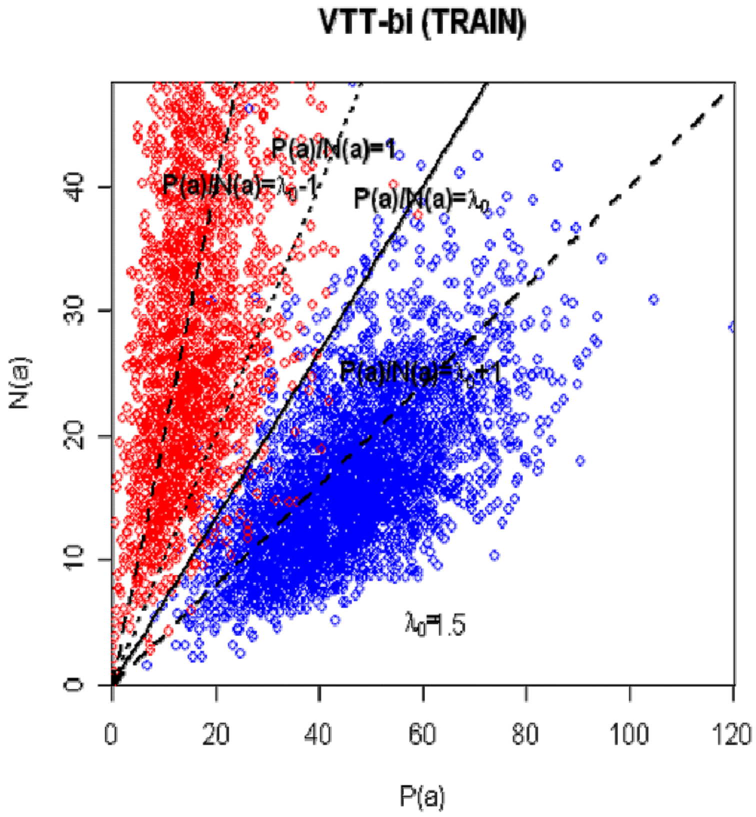
$$\cos\alpha(w) = \frac{p_{TP}(w)}{\sqrt{p_{TP}^2(w) + p_{TN}^2(w)}}$$



IAS: VTT Plots

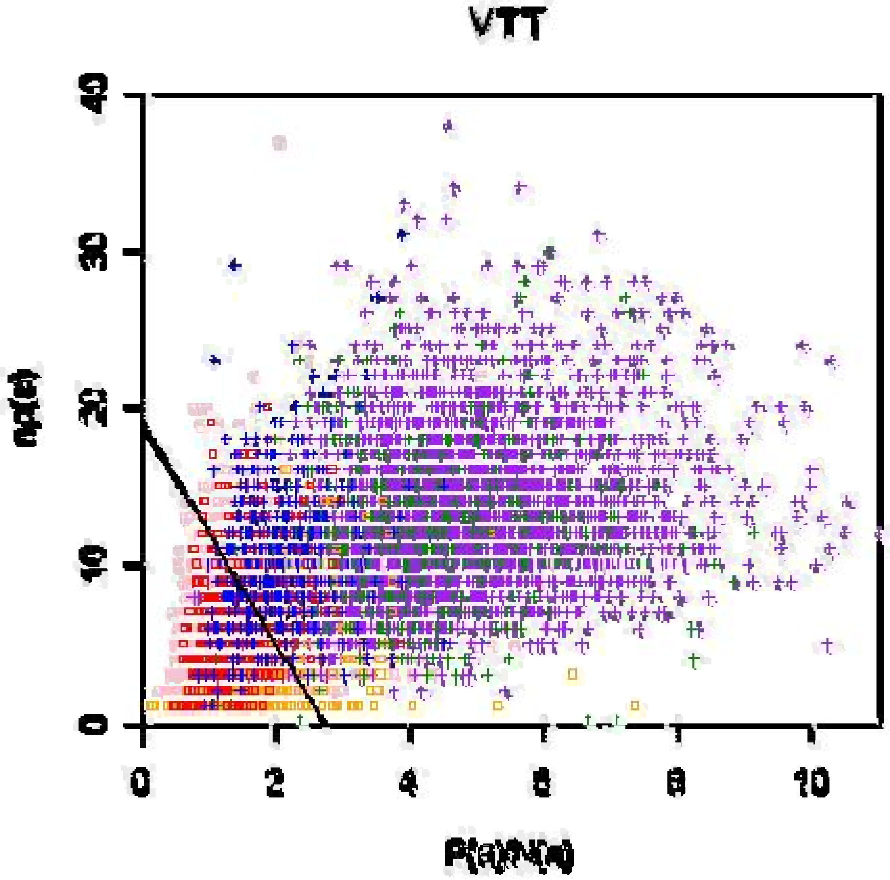
training data





IAS: VTT Plots

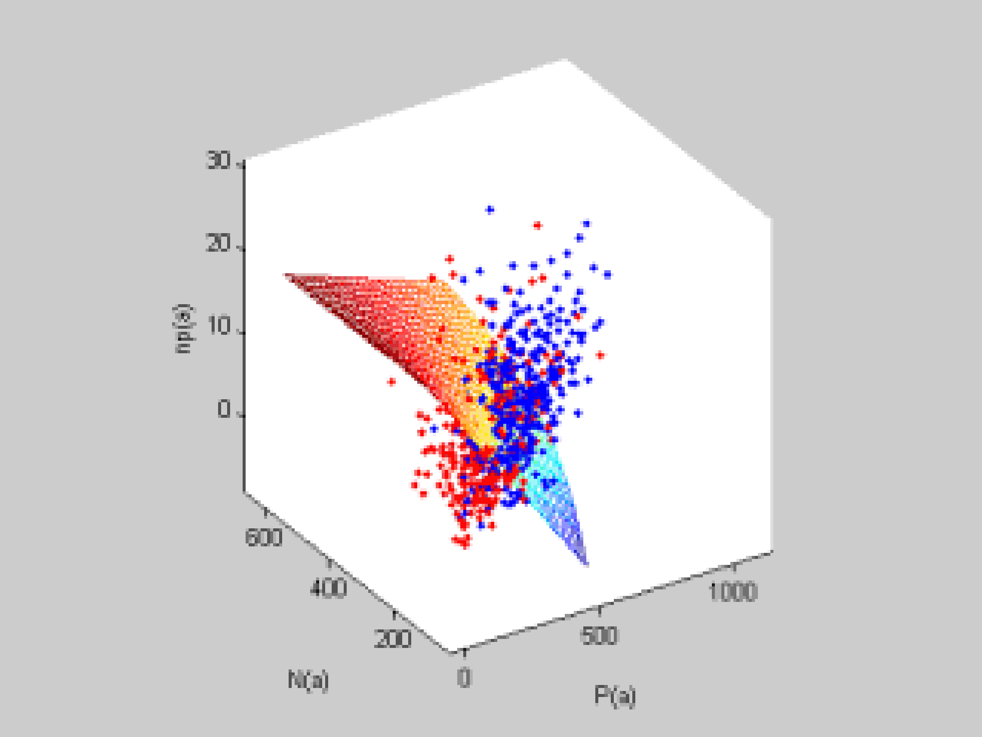
all data



Informatics
luis rocha 2007

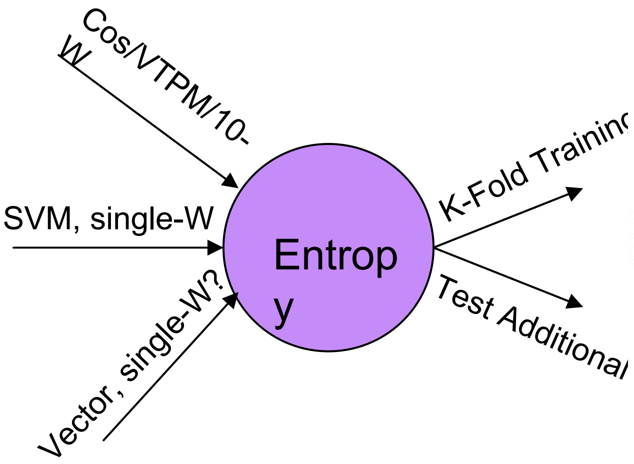
IAS: VTT video

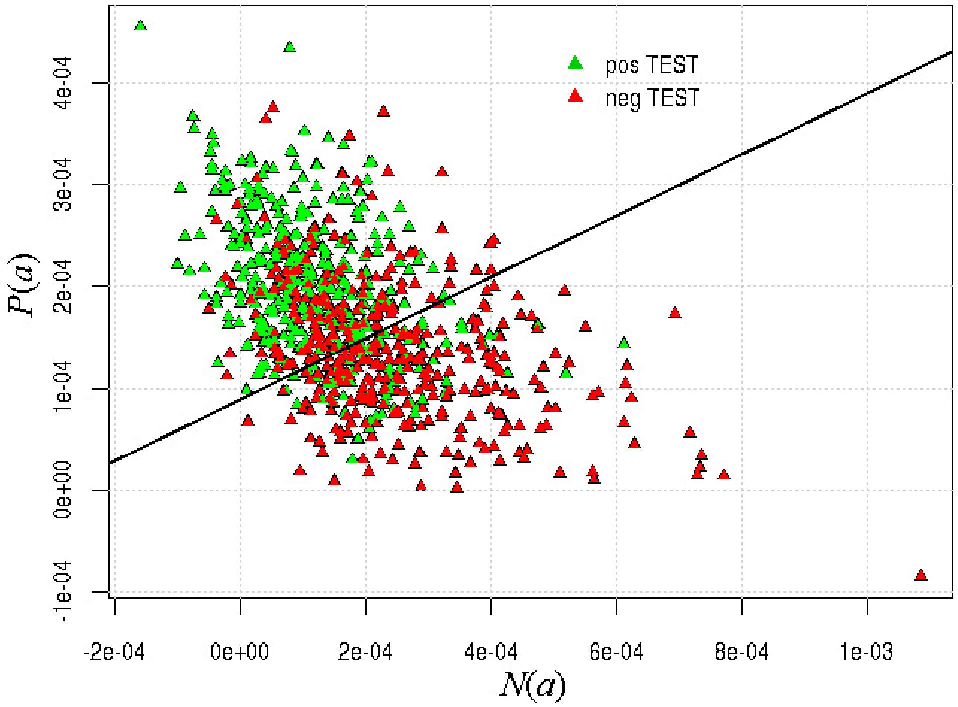
Test data



IAS: Run 3: SVD plus uncertainty integration

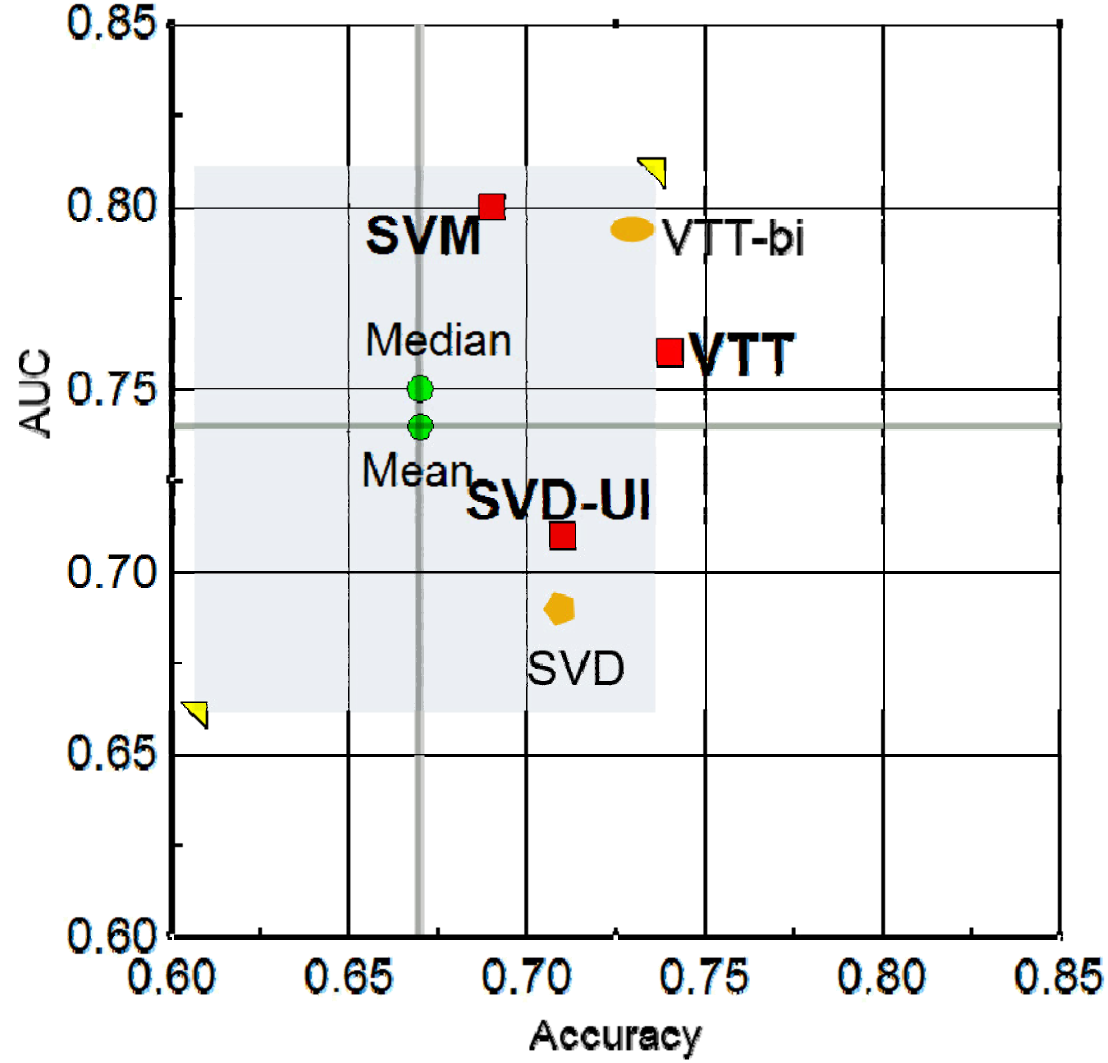
- Pool from 4 classification methods and integrate them via the "smallest neighborhood entropy" criteria on the space of words
  - SVD/LSA, VTT, VTT-bi, Fixed Threshold
  - Same feature set (650)
- Results
  - Same labeled prediction as SVD alone, different ranking
  - Our worst run (though still above the mean for accuracy)
  - No change with more features



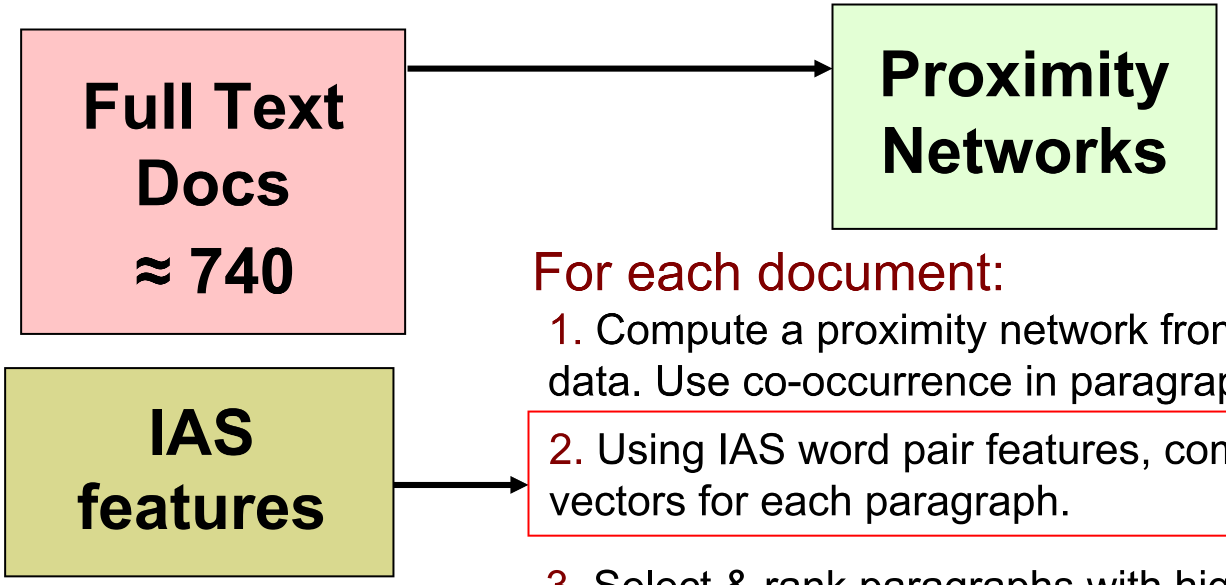


Informatics
luis rocha 2007

summary



ISS and IPS Subtasks



Full Text Docs

≈ 740

Proximity Networks

IAS features

For each document:

1. Compute a proximity network from co-occurrence data. Use co-occurrence in paragraph.

2. Using IAS word pair features, compute feature vectors for each paragraph.

3. Select & rank paragraphs with highest number of features with inverse frequency (protein mentions).

4. Select and rank protein interaction pairs in sentences of paragraphs in 3. Organisms restricted only by MeSH information. (ISS and IPS output)

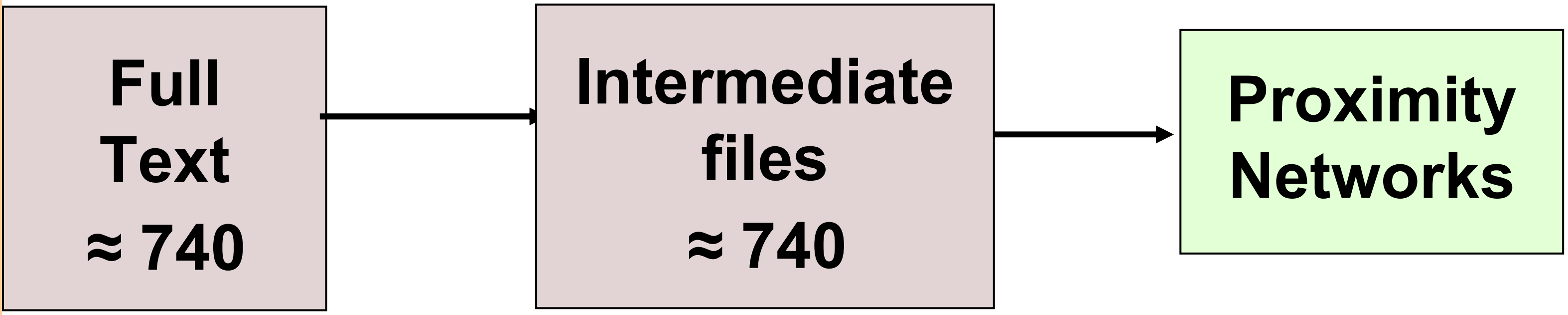
5. Expand protein pair sentences with closest words in proximity network (using biocreative 1 method).

6. Rank sentences obtained in 4, with (1) most word features, (2) same with expansion, (3) same with weighting factor. (ISS output).

Proximity networks

For each document:

Computed a proximity network from co-occurrence data. Used co-occurrence in paragraph. Removed stop words, stemmed text,TFIDF



Full Text ≈ 740

Intermediate files ≈ 740

Proximity Networks

$$R : P \times W, r_{i,j} \in \{0,1\}$$

P is the set of all *m* paragraphs in a document, and *W* is the set of all *n* words.

# paragraphs words w_i and w_j co-occur

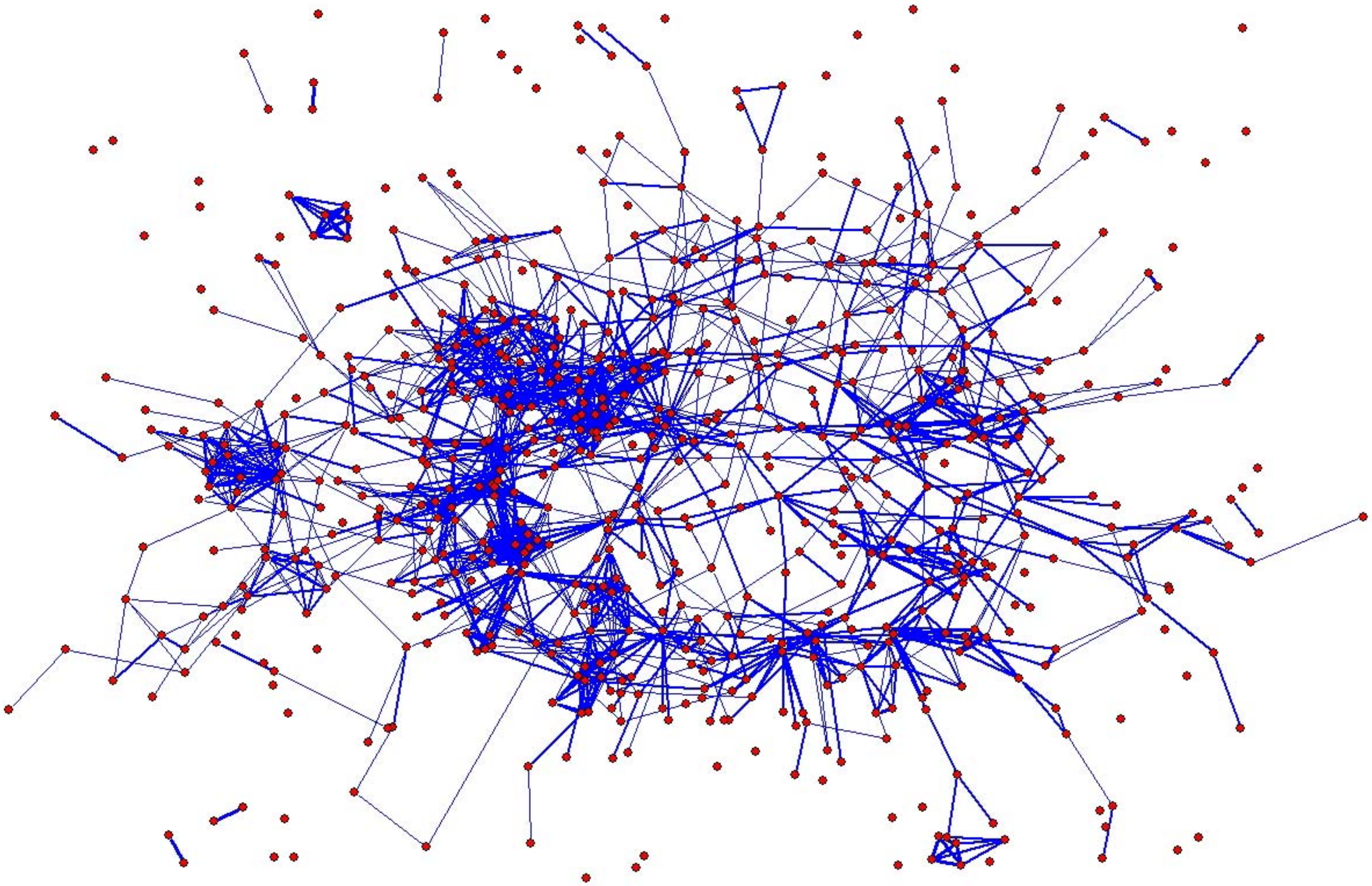
$$WPP\left(w_i, w_j\right) = \frac{\sum_{k=1}^{m} \left(r_{i,k} \wedge r_{j,k}\right)}{\sum_{k=1}^{m} \left(r_{i,k} \vee r_{j,k}\right)}$$

# paragraphs words w_i or w_j occur

INDIANA
UNIVERSITY

Proximity network

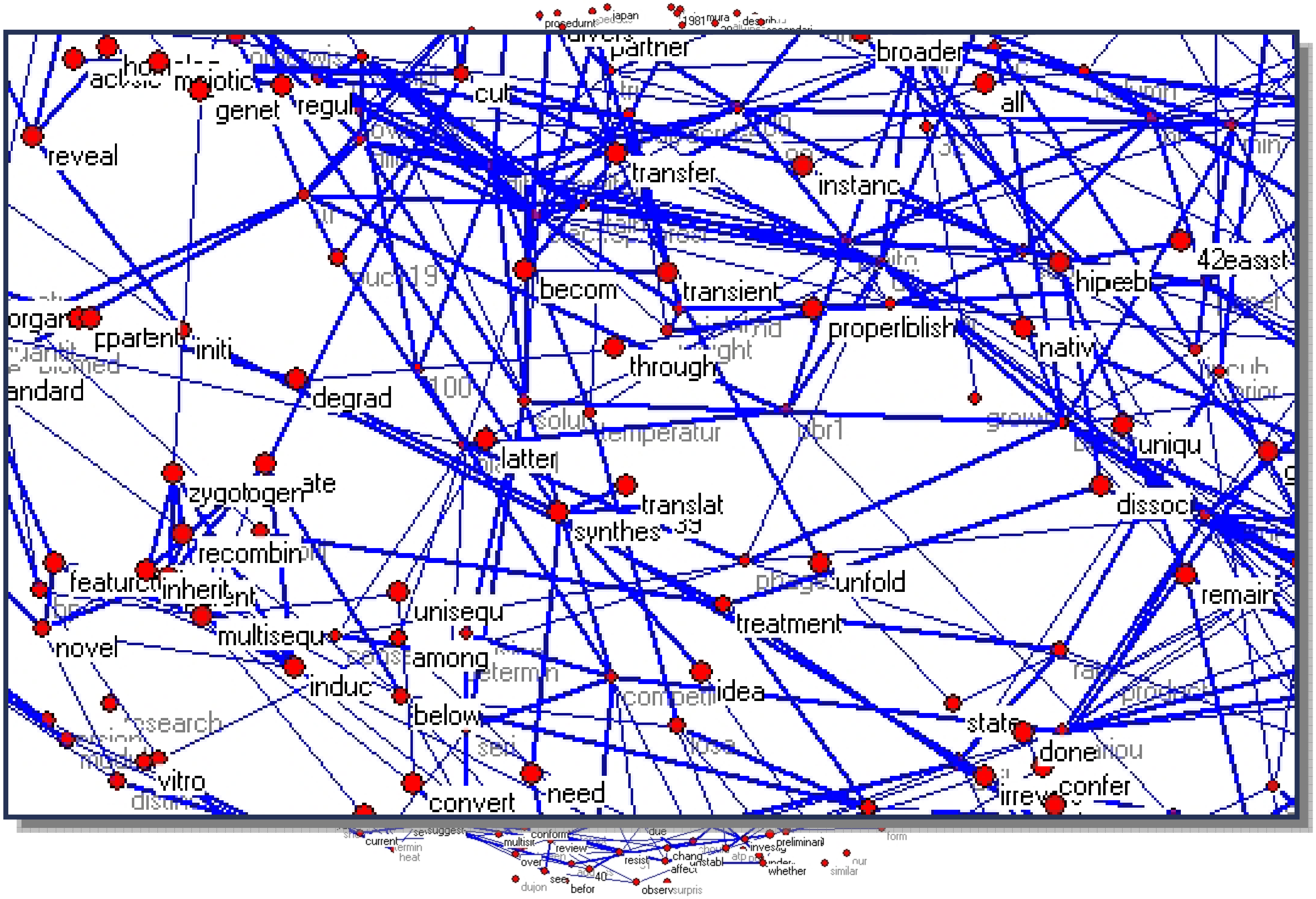
Document 10464305 (wpp>0.4)



Informatics
luis rocha 2007

Proximity network

Document 10464305 (wpp>0.4)



rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA
UNIVERSITY

IPS and ISS results

- IPS
  - No appreciative difference between three runs
  - recall was above the mean and median of all submissions (above one standard deviation). Precision very low
  - F-score near the mean and median
  - These results were true for both the identification of protein-protein interaction pairs
- ISS
  - Slight improvement with runs
  - Proximity expansion improved and so did weight factor with paragraph rank (from IPS) and protein mentions
  - Average performance
  - Again our results were in line with the averaged
    - matches (387) and unique matches (156) to previously selected above average (207.46 and 128.62)
    - we predicted many more passages (18371) and unique passages (5252) than the average (6213.54 and 3429.65, respectively), but with some cost to accuracy.
    - mean reciprocal rank of correct passages substantially higher than average (0.66 to 0.56)--- second group
- Both cases with higher Recall
  - Probably due to errors in feature calculation, and organism disambiguation

resources

- Web Resources
  - BLIMP: Biomedical and Literature (and text) Mining Publications
    - http://blimp.cs.queensu.ca/
  - BIONLP.ORG
    - http://www.bionlp.org/
    - http://www.ccs.neu.edu/home/futrelle/bionlp/
- Conferences
  - Pacific Symposium on Biocomputing
    - http://psb.stanford.edu/psb-online/
  - Intelligent Systems for Molecular Biology (ISMB) BioLink Special Interest Group
    - http://www.cs.queensu.ca/biolink05/
    - http://ismb2006.cbi.cnptia.embrapa.br/
  - BioCreative
    - http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
    - BMC Bioinformatics, **6** Suppl 1: http://www.biomedcentral.com/bmcbioinformatics/6?issue=S1
  - Linking Natural Language Processing and Biology (BioNLP'06)
    - http://compbio.uchsc.edu/BioNLP06/cfp.shtml
- Journals
  - Bioinformatics, BMC Bioinformatics, Jpurnal of Computational Biology, Nucleic Acids Research, PloS Biology, Journal of Biomedical Informatics, Nature Genetics, Genome Biology, Science STKE, etc.

important papers

- Overviews
  - H. Shatkay and R. Feldman [2003]. "Mining the biomedical literature in the genomic era: An overview". *Journal of Computational Biology*, **10**(6):821–856.
  - Jensen, L.J., J. Saric, and P. Bork [2006]. "Literature mining for the biologist: from information retrieval to biological discovery". *Nature Reviews Genetics* **7**, 119-129.
- Microarray automatic annotation tools
  - L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein [1999]. Med-Miner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**(6):1210–1214.
  - D. R. Masys, J. B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. [2001] "Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics, **17**(4):319–26.
  - T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig [2001]. "A literature network of human genes for high-throughput analysis of gene expression". *Nat. Genet.*, **28**(1):21–28.
  - P. Srinivasan [2001]. MeSHmap: a text mining tool for MEDLINE. *Proc AMIA* pp 642–646.
  - K. G. Becker, D. A. Hosack, G. Dennis, R. A. Lempicki, T. J. Bright, C. Cheadle, and J. Engel [2003]. "PubMatrix: a tool for multiplex literature mining". *BMC Bioinformatics*, **4**(1):61.
  - R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry. Gene clustering by Latent Semantic Indexing of MEDLINE Abstracts [2005]. *Bioinformatics*, **21**(1):104–115.
- Extraction of Gene-Disease Relations
  - Hristovski, D. , Peterlin, B. , Mitchell, J. A. & Humphrey, S. M. "Using literature-based discovery to identify disease candidate genes". *Int. J. Med. Inform.* **74**, 289–298 (2005).
  - H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii [2006]. "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning". *Pacific Symposium on Biocomputing* 11:4-15.

important papers

- **Validation of Literature Mining Techniques**
  - BioCreative Volume: *BMC Bioinformatics*, **6** Suppl 1.
  - *Proc. of the Second BioCreative Challenge Evaluation Workshop* (ISBN 84-933255-6-2).
- **Networks**
  - Marcotte, E. M. , Xenarios, I. & Eisenberg, D. Mining literature for protein–protein interactions. *Bioinformatics* **17**, 359–363 (2001).
  - Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I [2004]. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*. **20**(5):604-11
  - Hoffmann, R. & Valencia, A [2004]. A gene network for navigating the literature. *Nature Genet.* **36**, 664.
  - Rzhetsky, A. et al [2004]. *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data*. *J. Biomed. Inform*. **37**, 43–53.
  - Cooper, J. W. & Kershenbaum, A [2005]. "Discovery of protein–protein interactions using a combination of linguistic, statistical and graphical information". *BMC Bioinformatics* **6**, 143.
  - Ramani, A. K. , Bunescu, R. C. , Mooney, R. J. & Marcotte, E. M. "Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome". *Genome Biol.* **6**, R40.
  - Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A [2005]. "Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks". *Sci STKE*. 283:21
  - Hao, Y. , Zhu, X. , Huang, M. & Li, M [2005]. "Discovering patterns to extract protein–protein interactions from the literature: part II". *Bioinformatics 21*, 3294–3300.
  - Saric, J. , Jensen, L. J. , Ouzounova, R. , Rojas, I. & Bork, P. [2005]. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* **26**.
- **Protein Subcellular localization**
  - A. Hoglund, T. Blum, S. Brady, P. Donnes, J. San Miguel, M. Rocheford, O. Kohlbacher, and H. Shatkay [2006]."Significantly Improved Prediction of Subcellular Localization by Integrating Text and Protein Sequence Data". *Pacific Symposium on Biocomputing* 11:16-27.