Improving Information Extraction through Biological Correlation

Francisco Moreira Couto

Outline

- Motivation
- CAC method
- Case-Study
- Results
- Conclusions

Text Mining Motivation

- 400.000 new articles in PubMed each year
- Reading 10 articles per day:
  Takes **112** years to read those articles

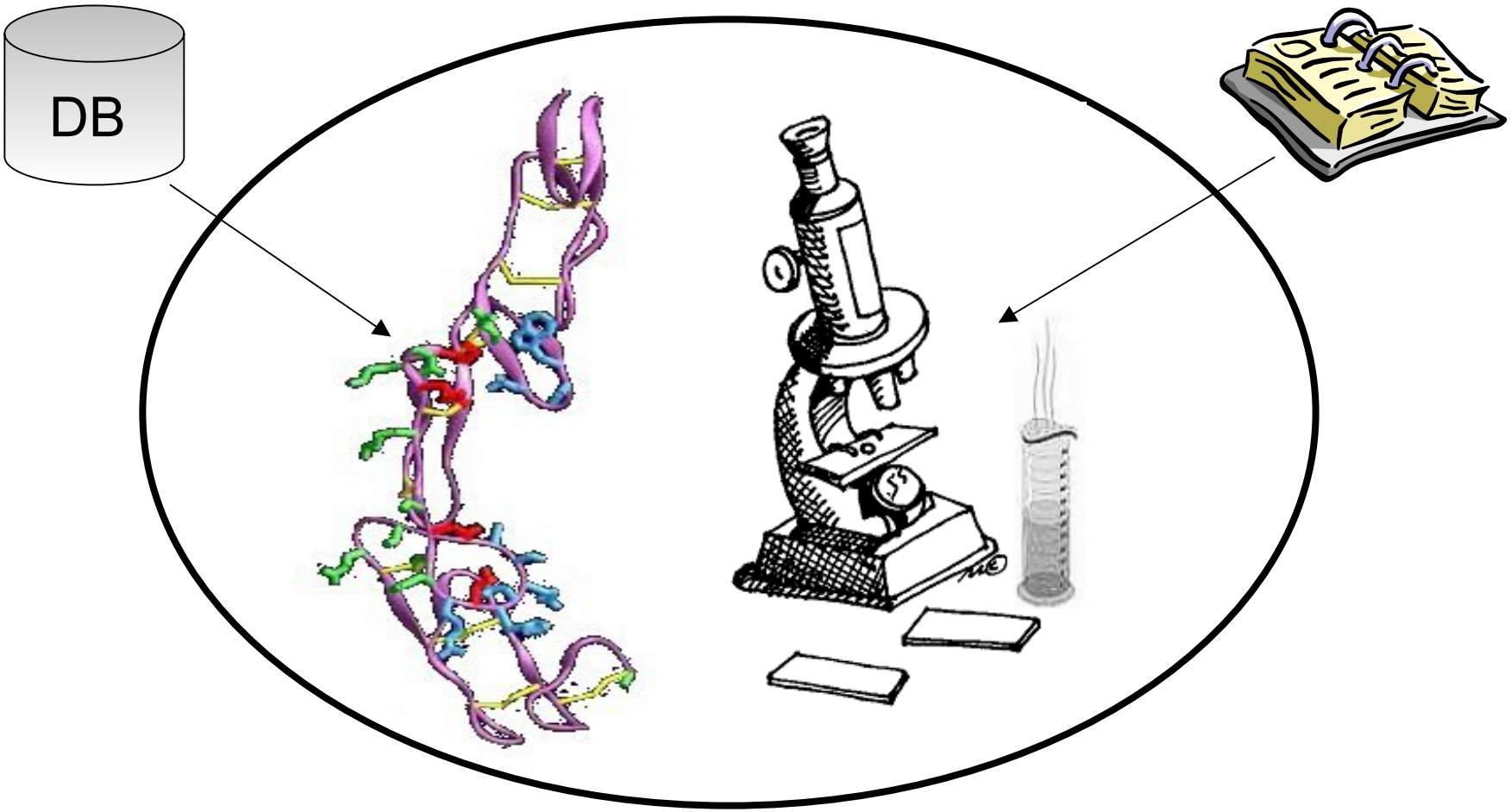
State-of-the-art

- Text mining in Molecular Biology produces weaker results than in other areas
- KDD Cup 2002:
  - Statistical text classification systems (Naïve Bayes, SVMs) without domain knowledge achieved poor results
  - Domain knowledge is crucial, but has high costs

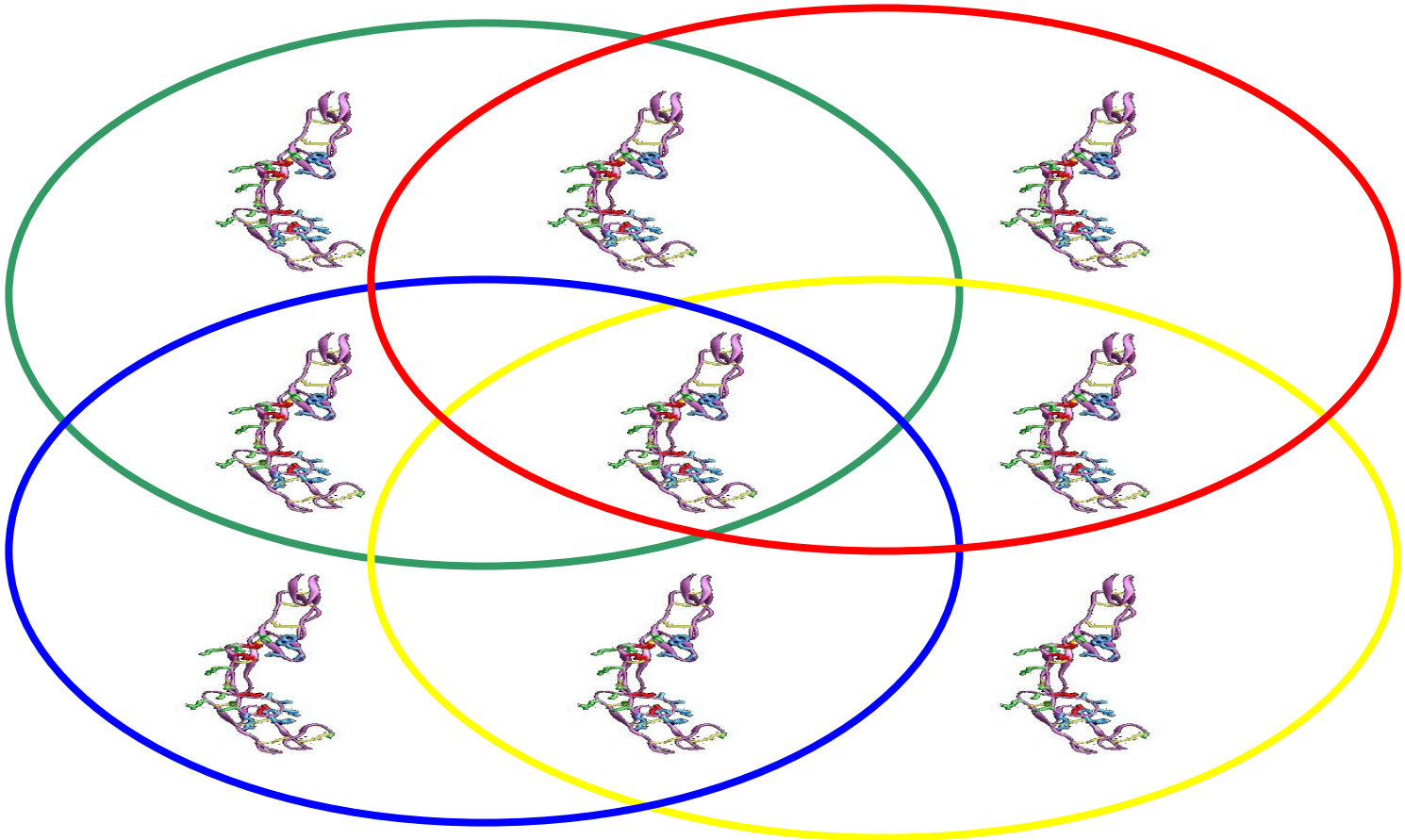
Outline

- Motivation
- CAC method
- Case-Study
- Results
- Conclusions

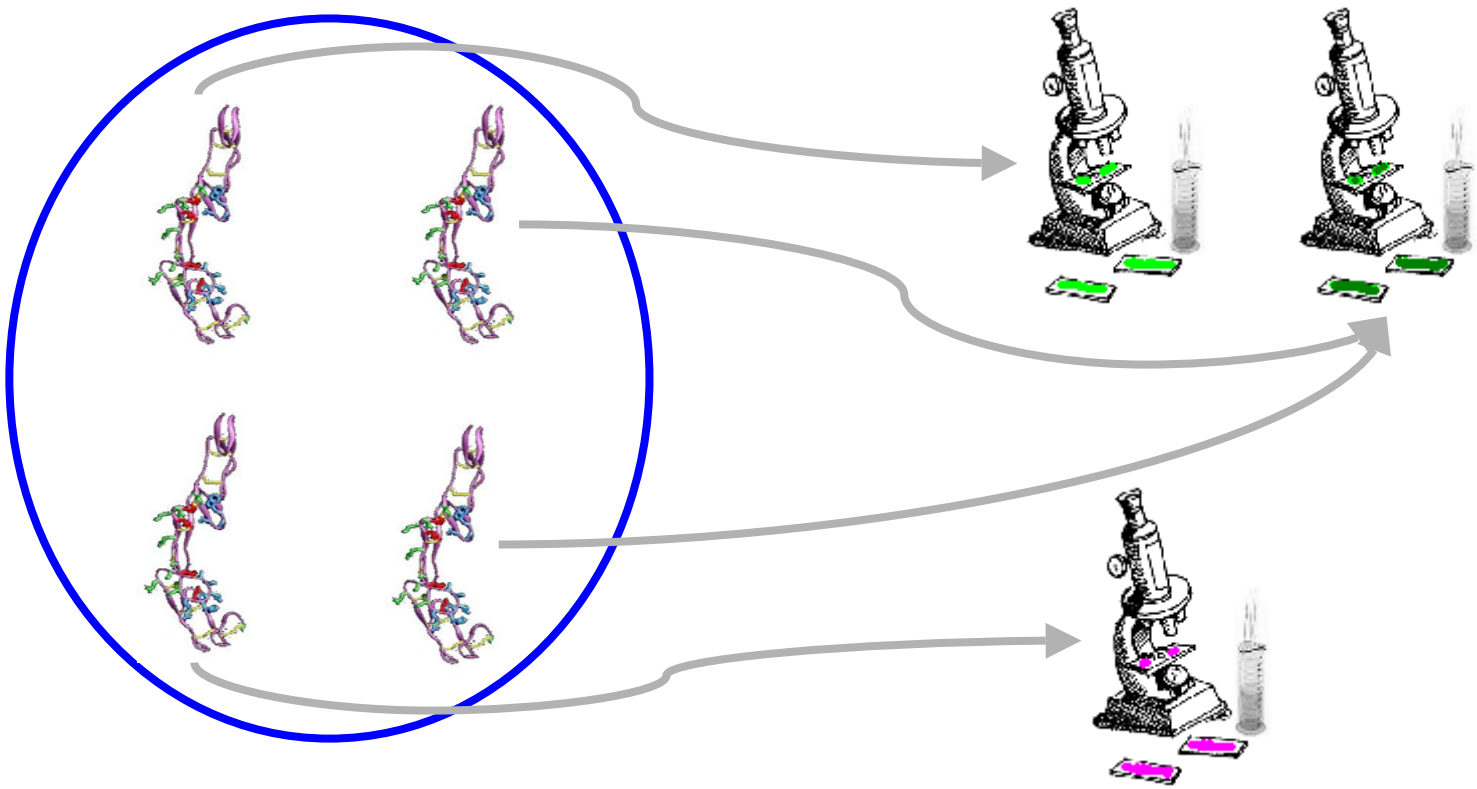
Annotation



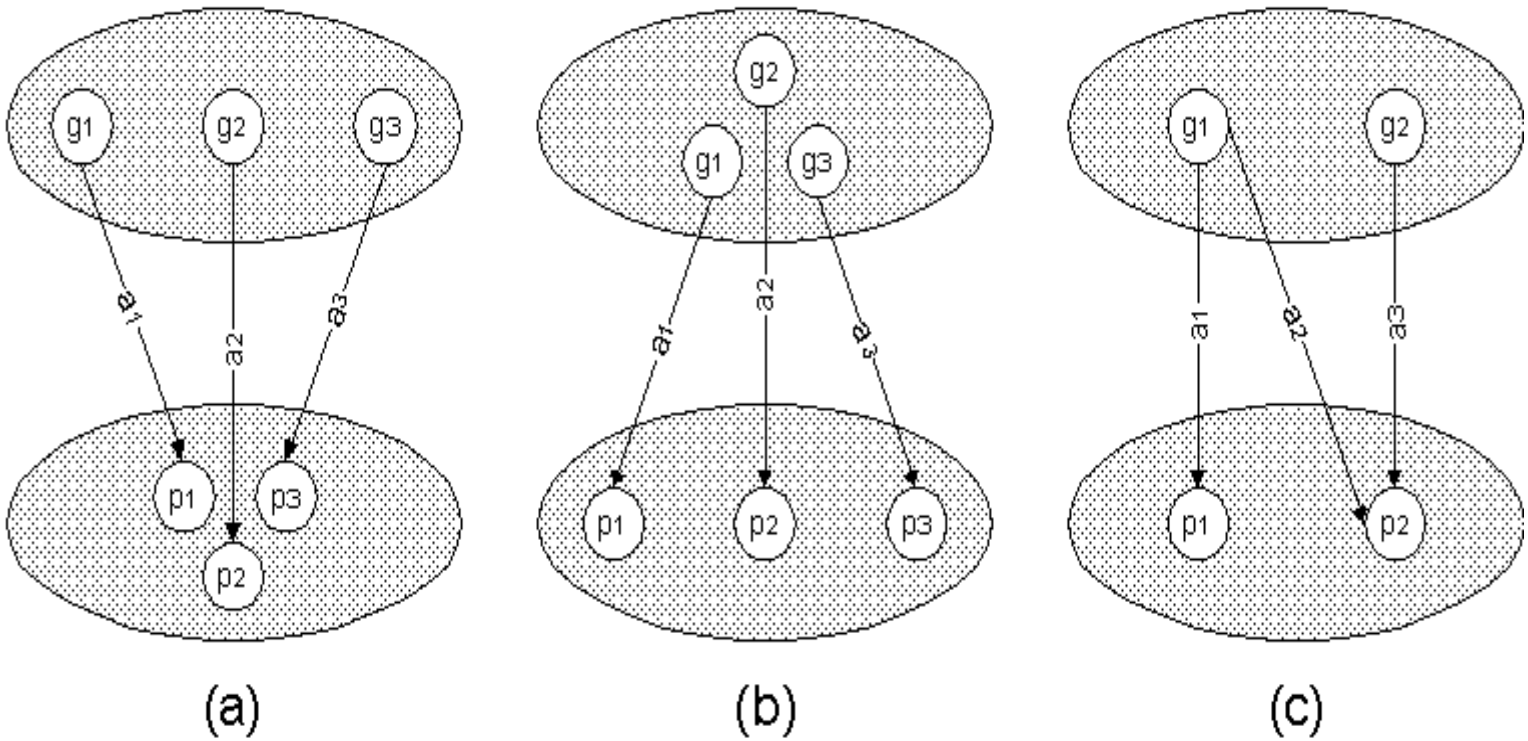
Families



Correlation



Convergence Examples

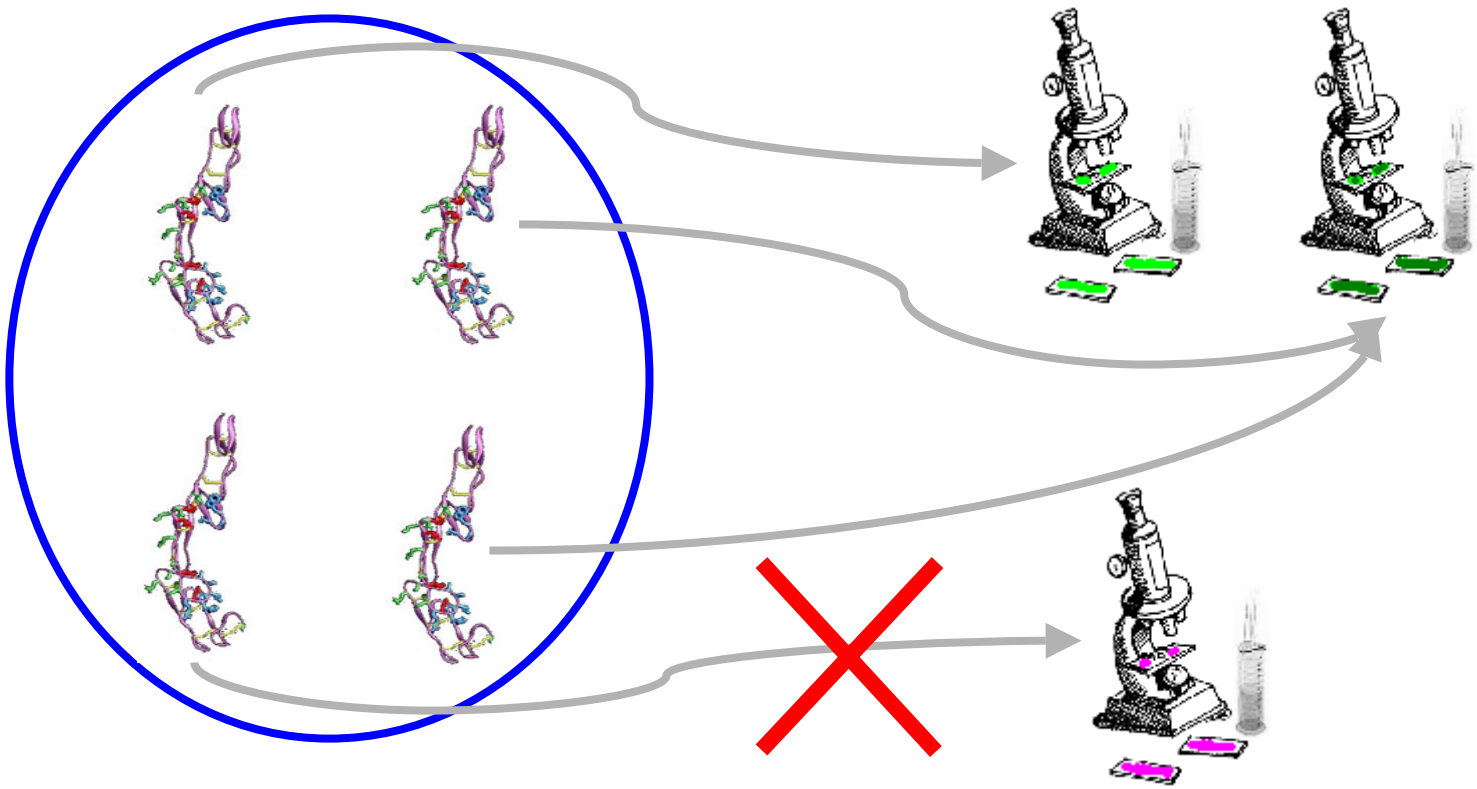


Convergence Formulas

$$\Gamma((g_1, p_1), (g_2, p_2)) = \frac{\Delta(g_1, g_2)}{\Delta(p_1, p_2)}$$

$$\mathcal{D}_h(a_0) = \#\{a_x : a_x \in \mathcal{A}_f \wedge \Gamma(a_0, a_x) \geq h\}$$

Selection of Annotations



Structural Distances

- Sequence

  (MRRTELSVRG... ~ MRRTELSVRG…)

  – BLAST

- Modular structure

  – Number of common modules

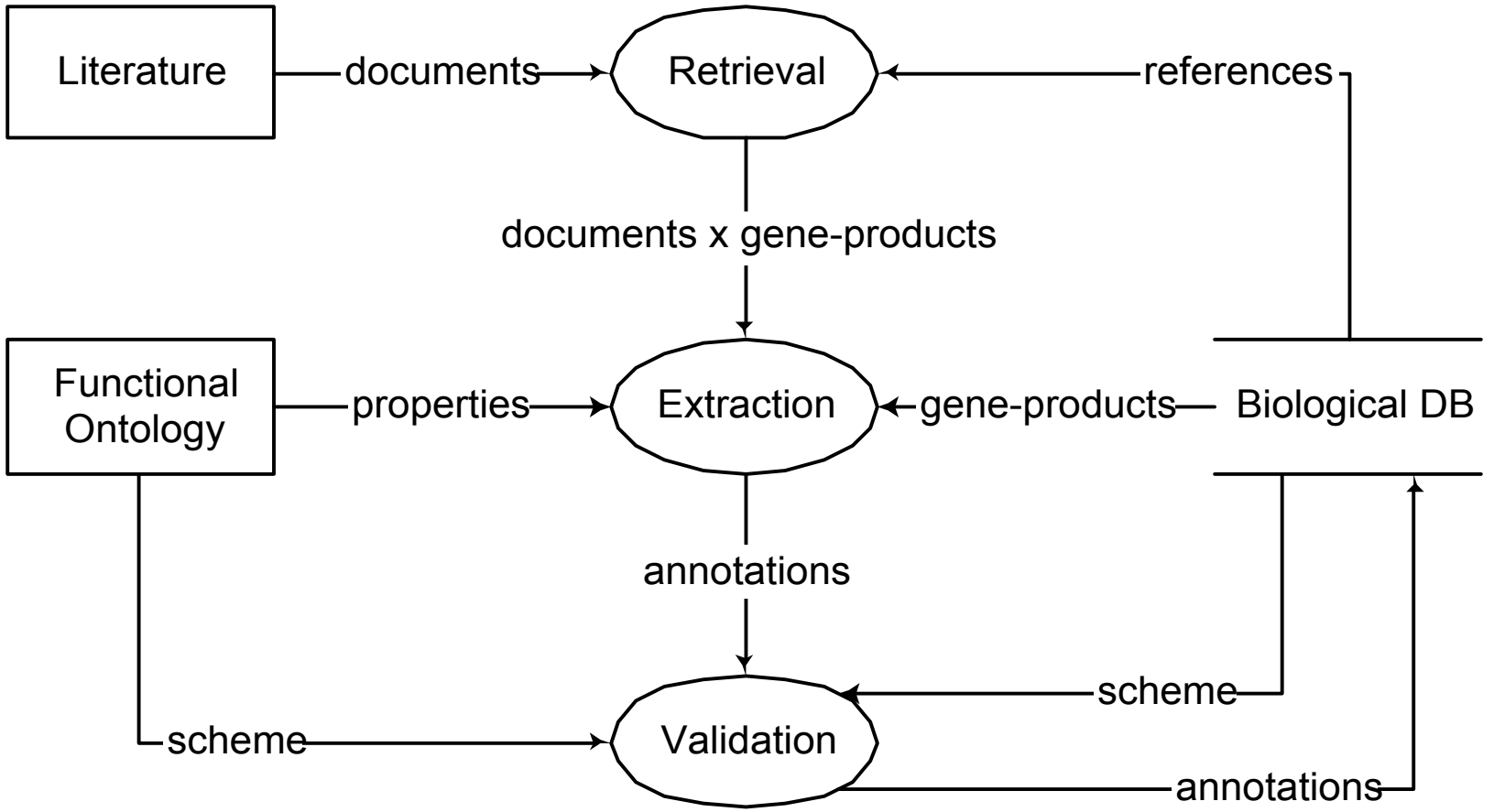
Functional Distances

- Semantic Similarity Measures
  - Information content
    - Inversely proportional to each term frequency
    - Similarity increases with the information content of a shared parent
  - Conceptual Distance
    - Depends on the path length between the two nodes

Outline

- Motivation
- CAC method
- Case-Study
- Results
- Conclusions

IE System


Literature
documents
Retrieval
references
documents x gene-products
Functional Ontology
properties
Extraction
gene-products
Biological DB
annotations
scheme
Validation
scheme
annotations

CAZy (Carbohydrate Active enZYmes)

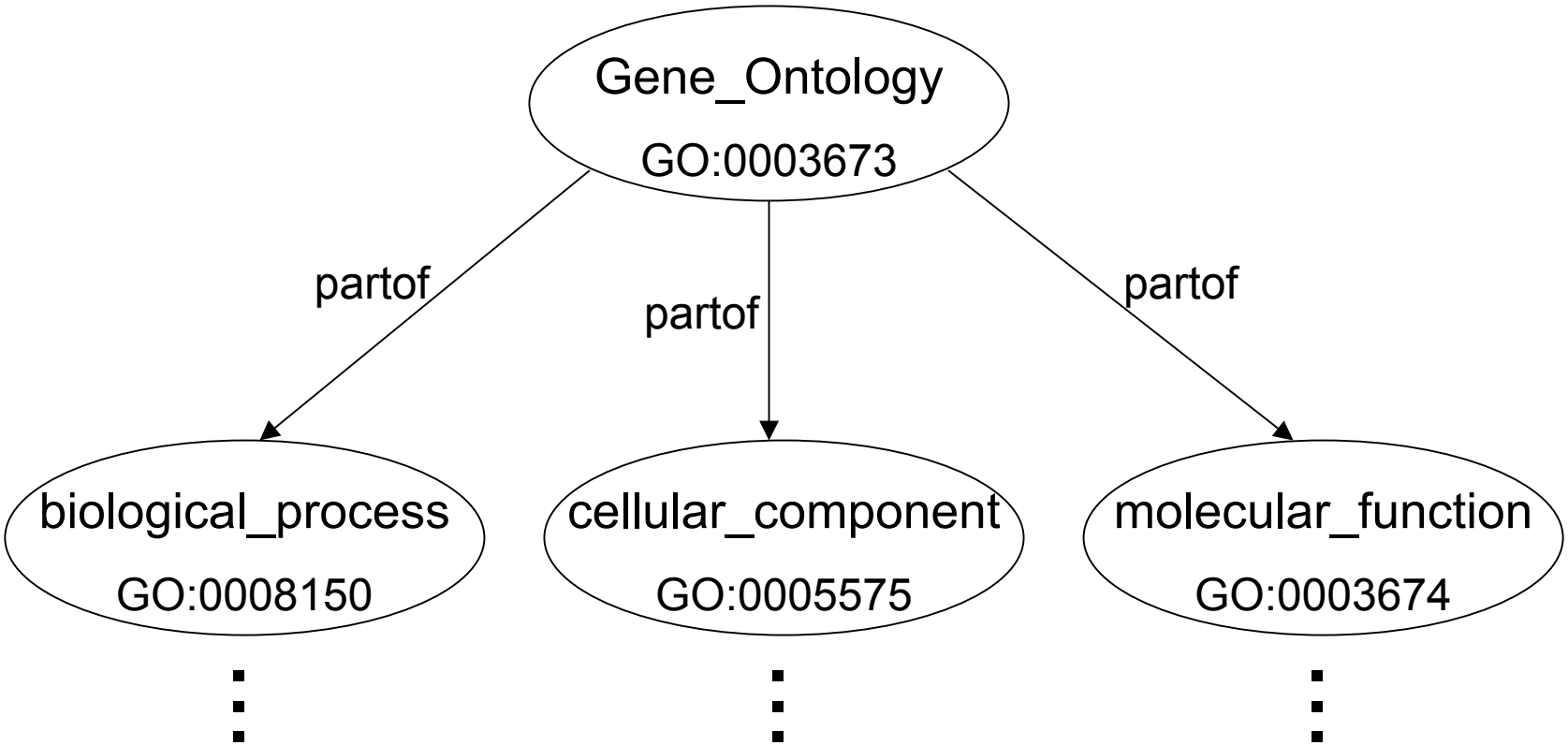
<< phpCAZy >> CAZyModO

CAZyModO - Glycoside Hydrolase Classification | Family GH48 | Modular Organization | Go

CAZy Family	Glycoside Hydrolase Family 48
Known Activities	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91).
Mechanism	Inverting
Catalytic Nucleophile/Base	Not known
Catalytic Proton Donor	Not known
3D Structure Status	Available (see PDB)
Note	formerly known as cellulase family L.
Relevant Links	InterPro; PRINTS
Statistics	CAZyModO(12)

#ac	Protein		Organism	Modular Organisation	Status
154	cellulase CelA		*Anaerocellum thermophilum*	1 GH9 443 445CBM3612 613 681 682CBM3834 835 880 881CBM31034 1035 1080 1081 GH48 1711	
1497	CelA		*Caldicellulosiruptor saccharolyticus*	1 23 24 GH9 466 467CBM3642 643 700 701CBM3857 858 903 904CBM31060 1061 1112 1113 GH48 1742	
1591	cellobiohydrolase B	*Cel48A*	*Cellulomonas fimi*	1 33 34 53 54 GH48 [illegible] 860FN3 [illegible] 904FN3 [illegible] 991FN3 [illegible] 989CBM2 1090	
10828	ORF CAC0911		*Clostridium acetobutylicum ATCC 824*	1 GH48 664 661DOC1726	
1716	CelF	*Cel48A*	*Clostridium cellulolyticum*	1 29 30 GH48 660 665DOC1722	3D
1730	exoglucanase S		*Clostridium cellulovorans*	1 32 33 GH48 643 644DOC1703	
1733	CelD		*Clostridium josui*	1 29 30 GH48 650 651DOC1719	

GO (Gene Ontology)



Outline

- Motivation
- CAC method
- Case-Study
- Results
- Conclusions

Information Retrieved

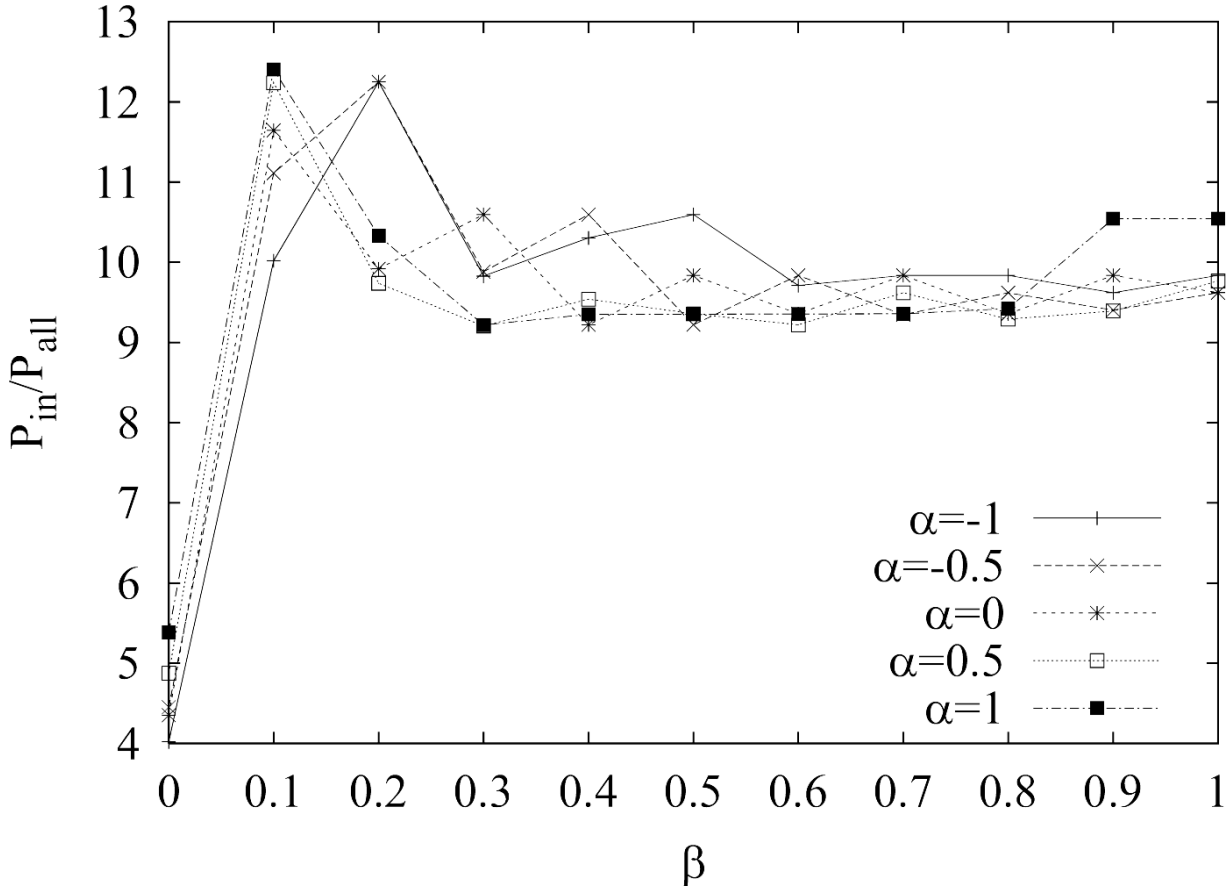
	References	Documents
GenBank(GenPept)	22849	4575
SwissProt	8998	4006
PDB	3561	785
Total		6377

13869 annotations; 6918 enzymes;
1342 terms

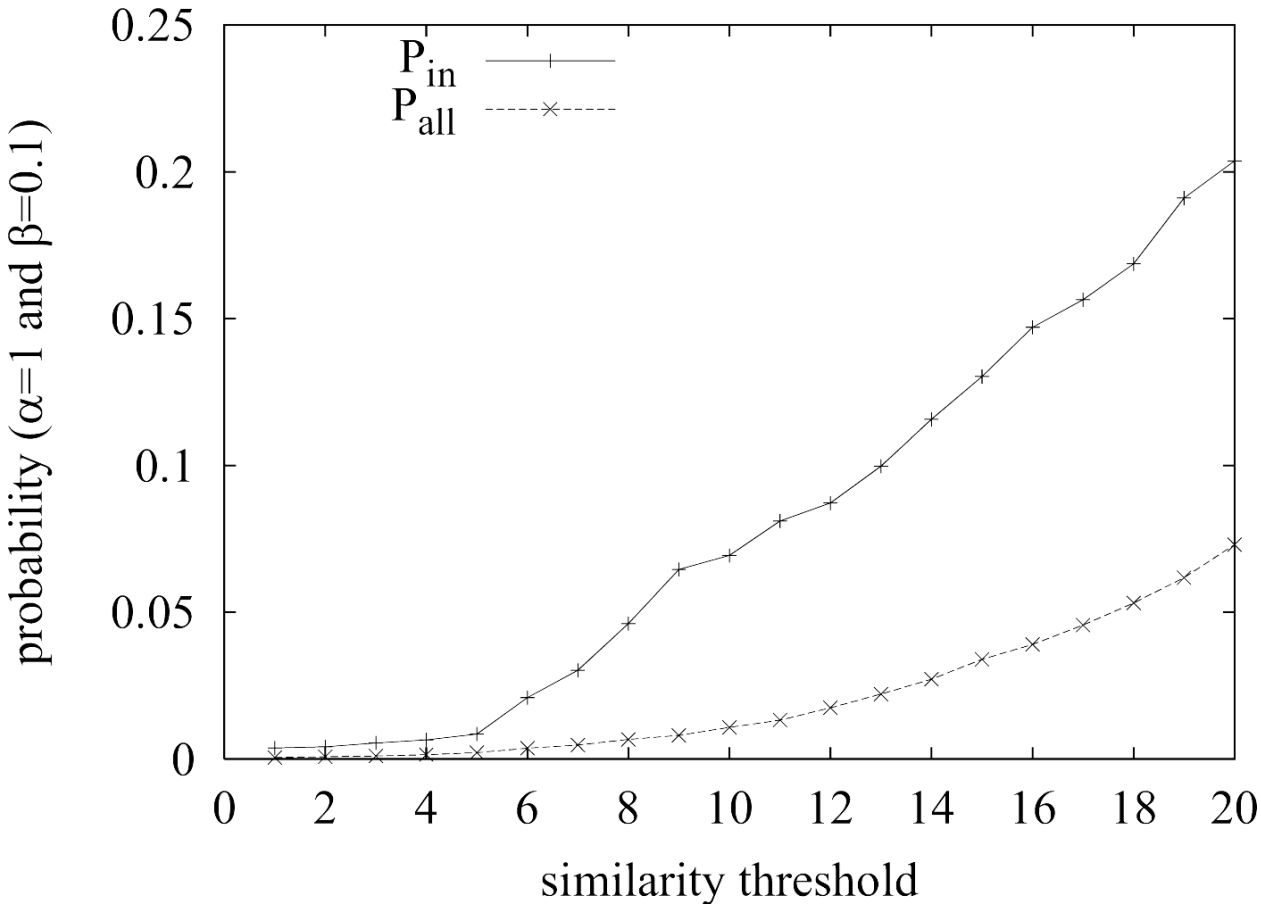
Probabilities

- Similar terms is when $\Delta(t_1,t_2)\leq k$
- Pin = probability of extracting similar terms in a family
- Pall = probability of extracting similar terms in general
- If Pin>>Pall then there is a correlation
- Computed for 90 families

Pin over Pall



Pin against Pall



Conceptual Dist. Parameters

- α controls the node depth contribution
- β controls the density factor contribution
- Assumptions:
  - Similarity increases with the node depth
  - Similarity increases with the number of nodes with the same parent node.

Outline

- Motivation
- CAC method
- Case-Study
- Results
- Conclusions

Contributions

- Correlation between structure and function was computed from automatically extracted annotations
- Conceptual distance integrated with information content achieved higher levels of correlation
- We used this correlation metric to propose an effective method that improves the precision of extracted information
- The method does not need training data or grammars expressing domain specific rules

Future Achievements