



Streaming Facts from Scientific Publications to the Scientist

International Symposium on Biomedical Informatics Barcelona

June 26th, 2007

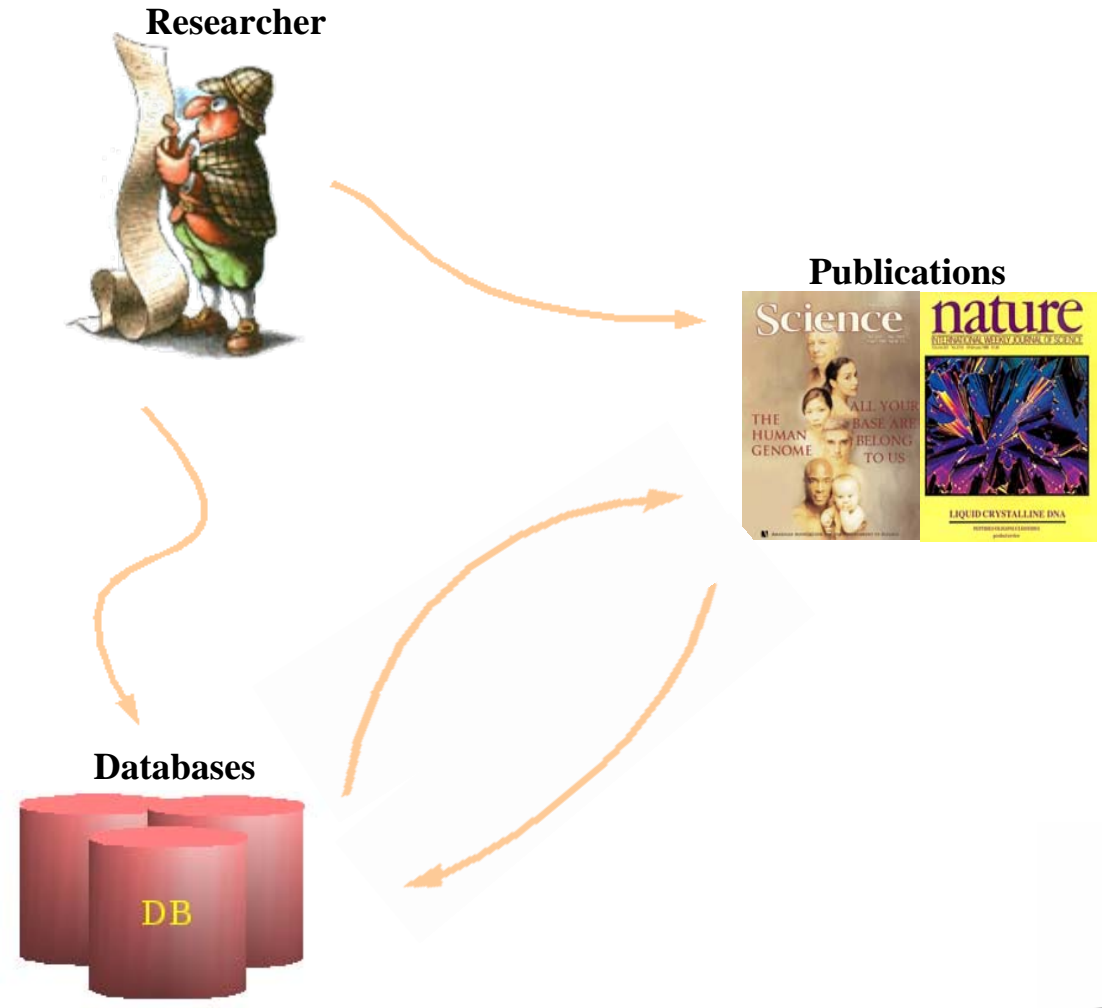
Dietrich Rebholz-Schuhmann, MD, PhD
EBI, WT Genome Campus
Hinxton, Cambridge, U.K.



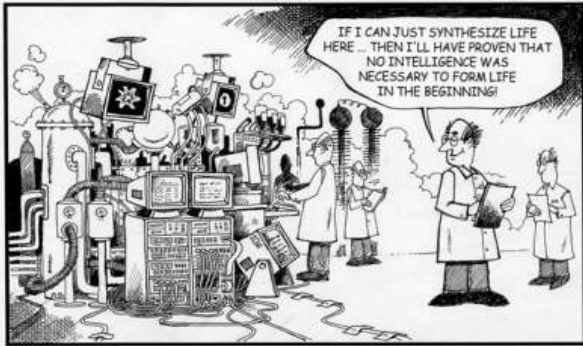
This is the future ...

- Scientists state their facts clearly in their publication.
- Publications are delivered right after the paper acceptance into a repository in the public domain.
- The repository is automatically fully integrated with all other data resources.
- Automatic agents grab the facts and deliver them to the scientist.
- Huge amounts of data is contained in the literature

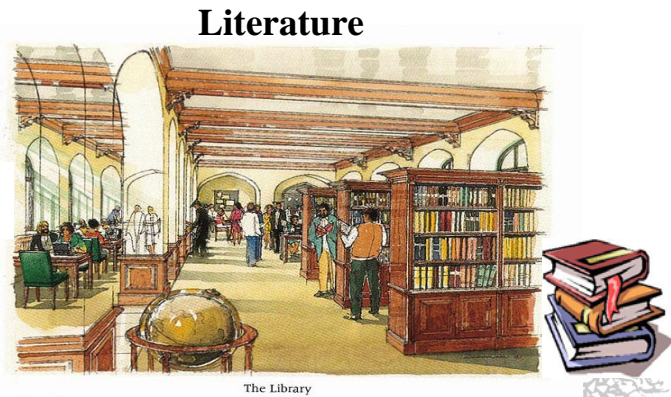




Experiments



Biologists link facts from the literature to content in electronic databases





What is required to make the future work?

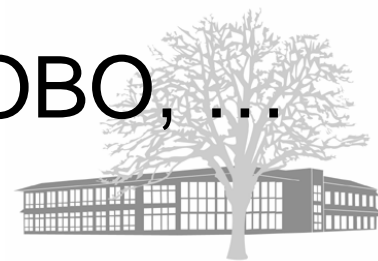
- Use large sets of terminologies
(=> text features).
- Make use of text processing engines.
- Harmonise document formats and processing techniques.
- Learn how concepts and facts are represented in the literature
(=> syntax and semantics).
- Integrate the engines into applications.
- Make the publishers realize ...

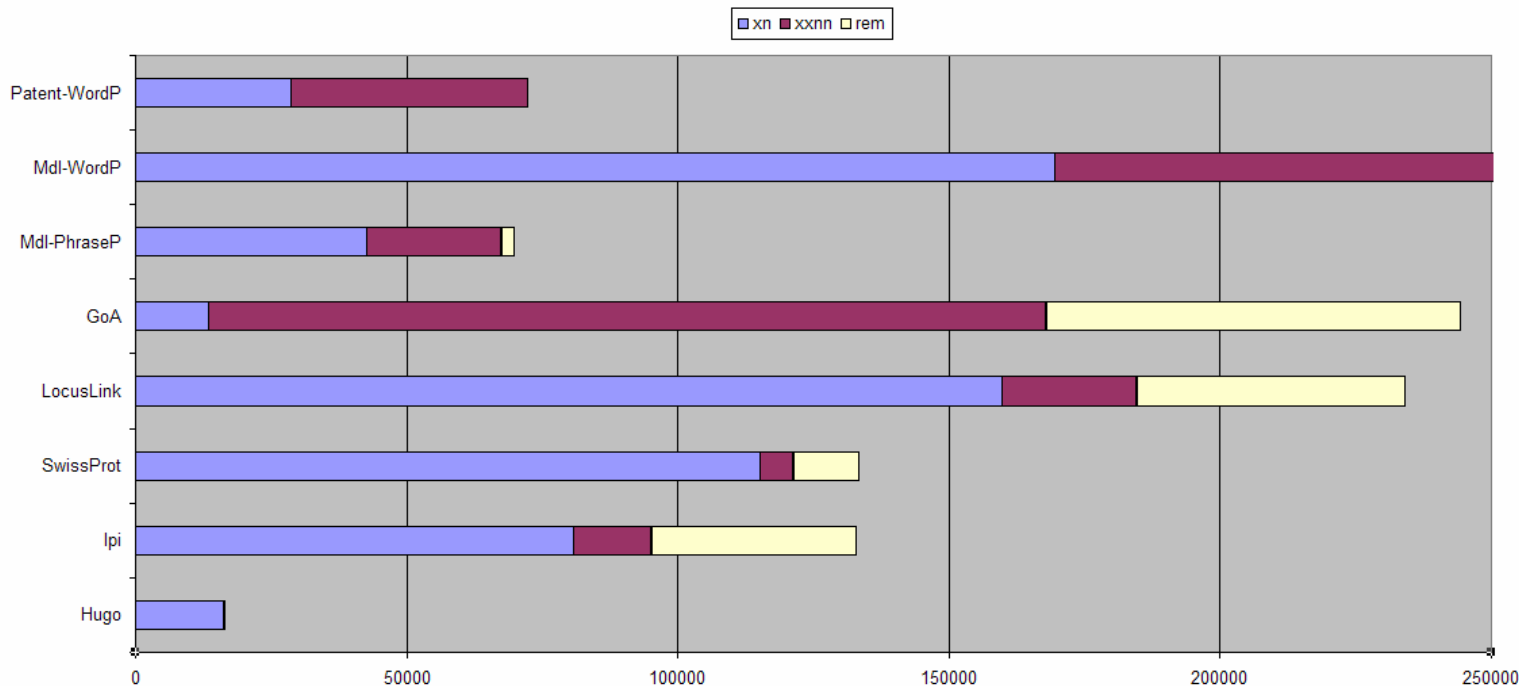




Terminologies ...

- Protein/gene names: UniProtKb, EntrezGene, ...
- Chemical Entities: ChEBI, OSCAR3, ...
- Diseases and Syndromes: UMLS, Snowmed-CT, ...
- Species: NCBI taxonomy
- Other concepts: Gene Ontology, OBO, ...





The number of unique PGNs (in lowercase only) in Hugo, Ipi, Swiss-Prot, LocusLink, GoA, and in the sets of PGNs from Medline (phrase patterns: Mdl-PhraseP; word patterns: Mdl-WordP), and from the European patent abstracts (Patent-WordP) is shown. GoA is clearly different from the other databases, since GoA contains mainly descriptive names.

xn: e.g. HZF1

xxnn: e.g. ErbA-related protein 3

rem: e.g. embryonal long terminal repeat-binding protein (ELP)-1



Statistics on the term repository

- Contents by semantic category

<i>Clusters</i>	
<i>SEMTYPE</i>	<i>Count</i>
ChebiChem	13,473
Enzyme	3,970
GeneProt	232,258
SpeciesTax	367,565
Verb terms	730

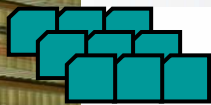
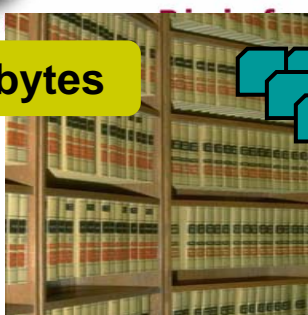
- Contents by morphological category

<i>Variants</i>	
<i>SEMTYPE</i>	<i>Count</i>
ChebiChem	57,581
Enzyme	11,202
GeneProt	1,931,786
SpeciesTax	441,993
Verb terms	To be estimated





Gigabytes



GO



Swiss-Prot
incl. Acronym
detection +
disambiguation



Species



Drugs

Link / Feed into
Public Data



By the Rebholz group

Whatizit

Your input:
Text or
PMIDs or
UniProtKb Ids

1 Place your text/query here:

Input
type

2 The text is:

- Plain Text
- Plain Text
- A Lucene Query
- A list of PMIDs
- A list of UniProt ids

3 Select a pipeline:

- whatizitSwissprotGo
- whatizitSwissprotGo
- whatizitEBIMed
- whatizitGO
- whatizitSwissprotDisease
- whatizitDisease
- whatizitDrugs
- whatizitOrganisms

Tag text

Annotation
modules

Whatizit is a text process... you to do te





















Name of pipeline for annotation

By the Rebholz group

Description

Up and running

Pipeline	Description	Available
 whatizitSwissprotGo	Swissprot protein names and Gene Ontology (GO) terms	
 whatizitEBIMed	Swissprot protein names, Go terms, Drugs from MedlinePlus and Organisms from the NCBI Taxonomy	
 whatizitGO	Gene Ontology (GO) terms.	
 whatizitSwissprotDisease	Experimental. Swissprot protein names and diseases	
 whatizitDisease	Experimental. Disease names.	
 whatizitDrugs	Drug names from MedlinePlus.	
 whatizitOrganisms	Organisms from the NCBI Taxonomy..	
 whatizitIntact	IntAct's database controlled vocabulary, with species, swissprot protein names, drug names from MedlinePlus and GO terms.	
 whatizitSwissprot	Swissprot Terms	





The activated **glucocorticoid receptor** forms a **complex** with **Stat5** and enhances Stat5-mediated transcriptional induction.

PMID: 98444965

noise
AR and **CBP** can physically **interact** in vitro as was shown in glutathione S-transferase pulldown assays.

PMID: 99041948

don't miss pair
At no T3 concentration did both **NCoR** and **SRC-1** **bind** to **wt TR**, indicating that their binding to TR was mutually exclusive.

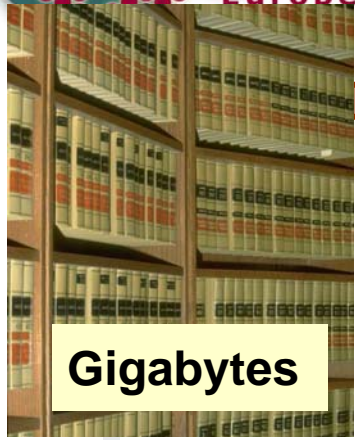
PMID: 99023934

Furthermore, **CREB-binding protein** also significantly increased activation by **GHF-1**, and **both** proteins **associated** in vitro.

PMID: 98434578

*difficult reference,
currently missed*





Gigabytes

```
<?xml ...?><textmine>...</textmine>
```

input filter



```
<abs><SENT>...</SENT>...</abs>
```

Paragrapher
sentencer



```
<gene>...</gene>  
<species>...</species>  
<disease>...</disease>
```

domain dictionaries

```
<tok>  
  <sur>globulins</sur>  
  <lem cat="n" mor=":m">globulin</lem>  
</tok>
```

POS tagger

```
<PPI>  
  <NP></NP><VERB></VERB><NP></NP>  
</PPI>
```

chunk parsing

```
<a href=...></a>  
<span style=...></span>
```

XSLT





Where do
I look?

Could I find
documents?

What do
they contain?

Results?



Medline
Patents
Own Text
UniProtId

Selection of
documents

Extraction
modules
GO terms
Genes/proteins
Diseases
Chemical entities
Protein interactions

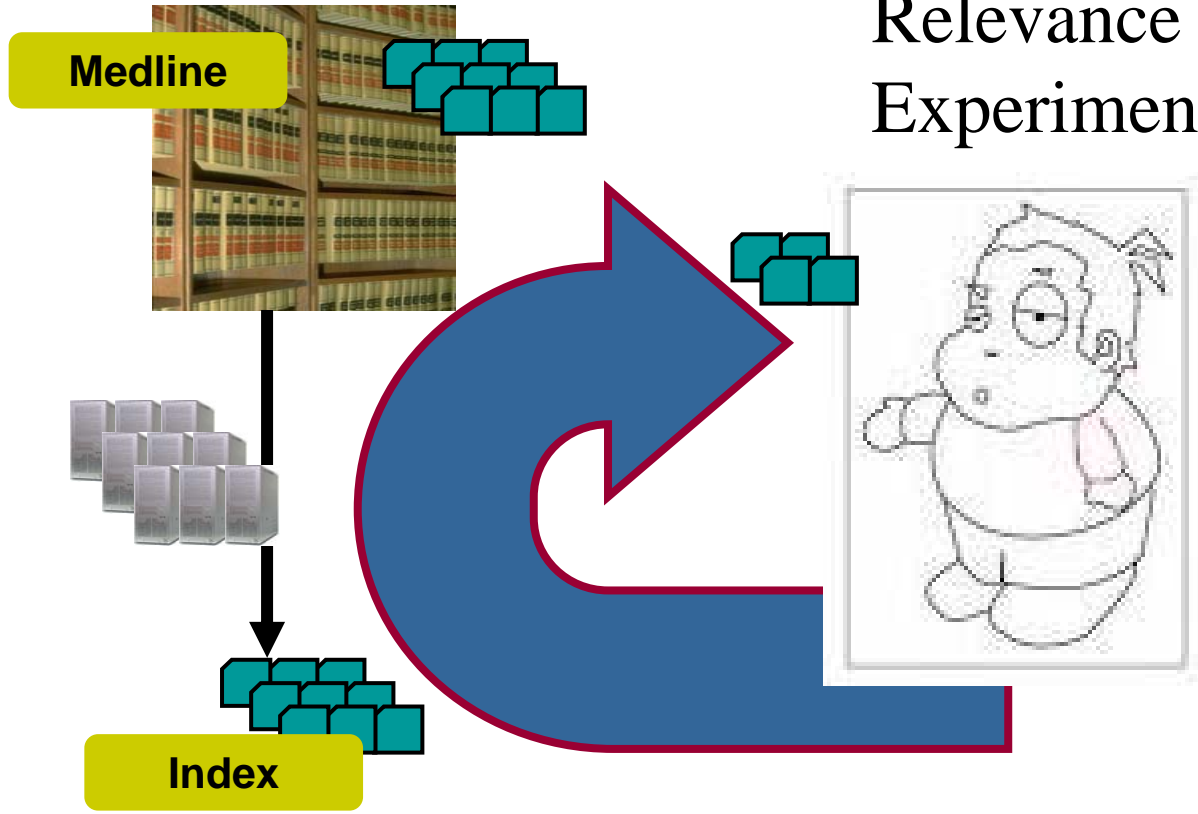
Text in
XML with
annotations

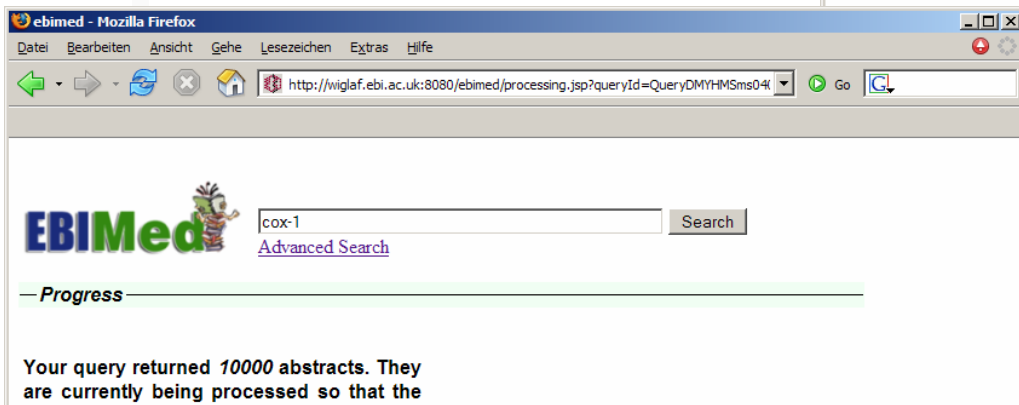


The engine

Query??

- Function of a protein?
- Relevance to a disease?
- Experimental method?





PMID, the PubMed (NLM's database that incorporates MEDLINE) unique identifier, is a 1 to 8 digit accession number with no leading zeros. It is present on all records exported to licensees and is the accession number for managing and disseminating records. PMIDs are not reused after records are deleted. [More](#).

Indexed fields: [PMID, AbstractText, ArticleTitle, AuthorList, MeshHeadingList] (searched by default), DateCreated, DateCompleted, DateRevised, PubDate, Language.

[European Bioinformatics Institute](#) 2002-2004. All Rights and Trademarks Reserved.

Your query returned 10000 abstracts. They are currently being processed so that the relevant information can be summarized in tabular form. This is a computing intense operation that will take a little while, please be patient.

81%

(Note: Your query returned a very large number of abstracts that have been restricted to the first 10000 by order of relevance. This is to avoid having you wait a potentially long time. Please consider refining your query to a more specific one. Thank you)



(Refreshing period 6 secs.)

There is no flying without wings.

I will refine the query



HitPair table

(538 HitPairs / 1343 sent. / 91 abstracts / 6.487 seconds)

Rows 1 to 5 (out of 57)

first << 1/12 >> last

Uniprot	Uniprot	Cellular component	Biological process	Molecular function	Drug	Species
COX-2 (score: 248)	COX-1 (50) cyclooxygenase-2 (18) COX (14) pain (8) PCR (5) HGF (10)		development (6) apoptosis (5) excretion (1) blood coagulation (1) gastric acid secretion (1)	binding (2) protein binding (1) E2 (1)	Vioxx (41) Celebrex (12) aspirin (9) ibuprofen or Sulindac or naproxen or diflunisal or	this (23) cancer (18) human (12) rats (4) mouse (3) MCF (2)



170 Medline Abstracts



Type	Hits	HitPairs
Protein/Gen	66	487
Cellular component	4	15
Biological process	21	102
Molecular function	4	19
Drug	27	283
Species	34	168
Total	156	1074

HitPair table

- You can explore a total of 487 permutations for this HitPair table arrangement. Click on the secondary columns' headers to rearrange the table.
- Rows 1 to 5 (out of 58).

first << 1/12 >> last

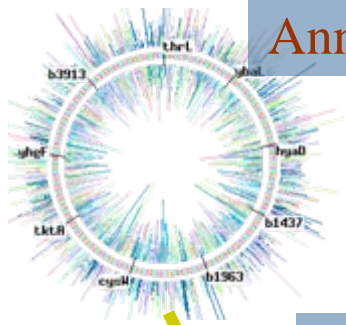
Protein/Gen	Protein/Gen	Cellular component	Biological process	Molecular function	Drug	Species
COX-2 <small>(score: 279)</small>	COX-1 (23/54) cyclooxygenase-2 (21/21) COX (11/14) pois (0/0) PCR (5/5) rats (5/5) gastrin (2/10) PGE (2/4) TNFalpha [®] (2/4) Bcl-2 (1/5) PPARgamma or peroxisome proliferator-activated receptor gamma [®] (1/4) Bax (1/4) IL-1beta (1/3) carbonic anhydrase (1/2) APC (1/2) tumor necrosis factor (1/1) com (1/1) ago (1/1) CA II (1/1) proton pump (1/1) epidermal growth factor (1/1)	---	development (6/6) apoptosis (3/5) excretion (1/1) blood coagulation (1/1) gastric acid secretion (1/1) transcription (1/1) homeostasis (1/1) reverse transcription (1/1) secretion (1/1) clotting or coagulation (1/1)	binding (1/2) protein binding (1/1) E2 (1/1)	Rofecoxib or Vioxx (56/87) celecoxib or celebrex (21/28) diclofenac or indomethacin or anti-inflammatory drugs (17/22) aspirin (9/11) Sulindac or ibuprofen or naproxen or difunisal or Naprosyn or nabumetone or Voltaren (7/10) meloxicam (3/5) acetic acid (3/3) acetaminophen (2/3) valdecoxib or Bextra (2/3) misoprostol (2/3) acetylsalicylic acid (2/2) prostacyclin (2/2)	cancer (10/19) human (8/12) mouse (2/3) MCF (2/2) mice (2/2) meta (1/1) dogs (1/1) glaucoma (1/1) flag (1/1) Parpagophytum (1/1) Chinese hamster (1/1)



EMBL-EBI Types of information from text

European Bioinformatics Institute

Functional
Annotation



Haplotype
Parameters

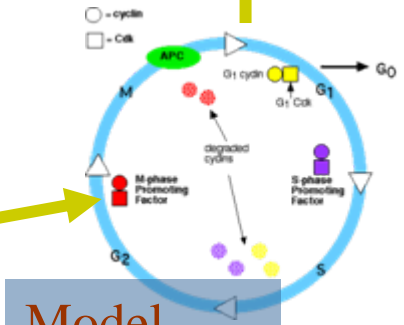
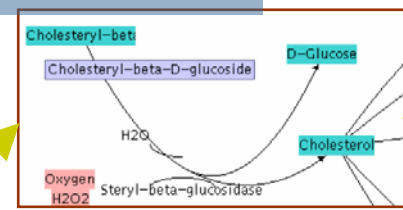
Genotype
Phenotype
Associations



Structure
Details

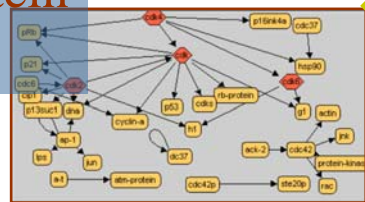


Molecular
Interaction
Parameters



Model
Parameters

Protein-Protein
Interactions





Functional annotation of proteins with GO terms based on the literature

- Identification + disambiguation of protein named entities
- Identification of GO terms
 - Molecular function
 - Biological process
 - Cellular component
- Linking both
- Improving quality by combining different types of data



* map kinase * activity 

GO terms concerned with MAP kinase

MAP kinase phosphatase activity
 MAP kinase activity
 MAP kinase 1 activity
 MAP kinase 2 activity
 MAP kinase kinase kinase activity
 MAP/ERK kinase kinase activity
 MAP kinase kinase kinase kinase activity

MAP kinase kinase activity
 MAP-kinase scaffold activity
 MAP-kinase anchoring activity
 activation of MAP/ERK kinase kinase
 MAP-kinase scaffold protein activity
 MAP-kinase anchor protein activity
 cytoplasmic translocation of MAP kinase

Exact matches in Medline can be found as follows:

map kinase activity
 map kinase kinase activity
 map kinase kinase kinase activity

map kinase phosphatase activity
 map kinases activity

A selection of noun phrases which contain a match to MAPK:

adhesion dependent map kinase activity
 alf4 induced map kinase activity
 attenuated oxytocin induced map kinase activity
 endothelin 1 stimulated p38 map kinase activity
 enhanced insulin induced map kinase activity
 epidermal growth factor stimulated map kinase activity
 map kinase phosphotransferase activity





Mapping of GO terms to text (**whatizitGO**)

$$s(z, t) = e(z, t)^\alpha \cdot I(t)^\beta \cdot pr(W, z)^\gamma$$

I = Information content of term t

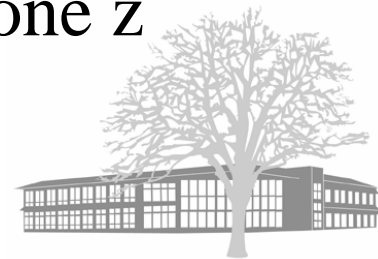
$$I(t) = \sum_{w \in tok(t)} -\log(p(w))$$

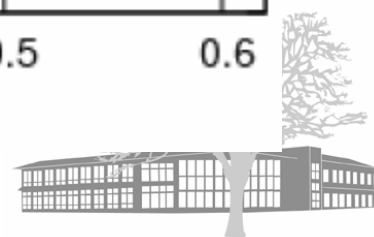
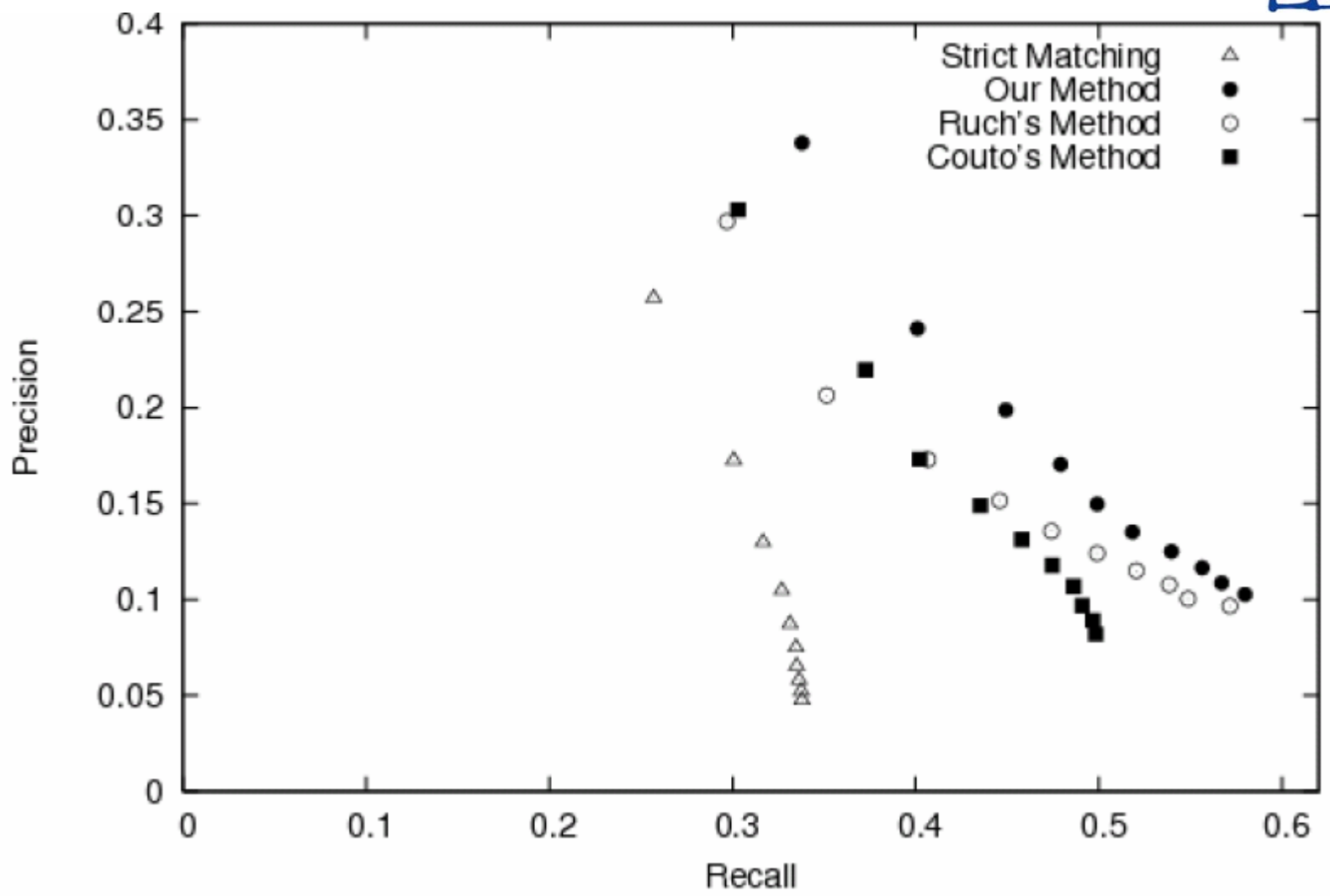
pr = Proximity of components of term t in zone z

$$pr(W, z) = \frac{\Sigma_{min}(W)}{\Sigma(W, z)}$$

E = Evidence for components of term t in zone z

$$e(z, t) = \frac{I(z \cap t)}{I(t)} = \frac{\sum_{w \in tok(z) \cap tok(t)} \log(p(w))}{\sum_{w \in tok(t)} \log(p(w))}$$





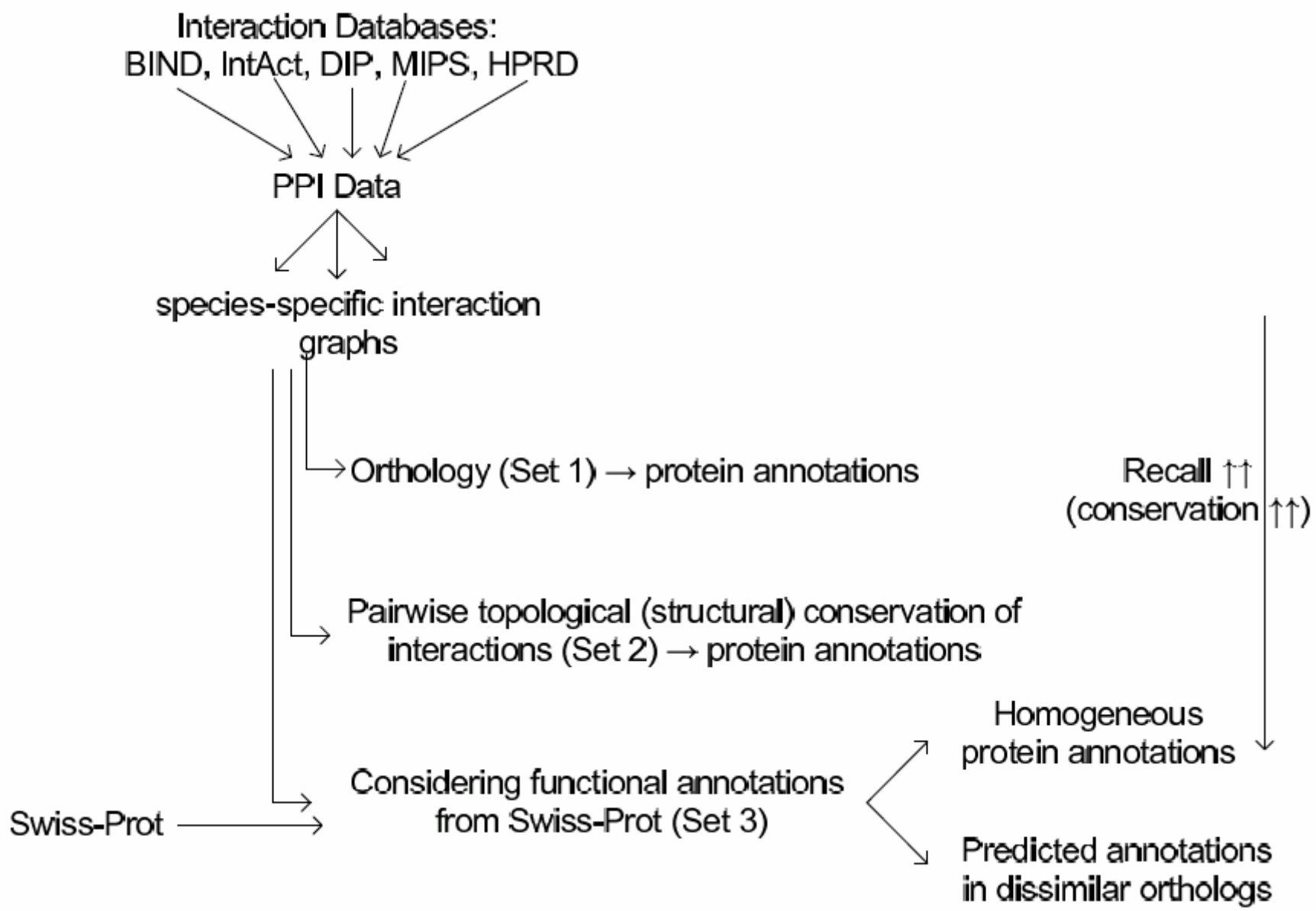
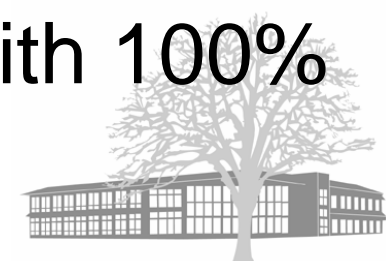


Table 4. Ontology specific consideration of confirmed known GO terms of set 3.

Extraction criteria	Recall – MF	Recall–BP	Recall–CC
Exact & Species	56/107 (52%)	31/85 (36%)	42/91 (46%)
Approx & Species	71/107 (66%)	41/85 (48%)	52/91 (57%)
Exact	83/107 (77%)	51/85 (60%)	67/91 (73%)
Approx	90/107 (84%)	69/85 (81%)	75/91 (82%)

Prediction of 34 new annotations with 100% precision



Conclusions:

- The information extraction process can be efficiently split up into single consecutive steps that can be performed in real time
- All extraction methods available to the public as Web services (NoE SemanticMining) for integration into bioinformatics solutions
- The use of good terminological resources is essential.
- Ontological terms are challenging but offer better semantics.



- UK node to public literature
- British Library, University of Manchester and the EBI
- In collaboration with NLM and the Funders



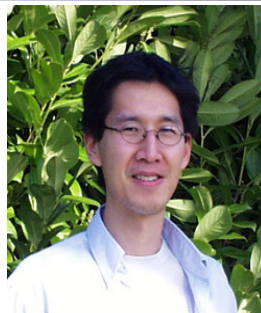
Dietrich
Rebholz-
Schuhmann



Sylvain
Gaudan



Miguel
Arregui



Kevin
Au-yeung



Jung-Jae
Kim



Vivian Lee



Nicolas
Rodriguez



Piotr Pezik



Antonio
Jimeno
Yepes



Alexander
Griekspoor

Special thanks to
Harald Kirsch

