Exploring the Semantics of Biomedical Ontologies

Francisco M Couto
12 April 2011
EBI External Seminar

How to compare?

- Manually
  - Before computers
  - based on catalogues
- Problems
  - Digital Era
  - Information overload
  - UniProtKB/TreMBL (14 million sequences)

New Entities

- Common step
  - **Compare** it with known entities
  - to transfer knowledge
- Common to any research area
  - Characterization of new proteins
  - Microarray Analysis
  - Information Retrieval in general

Computational Comparison

- Using digital representations
- Structural Similarity
  - Based on the syntax of the representations
  - Protein sequence (BLAST)
- Semantic Similarity
  - Based on the implicit and explicit semantics (Ontologies)
  - Protein molecular function (GO Analysis)

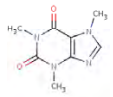
Structural Similarity

- Structure is "always" available
  - We normally start by identifying
    - how an entity look like
    - before finding what it does (meaning)
- Structure is less ambiguous
  - It's easier to establish an agreement on describing
    - how an entity look like
    - than on what it does
- Performance and Scalability
  - String matching techniques (BLAST)

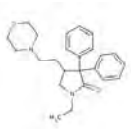
Semantic Similarity

- Solve the problems of structural similarity
  - Two entities that look similar
    - **Do not imply**
    - They have the same meaning
  - And two entities with similar meaning
    - **Do not imply**
    - that they look similar
- When we want to find similar entities
  - Based on their meaning
  - **Not on** how they look like

Example: Similar Semantics



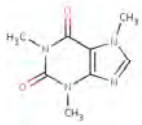
Caffeine (CHEBI:27732)



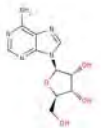
Doxapram (CHEBI:681849)

- Different structure but
- both central nervous system stimulants (CHEBI:35337)

Example: Similar Structure



caffeine (CHEBI:27732)



adenosine (CHEBI:16335)

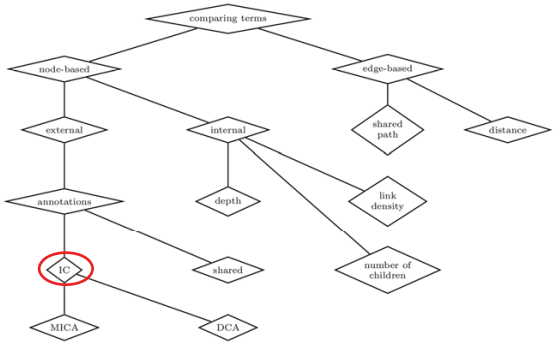
- Similar structure but
- different roles
  - adenosine is an anti-arrhythmia drug (CHEBI:38070)
  - nucleoside (CHEBI:18254)

4/12/2011

Semantic description

- Is not always available
- Semantic Web Initiative
  - inserting machine-readable metadata
- Ontologies
  - Serve as metadata vocabularies
  - to describe entities (annotations)
- Biomedical Ontologies
  - Large effort and involvement from the community
  - 101 ontologies at http://www.obofoundry.org/
  - 82 million annotations from UniProtKB-GOA

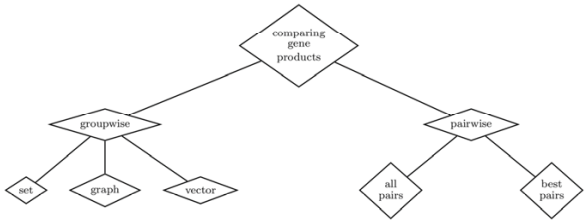
Comparing Concepts



Semantic similarity measures

- Input:
  - two ontology concepts
  - or two sets of terms annotating two entities
- Output:
  - a numerical value reflecting the closeness in meaning between them

Comparing entities



4/12/2011

Information Content

- measures how specific and informative a concept is

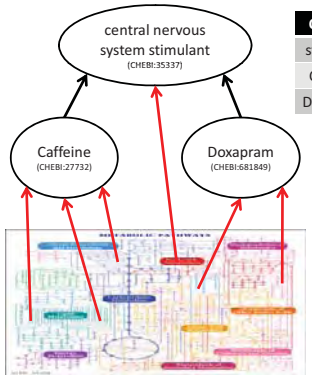
$$IC(c) = -\log(\frac{freq(c)}{maxFreq})$$

- Inversely proportional to frequency in a given corpus
- The frequency is also propagated to its ancestors
  - IC proportional to the depth of a concept
- Extrinsic IC
  - number of entities mapped to each concept
- Intrinsic IC
  - number of children

Similarity based on IC

- Similarity proportional to the
  - IC of the most informative common ancestor (MICA)
    - Shared information between two concepts
    - Resnik
  - Weighted Jaccard index of two sets of concepts
    - Shared information between two entities
    - SimGIC

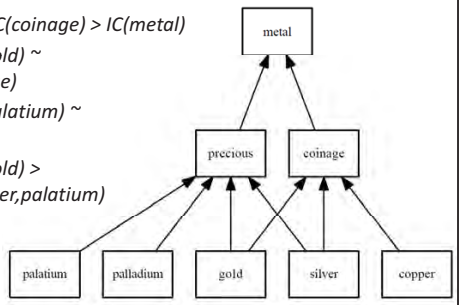
Example



Concept	freq	hfreq	IC
stimulant	1	6	0.0
Caffeine	3	3	0.3
Docapram	2	2	0.5

Example

- *IC(copper) > IC(coinage) > IC(metal)*
- *Sim(copper,gold) ~ IC(coinage)*
- *Sim(copper,palatium) ~ IC(metal)*
- *Sim(copper,gold) > Sim(copper,palatium)*



4/12/2011

Applications

- Chemical Classification
- Chemical entity recognition and mapping
- Ontology matching and extension
- Enzyme family coherency assessment
- Epidemic and VPH information retrieval

Classification Problems

- BBB:
  - Do compounds cross the blood brain barrier?
- P-gp:
  - Are compounds substrates to the P-glycoprotein?
- estrogen:
  - Are compounds ligands to an estrogen receptor?

CHEMICAL CLASSIFICATION

João D Ferreira, Francisco Couto 2010: Semantic Similarity for Automatic Classification of Chemical Compounds. PLoS Computational Biology 9(6), e1000937

Structural Similarity

- Fingerprints and Machine Learning

Testing set	Classification system	Accuracy	Reference
BBB	Artificial Neural Networks	75.7%	(Doniger et al., 2002)
	Random Forest	80.9%	(Svetnik et al., 2003)
	Support Vector Machines	81.5%	(Doniger et al., 2002)
P-gp	Four-point Pharmacophore	62.7%	(Penzotti et al., 2002)
	Support Vector Machines	79.4%	(Xue et al., 2004)
	Random Forest	80.6%	(Svetnik et al., 2003)
estrogen	Decision Forest	~80%	(Tong et al., 2003)
	Random Forest	82.8%	(Svetnik et al., 2003)

4/12/2011

Hybrid Approach

- Based on Semantic Similarity
  - CHEBI and KEGG pathways

$$\mathrm{sim}_{\mathrm{hybrid}} = \alpha \cdot \mathrm{sim}_{\mathrm{structural}} + (1 - \alpha) \cdot \mathrm{sim}_{\mathrm{semantic}}$$

Set	Chym Parameters	Accuracy
BBB_p	FP3, simGIC, all, $\alpha = 0.28$	90.9%
$P\text{-}gp_p$	FP4, simUI, all, $\alpha = 0.66$	87.7%
$estrogen_p$	FP4, simGIC, role, $\alpha = 0.42$	84.2%

New Predictions

Compound		Set	Coefficient
ID	Name		
1015	orthanilic acid	BBB_p	0.503
2654	aminoglutethimide	BBB_p	0.489
2089	O-methylserotonin	BBB_p	0.477
3638	chloroquine	BBB_p	0.475
2430	aconitine	$P\text{-}gp_p$	0.474
1883	4-hydroxystyrene	$estrogen_p$	0.577
5078	flavonol*	$estrogen_p$	0.577
5262	galangin	$estrogen_p$	0.577

* This compound is a false positive.

Structural vs. Semantic

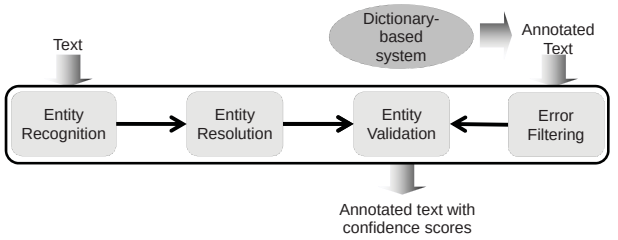
Alpha	BBB_p	$P\text{-}gp_p$	$estrogen_p$
0.0	81.0%	74.1%	73.3%
0.1	86.9%	74.1%	74.3%
0.2	88.9%	79.0%	74.3%
0.3	**90.2%**	76.5%	79.2%
0.4	88.2%	81.5%	**84.2%**
0.5	85.0%	84.0%	83.2%
0.6	83.0%	85.2%	78.2%
0.7	83.0%	**86.4%**	81.2%
0.8	81.0%	82.7%	76.2%
0.9	77.1%	84.0%	71.3%
1.0	71.9%	85.2%	79.2%

CHEMICAL ENTITY RECOGNITION AND MAPPING

Tiago Grego, Piotr Pezik, Francisco Couto, Dietrich Rebholz-Schuhmann, Identification of Chemical Entities in Patent Documents.3rd International Workshop on Practical Applications of Computational Biology and Bioinformatics (IWPACBB'09) 2009

4/12/2011

ICE: Identifying Chemical Entities



A common example of a chemical substance is pure water; it has the same properties and the same ratio of hydrogen to oxygen whether it is isolated from a river or made in a laboratory. Some typical chemical substances are diamond, gold, salt (sodium chloride) and sugar (sucrose).

Example

- "Despite a lack of data regarding their efficacy, both caffeine and doxapram have been recommended for treatment of hypercapnia in equine neonates with central nervous system damage." (PMID: 18371030)
- The fact that caffeine and doxapram are semantically similar
  - is an evidence for being correctly identified

Entity Validation

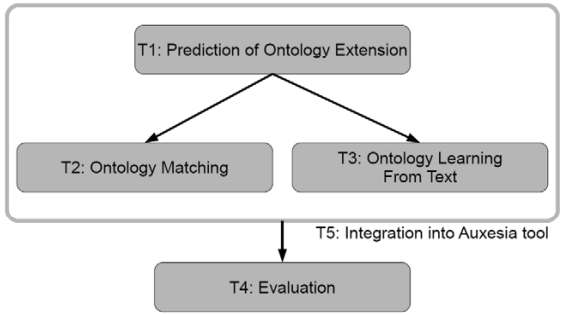
- The scope of a scientific publication is typically narrow
  - Specific protein, specific metabolic pathway, specific disease, etc
- Thus it is expected for the entities present in a document to be related
  - Measure semantic similarity between the entities

ONTOLOGY MATCHING AND EXTENSION

Catia Pesquita, Cosmin Stroe, Isabel F. Cruz, Francisco Couto, BLOOMS on AgreementMaker: results for OAEI 2010 - ISWC Workshop on Ontology Matching 2010

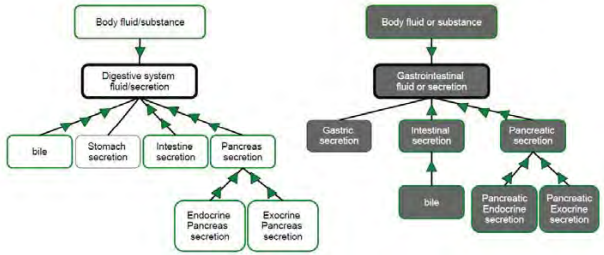
4/12/2011

Ontology Extension Approach



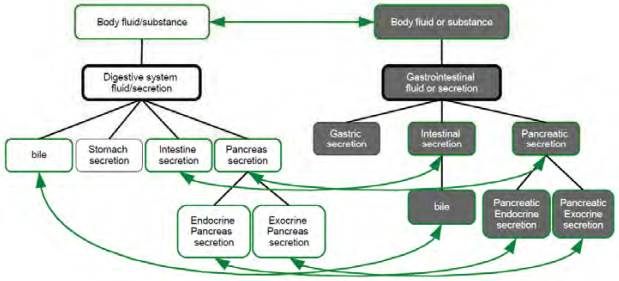
Structural level

- Similarity Propagation:

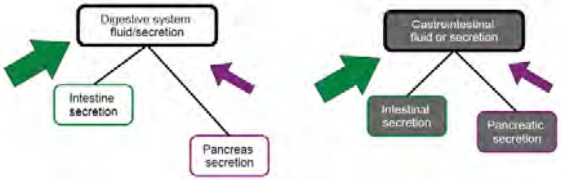


Ontology Matching

- String Matching:



Semantic Similarity in Propagation



4/12/2011

OAEI 2011 anatomy results

- Adult Mouse Anatomy (2744 classes)
- NCI Thesaurus (3304 classes) describing the human anatomy.

System	precision	recall	f-measure
AgreementMaker (F-measure)	90.3%	85.3%	87.7%
AgreementMaker (precision)	96.2%	75.1%	84.3%
BLOOMS-XLDB	96.7%	72.5%	82.9%

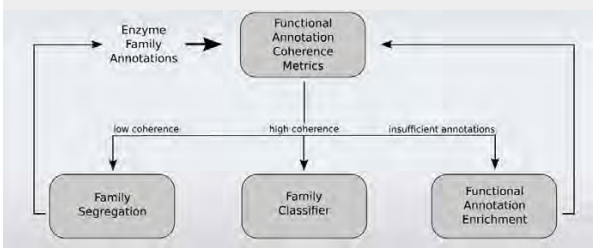
Goals

- Measure and Improve
  - functional annotation coherence of protein families.
- Enrich
  - under-annotated families with functional annotations.
- Classify
  - novel sequences into richly annotated protein families.

ENZYME FAMILY COHERENCY ASSESSMENT

Hugo Bastos, Tiago Grego, Francisco Couto, Pedro M. Coutinho, Enzyme family coherence assessment: validation and prediction. JB'2009 - Challenges in Bioinformatics p. 26-30, Lisbon, Portugal, November, 2009.c

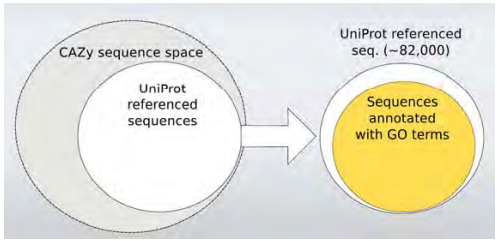
Approach



4/12/2011

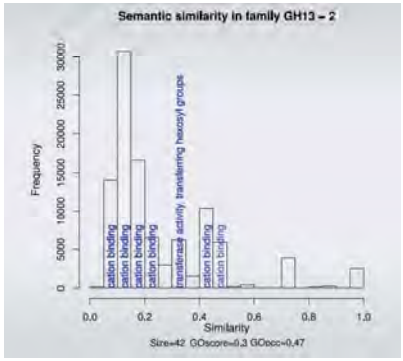
CAZy: case-study

- Database specialized in catalytic modules of enzymes that degrade, modify or create glycosidic bonds and modules associated to carbohydrate adhesion

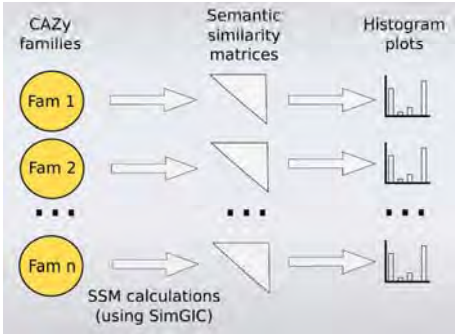


Plots

- Histograms plot frequency of protein pairs within SSM ranges
- Labels show most frequent term with highest IC for each similarity range



Semantic Similarity Analysis



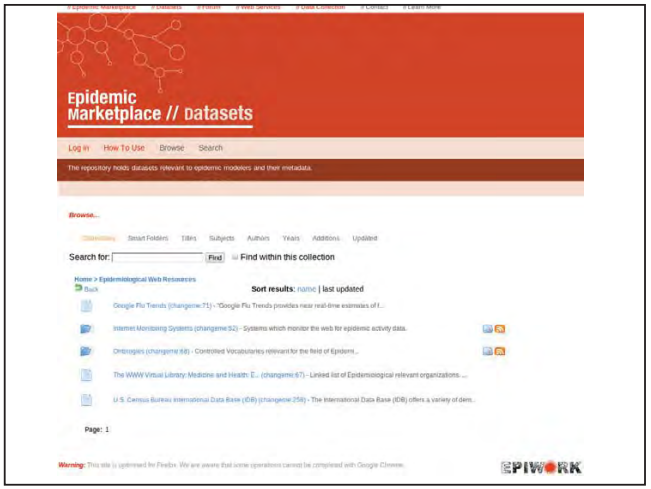
EPIDEMIC AND VPH INFORMATION RETRIEVAL

Luis Filipe Lopes, Fabrício A.B. Silva, Francisco Couto, João Zamite, Hugo Ferreira, Carla Sousa, Mário J. Silva, Epidemic Marketplace: An Information Management System for Epidemiological Data..Proceedings of the ITBAM - DEXA 2010 2010

4/12/2011

Epidemiological and Virtual Physiological Human

- Epidemic
  - Diseases
  - Symptoms
  - Anatomy
  - Phenotypes
  - Chemical substances
  - Proteins/Genes
  - Clinical Records
  - Geospatial
  - Therapeutics
  - Vaccine
  - Transmission modes
  - Organisms
  - …
- VPH
  - Diseases
  - Symptoms
  - Anatomy
  - Phenotypes
  - Chemical substances
  - Proteins/Genes
  - Clinical Records
  - Cellular Component
  - Gene Expression
  - Cell Type
  - Pathways
  - Radiology
  - ….



Information Retrieval

- Input
  - search keywords
- Output:
  - Ranked list of datasets and/or models
- Ranking Method
  - Semantic Similarity
  - Using multi-domain annotations
  - Mapping search keywords to ontology terms

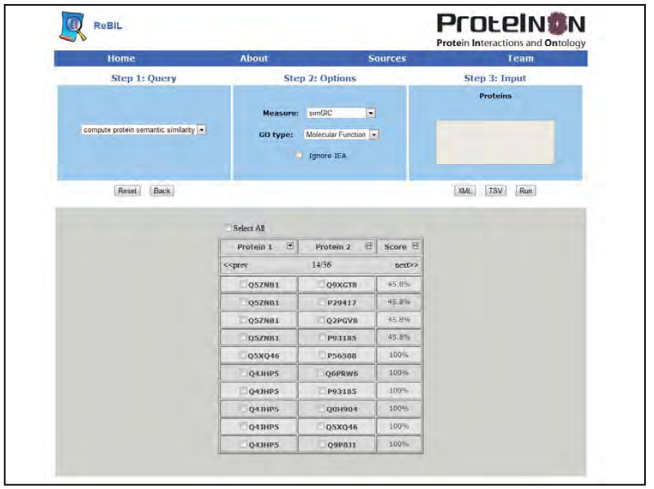
Challenges

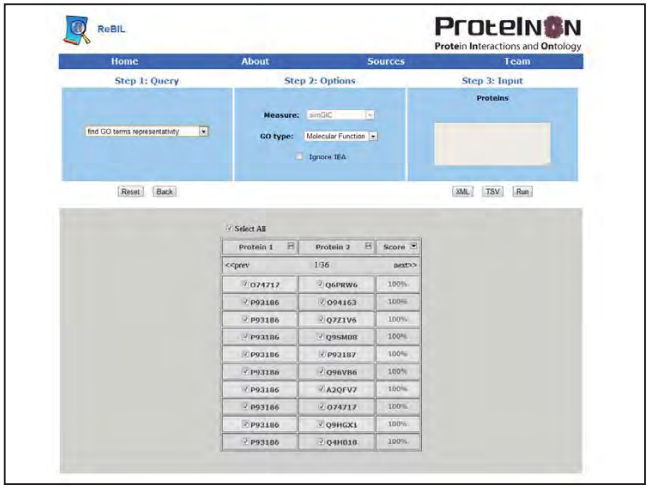
- Dataset and Model annotation
  - Ontology Selection
  - Manual or/and semi-automated
- Semantic Similarity
  - Extend to another ontologies
    - Nowadays: GO, CHEBI, HPO
    - Next: FMA
- Integrate multi-domain measures
  - Weights

4/12/2011

OUR WEB TOOLS

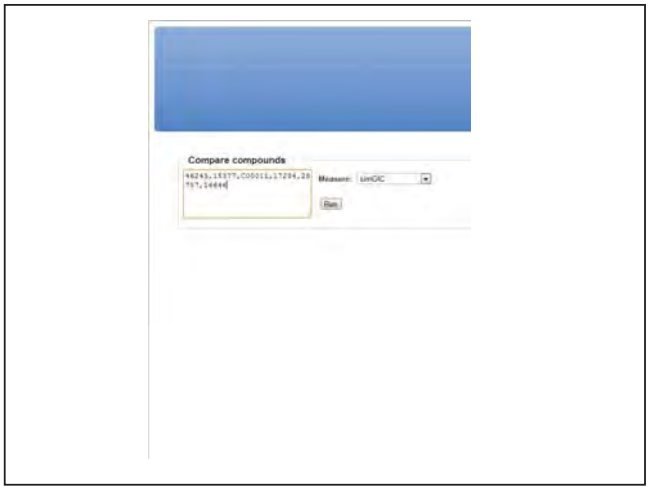
http://xldb.di.fc.ul.pt/wiki/BOA



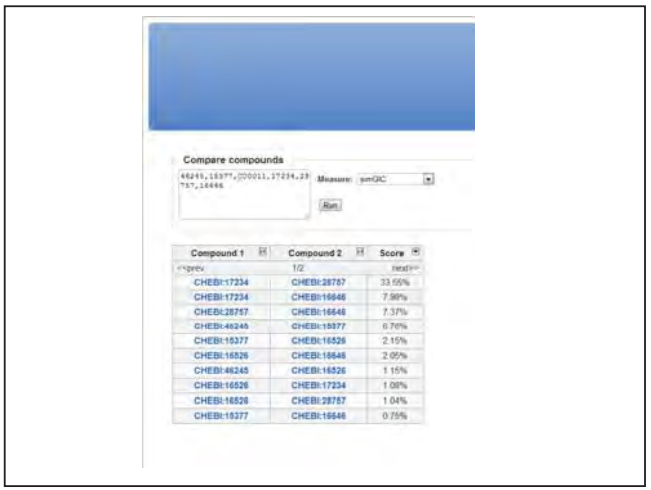


4/12/2011

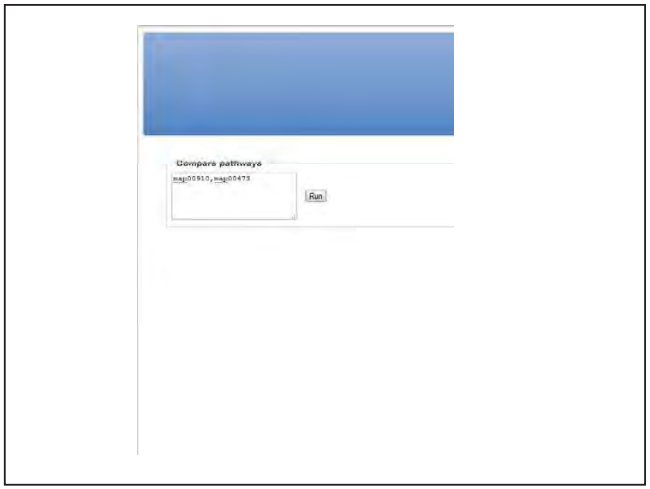

ReBIL
ProteInON
Protein Interactions and Ontology
Home
About
Sources
Team
Step 1: Query
Step 2: Options
Step 3: Input
Proteins
Reset
Back
Select All


Compare compounds
Measure:
Run


CMPSim
Compare compounds
Search
Tools
BOA


Compare compounds
Measure:
Run
Compound 1
Compound 2
Score
1/2
CHEBI:17234 CHEBI:28757 33.55%
CHEBI:17234 CHEBI:16646 7.96%
CHEBI:28757 CHEBI:16646 7.37%
CHEBI:46245 CHEBI:15377 6.76%
CHEBI:15377 CHEBI:16526 2.15%
CHEBI:16526 CHEBI:16646 2.05%
CHEBI:46245 CHEBI:16526 1.15%
CHEBI:16526 CHEBI:17234 1.08%
CHEBI:16526 CHEBI:28757 1.04%
CHEBI:15377 CHEBI:16646 0.75%

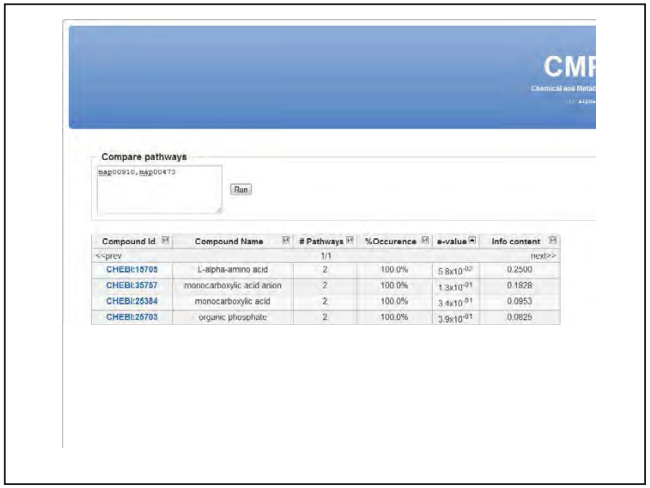
4/12/2011



CESSM

- Platform to evaluate GO semantic similarity measure
- Provides a list of 13,430 protein pairs
  - 1,039 distinct proteins
  - Blast E-value < 10^{-4}
  - At least one PFam and EC annotation
- Calculate their similarity with your measure
- Provides a quantitative comparison

Similarity	GI	JA	JB	JM	LA	LB	LM	RA	RB	RM	UI	tm
ECC	0.62	0.64	0.4	0.45	0.6	0.42	0.45	0.64	0.34	0.36	0.56	0.53
Pfam	0.64	0.62	0.44	0.18	0.57	0.45	0.18	0.56	0.33	0.13	0.49	0.55
SeqSim	0.72	0.59	0.5	0.12	0.67	0.47	0.12	0.61	0.3	0.1	0.55	0.7

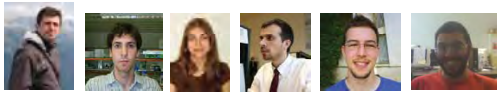


Web Services

Ontology/concept|instance/inputList/feature/parameter1/paramValue1/…

- Ontology:
  - GO: term (concept) and protein (instance)
  - CHEBI: compound (concept) and pathway (instance)
- Feature:
  - ssm (semantic similarity measures)
  - characterization (only available for GO_prontein)
- Examples:
  - http://10.10.4.28/~btavares/BOA/GO/term/GO:0050681,GO:0030331/ssm/measure/simGIC/IEA/false/
  - http://10.10.4.28/~btavares/BOA/GO/protein/Q13263,P35222/ssm/IEA/false/goType/molecular_function/measure/simUI
  - http://10.10.4.28/~btavares/BOA/Chebi/compound/46245,15377,C00011,C00049/ssm/measure/simUI
  - http://10.10.4.28/~btavares/BOA/Chebi/pathway/map00910,map00473/ssm/

Thanks for your attention!



Biomedical Informatics research line
at University of Lisbon

http://xldb.fc.ul.pt/wiki/Biomedical_Informatics