Enhancing the Semantics of Links and Paths in Life Sciences Data Sources

Louiqa Raschid
University of Maryland

Collaborators:
Felix Naumann, S. Heymann, P. Rieger, Humboldt University
George Mihaila, IBM
Maria Esther Vidal, Universidad Simon Bolivar
Adam Lee and Yao Wu, University of Maryland
Alex Lash, Sloan Kettering

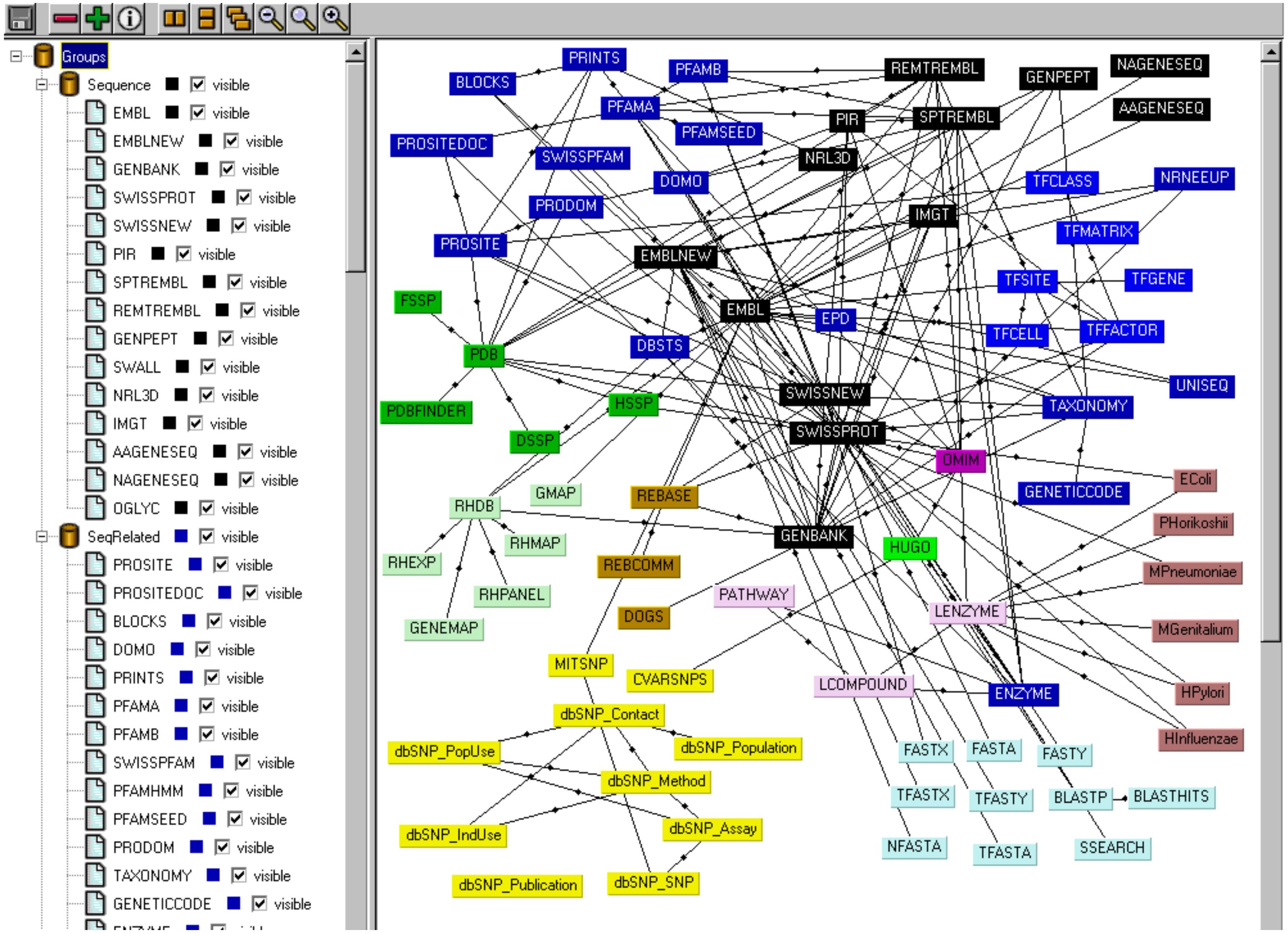
Acknowledgements:
Zoe Lacroix, ASU
David Lipman, NCBI/NIH
Barbara Eckman, P. Schwarz and J. Rice, IBM
National Science Foundation
National Institutes of Health

Outline

- Motivation
- Example of (automated) link extraction
- Example of link labeling (ontologies)
- Data model for enhanced links and paths
- Query language for enhanced links and paths
- Query evaluation
- Status

Motivation: the “SRS Universe”

[Thanks to Thure Etzold, DILS 2004]

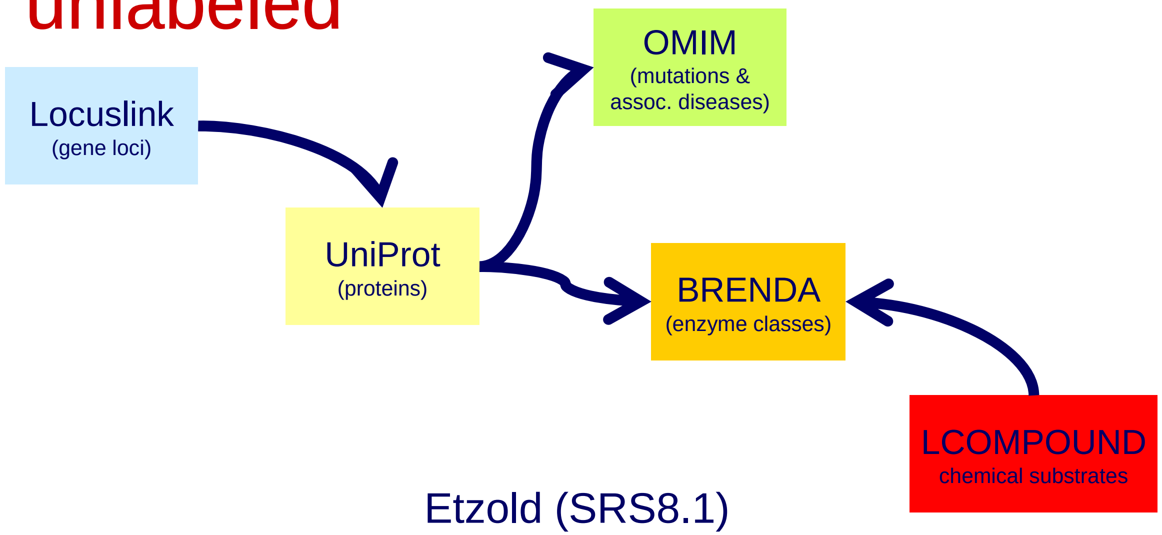


Groups

- Sequence ■ ☑ visible
  - EMBL ■ ☑ visible
  - EMBLNEW ■ ☑ visible
  - GENBANK ■ ☑ visible
  - SWISSPROT ■ ☑ visible
  - SWISSNEW ■ ☑ visible
  - PIR ■ ☑ visible
  - SPTREMBL ■ ☑ visible
  - REMTREMBL ■ ☑ visible
  - GENPEPT ■ ☑ visible
  - SWALL ■ ☑ visible
  - NRL3D ■ ☑ visible
  - IMGT ■ ☑ visible
  - AAGENESEQ ■ ☑ visible
  - NAGENESEQ ■ ☑ visible
  - OGLYC ■ ☑ visible
- SeqRelated 🟦 ☑ visible
  - PROSITE 🟦 ☑ visible
  - PROSITEDOC 🟦 ☑ visible
  - BLOCKS 🟦 ☑ visible
  - DOMO 🟦 ☑ visible
  - PRINTS 🟦 ☑ visible
  - PFAMA 🟦 ☑ visible
  - PFAMB 🟦 ☑ visible
  - SWISSPFAM 🟦 ☑ visible
  - PFAMHMM 🟦 ☑ visible
  - PFAMSEED 🟦 ☑ visible
  - PRODOM 🟦 ☑ visible
  - TAXONOMY 🟦 ☑ visible
  - GENETICCODE 🟦 ☑ visible

Labels for Links

unlabeled



Locuslink
(gene loci)

UniProt
(proteins)

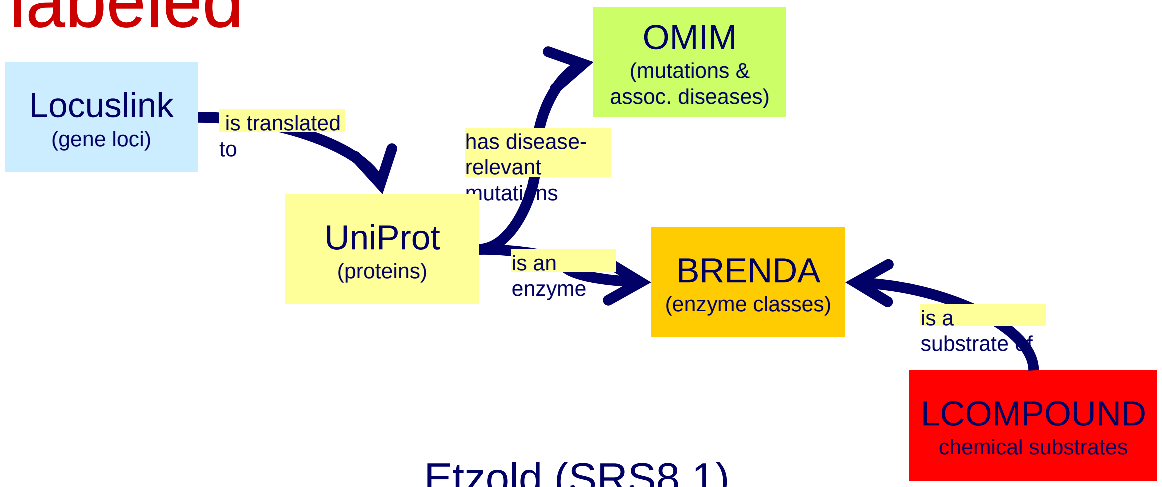
OMIM
(mutations & assoc. diseases)

BRENDA
(enzyme classes)

LCOMPOUND
chemical substrates

Etzold (SRS8.1)

labeled



Locuslink
(gene loci)

is translated to

UniProt
(proteins)

has disease-relevant mutations

OMIM
(mutations & assoc. diseases)

is an enzyme

BRENDA
(enzyme classes)

is a substrate of

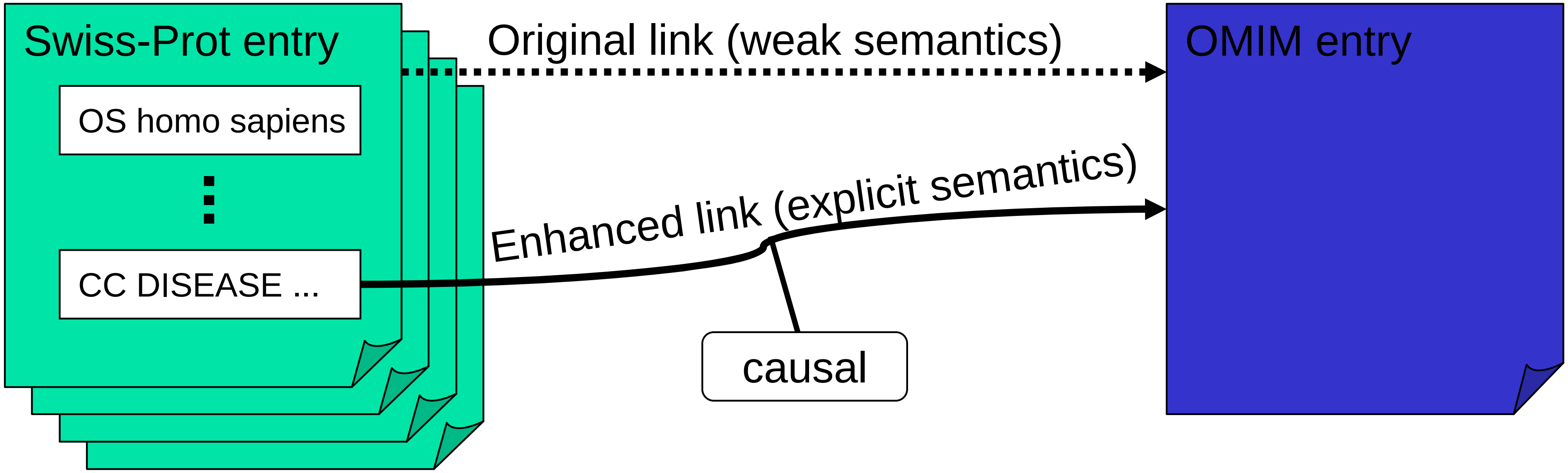
LCOMPOUND
chemical substrates

Etzold (SRS8.1)

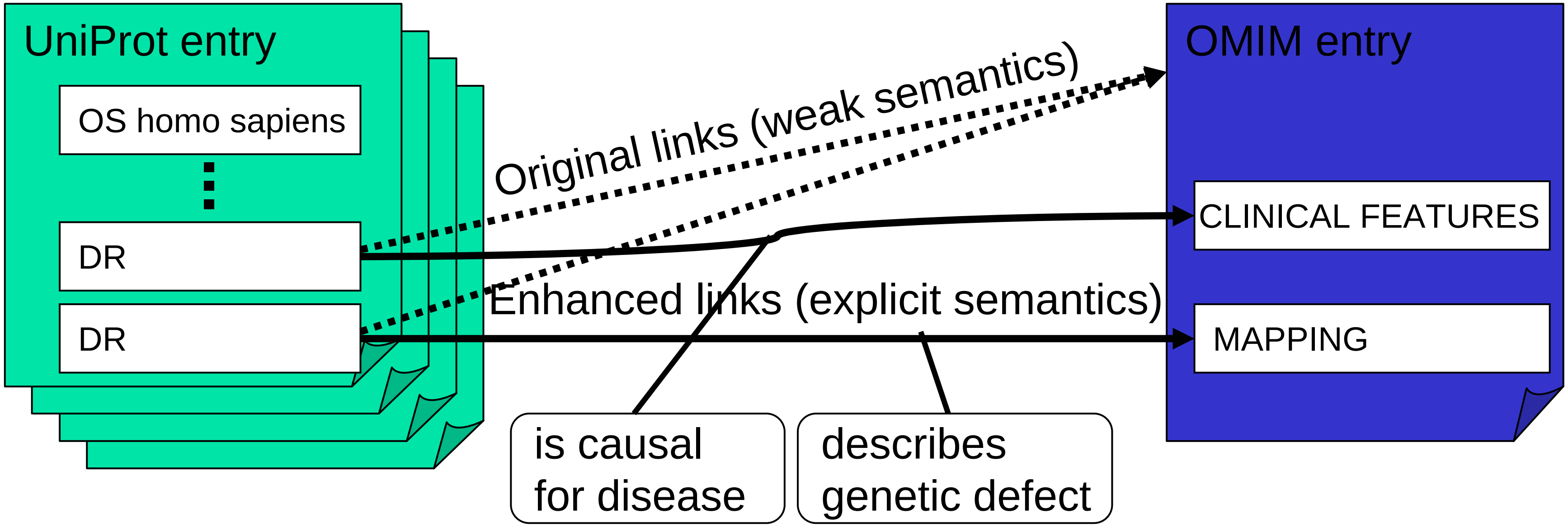
What’s in a link?

- Link may be added for different reasons.
  - Represents the result of an experiment protocol to test a hypothesis.
  - Data curators may add links following domain specific conventions.
  - A link may have been predicted by some (machine learning) software.....
- Current link implementation neither captures explicit semantics nor differentiates semantics.
  - RefSeq and LocusLink (NCBI); PDBSProtEC; ....
- Biologists can usually infer the meaning of a link and differentiate semantics but search engines and mediators cannot.

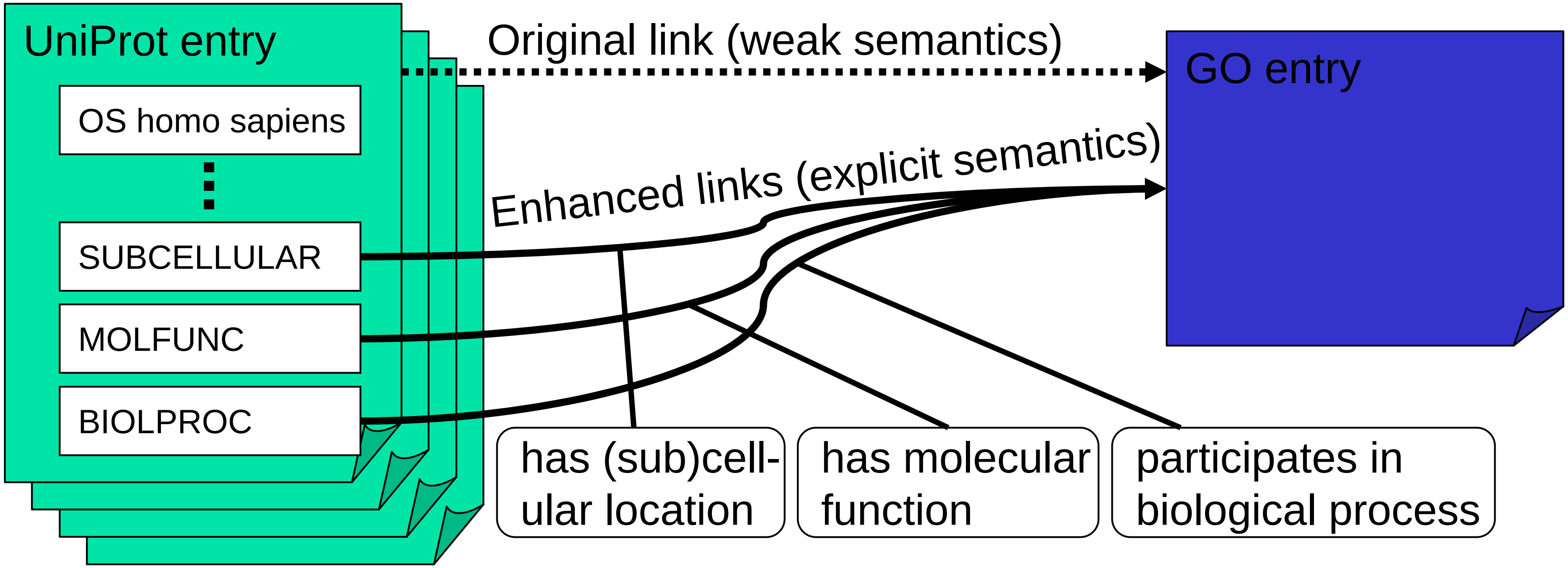
Enhancing semantics of links


Swiss-Prot entry
OS homo sapiens
CC DISEASE ...
Original link (weak semantics)
Enhanced link (explicit semantics)
causal
OMIM entry

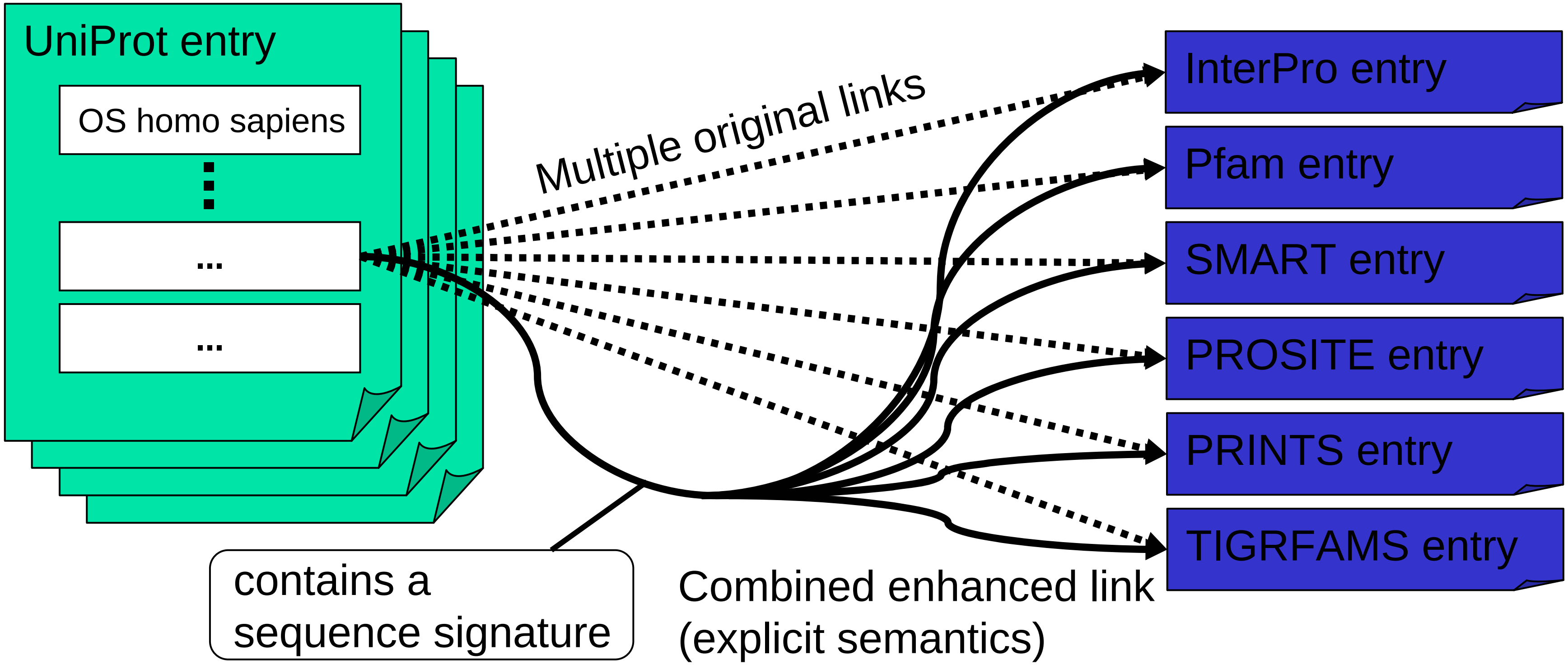
Example of Enhanced Links



Example of Enhanced Links


UniProt entry
OS homo sapiens
SUBCELLULAR
MOLFUNC
BIOLPROC
Original link (weak semantics)
GO entry
Enhanced links (explicit semantics)
has (sub)cellular location
has molecular function
participates in biological process

More Complex Example of Enhanced Link



enh-links

- Structure for enhanced links.
  - Source object and path to specific element of source object.
  - Target object and path to specific element of target object.
  - Label and meaning of links.
- Example: Link from UniProt to OMIM
  - **$source is causal for disease $target**

    UniProt ./DR[@name()="MIM” ./CC[@name()="DISEASE“

    OMIM ./ID && ./CLINICAL FEATURES
  - **$target describes genetic defects for $source**

    Uniprot ./DR[@name()="MIM"

    OMIM ./ID && ./MAPPING
- Semi-automated techniques to enhance link labels.

Outline

- Motivation
- Example of (automated) link extraction
- Example of link labeling (ontologies)
- Data model for enhanced links and paths
- Query language for enhanced links and paths
- Query evaluation
- Status

PubMed to Human Genome: Machine assisted extraction of links

- Map PubMed citations (with marker references) to the human genome and determine link semantics.
- Each PubMed citation with a genomic locus reference may have four attributes:
  - (a) start position (string)
  - (b) end position (string: will be empty if start is a point locus)
  - (c) trait (one term of an enumerated list: "hypermethylation", "loss", "gain", "mutation", "polymorphism" "phenotype")
  - (d) LOD score (floating point number)

  - Only (a) will be required, and all others may or may not be present.
  - None are dependent on the others except for the above.

  - Expected occurrence of attribute values:
  - (a) 100%
  - (b) <50%
  - (c) >75%
  - (d) >75%

Machine assisted extraction, generation and enhancement of links

- Two link types:
  - marker semantics;
  - "study" or "trait" semantics.
- Marker semantics can be further enhanced
  - single marker without score;
  - single marker with score;
  - interval between markers without scores;
  - interval between markers with scores ……
- Both source and (implicit) target entries contribute to the link semantics.

Example of assisted extraction

- Extraction protocol:
  - Search for "Fanconi" using Entrez Global Query. 105 OMIM hits; Select result #3: Fanconi Renotubular Syndrome.
  - Follow Entrez Link to PubMed. Select result #7, “Lichter-Konecki, U., Broman, K.W., Blau, E.B., Konecki, D.S.
  - Markers "D15S182” and “D15S537" appear in the abstract.
  - Use NCBI Map Viewer with parameters "Homo sapiens (human)“/organism and "D15S182 OR D15S537" /query box.
  - Results in 2 hits on Chromosome 15.
- Enhanced link from the disease “Fanconi Renotubular Syndrome” to a position interval on the human genome defined by markers D15S182 and D15S537 with log odds (LOD) scores of 4.44 and 4.68, respectively.
- How can this be used in a query? “Give me all genes associated with Fanconi Renotubular Syndrome” can use this enhanced path OMIM-PubMed-human genome concatenated with a link to your favorite gene source.

Parsing rules (incomplete?)

1) split abstract into sentences

2) for each sentence

2.a) find all semantic terms (below)

2.b) find all markers

2.c) find all words indicating intervals (between, from-to)

2.d) find all negation words

2.e) find all LOD scores

2.f) for each marker

2.f.i) associate marker to LOD scores (within 50 letters of one another?)

2.f.ii) associate two markers if interval words come between them

2.f.iii) associate all semantic terms in sentence to marker

2.f.iv) associate negation term to semantic term (within 30 letters of one another?)

2.g) for each interval

2.g.i) associate all semantic terms in sentence to markers in interval

Genomic marker semantics (ontology?)

1) EPIGENETIC ALTERATION
1.a) methylation
1.a.i) hypermethylation
1.a.ii) hypomethylation
1.b) histone moiety alteration
1.b.i) acetylation
1.b.ii) deacetylation

2) GENOMIC SEGMENT LOSS (synonym: loss)
2.a) genomic instability
2.a.i) microsatellite instability
2.a.ii) allelic imbalance (synonym: allelic loss, allelic reduction)
2.a.ii.1) loss of heterozygosity (synonym: LOH)
2.a.ii.2) hemizygosity
2.b) heterozygosity
2.c) homozygosity
2.d) haploinsufficiency

Genomic marker (ontology?)

3) GENOMIC SEGMENT GAIN (synonym: gain, amplification)

4) GENOMIC SEQUENCE ALTERATION
4.a) mutation
4.b) polymorphism
4.b.i) microsatellite
4.b.ii) restriction fragment length polymorphism (synonym: RFLP)
4.b.iii) single nucleotide polymorphism (synonym: SNP, SNiP)
4.c) translocation

5) PHENOTYPIC ASSOCIATION (synonym: phenotype, trait)
5.a) locus association (synonym: locus, loci)
5.a.i) linkage
5.a.ii) quantitative trait locus (synonym: QTL)
5.b) allelic association (synonym: allele)
5.b.i) linkage disequilibrium

Outline

- Motivation
- Example of (automated) link extraction
- Example of link labeling (ontologies)
- Data model for enhanced links and paths
- Query language for enhanced links and paths
- Query evaluation
- Status

Data Model

- Ontology Graph:
  - Logical Classes C (protein, gene, publication, disease,…)
  - Link Labels L (is causal for disease, describes genetic defect)
  - Link Types LT - (C_1, l, C_2) – l is a link label from L - between origin class C_1 and target class C_2

    (protein, is causal for disease, disease)
- Source Graph
  - Sources S and a mapping $ms: S \rightarrow C$
  - A set of source links L_S - (S_1, l, S_2) such that there exists (C_1, l, C_2) and S_1 maps to C_1 and S_2 maps to C_2 in ms
  - Sources are responsible to publish L_S

Data Model

- Object Graph
  - Objects O and mapping $mo: O \rightarrow S$
  - Links between objects L_O iff there exists (C_1, l, C_2) in LT and (S_1, l, S_2) in L_s
- Link composability matrix LL specifies meaningful concatenations of link types of LT
  - ☺ $(C_1, l_a, C_2) . (C_2, l_b, C_3)$
  - ☹ $(C_1, l_a, C_2) . (C_3, l_b, C_4)$
  - ☺ $l_a . l_b$
- ☹ A life science data model is a 10-tuple!

  $(C, L, LT, S, L_S, ms, mo, O, L_O, LL)$

Query Language

(C, L, LT, S, L_S , ms, mo, O, L_O ,LL)

- Identify **sources** in ***S*** that implement a given class.
- Identify **sources** in ***S*** that contain objects of a given class whose structure satisfies a given predicate.
  protein [in $s| $s contains Attr]
- Identify **objects** in ***O*** from a given source in ***S*** or from a given class in ***C*** whose attribute values satisfy a given predicate.

  protein [in S_n| o in S_n contains Attr with Value]
- Identify **paths** in ***S, L_S*** that satisfy a path regular expression and ***LL*** and can include source predicates.

  $p\ .^{lab-a}\ g\ .^{lab-b}\ d\ .^{lab-c}\ c$
- Identify **objects and paths** in ***O, L_O*** that satisfy a path regular expression and ***LL*** and can include source/object predicates.

Query Evaluation

$(C, L, LT, S, L_S, ms, mo, O, L_O, LL)$

- Path-labeled-link regular expression with predicates.
  - Validate regular expression against ***LT*** and ***LL***.
  - Enumerate paths in ***S, L_S*** .
  - Eliminate meaningless paths using ***LL***.
  - Rank and further select/eliminate paths.
  - Enumerate paths on ***O, L_O*** ***.***
- Naive evaluation in a mediator architecture.
- Statistics and optimization.
- Ranking results.

Outline

- Motivation
- Example of (automated) link extraction
- Example of link labeling (ontologies)
- Data model for enhanced links and paths
- Query language for enhanced links and paths
- Query evaluation
- Status
  - Automated extraction and labeling of links
  - Query evaluation