

Biomedical Computing Interest Group



May 19, 2005

From *terminology* integration to *information* integration

An example in the domain of genomics



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

◆ Background

- Terminology integration:
The Unified Medical Language System
- Information integration:
Genomics as an example

◆ Applications

- GenesTrace
- BioMeKe



Terminology integration

The Unified Medical Language System

Motivation

- ◆ Started in 1986
- ◆ National Library of Medicine
- ◆ “Long-term R&D project”
- ◆ Complementary to IAIMS

(Integrated Academic
Information Management Systems)

«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is **the variety of ways the same concepts are expressed** in different machine-readable sources and by different people.
- The second is the **distribution** of useful information among many disparate databases and systems.»



Source Vocabularies

(2005AA)

- ◆ 134 source vocabularies
 - 132 contributing concept names
- ◆ ~80 families of vocabularies
 - multiple translations (e.g., MeSH, ICPC, ICD-10)
 - variants (American-English equivalents, Australian extension/adaptation)
 - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIR-IV)
- ◆ Broad coverage of biomedicine
- ◆ Common presentation



Biomedical terminologies

◆ General vocabularies

- anatomy (UWDA, Neuronames)
- drugs (RxNorm, First DataBank, Micromedex)
- medical devices (UMD, SPN)

◆ Several perspectives

- clinical terms (SNOMED CT)
- information sciences (MeSH, CRISP)
- administrative terminologies (ICD-9-CM, CPT-4)
- data exchange terminologies (HL7, LOINC)

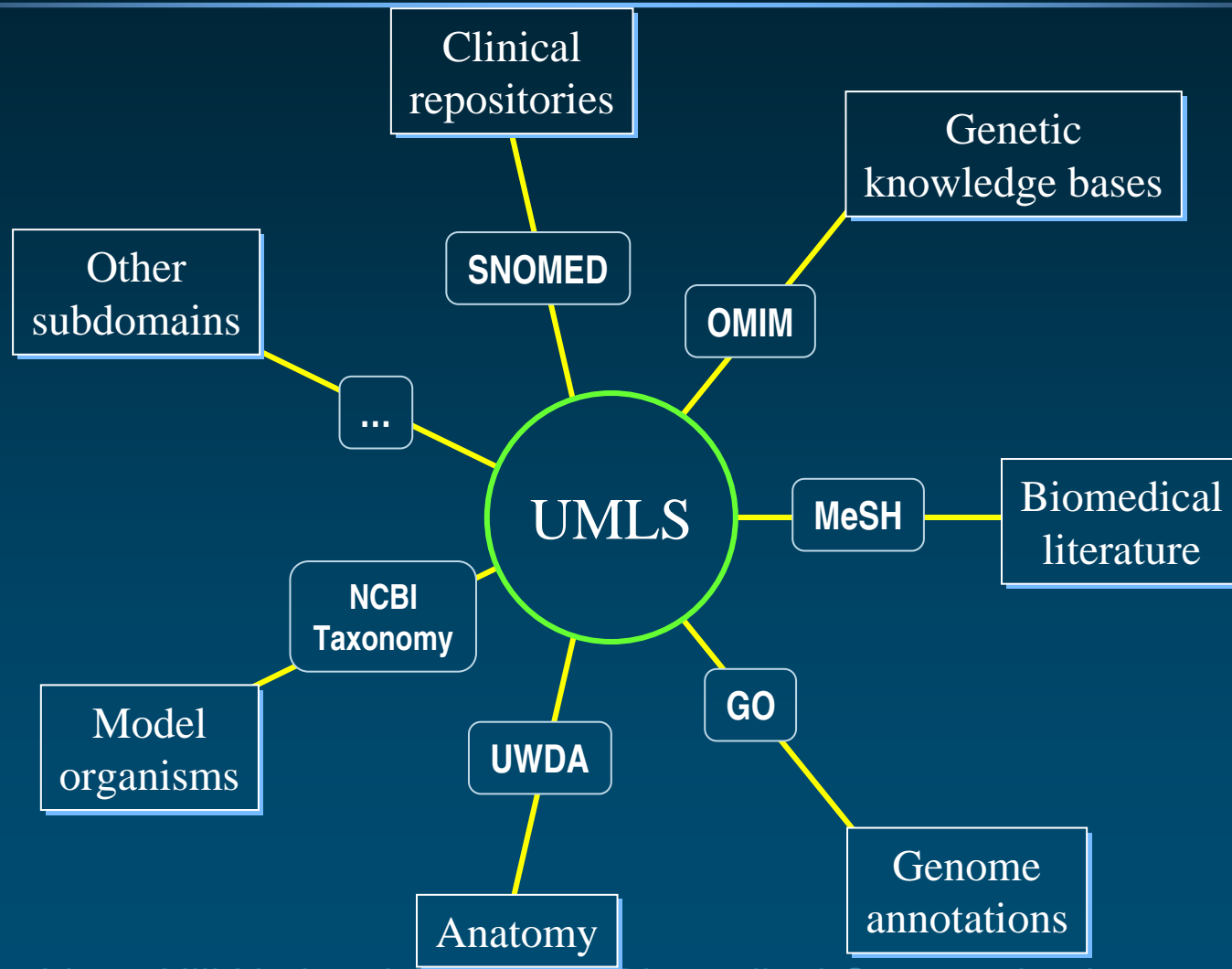


Biomedical terminologies (cont'd)

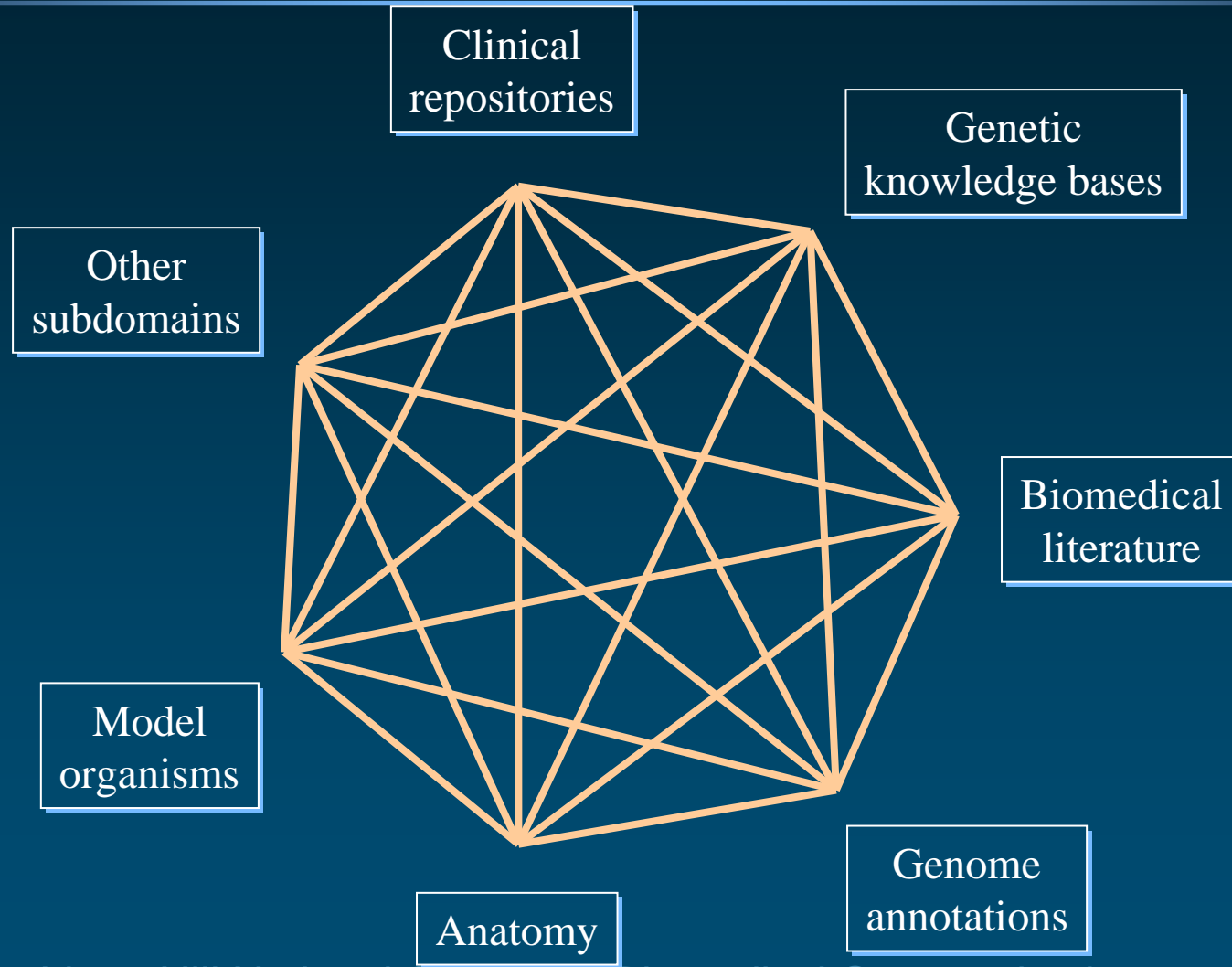
- ◆ Specialized vocabularies
 - nursing (NIC, NOC, NANDA, Omaha, PCDS)
 - dentistry (CDT)
 - psychiatry (DSM, APA)
 - adverse reactions (COSTART, WHO ART)
 - primary care (ICPC)
 - genomics (GO, OMIM, HUGO)
- ◆ Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)

Integrating subdomains



Integrating subdomains

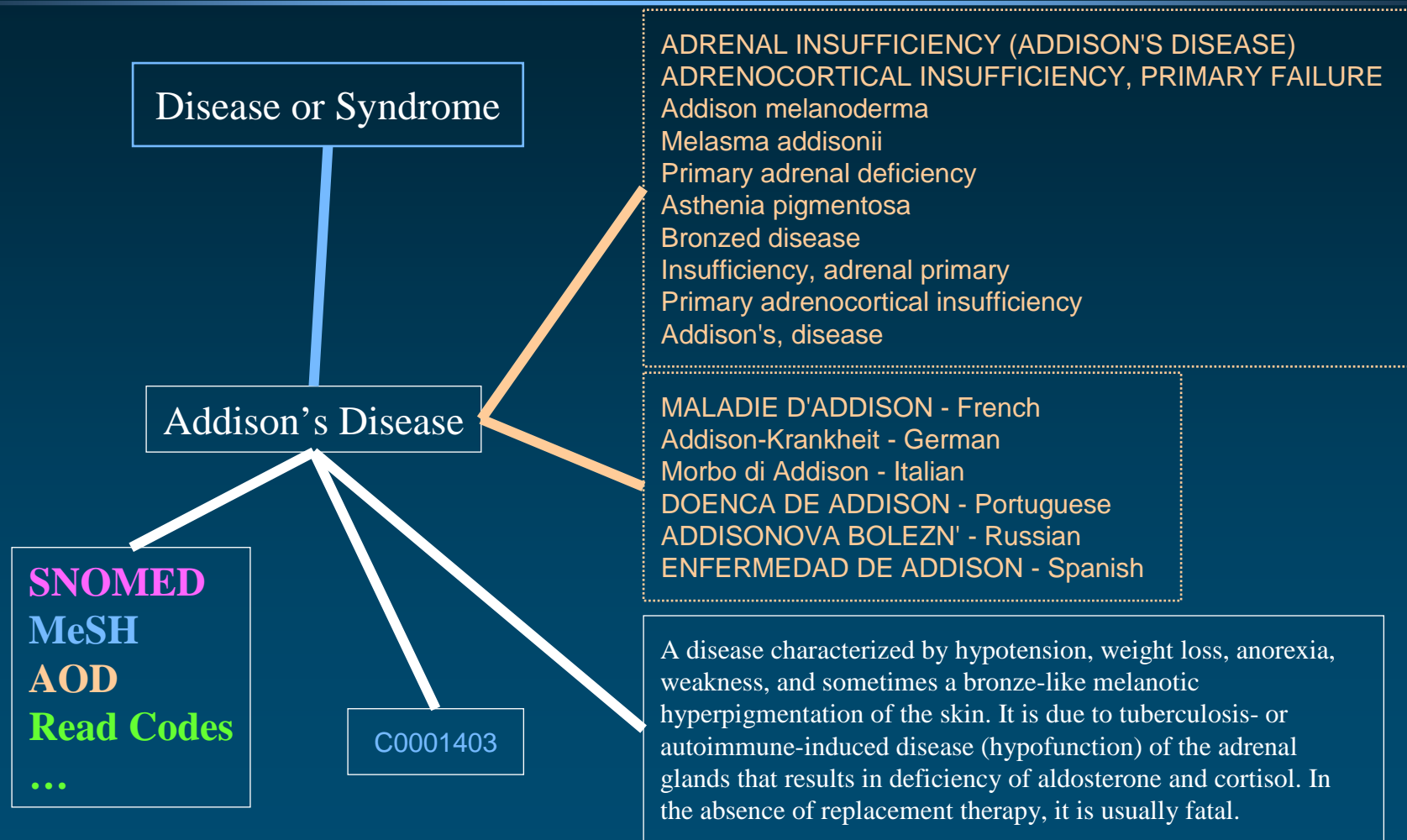


UMLS: 3 components

- ◆ Metathesaurus
 - Concepts
 - Inter-concept relationships
- ◆ Semantic Network
 - Semantic types
 - Semantic network relationships
- ◆ Lexical resources
 - SPECIALIST Lexicon
 - Lexical tools



Addison's Disease: Concept



Metathesaurus Concepts (2005AA)

- ◆ Concept (~ 1.2M) CUI
 - Set of synonymous concept names
- ◆ Term (~ 4.2 M) LUI
 - Set of normalized names
- ◆ String (~ 4.7M) SUI
 - Distinct concept name
- ◆ Atom (~ 5.5M) AUI
 - Concept name in a given source

A0000001 headache (source 1)
A0000002 headache (source 2)
S0000001

A0000003 Headache (source 1)
A0000004 Headache (source 2)
S0000002

L0000001

A0000005 Cephalgia (source 1)
S0000003

L0000002

C0000001



Cluster of synonymous terms

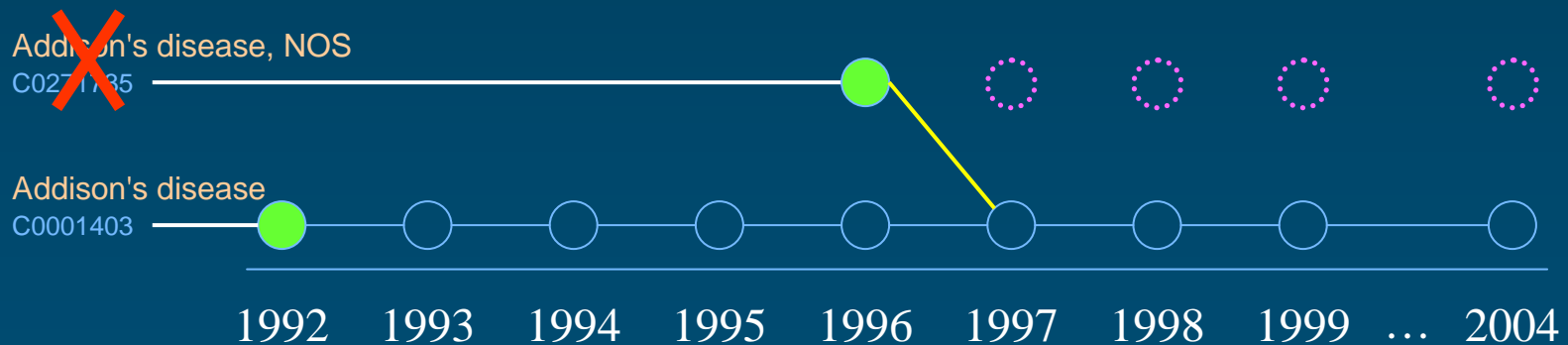
Concept
C0001621

Term L0001621	<p>S0011232 <i>Adrenal Gland Diseases</i></p> <p>S0011231 Adrenal Gland Disease</p> <p>S0000441 Disease of adrenal gland [...]</p> <p>S0481705 Disease of adrenal gland, NOS</p> <p>S0220090 Disease, adrenal gland</p> <p>S0044801 Gland Disease, Adrenal</p>
Term L0041793	<p>S0860744 <i>Disorder of adrenal gland, unspecified</i></p> <p>S0217833 Unspecified disorder of adrenal glands</p>
Term L0161347	<p>S0225481 <i>ADRENAL DISORDER</i> [...]</p> <p>S0627685 DISORDER ADRENAL (NOS)</p>
Term L0181041	<p>S0632950 <i>Disorder of adrenal gland</i> [...]</p> <p>S0354509 Adrenal Gland Disorders</p>
Term L0368399	<p>S0586222 <i>Adrenal disease</i> [...]</p> <p>S0466921 ADRENAL DISEASE, NOS</p>
Term L1279026	<p>S1520972 <i>Nebennierenkrankheiten</i> GER</p>
Term L0162317	<p>S0226798 <i>SURRENALE, MALADIES</i> FRE [...]</p>



Metathesaurus Evolution over time

- ◆ Concepts never die (in principle)
 - CUIs are permanent identifiers
- ◆ What happens when they do die (in reality)?
 - Concepts can merge or split
 - Resulting in new concepts and deletions



Metathesaurus Relationships

- ◆ Symbolic relations: ~9 M pairs of concepts
- ◆ Statistical relations : ~7 M pairs of concepts
(co-occurring concepts)
- ◆ Mapping relations: 100,000 pairs of concepts

-
- ◆ Categorization: Relationships between concepts and semantic types from the Semantic Network



Symbolic relations

◆ Relation

- Pair of “atom” identifiers
- Type
- Attribute (if any)
- List of sources (for type and attribute)

◆ Semantics of the relationship: defined by its *type* [and *attribute*]

Source transparency: the information
is recorded at the “atom” level



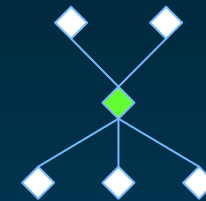
Symbolic relationships Type

◆ Hierarchical

- Parent / Child
- Broader / Narrower than

PAR / CHD

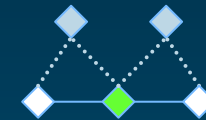
RB / RN



◆ Derived from hierarchies

- Siblings (children of parents)

SIB



◆ Associative

- Other

RO



◆ Various flavors of near-synonymy

- Similar
- Source asserted synonymy
- Possible synonymy

RL

SY

RQ

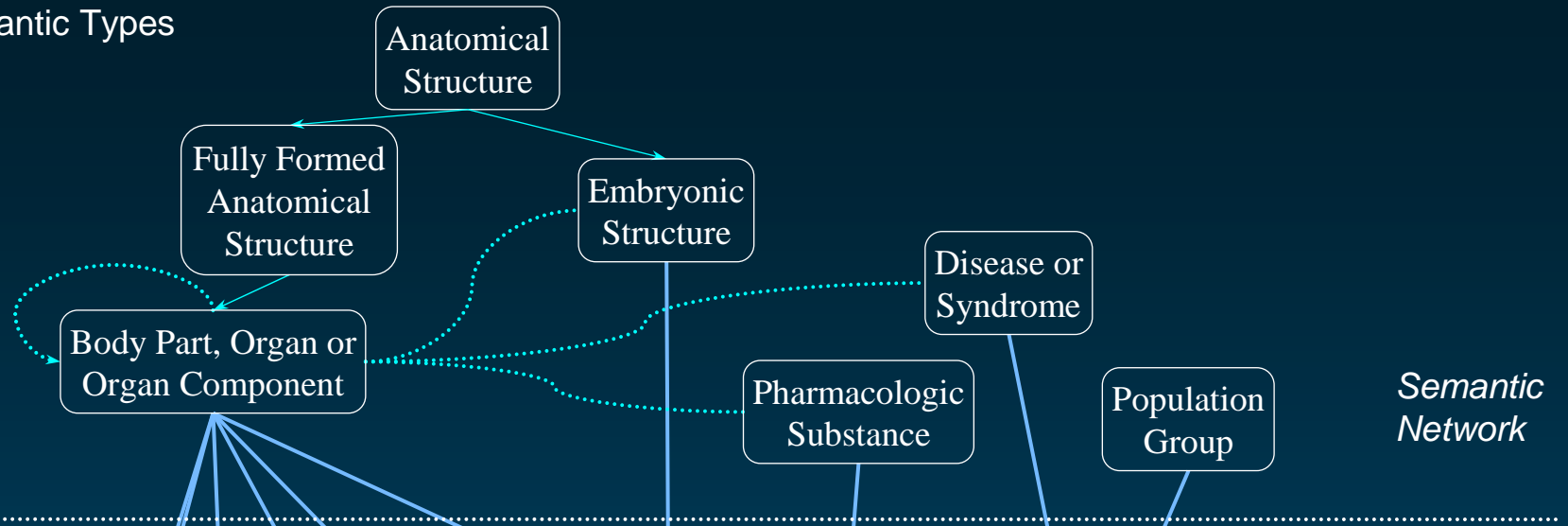


Symbolic relationships Attribute

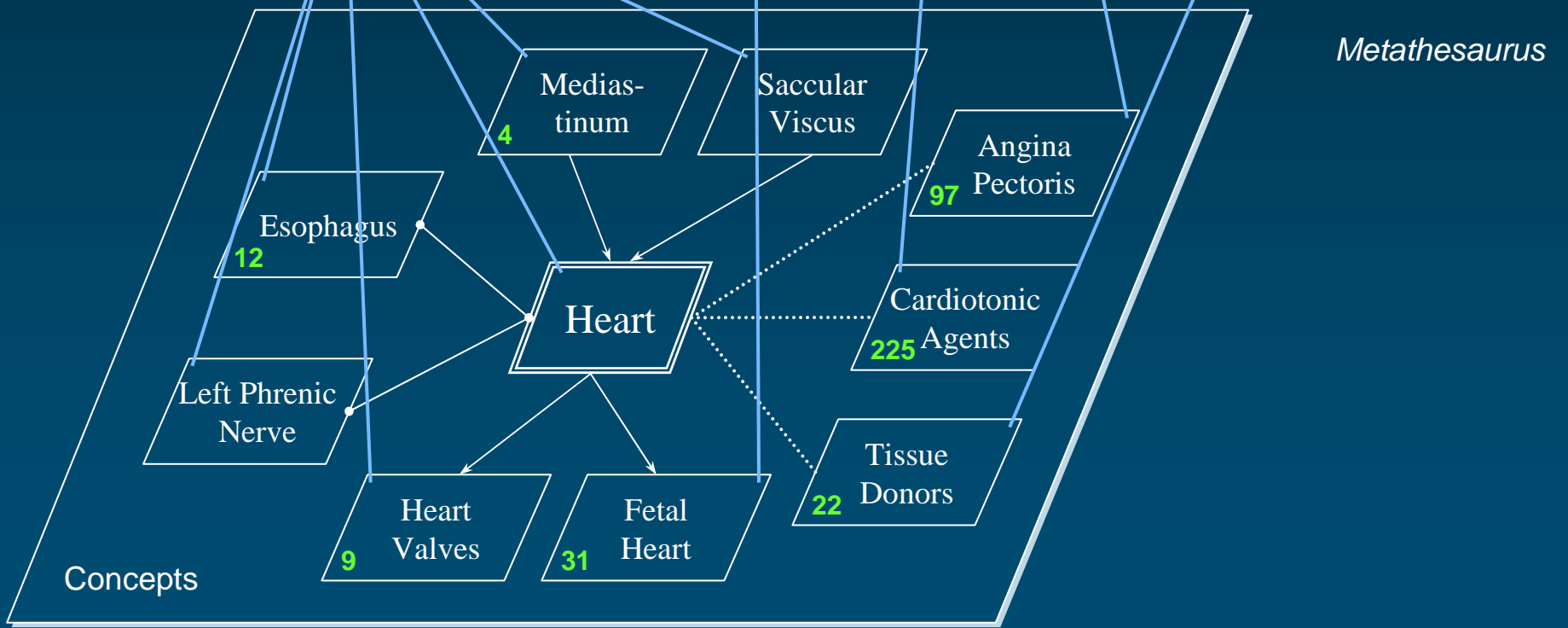
- ◆ Hierarchical
 - isa (is-a-kind-of)
 - part-of
- ◆ Associative
 - location-of
 - caused-by
 - treats
 - ...
- ◆ Cross-references (mapping)



Semantic Types



Semantic Network



Metathesaurus

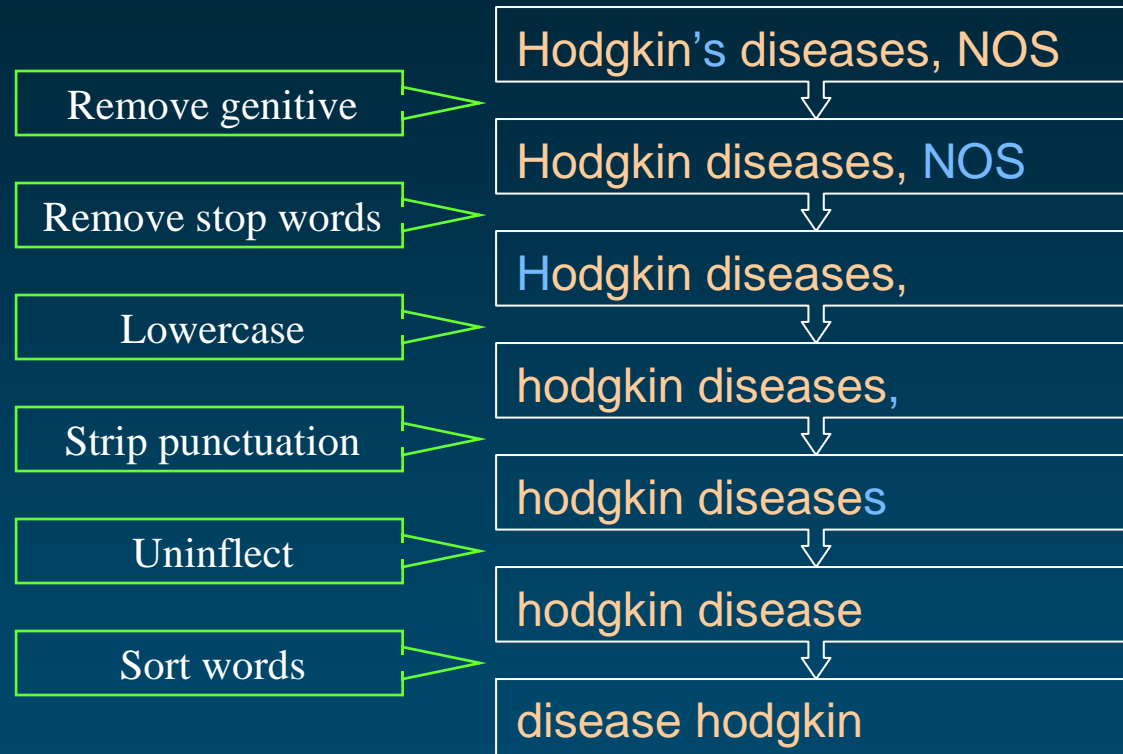
Concepts

Lexical tools

- ◆ To manage lexical variation in biomedical terminologies
- ◆ Major tools
 - Normalization
 - Indexes
 - Lexical Variant Generation program (lvg)
- ◆ Based on the SPECIALIST Lexicon
- ◆ Used by noun phrase extractors, search engines



Normalization



Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

normalize

disease hodgkin



Information integration

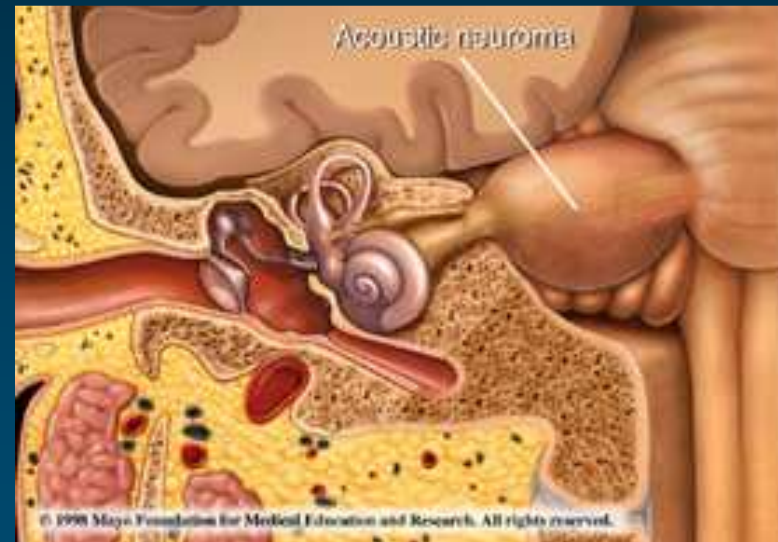
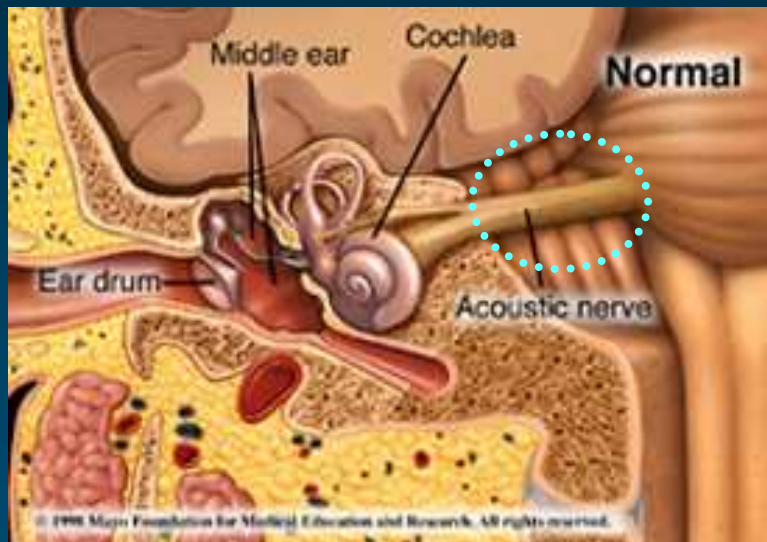
Genomics as an example

NF2 Gene, protein, and disease

Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the *NF2 gene* resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called *merlin* (but also neurofibromin 2 and schwannomin) and functions as a tumor suppressor.



Schwannoma (acoustic neuroma)



<http://www.mayoclinic.com>



{UMLS_2003} UMLS@ Semantic Navigator [2.10] - Netscape

{UMLS_2003} UMLS@ Semantic Navigator ...

Siblings

Disorders

- Cerebellopontine Angle Acoustic Neuroma ☒
- Diffuse neurofibroma ☒
- Melanocytic Vestibular Schwannoma ☒
- Neurofibromatosis (nonmalignant) ☒
- Neurofibromatosis 1 ☒
- neurofibromatosis 1 and 2 (NF1 and NF2) ☒
- Neurofibromatosis 3 ☒
- Neurofibromatosis type 3 ☒
- NEUROFIBROMATOSIS TYPE IV, OF RICCARDI ☒
- Neuroma, Acoustic, Unilateral ☒
- Segmental neurofibromatosis ☒

(11 siblings)

[direct children and narrower concepts of direct parents and broader concepts]

```

graph TD
    A[Tumor of acoustic vestibular nerve] --> D[Neurofibromatosis 2]
    B[Benign neoplasm of cranial nerves] --> D
    C[Neoplastic Syndromes, Hereditary] --> D
    E[Skin tumor of neural c...] --> D
    D --> F([Neuroma, Acoustic, Bilateral])
    D --> G([Schwannoma, Acoustic, Bilateral])
    
```

Other Related Concepts

Anatomy

- Acoustic Nerve ☒

Chemicals & Drugs

- Neurofibromin 2 ☒

Disorders

- Familial Acoustic Neuromas ☒
- Neoplasm of uncertain behavior NOS ☒
- Neurofibromatoses ☒
- Neurofibromatosis

- Nerve sheath Tumors [4] ☒
- Nervous System Neoplasms [6] ☒
- Neurolemmoma [35]
- Neurofibromatosis 1 [38] ☒
- Neuroma, Acoustic [26] ☒
- Peripheral Nervous System Diseases [3] ☒
- Peripheral Nervous System Neoplasms [6] ☒
- Postoperative Complications [9] ☒
- Retinal Diseases [6] ☒
- Skin Neoplasms [9] ☒

BCI **Neurofibromatosis 2** **LEGEND** *

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS_2003

Type of hierarchical rel.: All Parent/Child only Broader/Narrower only

Similar Concepts

(none)

Allegedly Synonyms

- Neurofibromatosis (nonmalignant) ☒

Closest MeSH Terms

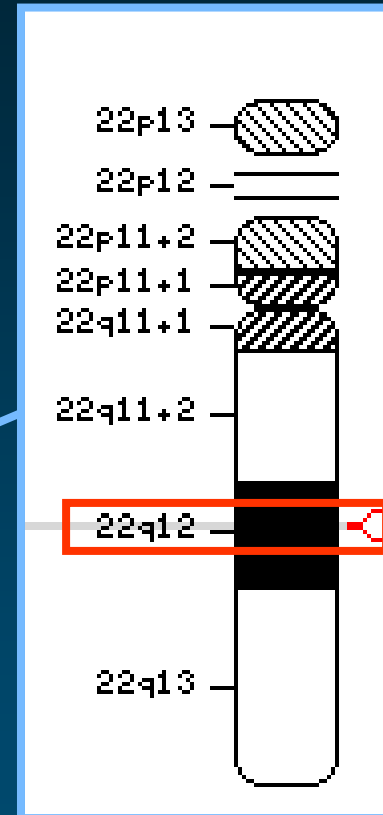
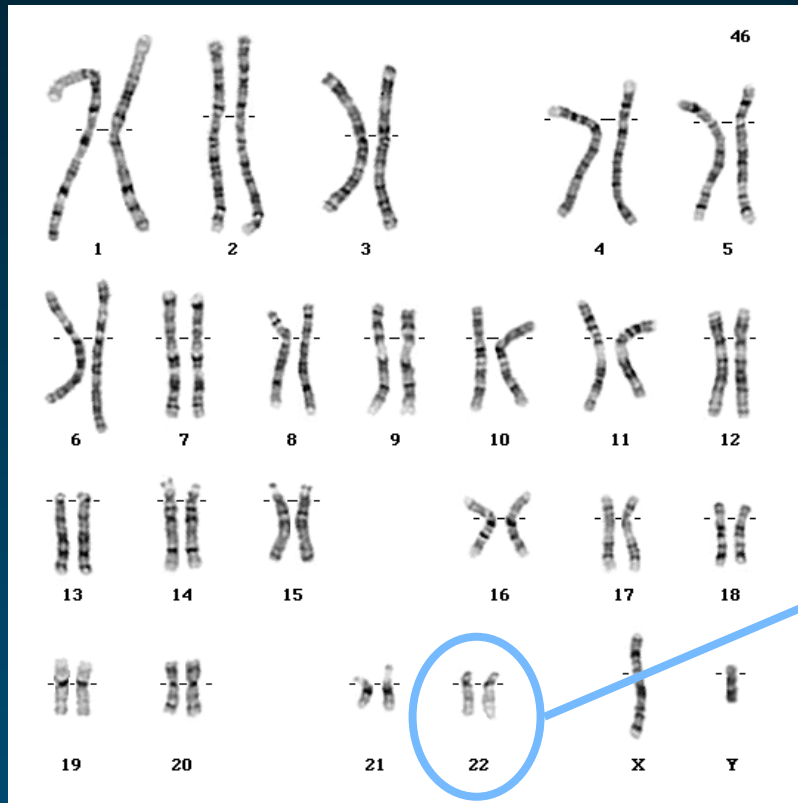
Main Headings

- Neurofibromatosis 2

Subheadings

Document: Done (1.328 secs)

NF2 gene

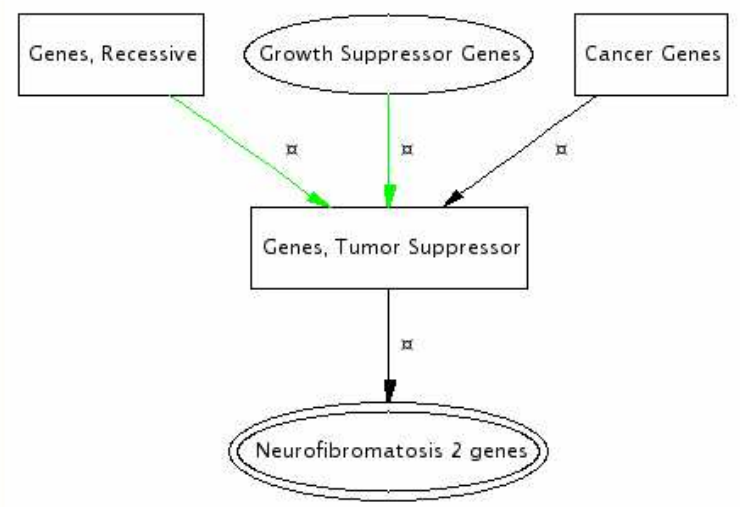


<http://staff.washington.edu/timk/cyto/human/>

<http://www.ncbi.nlm.nih.gov/mapview/>



- Siblings**
- Chemicals & Drugs**
- ADAM11 protein, human ☒
 - DLG5 protein, human ☒
 - DPM3 protein, human ☒
 - HCCS-1protein, human ☒
 - hssh3bp1 protein, human ☒
 - HUGL protein, human ☒
 - LAPSER1 protein, human ☒
 - mitochondria proteolipid-like protein, human ☒
 - MRG protein, human ☒
 - p53 gene/protein ☒
 - PLAGL1 protein, human ☒
 - RARRES3 protein, human ☒
 - SEZ6L protein, human ☒
 - TES protein, human ☒



- Genes & Molecular Sequences**
- APC Gene ☒
 - BAX Gene ☒
 - brca gene ☒
 - CDH1 gene ☒
 - CHES1 Gene ☒
 - cyclin-dependent kinase inhibitor 2A

- Other Related Concepts**
- Chemicals & Drugs**
- Neurofibromin 2 ☒
- Disorders**
- Neurofibromatosis 2 ☒
- (2 other related concepts)

- Closest MeSH Terms**
- Chromosome Deletion [7] ☒
 - Ependymoma [4] ☒
 - Glioma [4] ☒
 - Loss of Heterozygosity [7] ☒
 - Meningeal Neoplasms [25] ☒
 - Meningioma [30] ☒
 - mesothelioma <1> [4] ☒
 - Neoplasms [4] ☒
 - Neurilemmoma [20] ☒
 - Neurofibromatoses [64] ☒
 - Neurofibromatosis 2 [64] ☒
 - Neuroma, Acoustic [5] ☒
 - Spinal Cord Neoplasms [3] ☒

BCI **Neurofibromatosis 2 genes** **LEGEND ***

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS_2003

Type of hierarchical rel.: All Parent/Child only Broader/Narrower only

Similar Concepts (none)

Allegedly Synonyms (none)

Main Headings

- Genes, Neurofibromatosis 2

Subheadings

Merlin

◆ Synonyms

- Neurofibromin 2
- Schwannomin
- Schwannomerlin
- Neurofibromatosis-2

◆ 10 isoforms

◆ Annotations

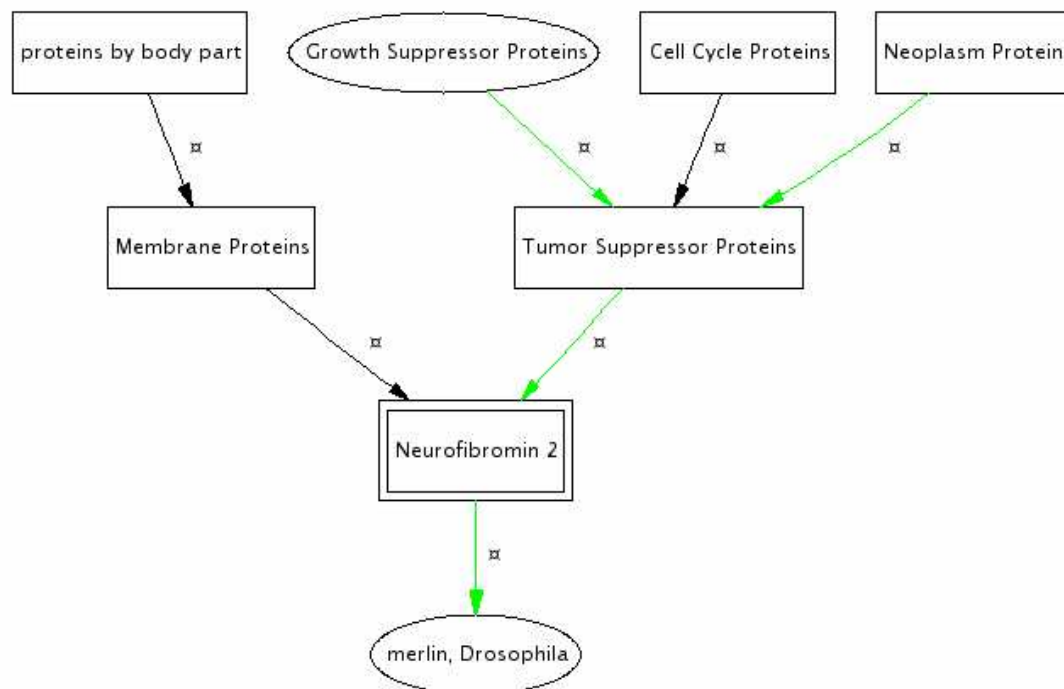
- Negative regulation of cell proliferation
- Cytoskeleton
- Plasma membrane



Siblings

Chemicals & Drugs

- (LA)12 peptide ✖
- (methyl)ammonium uptake carrier, Corynebacterium ✖
- 120-kDa hemocyte-specific membrane protein, flesh fly ✖
- 15a protein, Aedes aegypti ✖
- 22.6-kDa antigen, Schistosoma japonicum ✖
- 36-kDa vesicular integral membrane protein ✖
- 38L protein ✖
- 5-lipoxygenase-activated protein ✖
- 59 kDa dystrophin-associated protein ✖
- A-1 antigen ✖
- A-kinase anchor protein 149 ✖
- A-kinase anchor protein 15 ✖
- A-kinase anchor protein 200 ✖
- A-kinase anchor protein KL ✖
- A14.5L protein ✖
- A15 protein ✖
- ABC-me protein ✖
- ABU-1 protein, C elegans ✖
- AcfB protein ✖
- ACR3 protein ✖



Other Related Concepts

Disorders

- Neurofibromatosis 2 ✖

Genes & Molecular Sequences

- Neurofibromatosis 2 genes ✖

(2 other related concepts)

Co-occurring Concepts

Anatomy

- Arachnoid [1] ✖
- Cell Membrane [1] ✖
- Cerebellum [1] ✖
- Chromosomes, Human, Pair 22 [1] ✖
- Cytoplasm [1] ✖
- Cytoskeleton [2] ✖
- Microfilaments [1] ✖
- Purkinje Cells [1] ✖
- Schwann Cells [1] ✖
- Stem Cells [1] ✖

BCI

Neurofibromin 2

LEGEND *

Start again

Apply new parameters

Restrict to vocabulary:

Show all

Highlight vocabulary:

Nothing

UMLS data:

UMLS_2003

Type of hierarchical rel.:

All Parent/Child only

Broader/Narrower only

Similar Concepts

(none)

Allegedly Synonyms

(none)

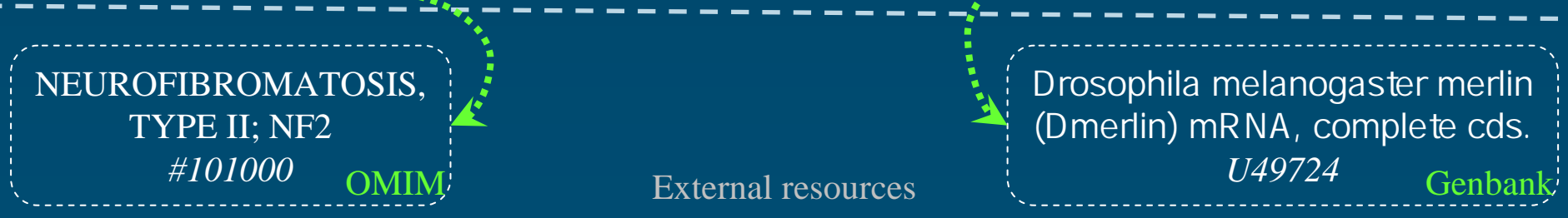
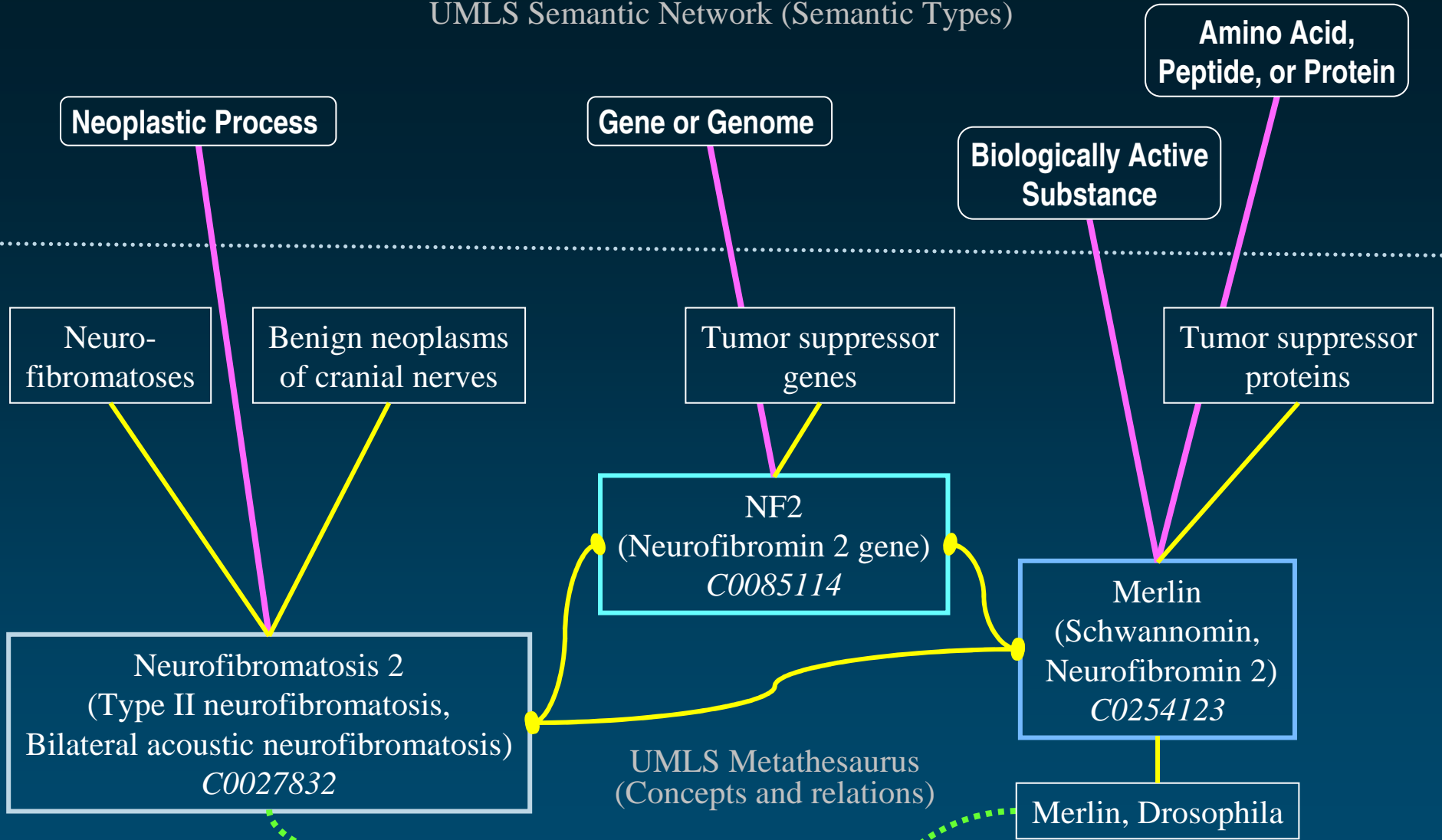
Closest MeSH Terms

Main Headings

- Neurofibromin 2

Subheadings

UMLS Semantic Network (Semantic Types)



Limitations

- ◆ Genes not systematically represented
 - Most gene products and diseases are
- ◆ Gene/Gene product-Disease relations
 - Not systematically represented
 - Not explicitly represented (e.g., co-occurrence)
- ◆ Cross-references not systematically represented
- ◆ Naming conventions (genes)



Applications (1)

*GenesTrace*TM

Lussier Lab
Columbia University

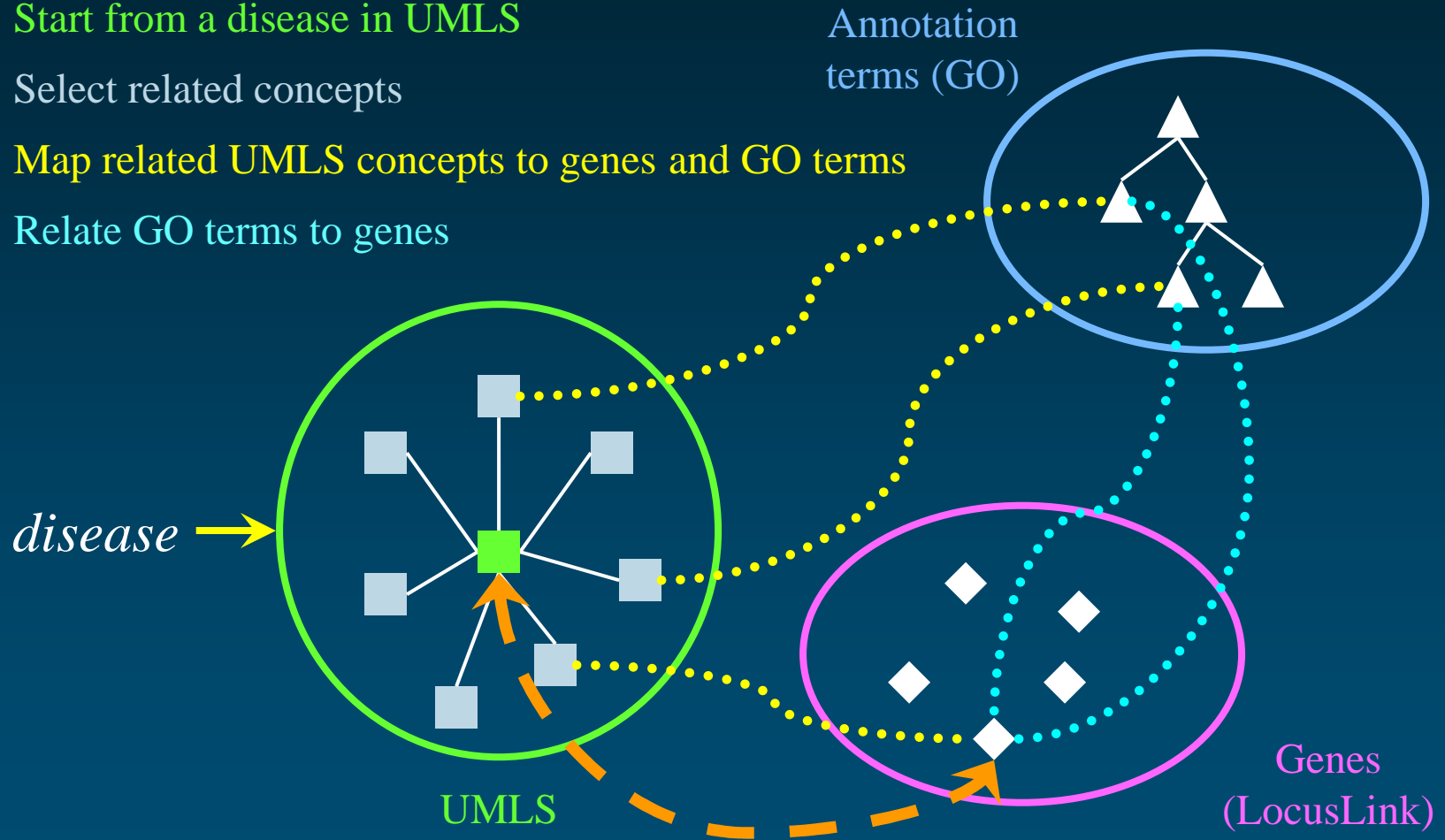
Objectives

- ◆ Relate diseases to genes through structured, integrated terminologies
- ◆ Biological Knowledge Discovery



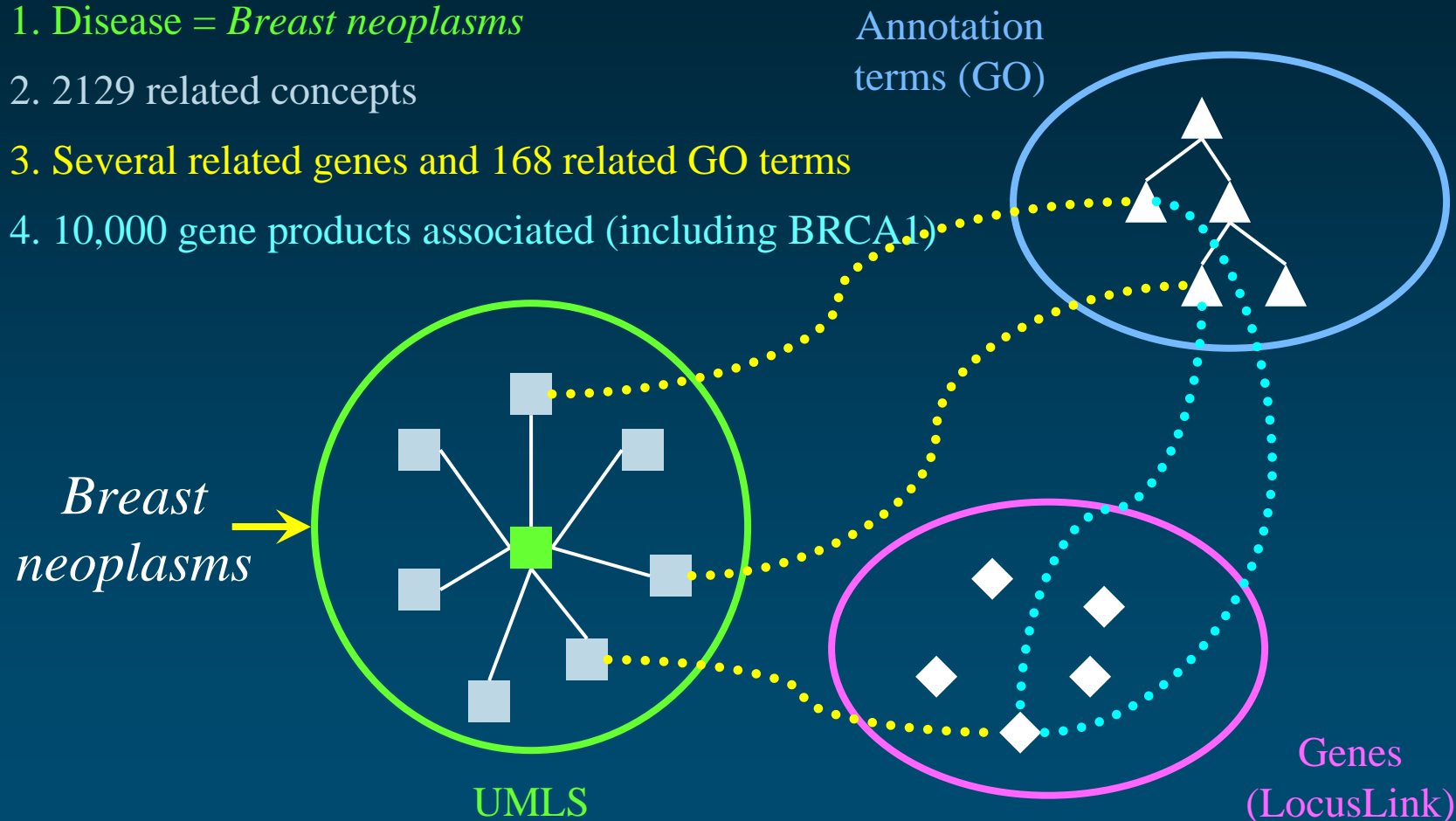
Resources and Methods

1. Start from a disease in UMLS
2. Select related concepts
3. Map related UMLS concepts to genes and GO terms
4. Relate GO terms to genes



Validation Breast cancer – BRCA1 association

1. Disease = *Breast neoplasms*
2. 2129 related concepts
3. Several related genes and 168 related GO terms
4. 10,000 gene products associated (including BRCA1)



Limitations

◆ Noise

- Too many non-specific GO terms associated (e.g., *nucleus*)
- Too many genes associated

◆ But

- Promising preliminary results
- Room for refinement



Lussier Lab

Columbia University

GenesTrace

- About
- Help
- People
- Grants

GenesTrace Online

[Home](#) | [About](#)

Search GenesTrace for diseases and genes sharing identical processes, functions or biological structures:

Disease or Gene.....

Query using.....

Please select the species database from below

- Fly(FB) Mouse(MGI) Worm(WB) Yeast(SGD) Swissprot(SPTR)

[GenesTrace - Columbia University](#)

[Help](#)

Applications (2)

BioMeKe

G. Marquet & al.

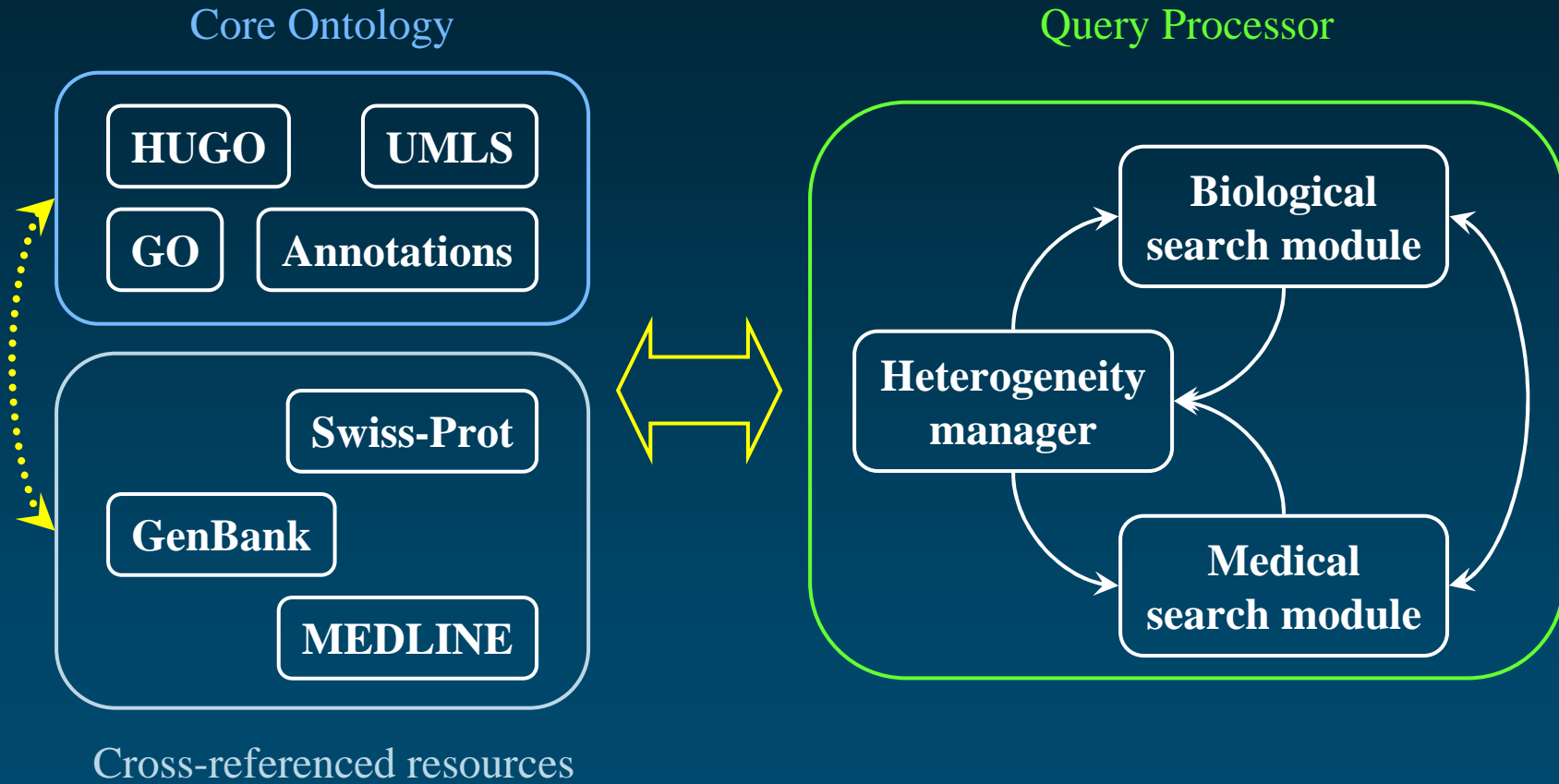
LIM, Univ. Rennes, France

Objectives

- ◆ To develop a knowledge warehouse for transcriptome analysis (liver diseases)
- ◆ Semantic interoperability
 - Medical knowledge bases
Clinical genomics
 - Molecular biology and genetics knowledge bases
Functional genomics



Components

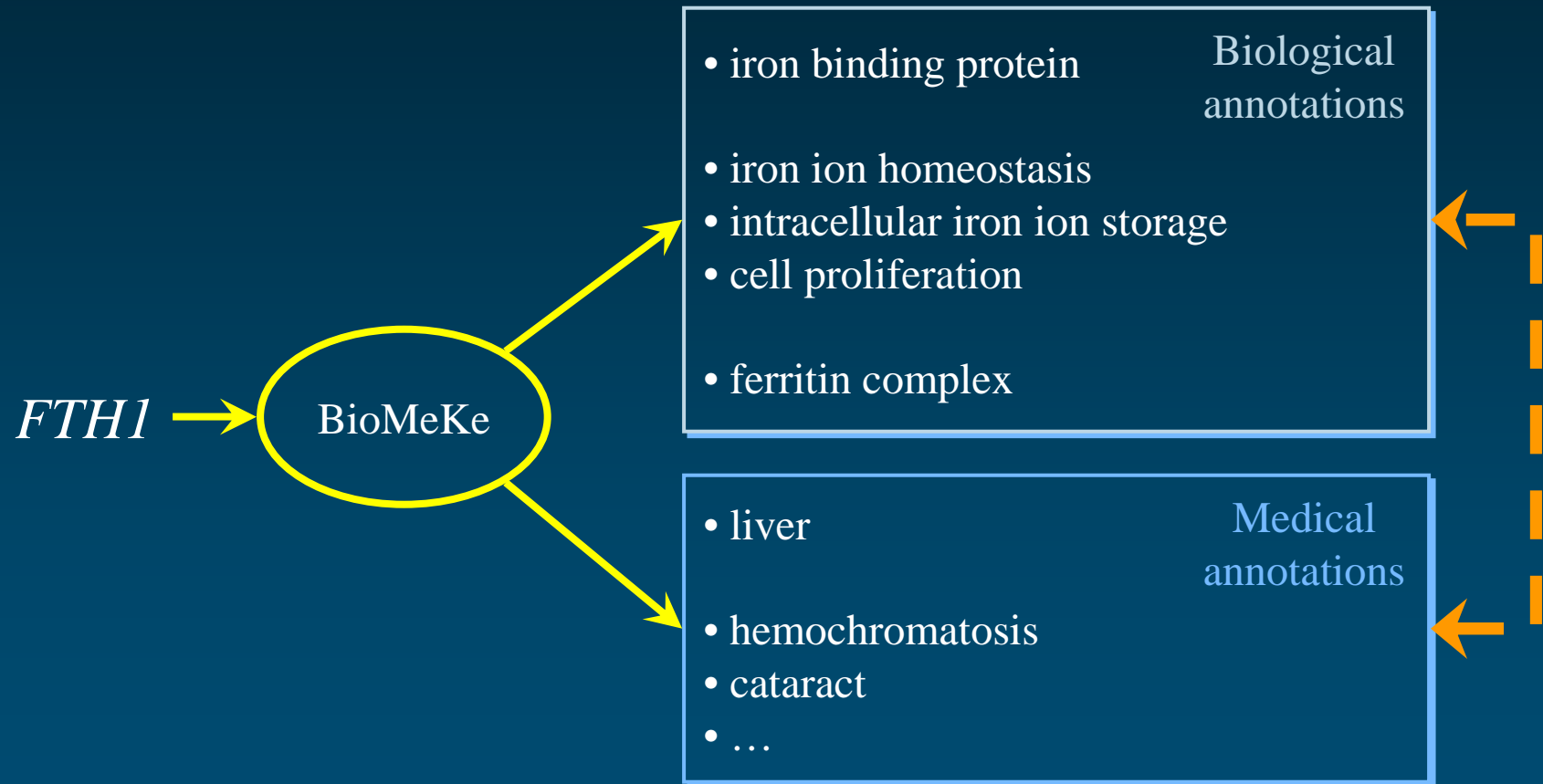


Example

- ◆ Input: *ferritin, heavy polypedpide 1*
- ◆ Mapping to biological resources
 - Not found in the Core ontology
 - Official name *Ferritin heavy chain* found through Xref
- ◆ Biological information obtained from GOA
- ◆ Mapping to medical resources
 - Not found in UMLS
 - Synonym *Ferritin H* found through Xref (Swiss-Prot)
- ◆ Medical information obtained through co-occurrence of MeSH index terms in MEDLINE



Results



Limitations

- ◆ Non-formal ontologies
 - Knowledge may be inconsistently represented
 - Knowledge may be implicit (mappings)
- ◆ Partial automation
 - User input required to select databanks, reformulate queries
- ◆ Semantic integration
 - Naming issues
 - Mappings must be updated regularly





Biological and Medical Knowledge Extraction system

BioMeKE has been achieved to **extract** and to **associate** Medical and Biological information such as Gene Ontology™, Unified Medical Language System®, Genew ...

BioMeKe is an **ontology-based tool** composed of two-part :

- Consultation of Ontology and link with Public Databank
- Biological and Medical annotation

Home

Registration

Download

Installation

Help

Information

Results

References

Genomic and Biological Resources in BioMeKE:

[Genew](#) : Lexical ressources providing official gene names and their synonyms and links to multiple others databases including Uniprot , LocusLink .

[Gene Ontology \(GO\)](#) : general "ontology" for molecular biology which provides a controlled vocabulary for annotationg sequences and genes products.

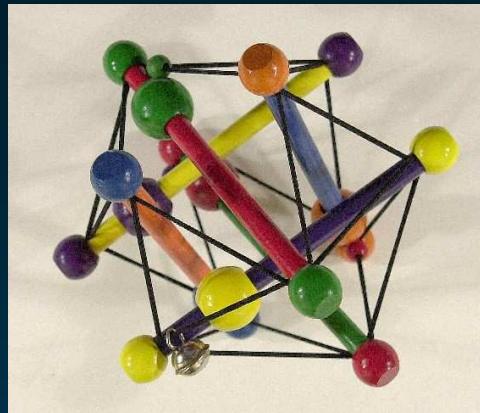
[GeneOntology Annotation@EBI \(GOA\)](#) : provides assignments of GO terms to genes products for all organisms, including human.

Conclusions

Conclusions

- ◆ Terminology integration provides some degree of information integration
- ◆ Most terminologies and the cross-referenced databases are readily available
- ◆ Lack of consistent representation
- ◆ Additional resources/techniques needed





Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Questions

- ◆ What do I need to do to get the UMLS?
- ◆ What is an ontology?
- ◆ How is ontology different from
 - Terminology? / Database? / Knowledge base?
- ◆ Is the UMLS an ontology?
- ◆ Does the UMLS use Protégé?
- ◆ I heard of OWL. Is that any good?
- ◆ What is the Semantic Web going to do for us?



References UMLS

◆ UMLS

umlsinfo.nlm.nih.gov

◆ UMLS browser

- Knowledge Source Server: umlsks.nlm.nih.gov
- Semantic Navigator:
<http://mor.nlm.nih.gov/perl/semnav.pl>
- (free, but UMLS license required)

◆ UMLS and information integration

- O. Bodenreider. The UMLS: Integrating biomedical terminology. *Nucl. Acids Res.* 2004;32(1) (*in press*)



References Applications

◆ GenesTrace

- Cantor MN, Sarkar IN, Bodenreider O, Lussier YA. GenesTrace: Phenomic knowledge discovery via structured terminologies. In: Pacific Symposium on Biocomputing 2005; 2005. (in press).
- <http://phene.cpmc.columbia.edu:8080/genesTrace/index.jsp>

◆ BioMeKE

- Marquet G, Burgun A, Moussouni F, Guerin E, Le Duff F, Loreal O. BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis. Stud Health Technol Inform. 2003;95:80-5.
- <http://www.med.univ-rennes1.fr/~marquet/>

