

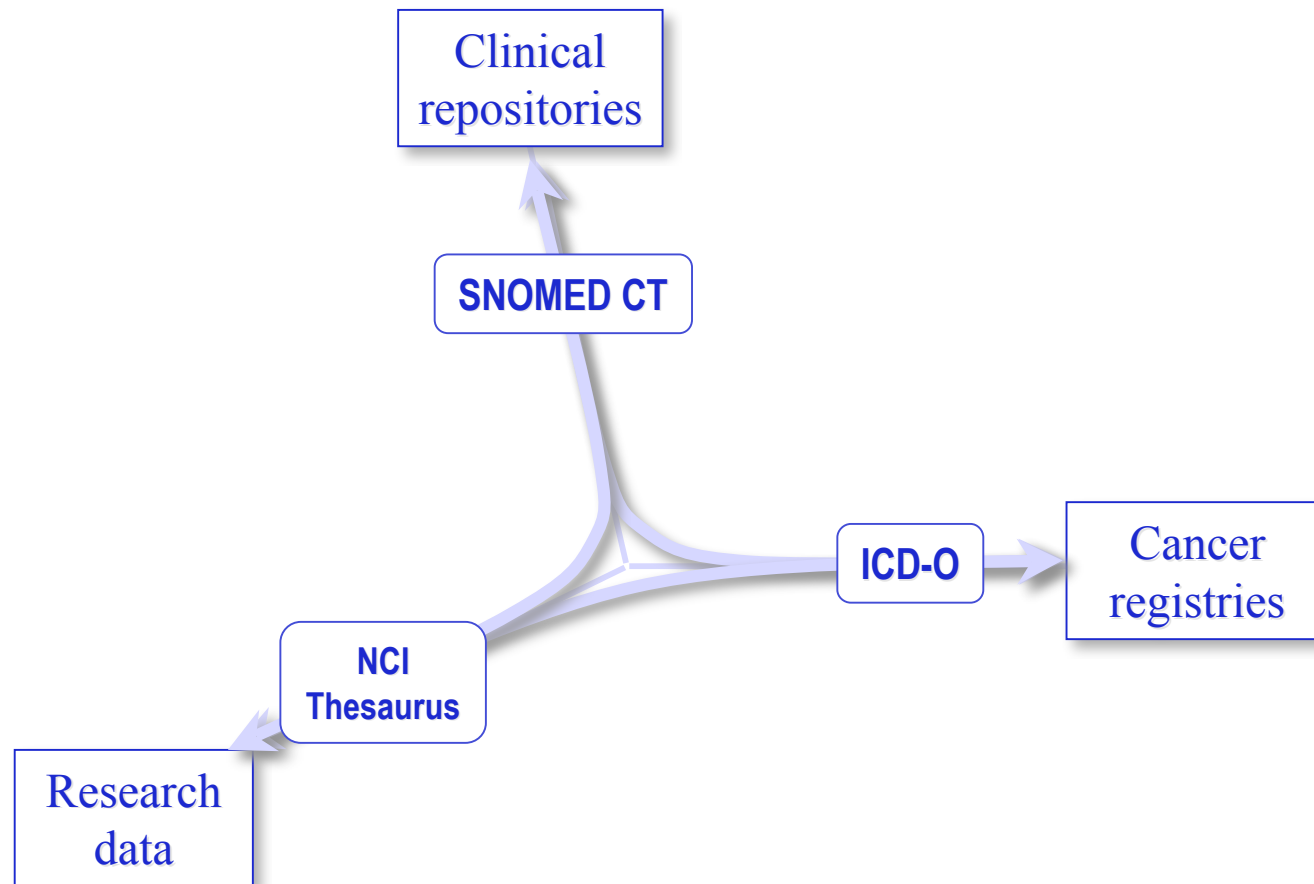
Disease ontologies and their relation to genes

Anita Burgun, MD, PhD

EA 3888 - IFR140 Faculté de Médecine – Université de Rennes1
- France

EBI Industry Programme Workshop
Hinxton, UK, 19 Jun 2008

Disease ontologies in human medicine



Cancers : their representation in the NCI Thesaurus



- Morphology
 - Neoplastic cell, Ringed sideroblast
- Location
- Stage: a measure of how much the cancer has grown and spread.
 - TNM (Tumor, Nodes, Metastases)
- Grade: degree of malignancy
 - Low grade tumors
 - the tumor cells tend to be slow growing, are well differentiated, tend to be less 'aggressive', and are less likely to spread quickly.
 - High grade tumors
 - the tumor cells tend to be fast growing, are poorly differentiated, tend to be more 'aggressive', and are more likely to spread quickly.
- Cytogenetic abnormalities, genes

Definition of Glioblastoma in NCIT (extract)

- ☒ Disease_Has_Finding **some** Mitotic_Activity
- ☒ Disease_Has_Finding **some** Necrotic_Change
- ☒ Disease_Has_Finding **some** Infiltrative_Pattern
- ☒ Disease_Is_Grade **only** Grade_4
- ☒ Disease_May_Have_Abnormal_Cell **some** Fibrillary_Neoplastic_Astrocyte
- ☒ Disease_May_Have_Abnormal_Cell **some** Gemistocytic_Neoplastic_Astrocyte
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q25-26
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q23
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Gain_of_Chromosome_7q
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_9p
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_10p
- ☒ Disease_May_Have_Finding **some** Seizure
- ☒ Disease_May_Have_Finding **some** Headache
- ☒ Disease_May_Have_Molecular_Abnormality **some** PTEN_Tumor-Suppressor_Gene_Inactivation

NEC

INF

- ☒ Disease_Excludes_Finding **only** Cellular_Origin_Unknown [from Neuroglial_Tumor;
- ☒ Disease_Excludes_Finding **some** Neuronal_Differentiation [from Neuroglial_Tumor;

Definition of Glioblastoma in NCIT (extract)

- ⊖ Disease_Has_Finding **some** Mitotic_Activity
- ⊖ Disease_Has_Finding **some** Necrotic_Change
- ⊖ Disease_Has_Finding **some** Infiltrative_Pattern
- ⊖ Disease_Is_Grade **only** Grade_4
- ⊖ Disease_May_Have_Abnormal_Cell **some** Fibrillary_Neoplastic_Astrocyte
- ⊖ Disease_May_Have_Abnormal_Cell **some** Gemistocytic_Neoplastic_Astrocyte
- ⊖ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q25-26
- ⊖ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q23
- ⊖ Disease_May_Have_Cytogenetic_Abnormality **some** Gain_of_Chromosome_7q
- ⊖ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_9p
- ⊖ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_10p
- ⊖ Disease_May_Have_Finding **some** Seizure
- ⊖ Disease_May_Have_Finding **some** Headache
- ⊖ Disease_May_Have_Molecular_Abnormality **some** PTEN_Tumor-Suppressor_Gene_Inactivation

NEC

INF

⊖ Disease_Excludes_Finding **only** Cellular_Origin_Unknown

[from Neuroglial_Tumor;

⊖ Disease_Excludes_Finding **some** Neuronal_Differentiation

[from Neuroglial_Tumor;

Definition of Glioblastoma in NCIT (extract)

- ☒ Disease_Has_Finding **some** Mitotic_Activity
- ☒ Disease_Has_Finding **some** Necrotic_Change
- ☒ Disease_Has_Finding **some** Infiltrative_Pattern

☒ Disease_Is_Grade **only** Grade_4

- ☒ Disease_May_Have_Abnormal_Cell **some** Fibrillary_Neoplastic_Astrocyte
- ☒ Disease_May_Have_Abnormal_Cell **some** Gemistocytic_Neoplastic_Astrocyte
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q25-26
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q23
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Gain_of_Chromosome_7q
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_9p
- ☒ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_10p
- ☒ Disease_May_Have_Finding **some** Seizure
- ☒ Disease_May_Have_Finding **some** Headache
- ☒ Disease_May_Have_Molecular_Abnormality **some** PTEN_Tumor-Suppressor_Gene_Inactivation

NEC

INF

☒ Disease_Excludes_Finding **only** Cellular_Origin_Unknown

[from Neuroglial_Tumor;

☒ Disease_Excludes_Finding **some** Neuronal_Differentiation

[from Neuroglial_Tumor;

Definition of Glioblastoma in NCIT (extract)

- ☺ Disease_Has_Finding **some** Mitotic_Activity
- ☺ Disease_Has_Finding **some** Necrotic_Change
- ☺ Disease_Has_Finding **some** Infiltrative_Pattern
- ☹ Disease_Is_Grade **only** Grade_4
- ☺ Disease_May_Have_Abnormal_Cell **some** Fibrillary_Neoplastic_Astrocyte
- ☺ Disease_May_Have_Abnormal_Cell **some** Gemistocytic_Neoplastic_Astrocyte
- ☺ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q25-26
- ☺ Disease_May_Have_Cytogenetic_Abnormality **some** del_10q23
- ☺ Disease_May_Have_Cytogenetic_Abnormality **some** Gain_of_Chromosome_7q
- ☺ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_9p
- ☺ Disease_May_Have_Cytogenetic_Abnormality **some** Loss_of_Chromosome_10p
- ☺ Disease_May_Have_Finding **some** Seizure
- ☺ Disease_May_Have_Finding **some** Headache
- ☺ Disease_May_Have_Molecular_Abnormality **some** PTEN_Tumor-Suppressor_Gene_Inactivation

NEC

INF

- ☹ Disease_Excludes_Finding **only** Cellular_Origin_Unknown
- ☺ Disease_Excludes_Finding **some** Neuronal_Differentiation

[from Neuroglial_Tumor;

[from Neuroglial_Tumor;

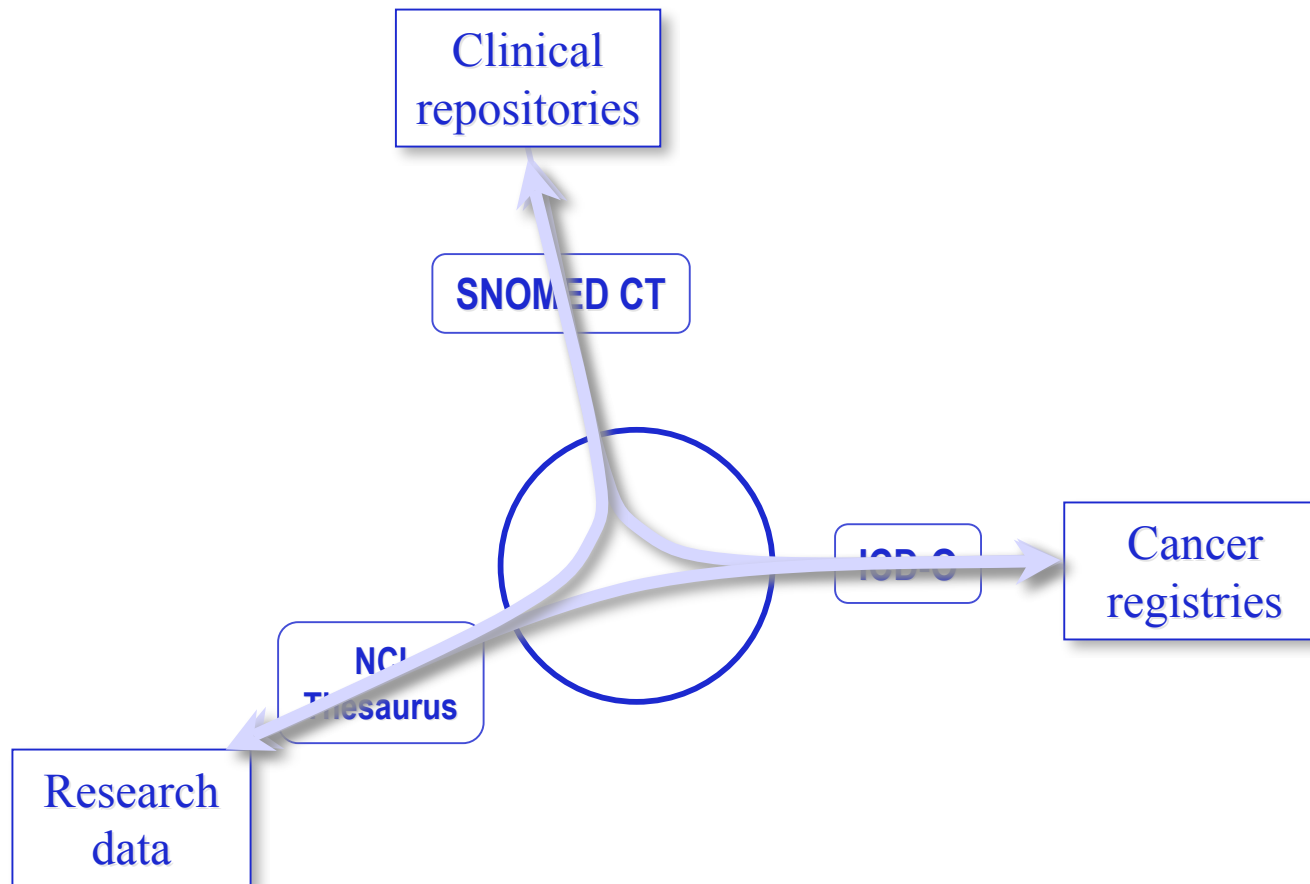
Limitations

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automatize acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

Issues

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automate acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

Mapping between DO



Terminology integration NCI Metathesaurus

The screenshot displays the NCI Metathesaurus interface. At the top left is the National Cancer Institute logo. In the center is the EVS (Enterprise Vocabulary Server) logo. On the right, there are tabs for 'Basic' and 'Advanced'. A search bar contains the word 'prostate' and a 'Search' button. Below the search bar, a navigation menu includes links for Concept, Definitions, Synonyms, Sources, Broader Concepts, Narrower Concepts, Related Concepts, Medications, Procedures, Laboratory, Diagnosis, Open NCI Hierarchy, and View Hierarchy Location. The main content area shows the concept 'C0033572: Prostate Gland' with the description 'Body Part, Organ, or Organ Component'. Below this is a table titled 'Sources of Prostate Gland' listing various terminologies and their corresponding terms. A dotted line highlights the 'ICDO3' entry 'Prostate, NOS' in the table, which is linked to the 'C0033572: Prostate Gland' concept header.

NATIONAL CANCER INSTITUTE

EVS ENTERPRISE VOCABULARY SERVER

Basic Advanced

prostate Search

The basic search is enabled. Click on the "Advanced" tab to customize your search criteria. You can mouse-over each advanced search item for help in utilizing the advanced features.

[Concept](#) | [Definitions](#) | [Synonyms](#) | [Sources](#) | [Broader Concepts](#) | [Narrower Concepts](#) | [Related Concepts](#) | [Medications](#) | [Procedures](#) | [Laboratory](#) | [Diagnosis](#) | [Open NCI Hierarchy](#) | [View Hierarchy Location](#)

C0033572: Prostate Gland ⓘ

Body Part, Organ, or Organ Component

Sources of [Prostate Gland](#)

ICDO3	HT	C61	PROSTATE GLAND
ICDO3	PT	C61.9	Prostate gland
ICDO3	SY	C61.9	Prostate, NOS
LNC217	LS	NOCODE	PROSTATE
MDBCAC2005_12	PT	0602	Prostate
MSH2007_2006_08_08	PM	D011467	Prostates
MSH2007_2006_08_08	MH	D011467	Prostate
MTH2006AD	PN	U002195	Prostate
NCI-GLOSS_0706D	PT	CDR0000046539	prostate
NCI2007_06D	PT	C12410	Prostate Gland

About
Browse
Copyright
New Term
Sources
User's Guide

Center for Bioinformatics

Office of Communications

FIRST GOV
Year First Click to the U.S. Government

Assess shared descriptions

- Associations of NCI Metathesaurus concepts
- 366 associations strictly equivalent
- ICD-O (22,881 associations) to NCIT
 - <2% complete matching
 - 87% partial matching
 - 11% none of the entities is found in NCIT
- NCIT (19,028 associations) to ICD-O
 - <2% complete matching
 - 53% partial matching
 - 45% none of the entities is found in ICD-O

Burgun A, Bodenreider O. Issues in Integrating Epidemiology and Research Information in Oncology: Experience with ICD-O3 and the NCI Thesaurus. AMIA Conf, Chicago, IL - Nov, 2007

Bodenreider O, Burgun A. Oncologic pathology in biomedical terminologies. Challenges for data integration. APIII conference on "*Anatomic Pathology Informatics and Imaging Support for Translational Medicine*" Pittsburgh, PA - Sept 10, 2007

Issues

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automate acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

Candidate ontologies

KEGG Orthology (KO) hierarchy



- Organization of metabolic pathway and disease maps
- DAG, four levels

KEGG KEGG Orthology (KO) Hierarchy

[1st Level | 2nd Level]

- ▶ **01100 Metabolism**
- ▶ **01200 Genetic Information Processing**
- ▶ **01300 Environmental Information Processing**
- ▶ **01400 Cellular Processes**
- ▶ **01500 Human Diseases**

[[BRITE](#) | [KEGG2](#) | [KEGG](#)]
Last updated: June 28, 2011

- ▼ **01100 Metabolism**
 - ▼ **01110 Carbohydrate Metabolism**
 - ▶ 00010 Glyc
 - ▶ 00020 Citr
 - ▶ 00030 Pent
 - ▶ 00040 Pent
 - ▶ 00051 Fruc
 - ▶ 00052 Gala
 - ▶ 00053 Asco
 - ▶ 00500 Star
 - ▶ 00530 Amin
 - ▶ 00520 Nucl
 - ▶ 00620 Pyru
 - ▶ 00630 Glyc
 - ▶ 00640 Prop
 - ▶ 00650 Buta
 - ▶ 00660 C5-B
 - ▶ 00031 Inos
 - ▶ 00562 Inos
 - ▼ **00010 Glycolysis / Gluconeogenesis [PATH:ko00010]**
 - [K00845](#) E2.7.1.2, glk; glucokinase [EC:2.7.1.2]
 - [K00844](#) E2.7.1.1; hexokinase [EC:2.7.1.1]
 - [K01084](#) E3.1.3.9, G6PC; glucose-6-phosphatase [EC:3.1.3.9]
 - [K01810](#) E5.3.1.9, pgi; glucose-6-phosphate isomerase [EC:5.3.1.9]
 - [K06859](#) E5.3.1.9A; glucose-6-phosphate isomerase [EC:5.3.1.9]
 - [K00850](#) E2.7.1.11, pfk; 6-phosphofructokinase [EC:2.7.1.11]
 - [K01086](#) E3.1.3.11; fructose-1,6-bisphosphatase [EC:3.1.3.11]
 - [K03841](#) FBP1, FBP2, fbp; fructose-1,6-bisphosphatase I [EC:3.1.3.11]
 - [K02446](#) GLPX; fructose-1,6-bisphosphatase II [EC:3.1.3.11]
 - [K04041](#) FBP3, fbp; fructose-1,6-bisphosphatase III [EC:3.1.3.11]
 - [K01622](#) E4.1.2.13; fructose-bisphosphate aldolase [EC:4.1.2.13]
 - [K01624](#) E4.1.2.13B, fbaA; fructose-bisphosphate aldolase, class II [EC:4.1.2.13B]
 - [K01623](#) E4.1.2.13A, fbaB; fructose-bisphosphate aldolase, class I [EC:4.1.2.13A]
 - [K01803](#) E5.3.1.1, tpiA; triosephosphate isomerase (TIM) [EC:5.3.1.1]
 - [K00134](#) E1.2.1.12, GAPD, gapA; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]
 - [K00927](#) E2.7.2.3, pgk; phosphoglycerate kinase [EC:2.7.2.3]
 - [K01834](#) E5.4.2.1, gpm; phosphoglycerate mutase [EC:5.4.2.1]
 - [K01689](#) E4.2.1.11, eno; enolase [EC:4.2.1.11]

Candidate ontologies

Gene Ontology (GO)



3 hierarchies :

- Molecular Function
- Cellular Component
- Biological Process

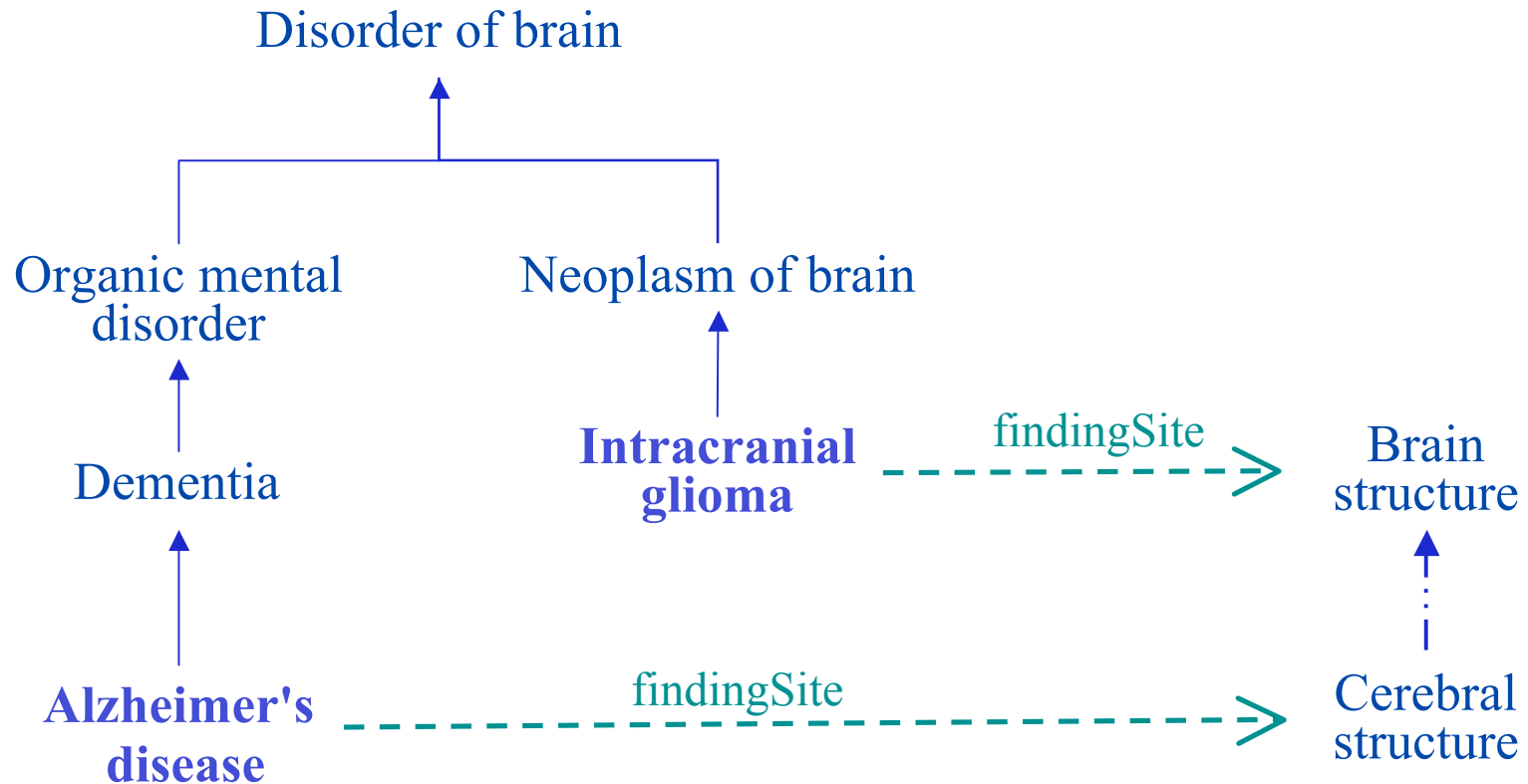
- [-] GO:0008150 : biological_process (168536)
 - [+] GO:0007610 : behavior (4588)
 - GO:0000004 : biological_process unknown (37303)
 - [-] GO:0009987 : cellular_process (113158)
 - [-] GO:0007154 : cell communication (17444)
 - [+] GO:0007155 : cell adhesion (2134)
 - [+] GO:0007267 : cell-cell signaling (1996)
 - [+] GO:0030260 : entry into host cell (98)
 - [+] GO:0009875 : pollen-pistil interaction (7)
 - [+] GO:0009991 : response to extracellular stimulus (273)
 - [-] GO:0007165 : signal transduction (14009)

Candidate ontologies

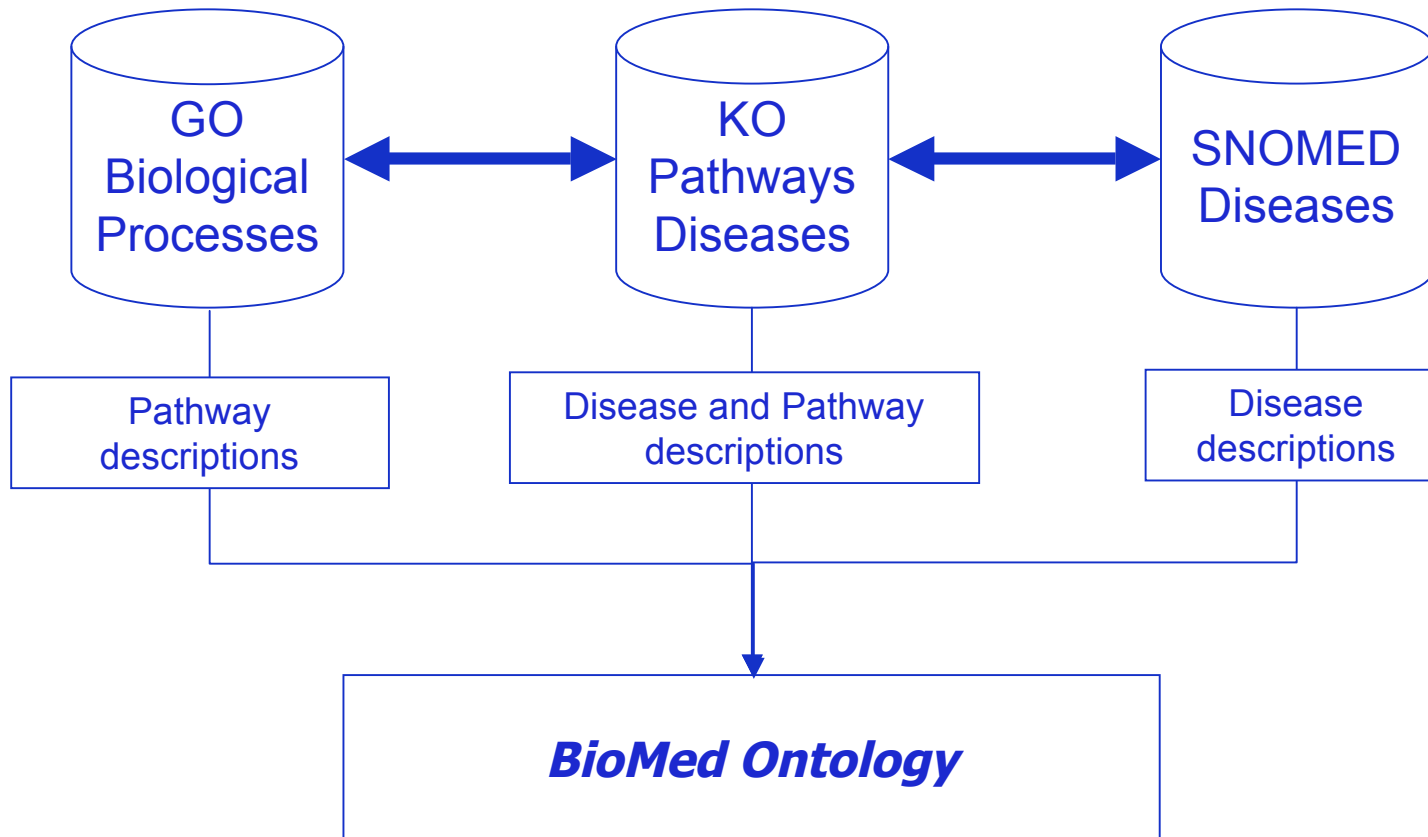


SNOMED-CT

Location
Morphology
Clinical manifestations
Etiology (e.g., virus)



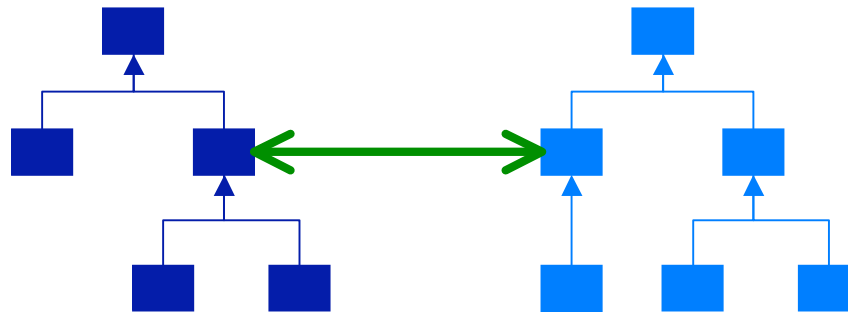
Overview



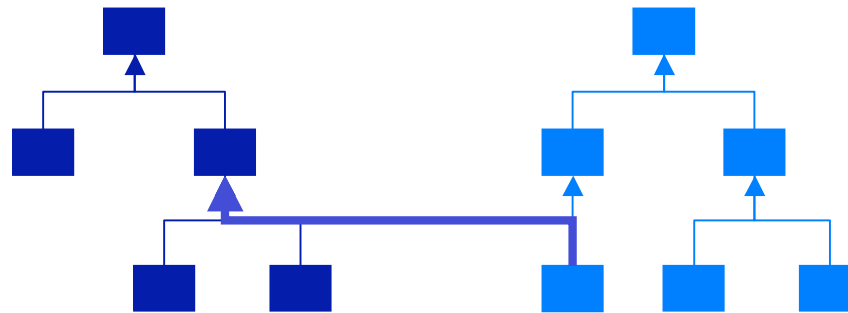
Chabalier J, Mosser J, Burgun A. Integrating biological pathways into disease ontologies, Medinfo 2007, 791-5

Integration

- **Mapping:** equivalence relationships between terms



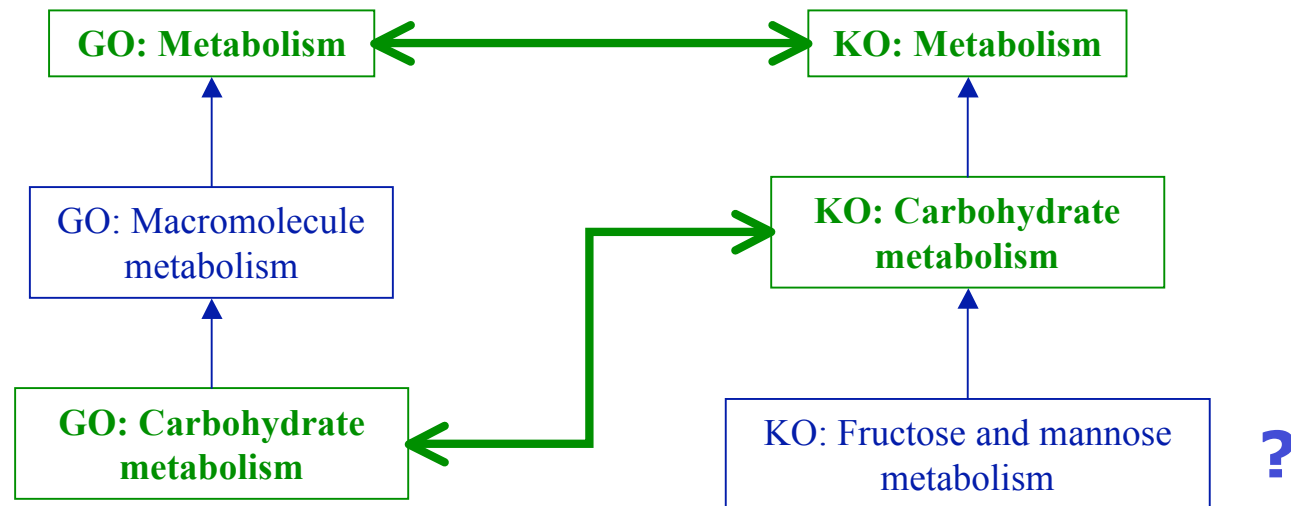
- **Aligning:** define relationships between terms (is-a, part-of, etc.)



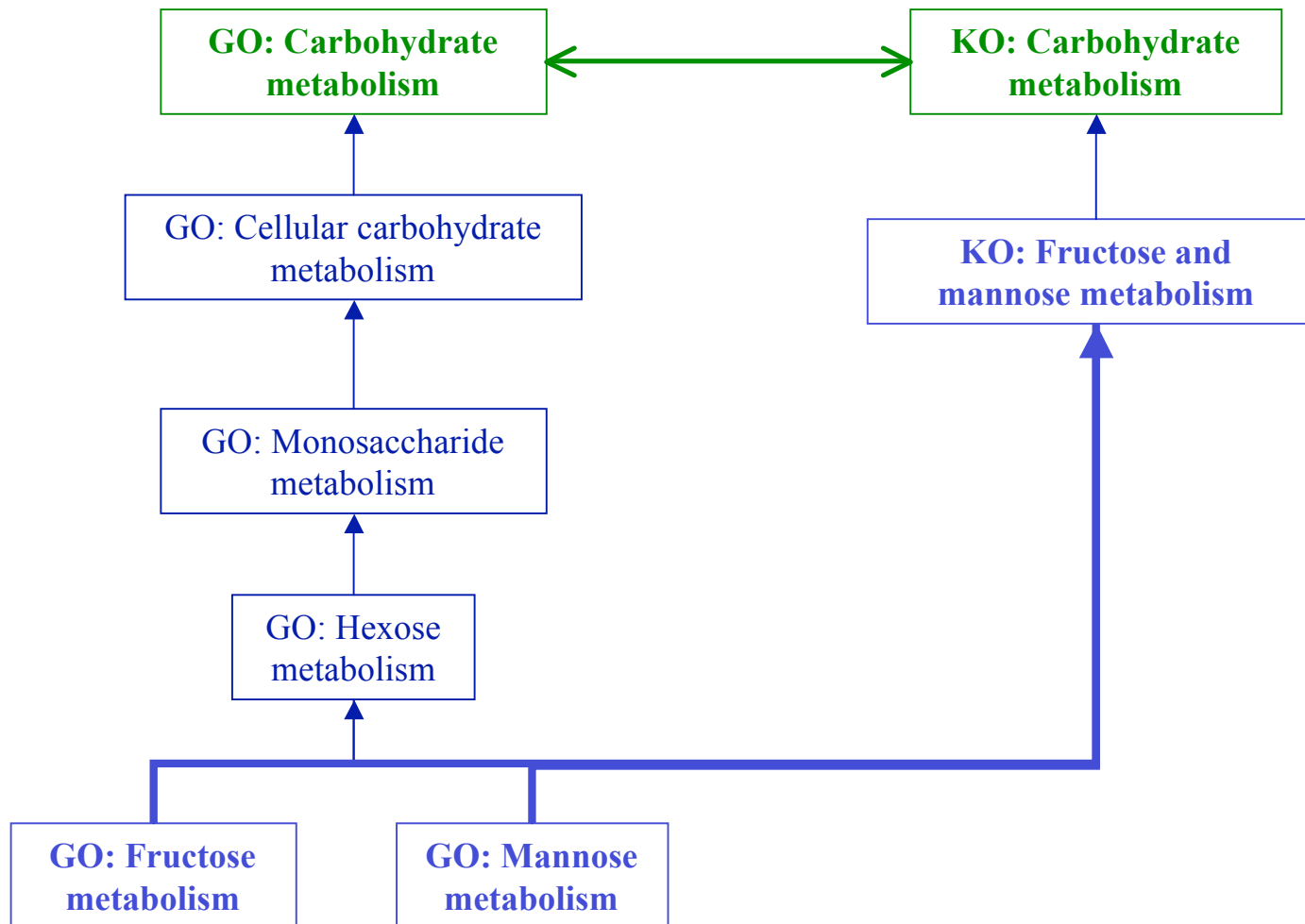
Mapping GO processes – KO pathways



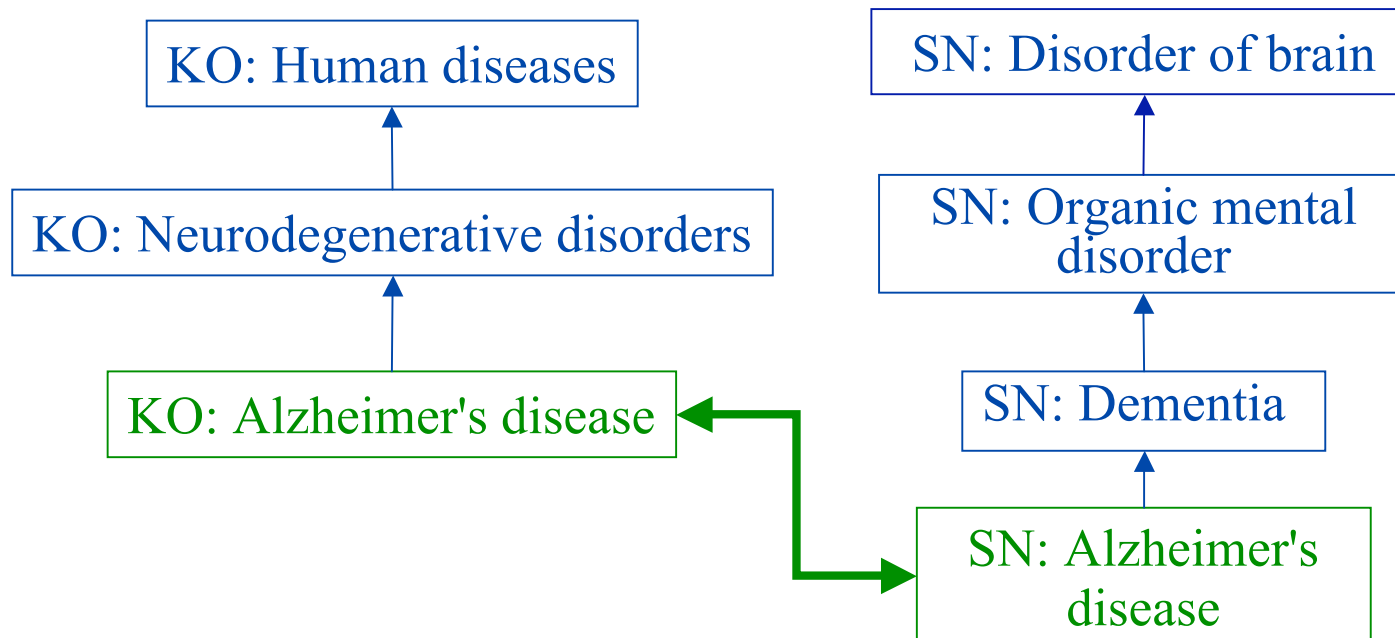
MetaMap Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc. 2001 AMIA Symp., 17-21



Mapping & aligning GO processes – KO pathways



Mapping KO diseases to SNOMED diseases



Integration

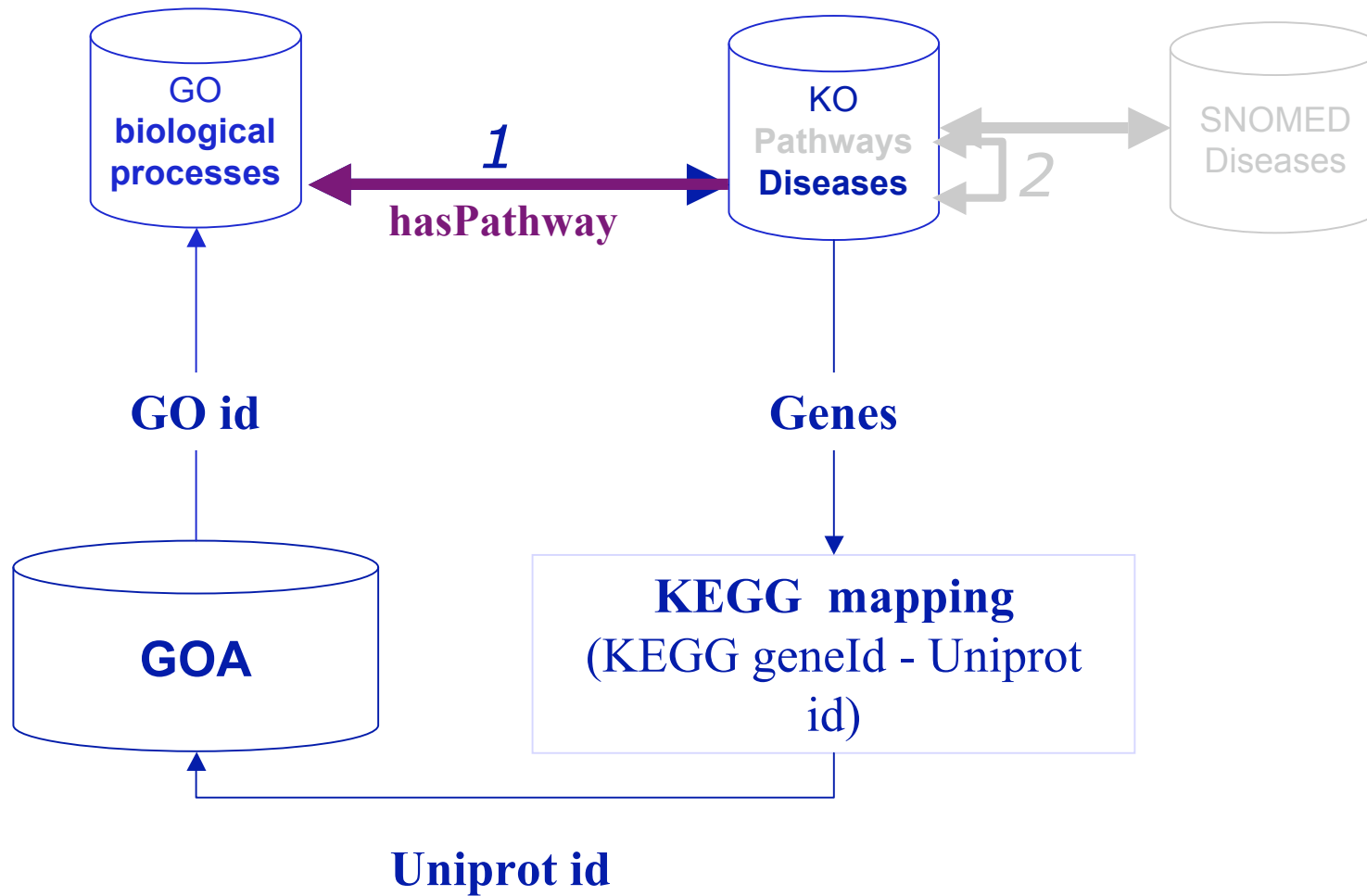


- **Add relations between diseases and processes**

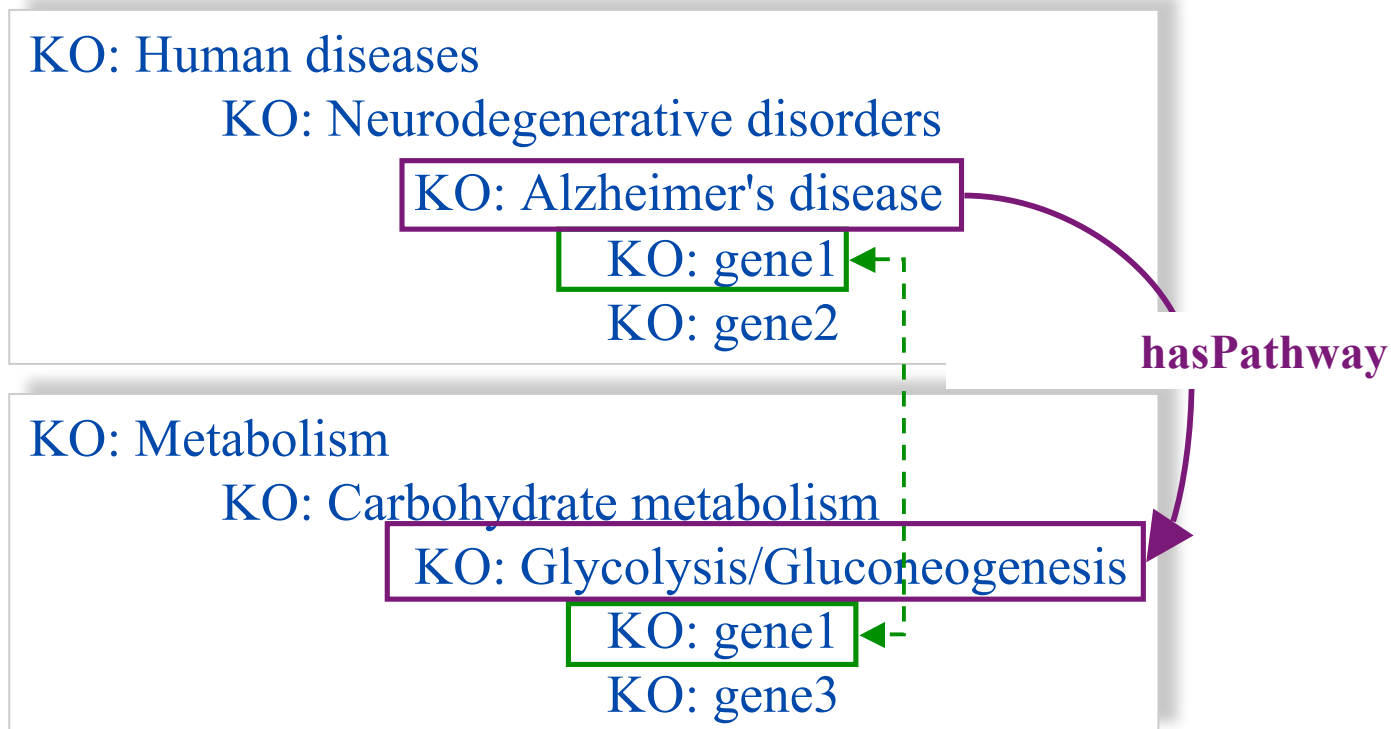
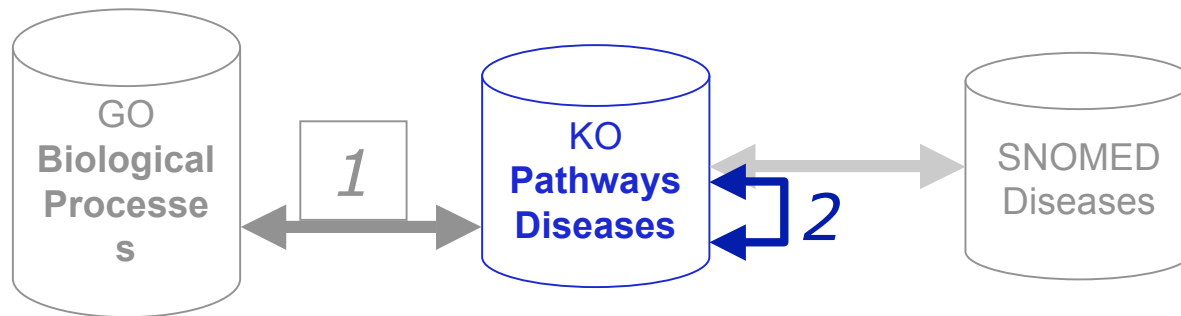
- **Condition** : at least one gene is involved in both a disease D and a pathway P :



Integration



Integration



Integration : results

BioMed Ontology

13,982 classes:

- 13,555 classes from GO
- 281 classes from KO
 - 252 pathways classes
 - 19 disease classes
- 146 classes from SNOMED

http://www.ea3888.univ-rennes1.fr/biomed_ontology/



Limitations

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automate acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

Query results

- Which pathways are shared by 2 neurological disorders : glioma & Alzheimer's disease?

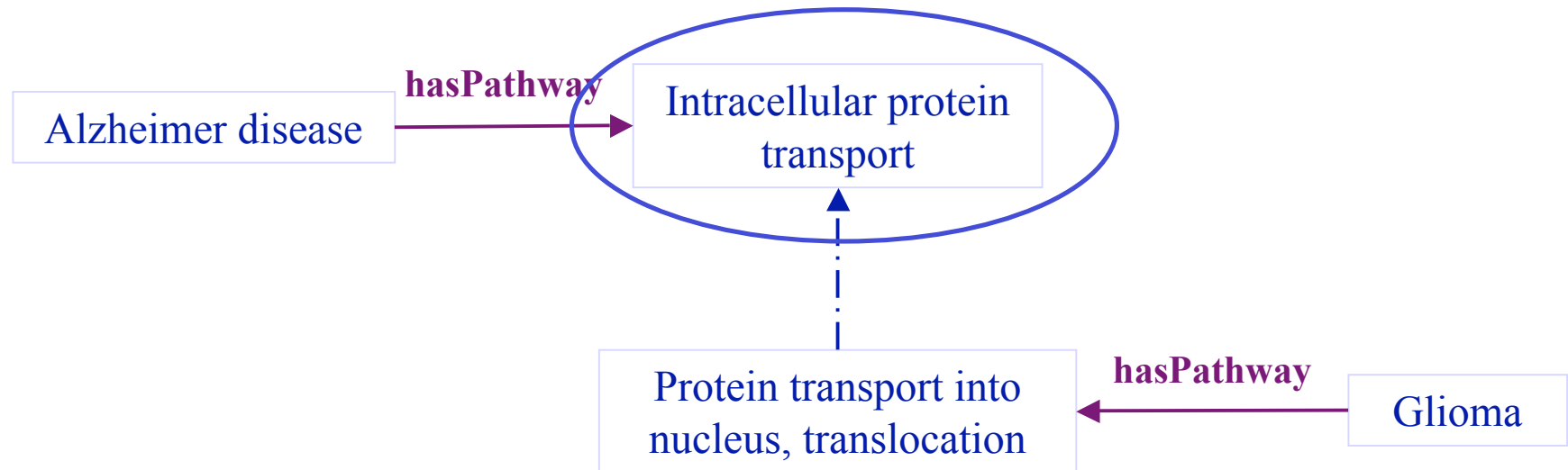
- → 37 pathways:

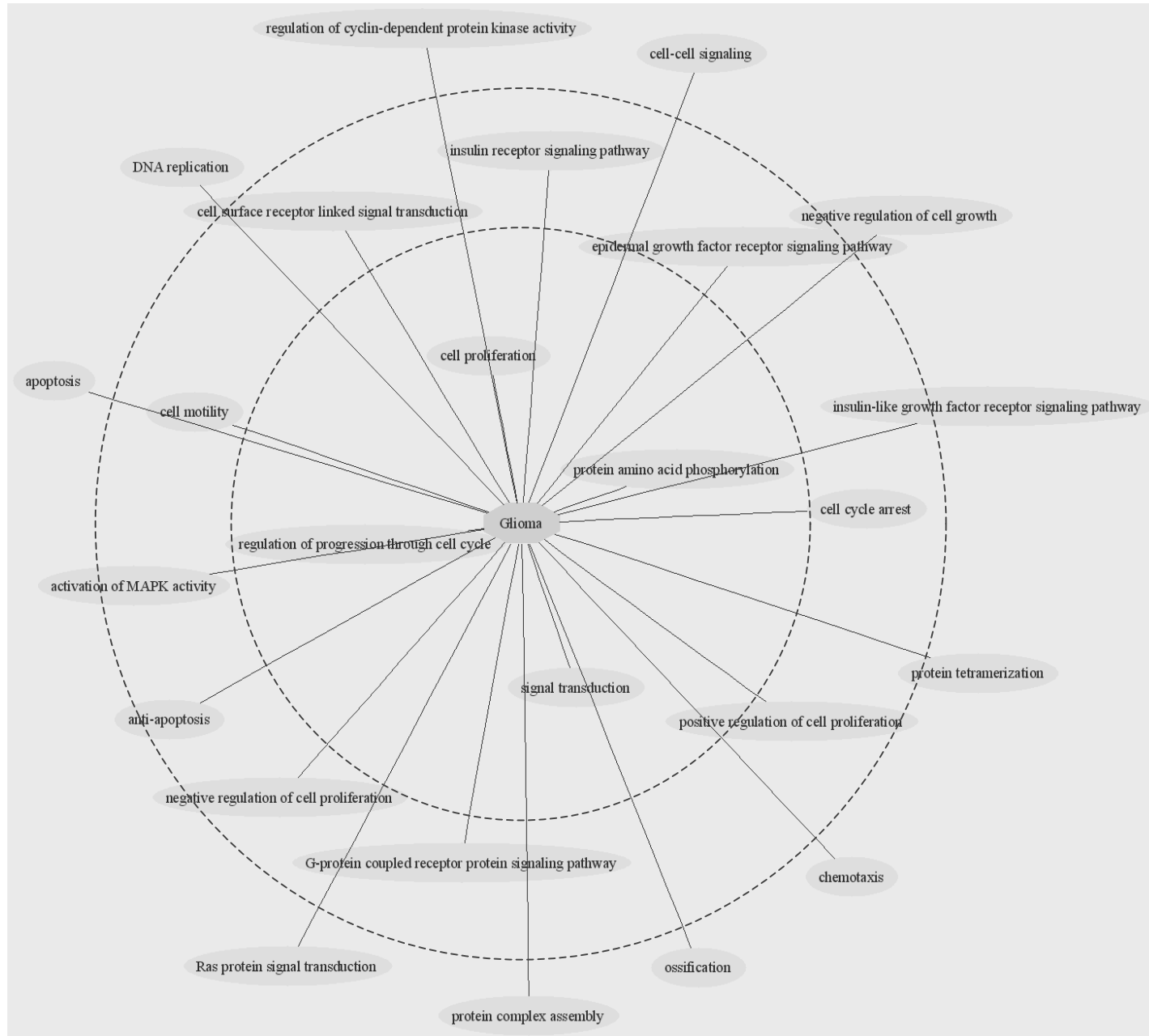
- MAPK signaling pathway
- Focal adhesion
- Insulin signaling pathway
- Melanogenesis
- B cell receptor signaling pathway
- heart development
- central nervous system development
- axon guidance
- peptidyl-serine phosphorylation
- protein amino acid phosphorylation
- cell cycle
- cell-cell signaling
- cell cycle arrest
- lipid catabolic process
- lipid metabolic process
- ubiquitin cycle
- transport

- ErbB signaling pathway
- Wnt signaling pathway
- protein tetramerization
- intracellular signaling cascade
- protein modification process
- glycogen metabolic process
- anagen
- induction of apoptosis
- negative regulation of apoptosis
- apoptosis
- anti-apoptosis
- Natural killer cell mediated cytotoxicity
- cell proliferation
- DNA replication
- chromosome organization and biogenesis
- calcium ion homeostasis
- signal transduction
- response to UV
- negative regulation of cell growth
- cytoskeleton organization and biogenesis

Benefits of ontologies

- pathway hierarchy





Pathways related to glioma

Limitations

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automatize acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [A. Butte]

- Objective: finding trends between genes and clinical measurements
- Subset of GEO datasets related to a given disease

Gene Expression Data

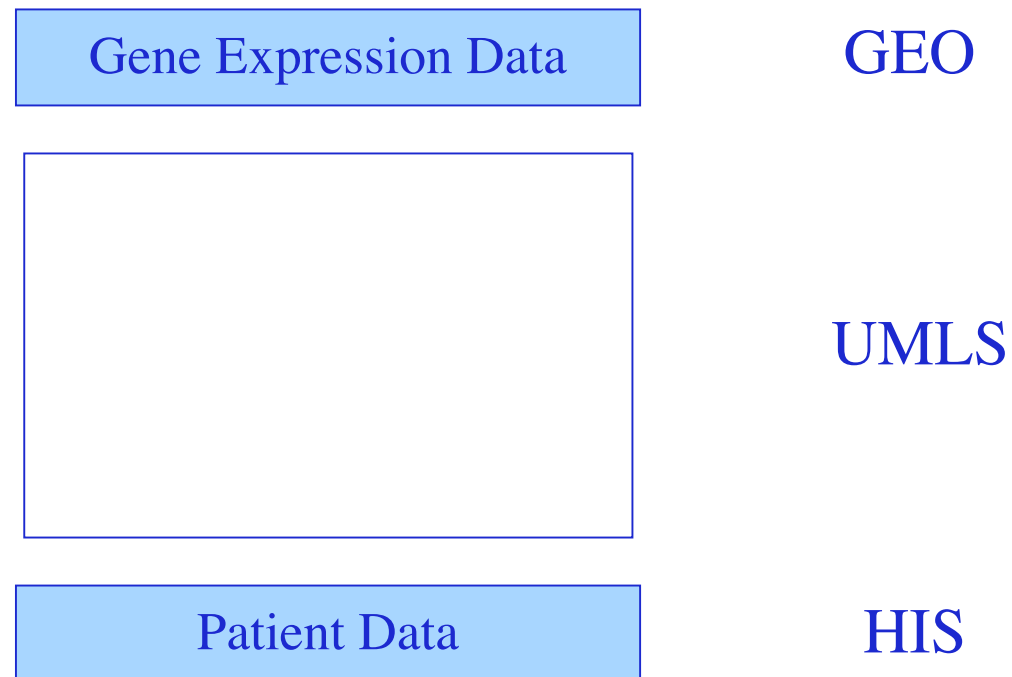
GEO

Patient Data

HIS

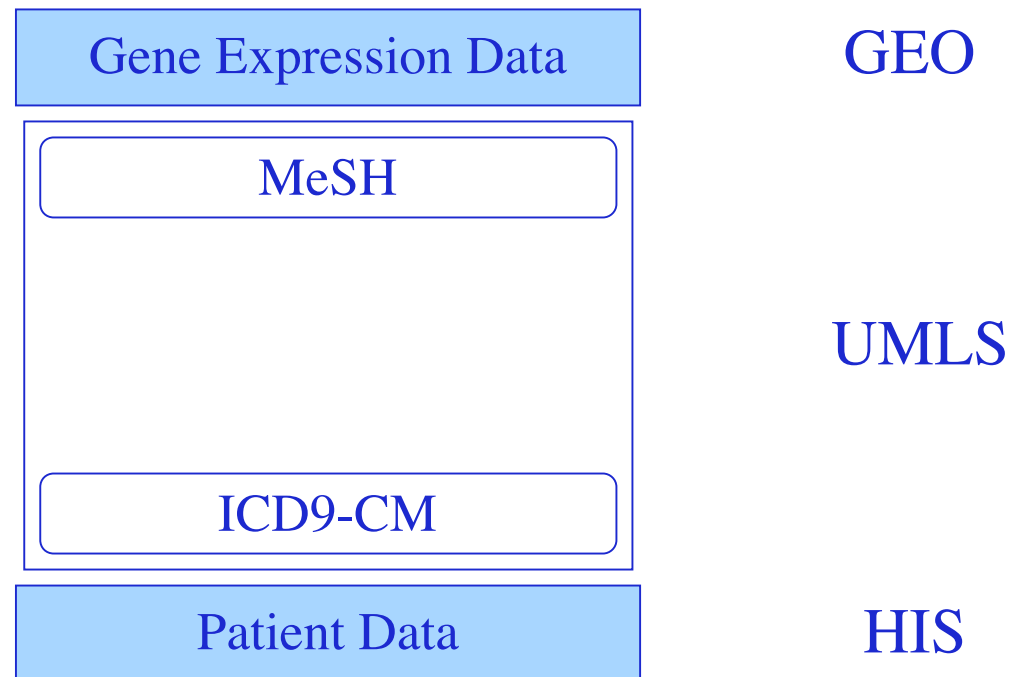
UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [A. Butte]

- Objective: finding trends between genes and clinical measurements
- Subset of GEO datasets related to a given disease



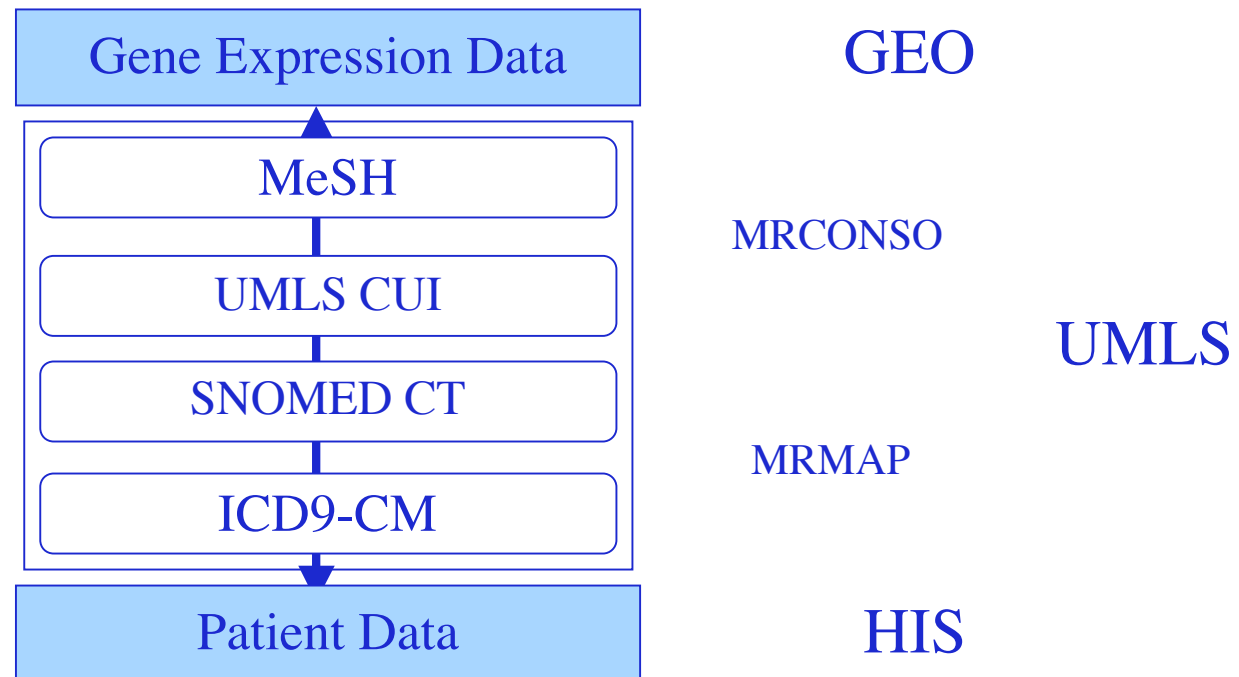
UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [A. Butte]

- Objective: finding trends between genes and clinical measurements
- Subset of GEO datasets related to a given disease



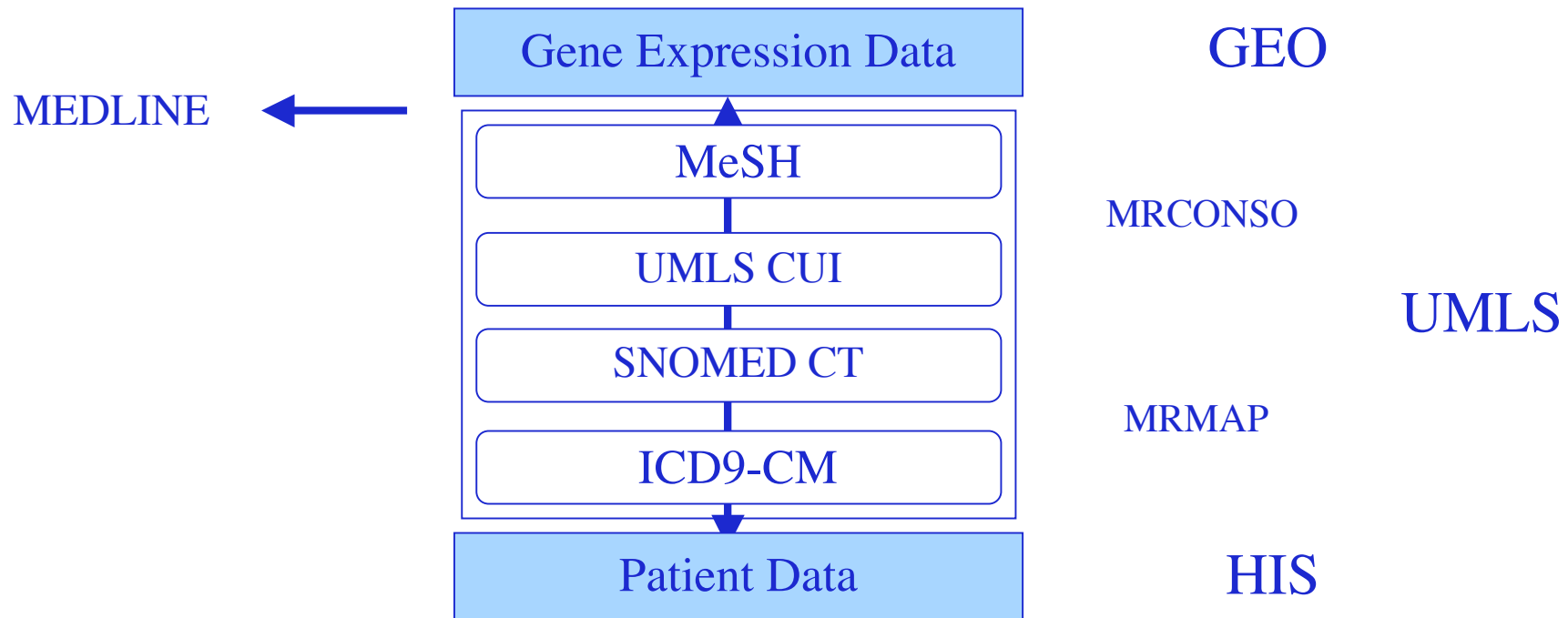
UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [A. Butte]

- Objective: finding trends between genes and clinical measurements
- Subset of GEO datasets related to a given disease



UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [A. Butte]

- Objective: finding trends between genes and clinical measurements
- Subset of GEO datasets related to a given disease



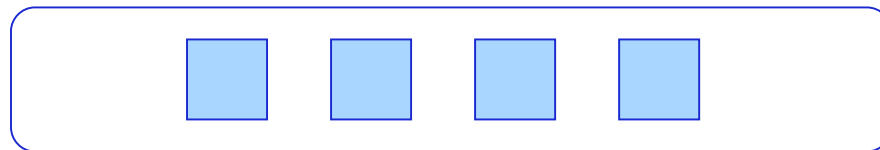
UMLS-based integration of patient data and GEO information : Stanford Translational Research Integrated Database Environment [J. Dudley, DP Chen, A. Butte]

- Results
- 737 GEO datasets (out of 7,264 experiments) related to human diseases
- 238 disease concepts
- 13,452 patients (out of 49,000) mapped to 211 disease concepts
- Integrate patient clinical data with microarray data at a population level

Dudley J., Chen D.P., Butte A. Using SNOMED CT for translational genomics data integration, KR-MED 2008, Phoenix, AZ, 31 May-2 June 2008

UMLS-based medical annotation of genes : BioMeKE [G. Marquet, A. Burgun]

- Objective : Adding medical (disease) annotation to GO annotation of genes
- UMLS Metathesaurus relations



Microarray experiments

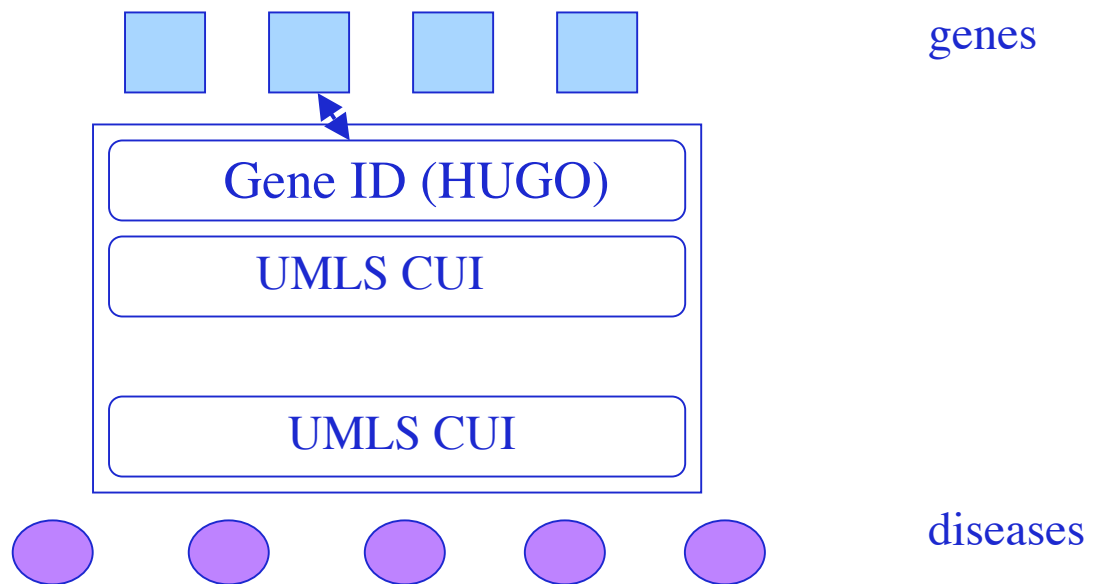
genes



diseases

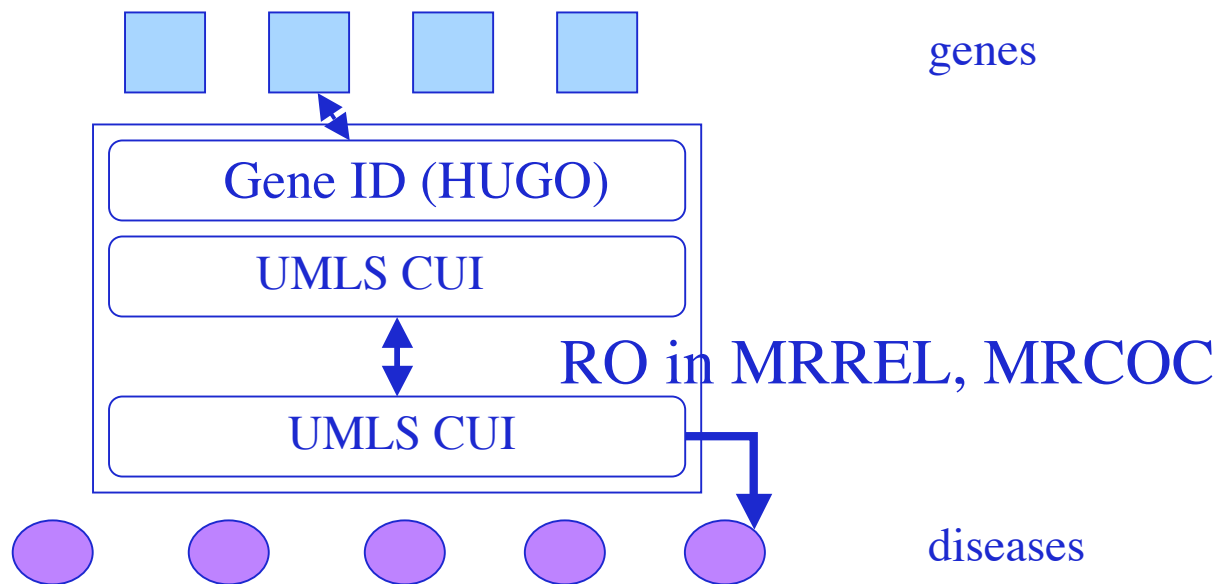
UMLS-based medical annotation of genes : BioMeKE [G. Marquet, A. Burgun]

- Objective : Adding medical (disease) annotation to GO annotation of genes
- UMLS Metathesaurus relations



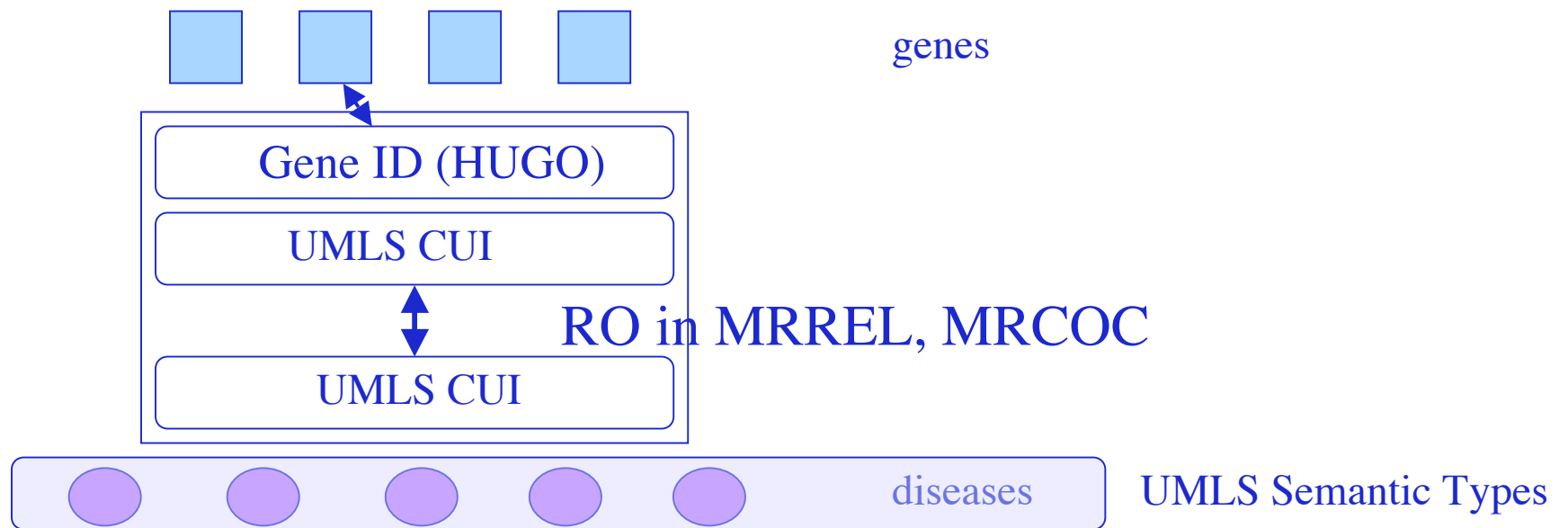
UMLS-based medical annotation of genes : BioMeKE [G. Marquet, A. Burgun]

- Objective : Adding medical (disease) annotation to GO annotation of genes
- UMLS Metathesaurus relations



UMLS-based medical annotation of genes : BioMeKE [G. Marquet, A. Burgun]

- Objective : Adding medical (disease) annotation to GO annotation of genes
- UMLS Metathesaurus relations



UMLS-based medical annotation of genes : BioMeKE [G. Marquet, A. Burgun]

- Results
 - Validation on 43 genes involved in iron metabolism
 - 17% HGNC identifiers have « BioMeKE annotation »
 - Average 5.7 MTH concepts (1-336)
- MRCOC : terms that index the same articles
- Text mining

Guérin E, Marquet G, Burgun A et al. Integrating and warehousing liver gene expression data, LNCS 2005, 3615, 158-174

Limitations

- Knowledge representation
 - Description Logics : grades, relation D-G
 - Genes + physio-pathological mechanisms
- Limited usage of NCIT
 - NCI Thesaurus : cancer, clinical trials, DL
 - SNOMED CT : medicine, patient records, DL
 - ICD-O : cancer, registries, terminology
 - Mapping between NCIT and other DO
- Coverage
 - Automate acquisition (link between disease and gene)
 - Integrate DO and Pathway Ontologies through KB
 - Integrate Disease information and Gene information through ontologies

Perspectives

- Address Knowledge Representation issues
 - Relations between diseases and genes
 - Associate a degree of confidence to the Disease/Pathway relationships
- Mapping, aligning terminologies
 - UMLS
 - Ontology alignment
- Cross validation
 - Combination of resources (e.g. OMIM, BioPax, Medline)
 - Combination of techniques (alignment, mediation, text mining)

