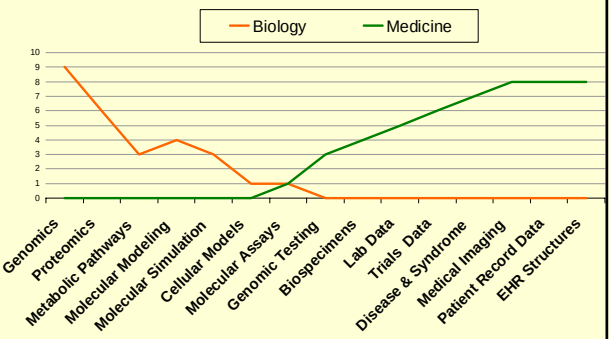
Reference
Ontologies in Biomedicine

Christopher G. Chute, MD DrPH
Professor and Chair, Biomedical Informatics
Mayo Clinic College of Medicine
Rochester, Minnesota, USA

AMIA, Washington DC, 2005

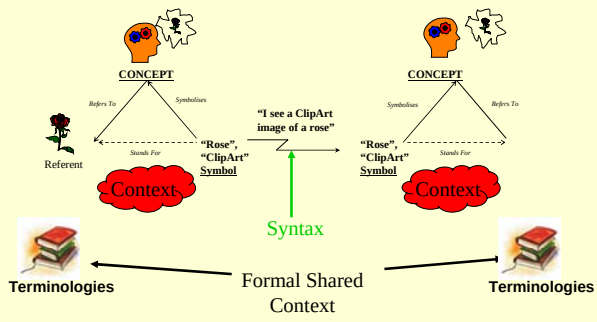
The Continuum Of Biomedical Informatics
Bioinformatics meets Medical Informatics



Big Science Collaboration and Semantics

- Communication of findings and results
  - Human publication
  - Sharing of data resources as building blocks
  - Foundation for incremental, big-science
- Harnessing computing requires formalization
  - Data format and structures
  - Discrete terms, vocabulary, ontology
  - Non-ambiguous concepts, non-overlapping terms
  - Information models and problem architectures
  - Standards, conventions, and shared context

Making Shared Context Explicit



[From Solbrig]

If science is communication what is its language?

- Most ontologies and vocabularies are created to meet a specific application or use-case
- Despair their re-use in alternative contexts
- Virtually all terminologies invoke concepts "out of domain"
  - LOINC – drugs
  - SNOMED, MeSH – anatomy, drugs
- Identifying common "atoms" an elusive goal
- Fraught with composition
  - Micro-information models (sentences and ¶'s)

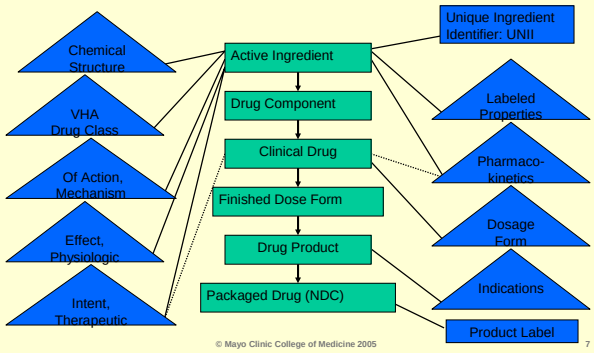
Science Interlingua
Interlocking Reference Terminologies

Reference Terminology

- A coherent organization of concepts about a well-characterized domain, e.g.
  - Units of Measure
  - Pharmaceuticals
  - Anatomy
  - Cellular processes
- Reference concepts that underpin scientific expression
- Abstract concept space that serves no application need (with or without the real world)

The VA, FDA, NLM Drug Ontology Projects

[modified from Steve Brown]



Science Interlingua
Babble vs Babel (and other dyslexiae)

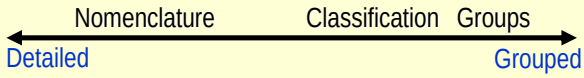
- Will an Esperanto of Science fail out of the box?
  - Is it language for humans?
  - It is language for computers?
- Is "the" science ontology [or interlocking plurality]
  - a period table?
  - a body of weights and measures?
  - a dictionary?
  - a thesaurus?
  - a language?
- Whither cognition?
  - Inference is not understanding

What are factors that erode shared context in biomedicine

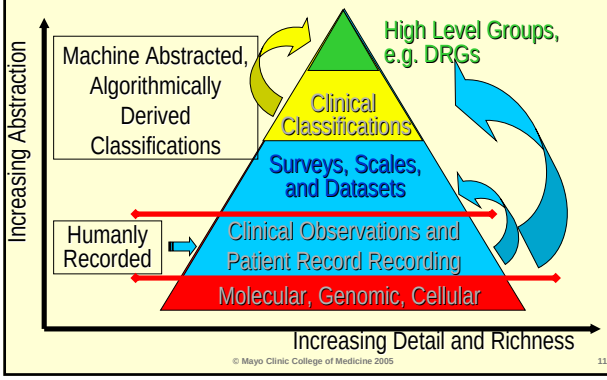
- Concept granularity (specificity)
- Vertical scope (molecules to society)
- Divergent concept content (codes)
- Divergent information models
- Terminology – Information model boundary
- Human use of "machine" concepts
- Human orneriness

Familiar Points Along Continuum
Modern Health Vocabularies

- Nomenclature – Highly Detailed Descriptions (SNOMED)
- Classification – Organized Aggregation of Descriptions into a Rubric (ICDs)
- Groupings – High Level Categories of Rubrics (DRGs)



Abstraction Layers
Narrow Boundary of Ideal, Human Recording



Blois, 1988
Medicine and the nature of vertical reasoning

- Molecular: receptors, enzymes, vitamins, drugs
- Genes, SNPs, gene regulation
- Physiologic pathways, regulatory changes
- Cellular metabolism, interaction, meiosis,…
- Tissue function, integrity
- Organ function, pathology
- Organism (Human), disease
- Sociology, environment, nutrition, mental health…

Domain-specific expansion of "MS"
Semantics by domain context

Cardiology	mitral stenosis
Neurology	multiple sclerosis
Anesthesia	morphine sulfate
Obstetrics	magnesium sulfate
Research science	manuscript
Physics	millisecond
Education	Master of Science
U.S. Postal Service	Mississippi
Computer science	Microsoft
Correspondence	female name prefix

Content vs. Structure
Contest or Synergy? Computer Equivalent?

Family History of Breast Cancer
Family History of Heart Disease
Family History of Stroke

Terminologic Model

Family History

Breast Cancer
Heart Disease
Stroke

Information Model

Equivalent Content

[adapted from Rossi-Mori]

Information Model (HL7)
Terminology Model (SNOMED)

HL7 RIM	SNOMED CT Attribute
targetSiteCode(Observation)	"finding site"
targetSiteCode(Procedure)	"procedure site"
methodCode (Observation & Procedure)	"method"
approachSiteCode(Procedure)	"approach," "access"
priorityCode(Act)	"priority"

[adapted from Markwell]

Context mis-match
Challenges for Reference concepts

- Cultural aggregation and splitting
  - 7 words for "rice" in Thai
  - 17 works for "snow" in Inuktitut
  - ☞ Must the reference enumerate the superset?
- Composition vs Terms
  - Pre-coordinated terms – "Colon Cancer"
  - Neologisms for "sentence concepts"
  - ☞ Should the reference include conceptual atoms and molecules?
  - Composition is in the eye of the beholder

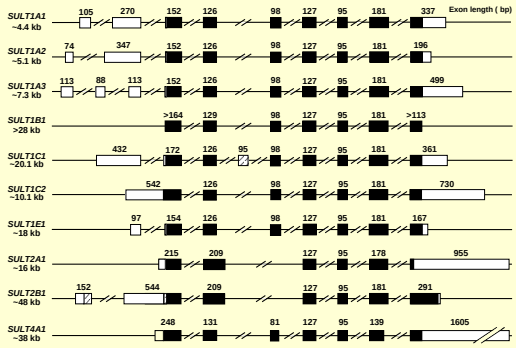
Reference Truth: Variations in Identity
On orthologs, paralogs, and SNPs

- Identity in context of resolving to same concept in a reference terminology
- Enzymes that share function: Sulfotransferase
- Orthologs across primates
- Paralogs (including pseudogenes) in humans
- Polymorphisms between individuals

SULT6B1 Sequence
Differences Between Primates

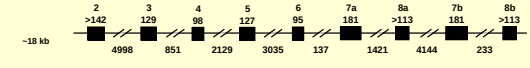
Location in Human SULT6B1		Nucleotide Sequence		
Nucleotide	Amino Acid	Human	Gorilla	Chimpanzee
A148G	LYS50GLU	A	A	G
G274A	GLU92LYS	G	A	A
A314C	LYS105THR	A	C	A
A321G	THR107THR	A	G	G
G336C	LEU112PHE	G	C	C
C390G	PHE130LEU	C	G	C
T400C	PHE134LEU	T	T	C
G429C	ARG143SER	G	C	C
GCT(538-540)CCC	ALA180PRO	GCT	CCC	GCT
C609A	ALA203ALA	C	A	A
C636T	HIS212HIS	C	C	T
A651C	PRO217PRO	A	C	A
A697G	SER233GLY	A	A	G

Human SULT Genes

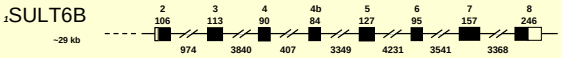


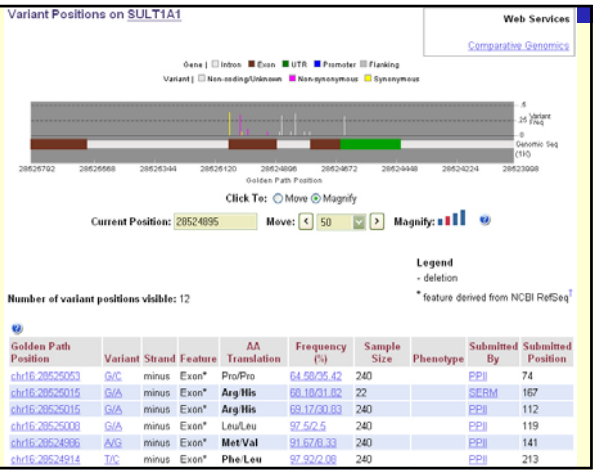
SULT Data Mining Pseudogenes

SULT1C3



SULT6B





Golden Path Position	Variant	Strand	Feature	AA Translation	Frequency (%)	Sample Size	Phenotype	Submitted By	Submitted Position
chr16:28525053	G/C	minus	Exon*	Pro/Pro	64.58/35.42	240		PPII	74
chr16:28525015	G/A	minus	Exon*	**Arg/His**	68.18/31.82	22		SERM	167
chr16:28525015	G/A	minus	Exon*	**Arg/His**	69.17/30.83	240		PPII	112
chr16:28525008	G/A	minus	Exon*	Leu/Leu	97.5/2.5	240		PPII	119
chr16:28524966	A/G	minus	Exon*	**Met/Val**	91.67/8.33	240		PPII	141
chr16:28524914	T/C	minus	Exon*	**Phe/Leu**	97.92/2.08	240		PPII	213

Discriminate Differences of Kind Reference Terminology

- At what level of granularity?
- What is a Sulfotransferase?
- What level of detail should a reference terminology of enzymes convey?
- Is the logical limit of all reference terminologies a basis on types of quark?

LexGrid as a *Terminology* Interchange

- Proliferation of ontologies and vocabularies
- Varieties of formats and terminology models
- Various versions over time
- Hard to find appropriate resource
- Establish “web of terminology” to link content
- Extension of Semantic Web concept
- Common tools, formats, and interfaces

LexGrig.org & HL7 CTS integrated into cBIO