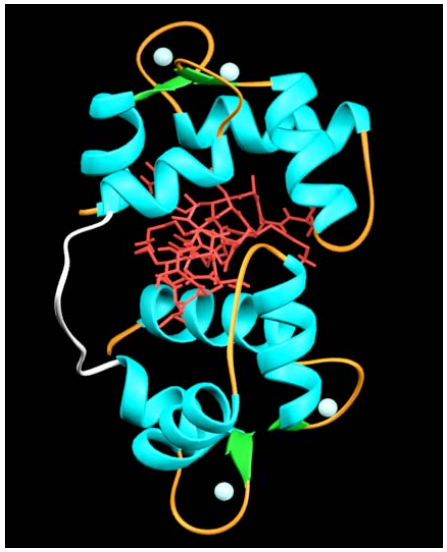
Exploiting contextual similarity for semantic prediction in the Biology domain

Karin Verspoor

Los Alamos National Laboratory

verspoor@lanl.gov

October 02, 2007



LAURs 05-4778, 07-1170, 07-4571, 07-2538, 07-6572

let's talk biology

- Biologists are working to understand the functions and roles of genes and proteins in biological processes
- Proteins perform most life functions and make up the majority of cellular structures
- The proteome is extremely dynamic, depending on associations between proteins and reactions among them

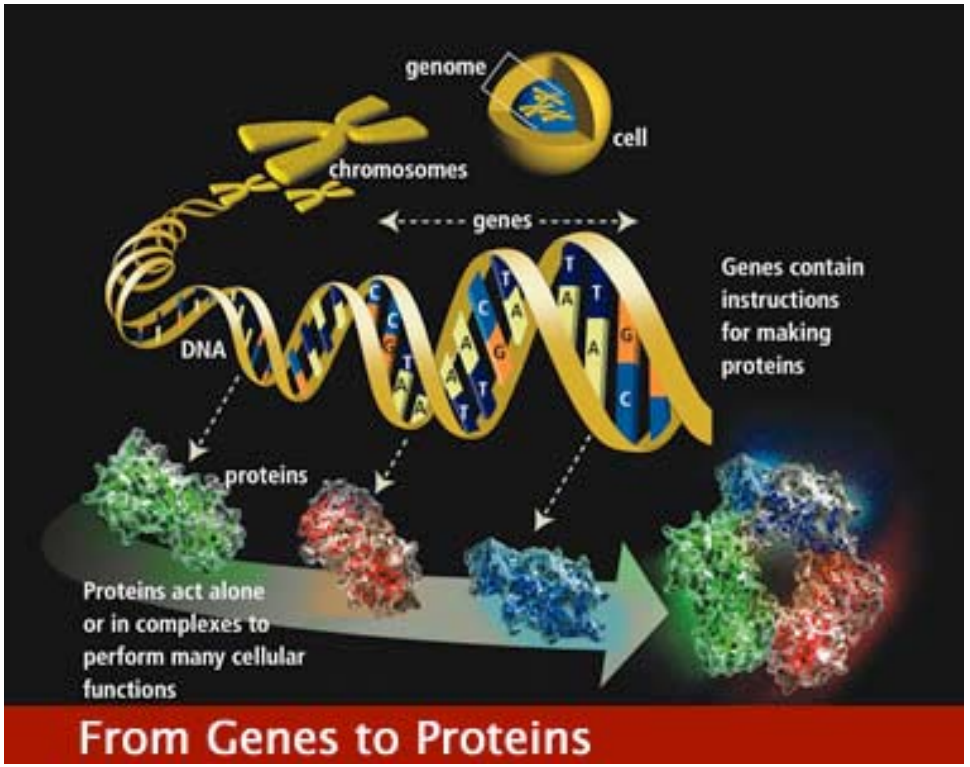
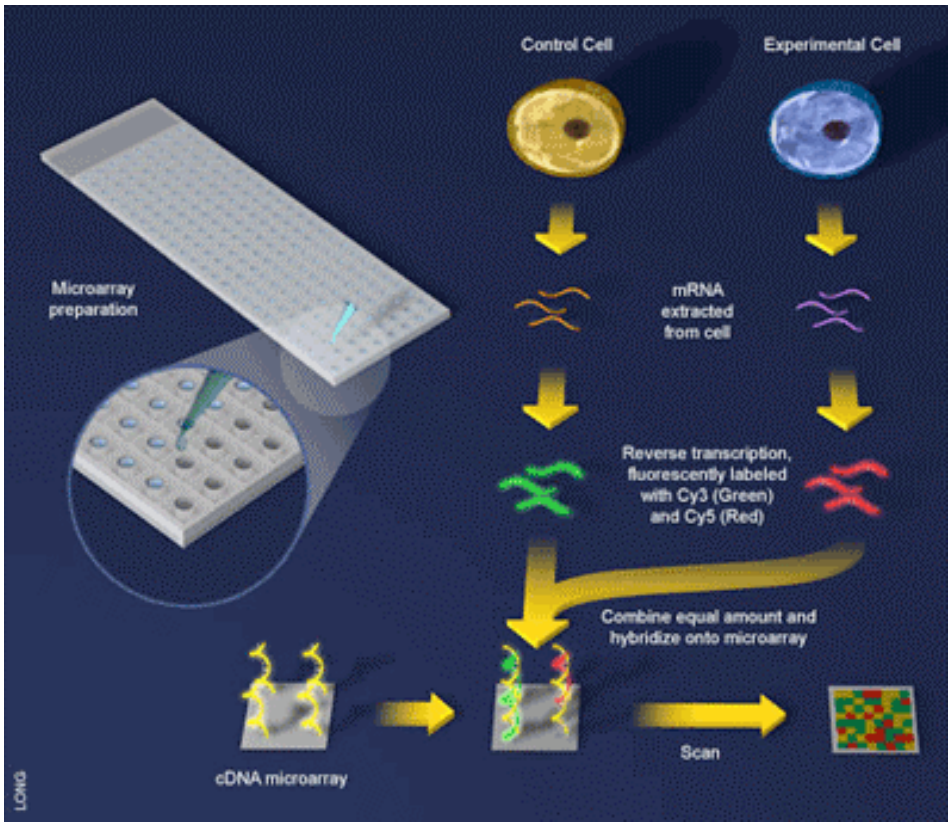


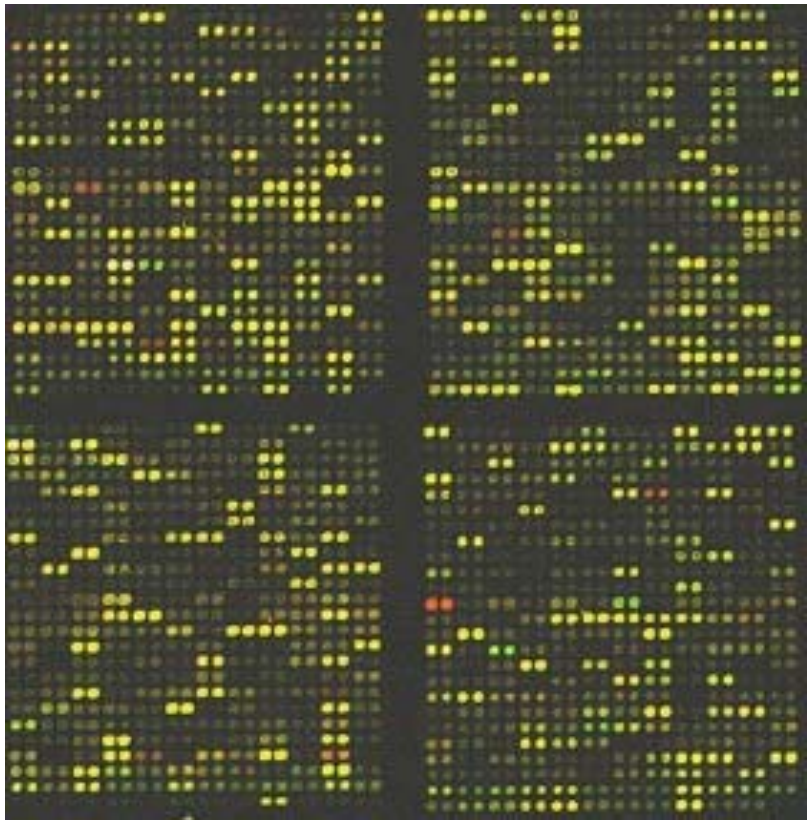
Image from http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

SNPs and chips (-or- what data biologists are gathering)

- Genome sequences
  - the entire DNA sequence of an organism
  - The Human Genome project characterized approximately 30,000 genes
- Protein structures
  - 3D models, protein folding
  - a protein's shape is key to understanding its function and roles
- Single Nucleotide Polymorphism (SNP)
  - a small genetic variation within a DNA sequence
  - a single nucleotide replaces another nucleotide
- Genetic expression profiles (microarray chips)
  - measure the expression levels of genes in cells through measurement of production of mRNA
  - generally used with control and variant DNA to compare expression

microarrays

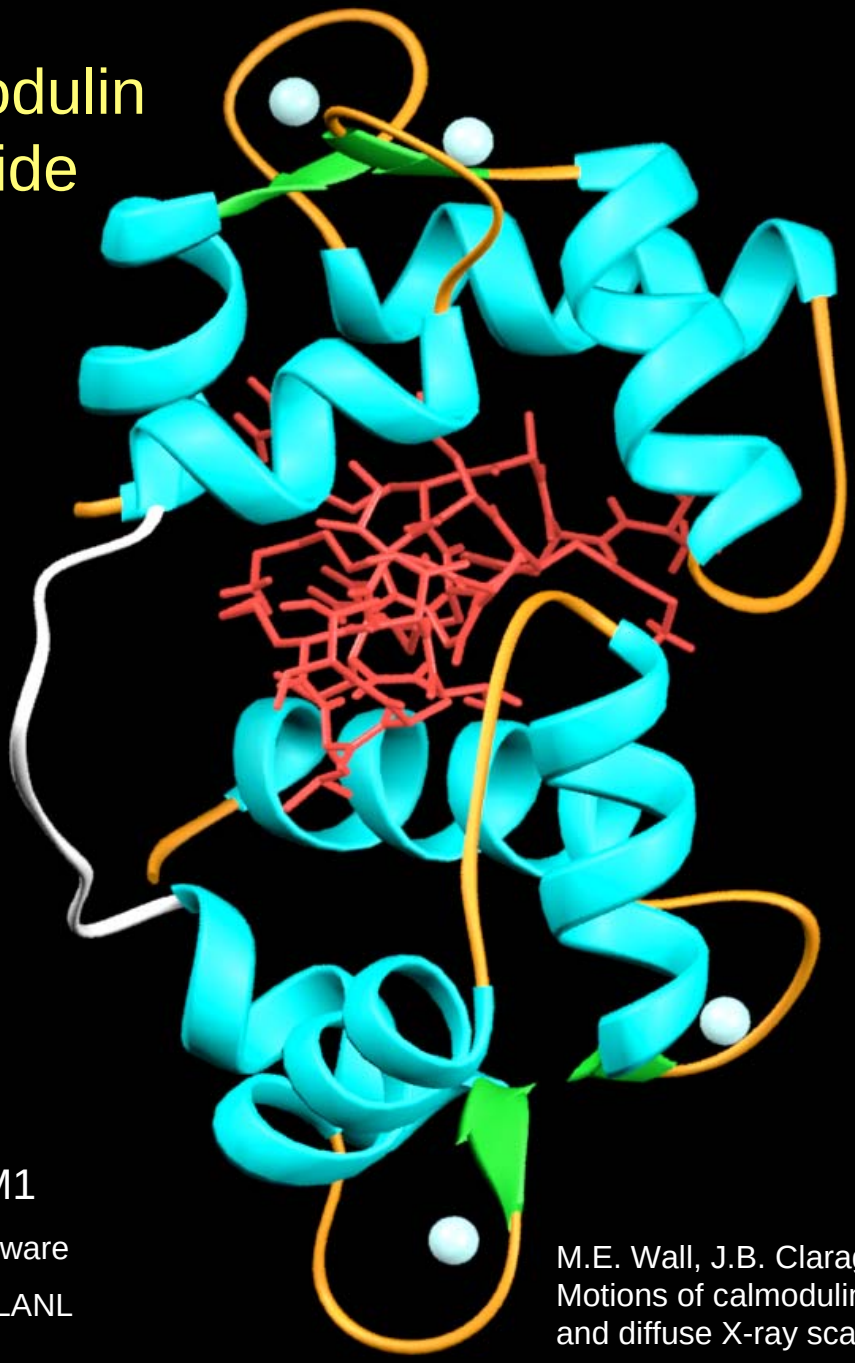




From http://www.scq.ubc.ca/?p=272

From http://science.nasa.gov/headlines/y2004/10sep_radmicrobe.htm

Complex of Calmodulin with CaMK-II peptide



Wall *et al*, PDB entry 1CM1

Image created with RIBBONS software

Slide courtesy of Michael Wall @ LANL

M.E. Wall, J.B. Clarage and G.N. Phillips, Jr. 1997. Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering. Structure 5:1599-1612.

biological data mining: questions biologists ask

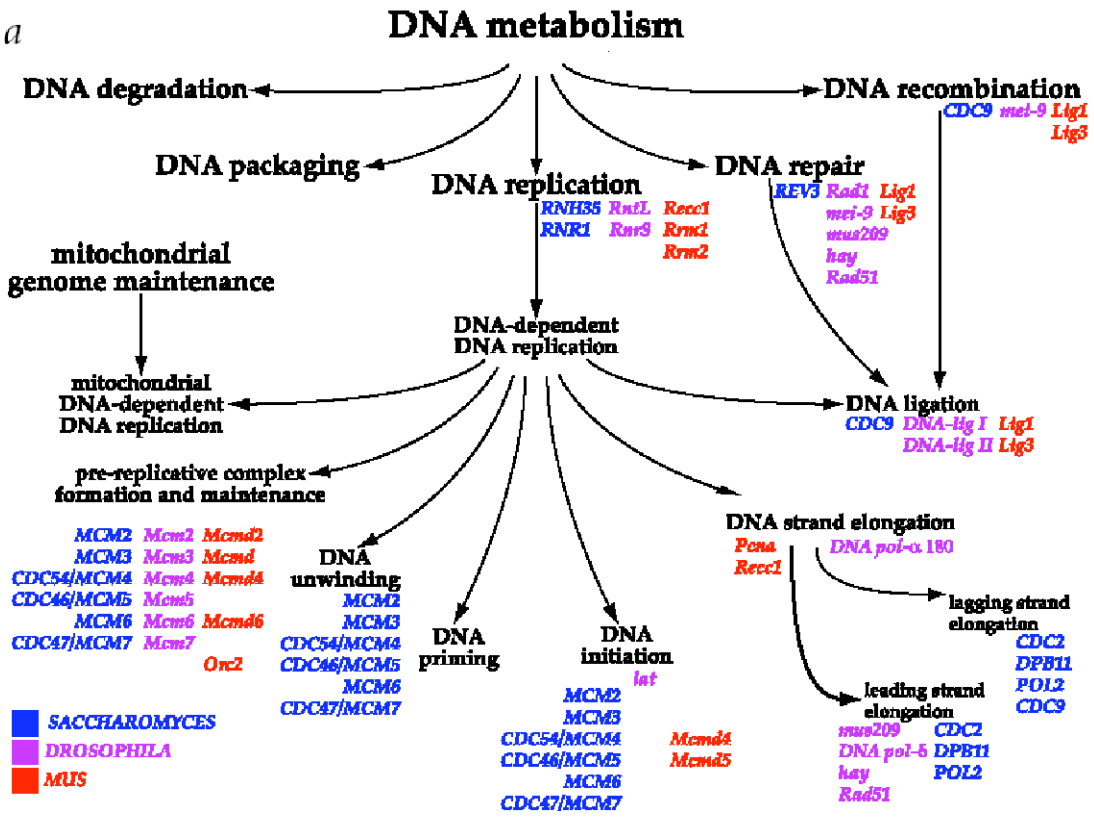
- What is known about this gene/protein/sequence?
  - Literature search
  - Database search
- What does this gene/protein do?
  - What sequences are like the sequence I'm studying? (homology search)
  - Hypothesize gene function (annotation, structure, …) of new gene through similarity to known gene
  - Tools: BLAST, FASTA, PSI-BLAST
    sequence alignment through edit distance
- Which genes are involved in this biological process?
  - Find genes with similar expression profiles (expression activity over time)
  - Tools: Microchips, Clustering analysis

characterization of bioentities

- Sequence data
- Expression experiments
- Protein structures (3-d)
- Formal representation of results of experimentation and analysis:
  - biological processes
  - molecular function
  - disease implication
  - molecular pathways
- Common vocabularies for description facilitate statistical and systems biological analysis
  - Tie sample-specific data to broader protein/gene knowledge
  - Link genomic and therapeutic data
- Ontologies define domain-specific concepts, together with how they are related semantically

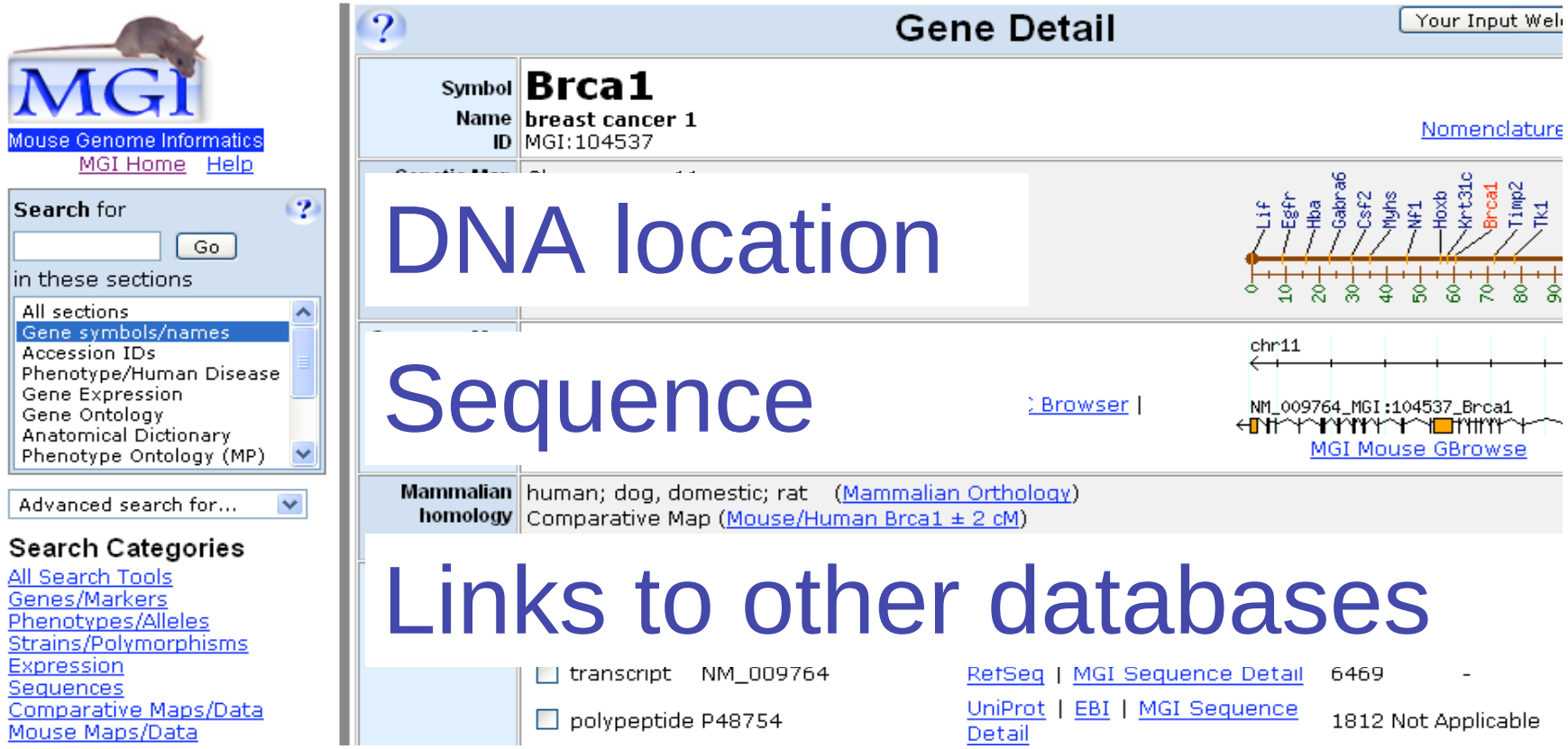
Gene Ontology (GO)

- Taxonomic controlled vocabulary
- ~ 20*K* nodes P_{GO} populated by genes, proteins
- Two orders on P_{GO}: $\leq_{isa}, \leq_{has}$



Gene Ontology Consortium (2000): "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, 25:25-29

Mouse Genome database

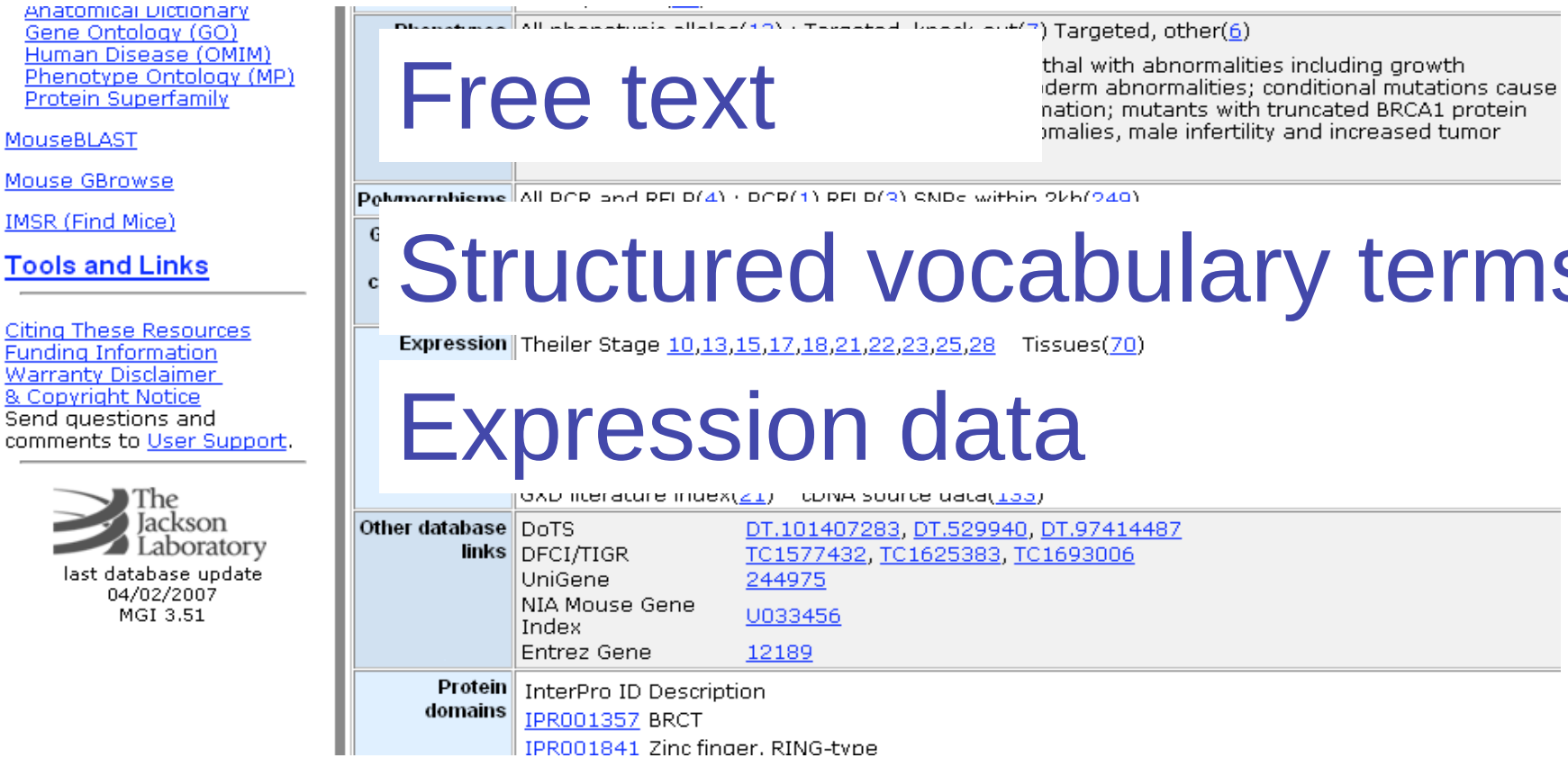


Los Alamos NATIONAL LABORATORY EST.1943

The World's Greatest Science Protecting America

NNSA

lots of structured biological data



Structured vocabulary

Gene Ontology (GO) classifications	Process	carbohydrate metabolic process, cell cycle...
	Component	centrosome, condensed chromosome...
	Function	damaged DNA binding, DNA binding...
	All GO classifications(36)	

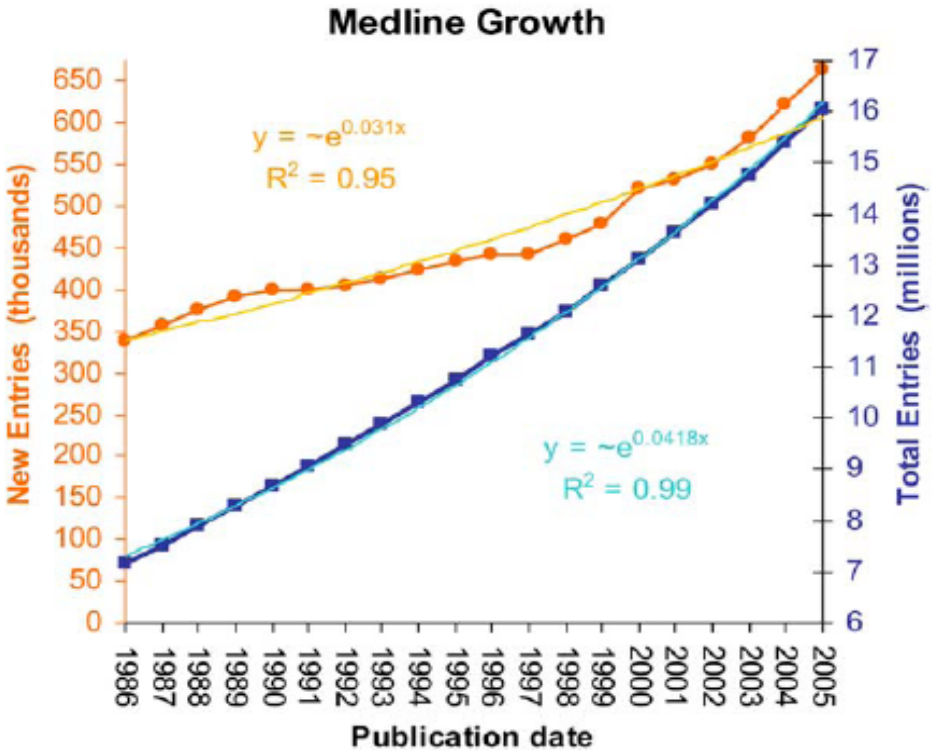
Free text (phenotypes)

Homozygous null mutants are embryonic lethal with abnormalities including growth retardation, neural tube defects, and mesoderm abnormalities; conditional mutations cause genetic instability and enhanced tumor formation; mutants with truncated BRCA1 protein survive, have a kinky tail, pigmentation anomalies, male infertility and increased tumor incidence.

122 references...

References	(Earliest) J:31493 Hall JM *et al.*, "Linkage of early-onset familial breast cancer to chromosome 17q21." Science 1990 Dec 21;250(4988):1684-9 (Latest) J:117113 Clark-Knowles KV *et al.*, "Conditional inactivation of Brca1 in the mouse ovarian surface epithelium results in an increase in preneoplastic changes." Exp Cell Res 2007 Jan 1;313(1):133-45 All references(122)

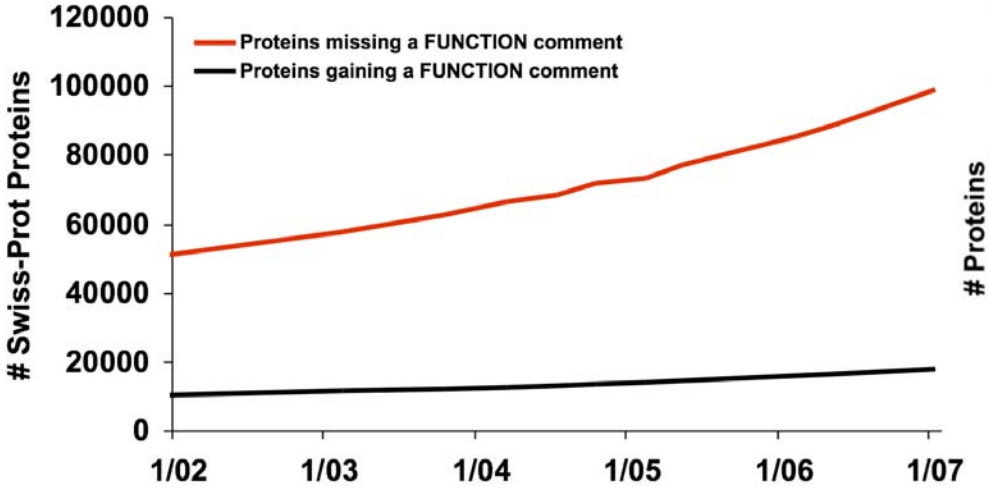
Why we care to produce *automated* methods for semantic characterization of bioentities

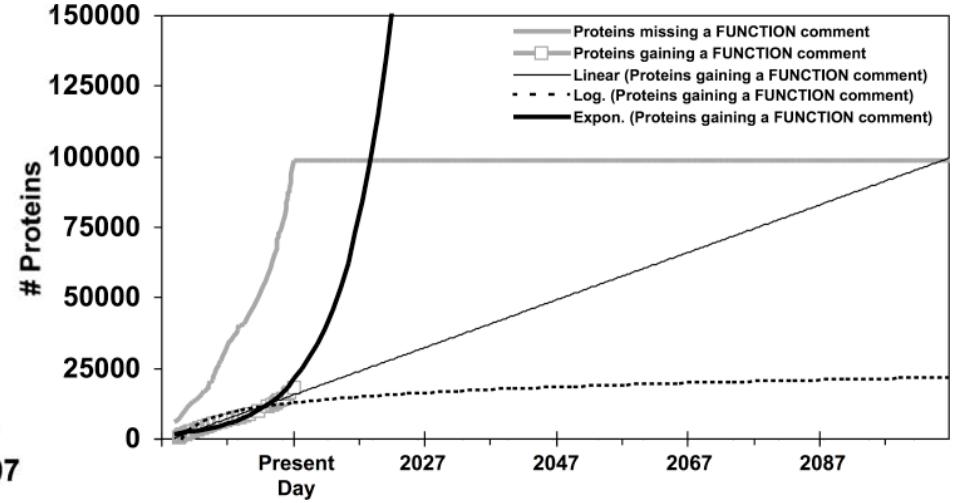


New entries in Medline with publication date in Jan-Aug 2005: 431,478 (avg. 1775/ day) (Hunter and Cohen 2006)

“Manual curation is not sufficient for annotation of genomic databases”

- Baumgartner et al, ISMB 2007

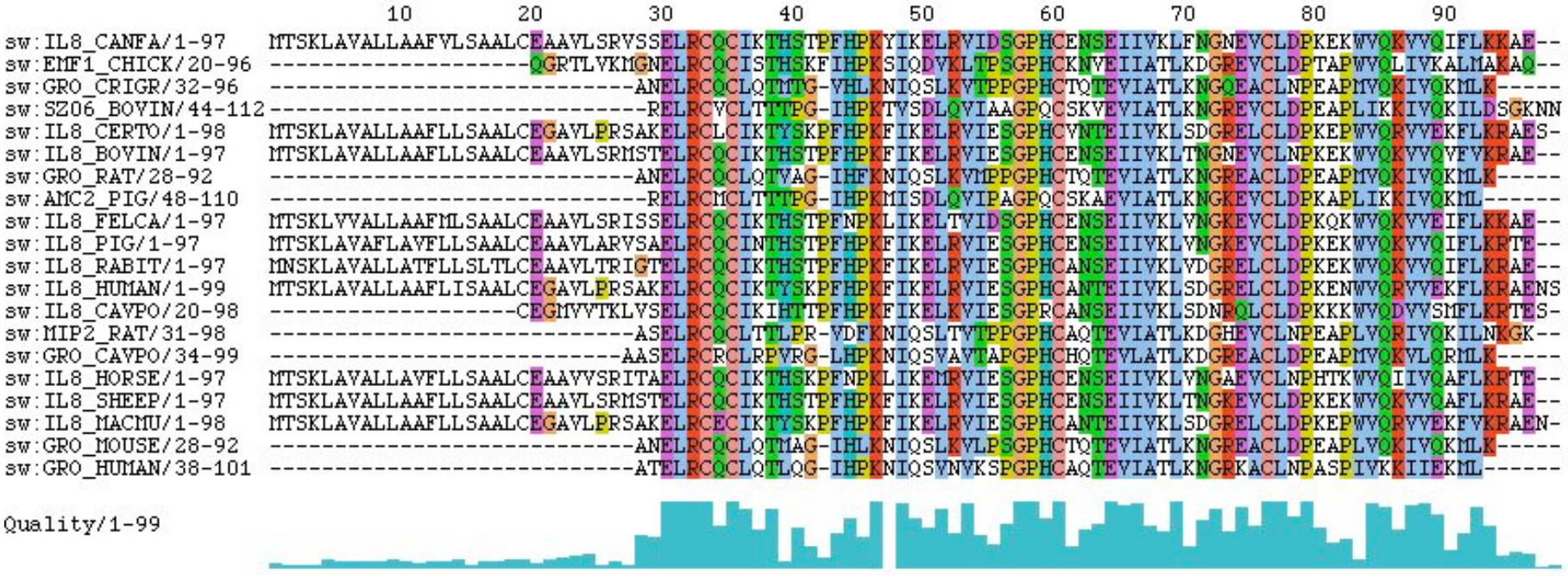




Graphs courtesy of William Baumgartner, U. Colorado

producing automated methods for semantic characterization of bioentities

- Modern bioinformatics techniques
  - statistical analysis
  - bootstrap from existing data based on similarity (“IEA”)



"The special sauce": Add Formal Semantics

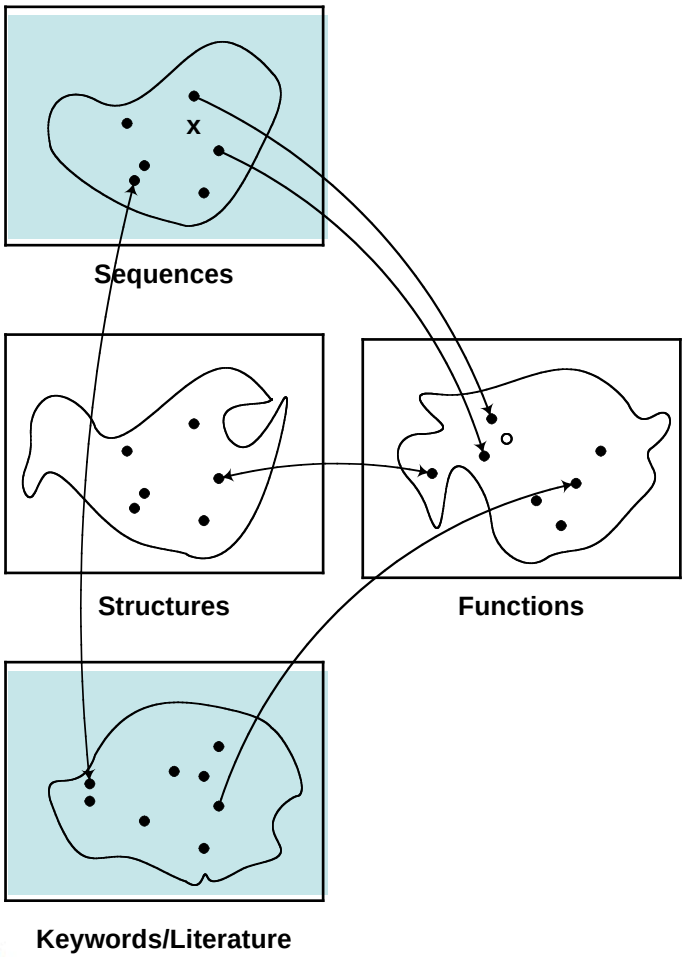
- exploit textual data sources
- take advantage of the structure and meaning of the data, whether human-specified or inferred
- use the semantics to organize, integrate and explore the data
- define similarity of entities in terms of semantic structure

Luckily for us, the biology domain is rich in semantic resources.

LANL Ontological methods for bioinformatic analysis

- Automated Protein Function Annotation using GO-space
- Formal concept analysis for semantic integration of cancer genome data

Automated Protein Function Annotation



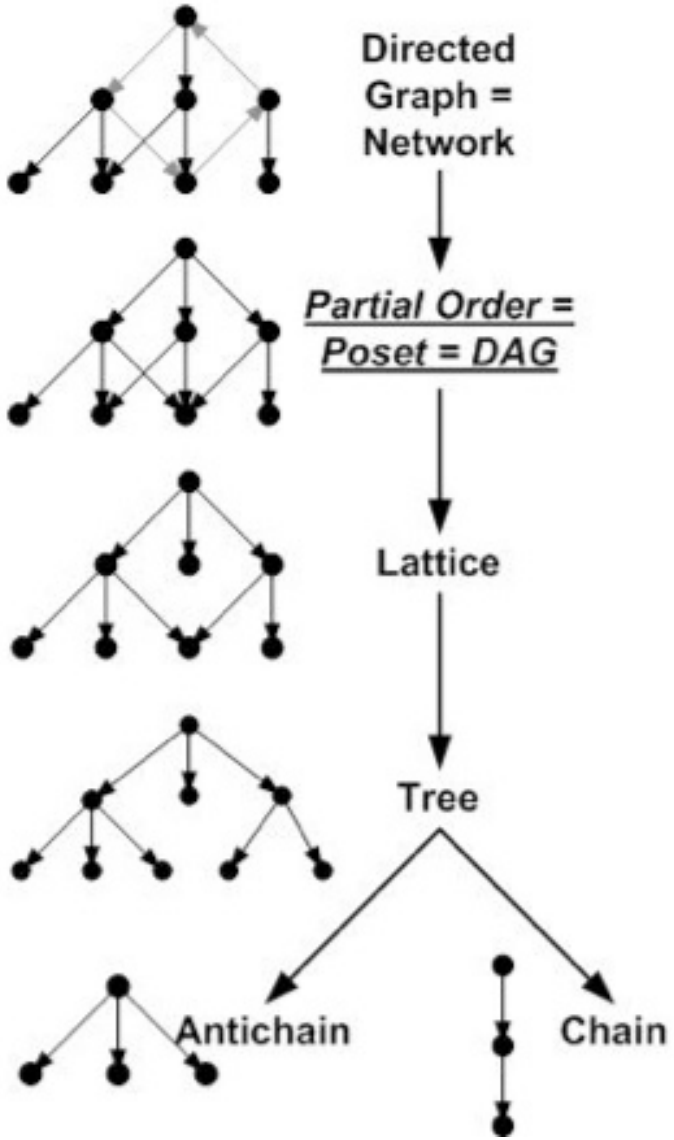
- Mappings
  - *From* regions of sequence, structure, keyword spaces
  - *Into* regions of **biological function** space:
    - taxonomic bio-ontologies of molecular function
- Characterize *formal* structure of bio-ontologies:
  - Order theoretical approaches
  - Combinatorial algorithms

POSOLE: POSet Ontology Laboratory Environment

- **POSOLE:** a general environment for ontology experimentation
  - Graph representation of an ontology as a POSet
  - POSet statistics analysis (e.g. depth, width, average rank)
  - Algorithms for node categorization utilizing the structure of the ontology
- **Deployment:** Ontology categorization for automated protein function annotation
  - Function: Gene Ontology node
  - Protein: target sequence or Swiss-Prot identifier
  - Map proteins to sets of potential Gene Ontology nodes
  - Ontology categorization: “clustering” nodes in ontology space to identify the most likely node assignment
- **Dual Queries:** Text and sequence neighborhoods

developed with Judith Cohn, Sue Mniszewski, Cliff Joslyn @ LANL

Hierarchies as Partially Ordered Sets



- **Partial Order:** Set P; relation $\leq \subseteq P^2$: reflexive, anti-symmetric, transitive
- **Poset:** $\mathcal{P} = \langle P, \leq \rangle$
- Simplest mathematical structures which admit to descriptions in terms of "levels" and "hierarchies"
- More specific than graphs or networks: no cycles, equivalent to Directed Acyclic Graphs (DAGs)
- More general than trees, lattices: single nodes, pairs of nodes can have multiple parents
- Ubiquitous in knowledge systems: constructed, induced, empirical

Basic POSET concepts

Poset: $\mathcal{P} = \langle P, \leq \rangle$

Comparable Nodes: $a \sim b := a \leq b$ or $b \leq a$

Up-Set: $\uparrow a = \{b \geq a\}$, **Down-Set:** $\downarrow a = \{b \leq a\}$

Chain: Collection of comparable nodes: $a_1 \leq a_2 \leq \ldots \leq a_n$

Height: Size maximal chain $\mathcal{H}(\mathcal{P})$

Noncomparable Nodes: $a \not\sim b$

Antichain: Collection of noncomparable nodes: $A \subseteq P, a \not\sim b, a, b \in A$

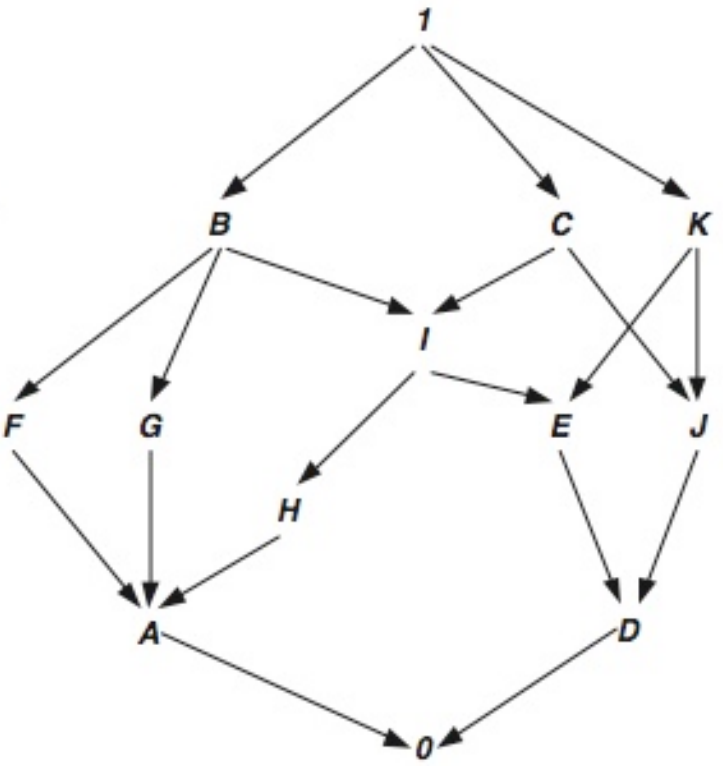
Width: Size maximal antichain $\mathcal{W}(\mathcal{P})$

Interval: $[a, b] := \{c \in P : a \leq c \leq b\}$, a bounded sub-poset of $\mathcal{P}$

Join, Meet: $a \vee b, a \wedge b \subseteq P$

Lattice: Then $a \vee b, a \wedge b \in P$

Bounded: Min $0 \in P$, Max $1 \in P$



Chain decomposition

Comparable Nodes: e.g. $D \leq 1 \in P$

Chain Decomposition: Set of all chains connecting them:

$$\begin{aligned} \mathcal{C}(D,1) &= \{C_j\} \\ &= \{D \prec E \prec I \prec B \prec 1, D \prec E \prec I \prec C \prec 1, \\ &\quad D \prec E \prec K \prec 1, D \prec J \prec C \prec 1, \\ &\quad D \prec J \prec K \prec 1\} \subseteq 2^P \end{aligned}$$

Chain Lengths: $h_j := |C_j| - 1$

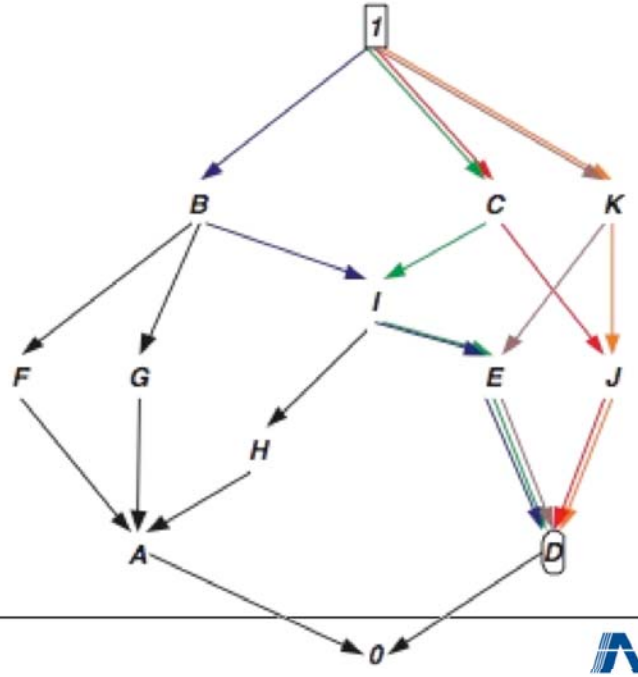
Vectors of Chain Lengths:

$$\vec{h}(a,b) := \left\langle h_j \right\rangle_{j=1}^{M} = \langle 4,4,3,3,3 \rangle$$

Extremes:

$$h_*(a,b) = \min_{h_j \in \vec{h}(a,b)} h_j = 3$$

$$h^*(a,b) = \max_{h_j \in \vec{h}(a,b)} h_j = 4$$



Pseudo-distances

Pseudo-Distance: Some aggregate measure of the number of "hops" between two comparable nodes: $\delta: P^2 \mapsto \mathbb{R}$ where $\forall a \leq b \in P, h_*(a,b) \leq \delta(a,b) \leq h^*(a,b)$

Normalized: $\bar{\delta} := \delta/(\mathcal{H}-1) \in [0,1]$

Minimum Chain Length: $\delta_m(a,b) := h_*(a,b), \bar{\delta}_m(a,b) := \bar{h}_*(a,b)$

Maximum Chain Length: $\delta_x(a,b) := h^*(a,b), \bar{\delta}_x(a,b) := \bar{h}^*(a,b)$

Average of Extreme Chain Lengths:

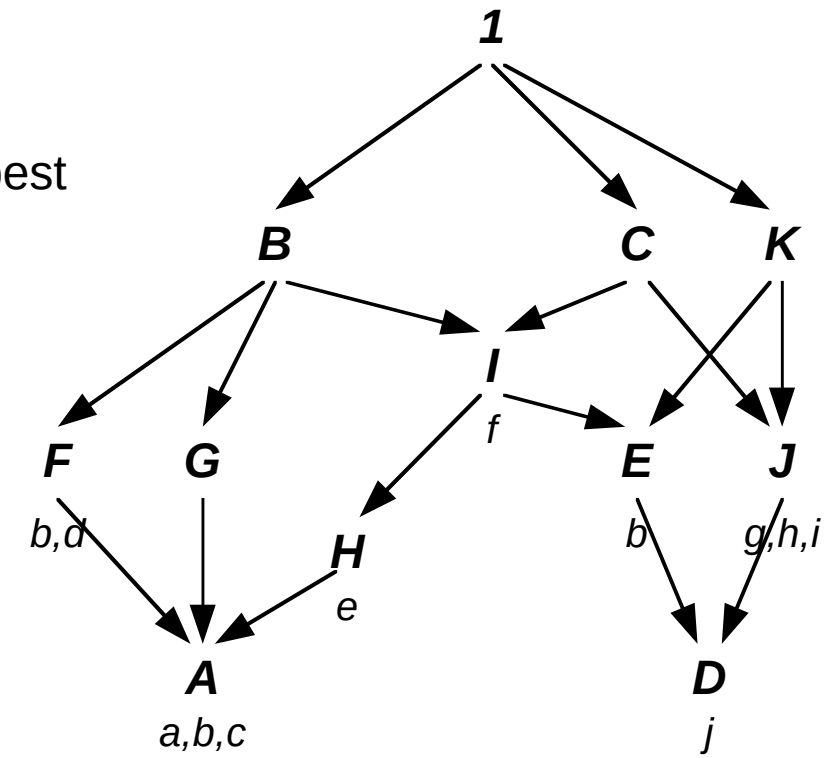
$$\delta_{ax}(a,b) := \frac{h_*(a,b)+h^*(a,b)}{2}, \quad \bar{\delta}_{ax}(a,b) := \frac{\bar{h}_*(a,b)+\bar{h}^*(a,b)}{2}$$

Average of All Chain Lengths:

$$\delta_{ap}(a,b) := \frac{\sum_{h_j \in \vec{h}(a,b)} h_j}{M}, \quad \bar{\delta}_{ap}(a,b) := \frac{\sum_{\bar{h}_j \in \vec{\bar{h}}(a,b)} \bar{h}_j}{M}$$

Order Theoretical Categorization Method

- Represent GO as labeled, finite ordered set
- Given labels (genes) *c, e, i* . . .
- What node(s) *A,B, C, . . . ,K* are best to attend to?
  - *C*
  - {*H*, *J*}
  - {*A*, *H*, *J*}



Scoring function

$$S_Y(p) = \sum_{x \in Y} \sum_{p' \in F(x):\, p' \le p} (\delta^r(p', p) + 1)^{-1}$$

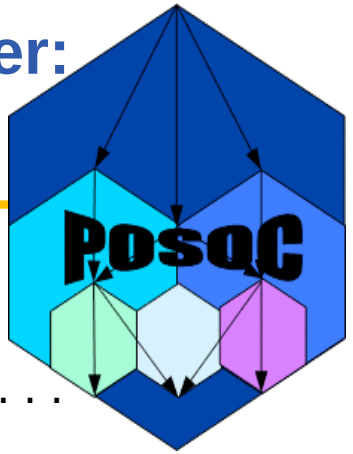
$$s \in \{\ldots, -1, 0, 1, 2, 3, \ldots\}$$

$$r = 2^s$$

POSOLE strategy

- Function Prediction as ***Categorization of Nearest Neighbors***
  - Application of POSOC categorization methodology utilizing the Gene Ontology structure to find the best covering nodes given a set of node “hits”
  - “Hits” are based on (application-dependent) mappings from neighbors of an input protein to Gene Ontology nodes
  - Covering nodes are function annotation predictions

Partially Ordered Set Ontology Categorizer: "Cluster" Genes in Ontology Space

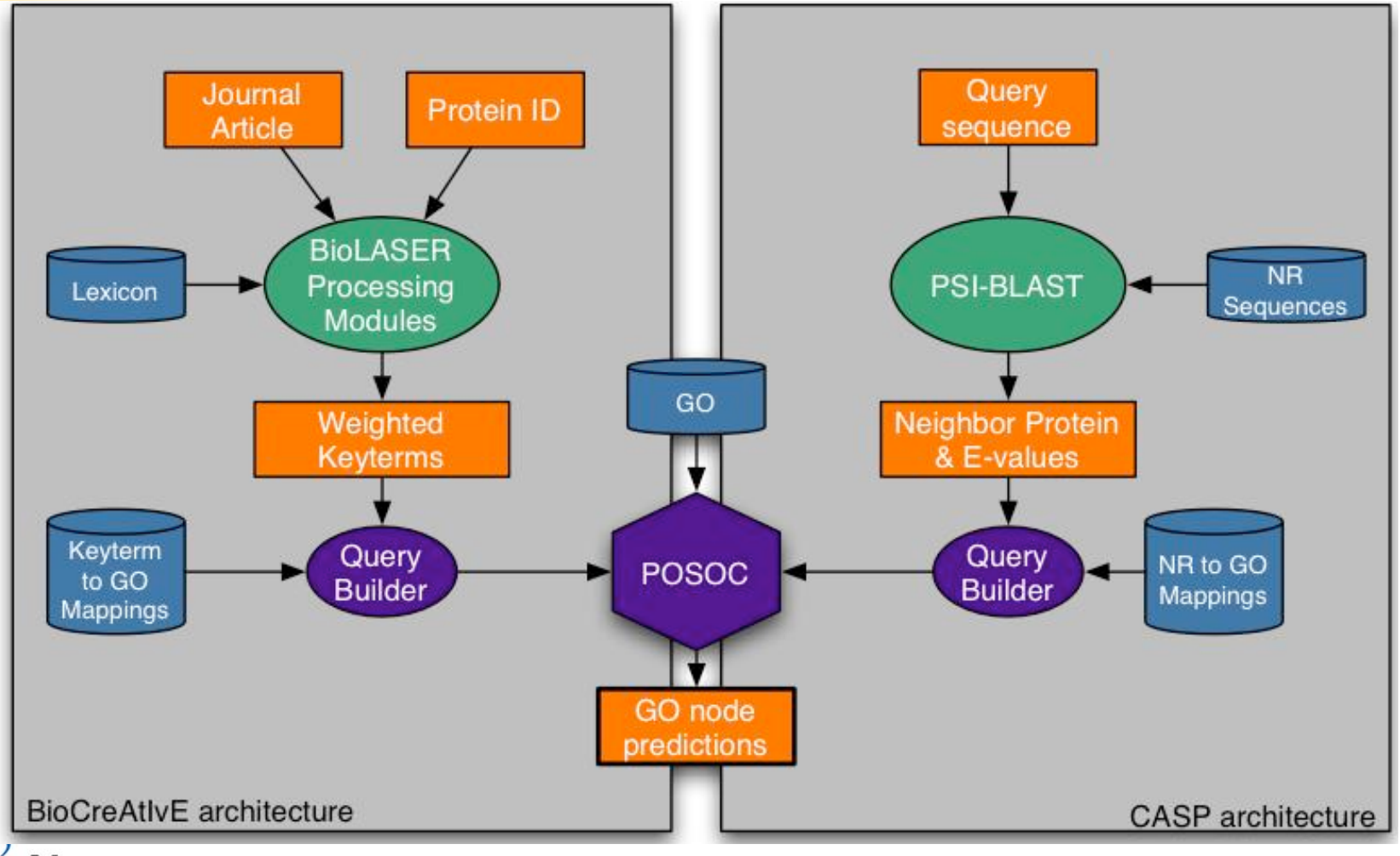


http://www.c3.lanl.gov/posoc/

- Given the Gene Ontology (GO) . . . And mappings to GO nodes . . .
- "Splatter" them over the GO . . . Where do they end up?
  - Concentrated? -- Dispersed?
  - Clustered? -- High or low?
  - Overlapping or distinct?
- Pseudo-distances between comparable nodes to measure vertical separation
- POSOC traverses the structure of the GO, percolating hits upwards, and calculating scores for GO nodes.
- Scores to rank-order nodes with respect to gene locations, balancing:
  - **Coverage:** Covering as many genes as possible
  - **Specificity:** But at the "lowest level" possible
- "Cluster" based on non-comparable high score nodes

Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy; and Heaton, Gary: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177

POSOLE applications



Application: BioCreAtIvE 2004 Task 2

Critical Assessment of Information Extraction in Biology

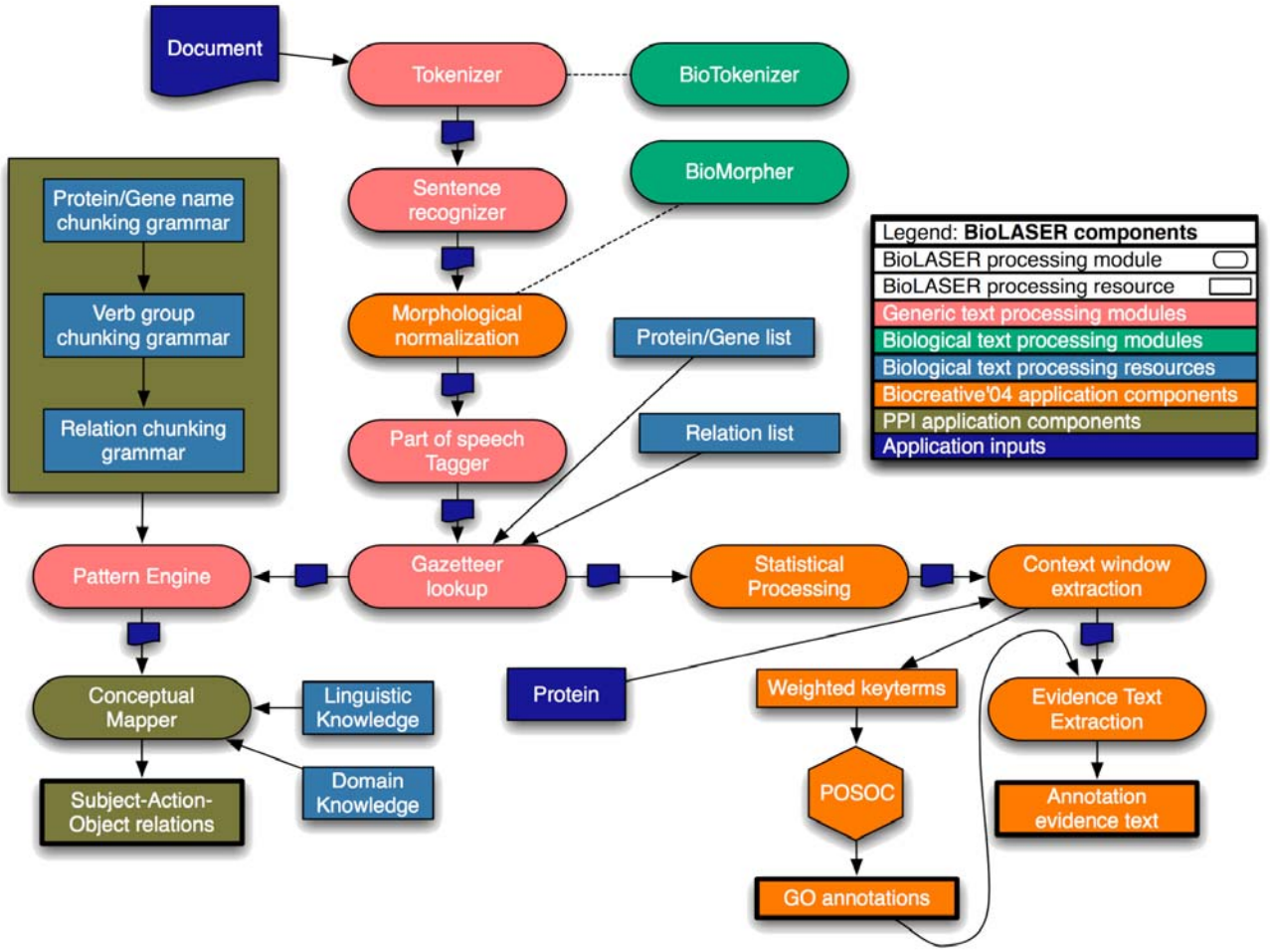
- Automatic assignment of Gene Ontology annotations to human proteins based on a journal publication
  - Given a Swiss-Prot/TrEMBL protein ID and a document, predict a GO node to which the protein should be annotated
  - Also return the evidence text from the document supporting the annotation
- Strategy: Annotation as Categorization of Document Neighborhood
  - Application of POSOC categorization utilizing the Gene Ontology structure to find the best covering nodes given a set of node “hits”
  - “Hits” in this case are based on overlaps between input terms and GO node terms (in labels, definitions)

POSOC as applied to context terms

- Collect all terms in a context window of *n* sentences around any reference to the protein of interest
- Transform an input query into a set of node hits:
  - Morphologically normalize GO node labels
  - Look for any overlaps between input terms and terms in the normalized node labels
  - An overlap = a node hit, with strength based on the input weight of the term (from TFIDF)
  - Multiple overlaps on a given node count as multiple hits
- POSOC returns a set of GO nodes representing cluster heads for weighted term input set, and data on which input terms contributed to the selection of each cluster head: *Annotation predictions*

BioLASER:
Los Alamos Semantic Event Recognizer for Biology

- Text analysis environment:
  - Relation extraction
  - Term vector analysis
- Domain-specific and application-specific components
- Markup workflow implementation
  - Using UIMA platform
  - GATE modules

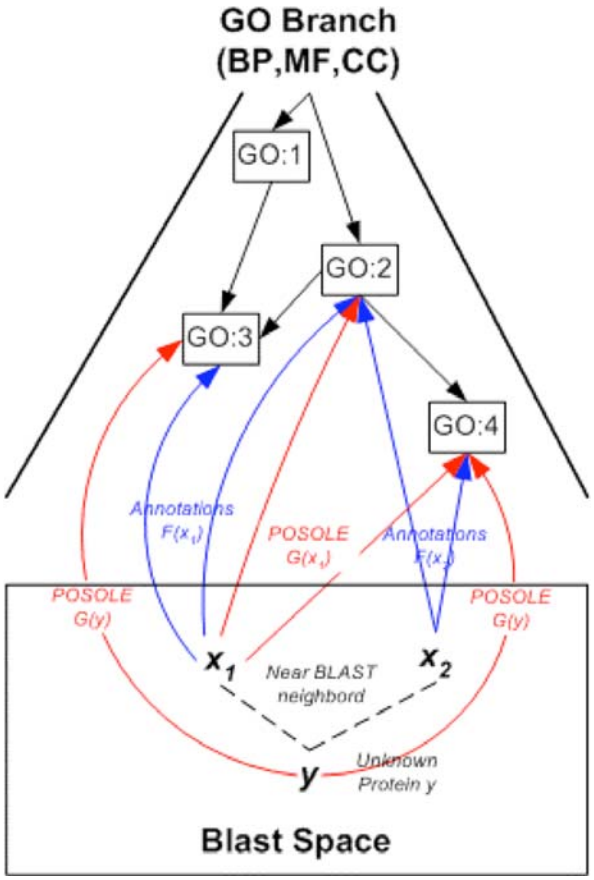


Application: CASP-6 Function Prediction

Critical Assessment of Structure Prediction evaluation

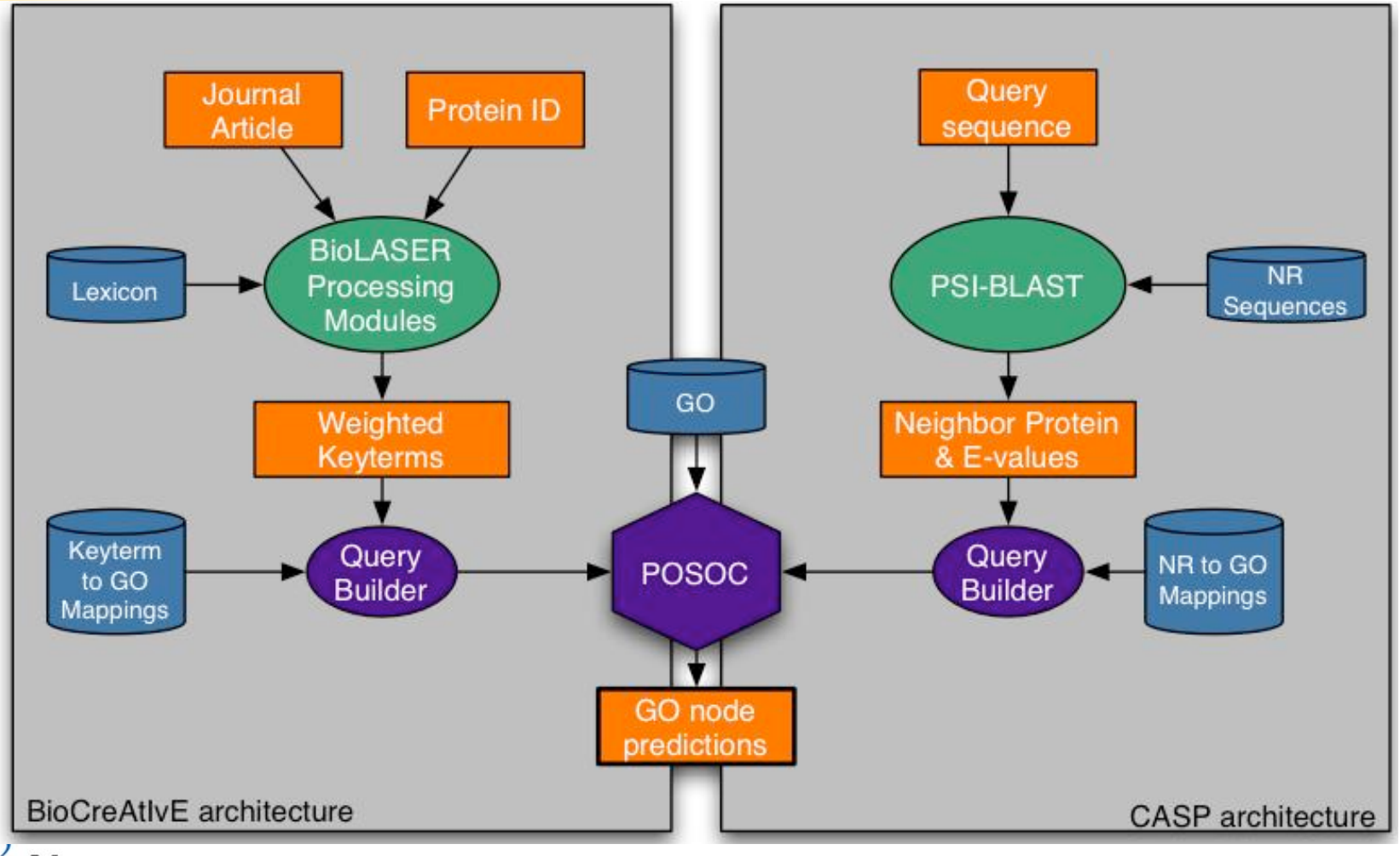
Function Prediction subtask

- Automatic assignment of Gene Ontology annotations to target protein sequences
- Strategy: Annotation as Categorization of Sequence Neighborhood
  - Application of POSOC categorization utilizing the Gene Ontology structure to find the best covering nodes given a set of node “hits”
  - “Hits” in this case are based on known mappings from proteins in the sequence neighborhood (BLAST neighborhood) of the target to Gene Ontology nodes

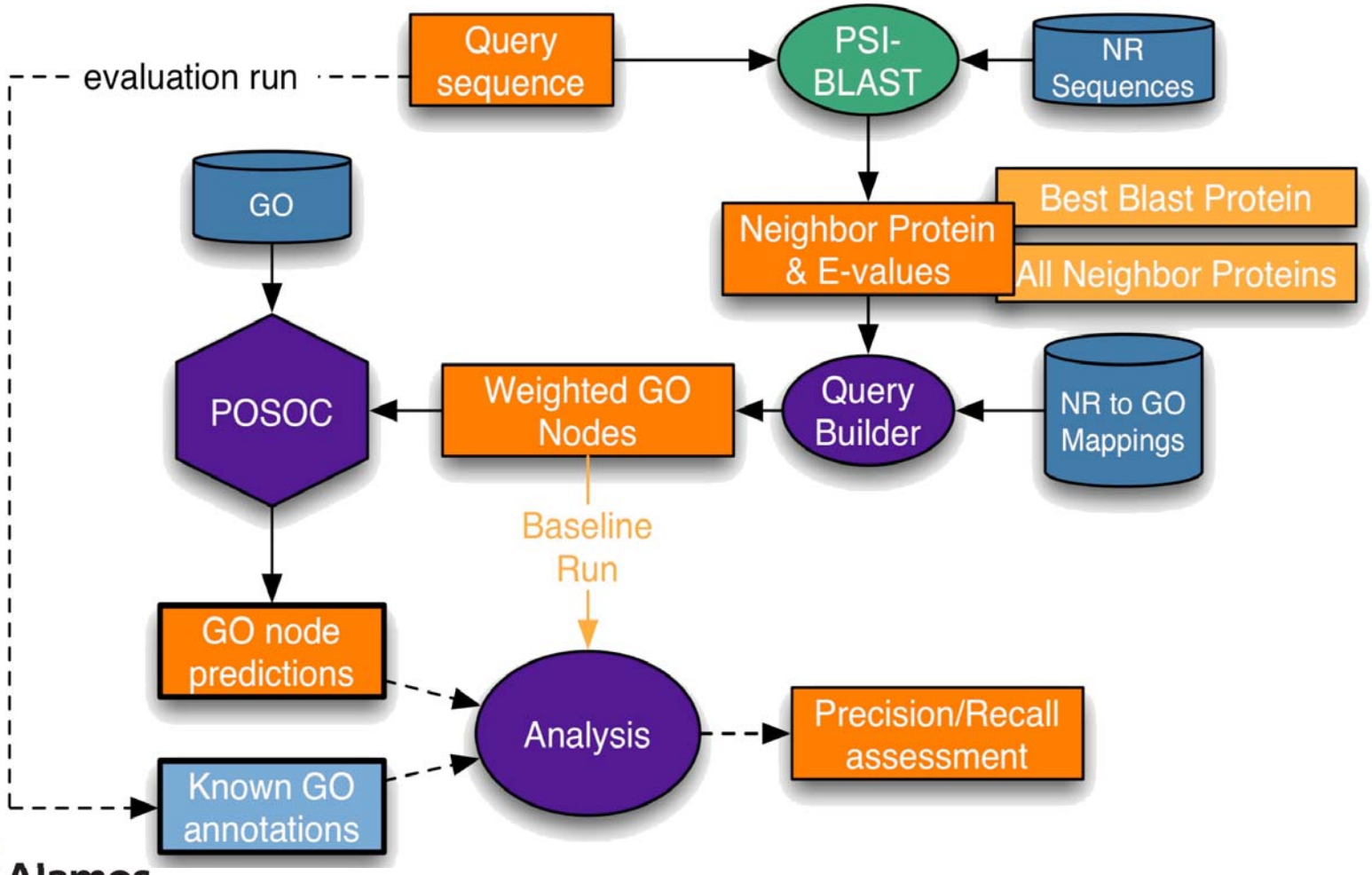


Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA (2006). “Categorization Approach to Automated Ontological Function Annotation”, *Protein Science*, v. **15**, pp. 1544-1549

POSOLE applications



CASP architecture



CASP Evaluation

- Test set
  - proteins with known Gene Ontology mappings
  - 4530 SwissProt protein sequences
  - Protein to GO Mappings derived from UniProt database
- Eliminate PSI-BLAST identity matches from mappings used in prediction
- Goal: compare function predictions made by the system with known functions assigned to each input protein

CASP Evaluation runs

POSOC: Full Neighborhood	POSOC: Best Blast
Baseline: Full Neighborhood	Baseline: Best Blast

- **Baseline Best Blast**: Predictions are the GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis. All predicted GO nodes are considered to be at rank 1.
- **Baseline Full Neighborhood**: Predictions are the GO nodes associated with ***all*** proteins matched in the PSI-BLAST analysis (with evalue < 10). The predictions are ranked according to the evalue of the corresponding PSI-BLAST match.
- **POSOC Best Blast**: Inputs to POSOC are the GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis, weighted by evalue of the match. POSOC categorizes and ranks these inputs to produce the predictions.
- **POSOC Full Neighborhood**: Inputs to are the GO nodes associated with ***all*** proteins matched in the PSI-BLAST analysis, weighted by evalue of the match. POSOC categorizes and ranks these inputs to produce the predictions.

Evaluation analysis

- Precision/Recall
  - Precision = % of predictions that are correct

$$P = \frac{\left|F(x) \cap G(x)\right|}{\left|G(x)\right|}$$

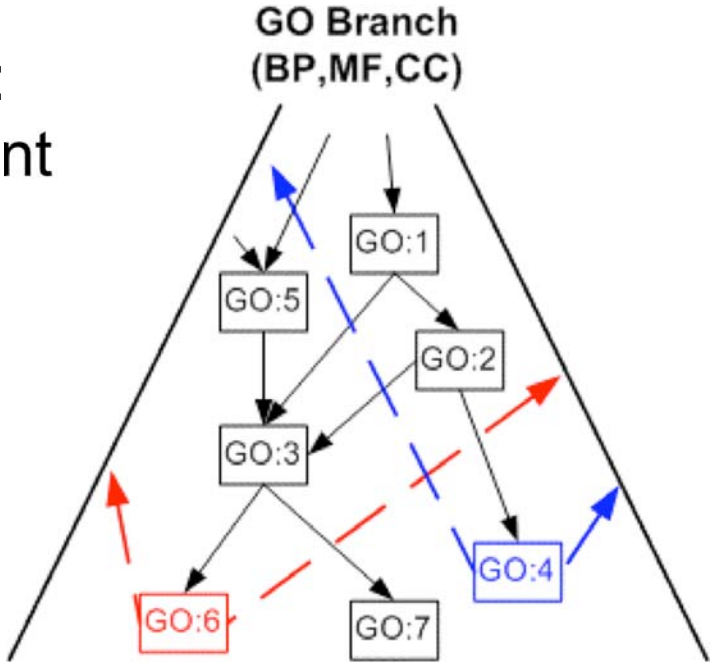
  - Recall = % of known predictions that are recovered

$$R = \frac{\left|F(x) \cap G(x)\right|}{\left|F(x)\right|}$$

- Extension to ranked list of predictions
  - Consider precision/recall at different ranks

Evaluation of Ontological predictions

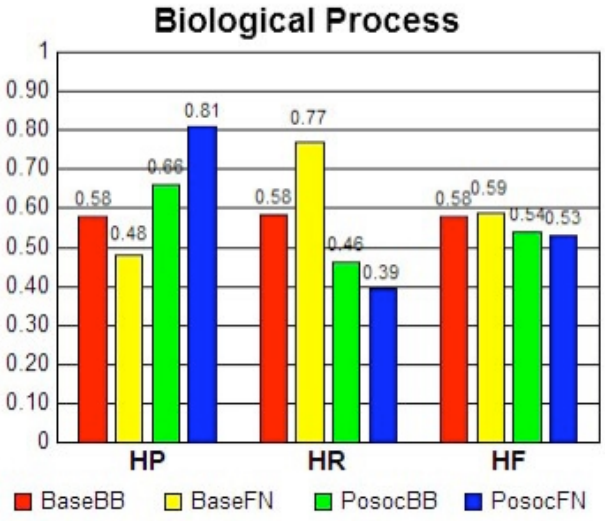
- Extension to ontological predictions: when does a GO node p in F(x) count as a "match" against a q in G(x)?
  - What about siblings? Ancestors?
  - Partial credit?
    - Based on proximity
    - Based on specificity
- Adapt hierarchical precision/recall measure from Kiritchenko et al 2005

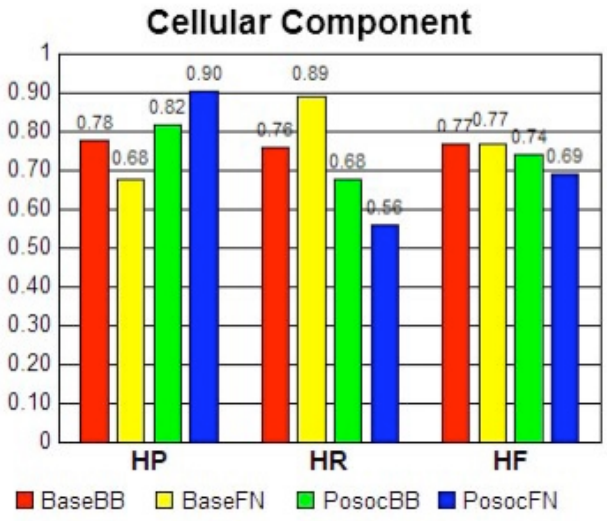


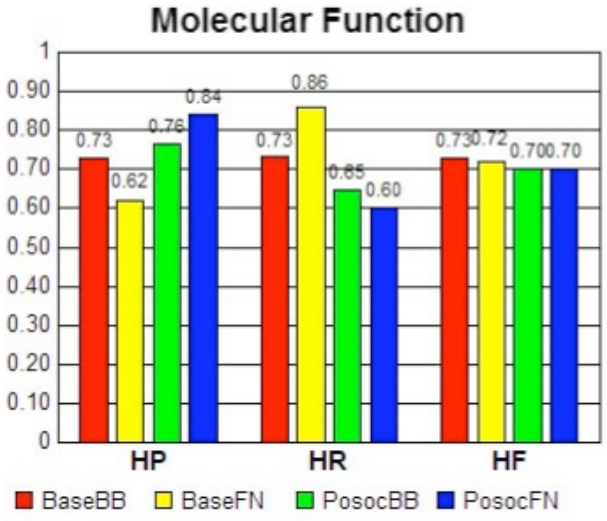
$$P = \sum_{q \in G(x)} \max_{p \in F(x)} \frac{\left| \uparrow p \cap \uparrow q \right|}{\left| \uparrow q \right|} \qquad R = \sum_{p \in F(x)} \max_{q \in G(x)} \frac{\left| \uparrow p \cap \uparrow q \right|}{\left| \uparrow p \right|}$$

Kiritchenko, S; Matwin, S; Famili, AF (2005). "Functional Annotation of Genes Using Hierarchical Text Categorization", *Proc. BioLINK SIG on Text Data Mining.*

CASP results







	P	R	F
BP	0.20	0.14	0.16
CC	0.36	0.25	0.29
MF	0.39	0.28	0.33

Non-hierarchical BaseBB

LANL ontological methods for bioinformatic analysis

- Automated Protein Function Annotation using GO-space
- Formal concept analysis for semantic integration of cancer genome data (with Damien Gessler at NCGR)

Towards a cure for cancer

- Sjoblom et al 2006:
  - genomic and informatic filtering of 816,986 putative nucleotide changes across 13,023 genes in breast and colorectal tumors
  - analysis yielded 189 variants directly implicated in breast and colorectal cancer
- How can we understand the roles of implicated genes?
  - complement quantitative validation methods with semantic info
  - develop formal methods to *integrate* curated semantic information into discovery and validation processes
- Strategy: **Formal Concept Analysis** of semantic data objects

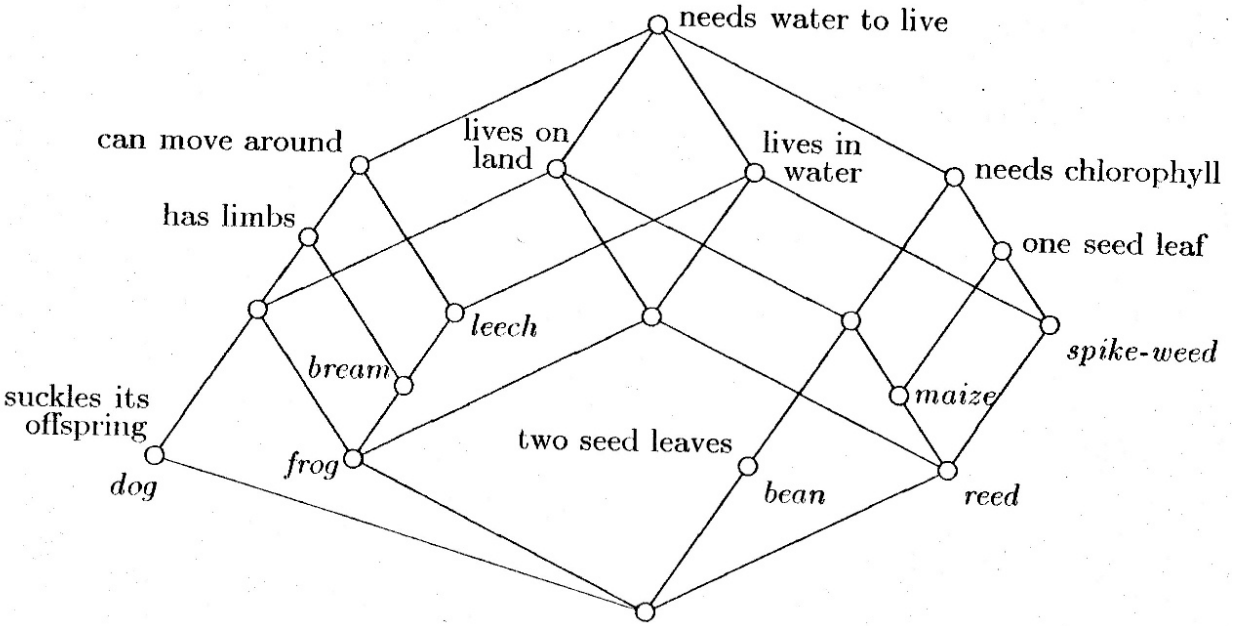
Sjoblom, Tobias; Jones, Sian; Wood, Laura D; Parsons, DW; et al (2006). "Consensus Coding Sequences of Human Breast and Colorectal Cancers", *Science,* v. **314**:5797, pp. 268-284.

Formal Concept Analysis

- Semantic hierarchy derived from relational data
- Visualization of relationships
- Hypothesis and rule generation and evaluation

		a	b	c	d	e	f	g	h	i
1	Leech	×	×					×		
2	Bream	×	×					×	×	
3	Frog	×	×	×				×	×	
4	Dog	×		×				×	×	×
5	Spike – weed	×	×		×		×			
6	Reed	×	×	×	×		×			
7	Bean	×		×	×	×				
8	Maize	×		×	×		×			

Figure 1.1 Context of an educational film "Living Beings and Water". The attributes are: a: needs water to live, b: lives in water, c: lives on land, d: needs chlorophyll to produce food, e: two seed leaves, f: one seed leaf, g: can move around, h: has limbs, i: suckles its offspring.

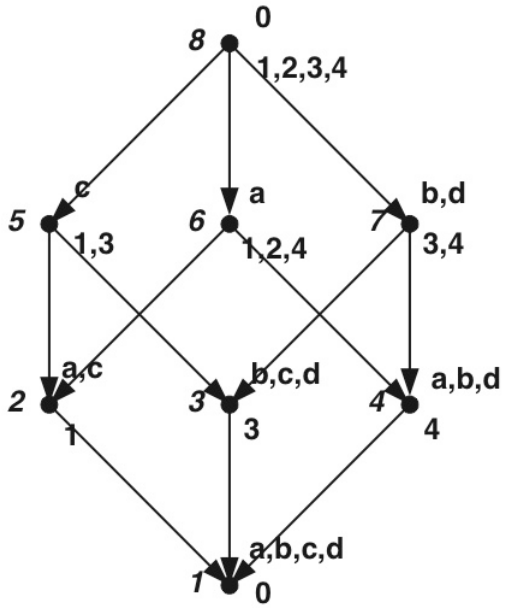


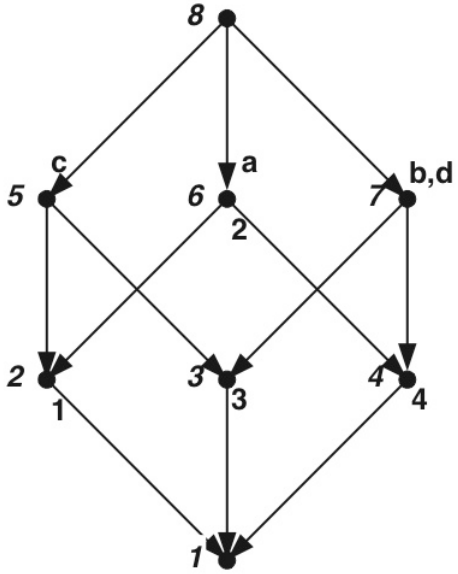
Ganter, Bernhard and Wille, Rudolf (1999). *Formal Concept Analysis*, Springer-Verlag.

FCA example

	a	b	c	d
1	√		√	
2	√			
3		√	√	√
4	√	√		√

Concept #	Concept
1	abcd/0
2	ac/1
3	bcd/3
4	abd/4
5	c/13
6	a/124
7	bd/34
8	Ø/1234





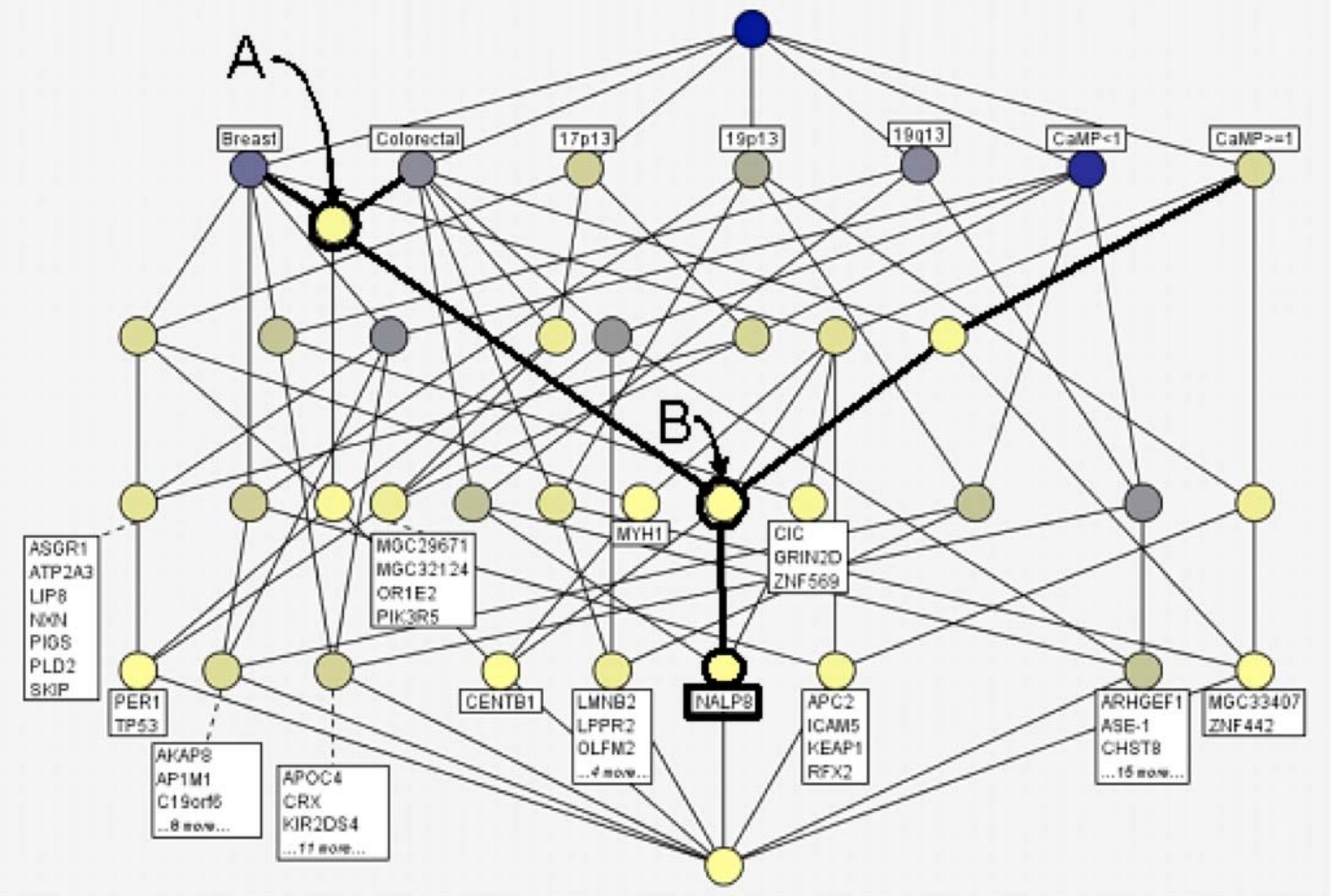
Cancer Genomics

- Variational mutation information from breast/colorectal tumors
- 75 genes in top three chromosome locations
- CaMP score: ≤ 1 not implicated (Low) , > 1 implicated (High)
- Formal context:

	Colorectal	19q13	CaMP<1	CaMP>=1	17p13	19p13	Breast
ASGR1			X		X		X
ATP2A3			X		X		X
LIP8			X		X		X
NXN			X		X		X
PIGS			X		X		X
PLD2			X		X		X
SKIP			X		X		X
NALP8	X	X		X			X
APC2				X		X	X
ICAM5				X		X	X
KEAP1				X		X	X
RFX2				X		X	X
APOC4		X	X				X
CRX		X	X				X
KIR2DS4		X	X				X

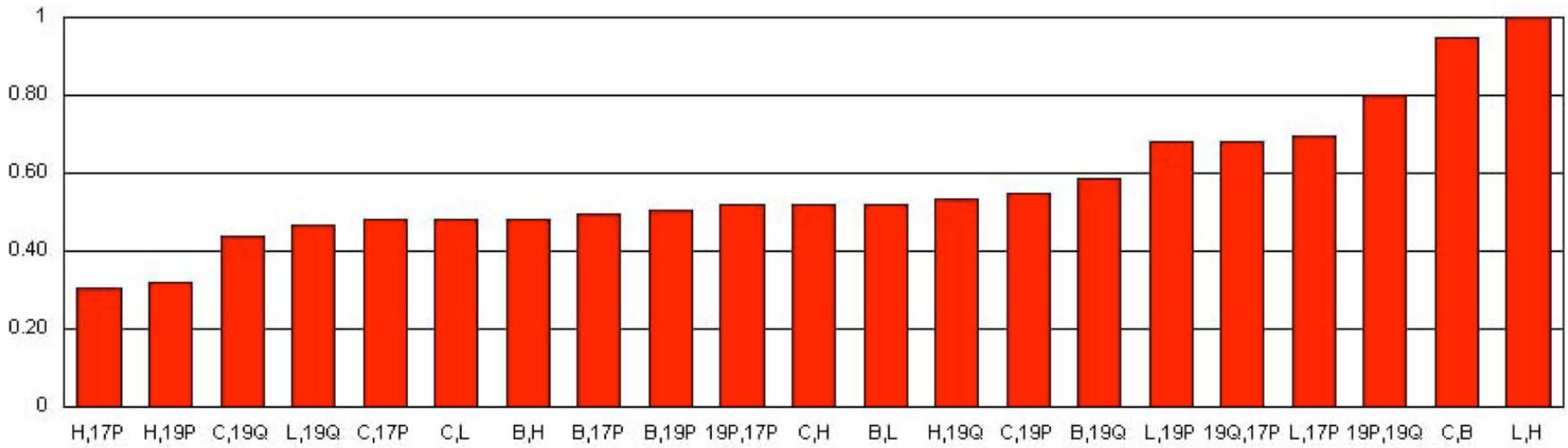
Concept Lattice

tumor type / chromosome location / CaMP score

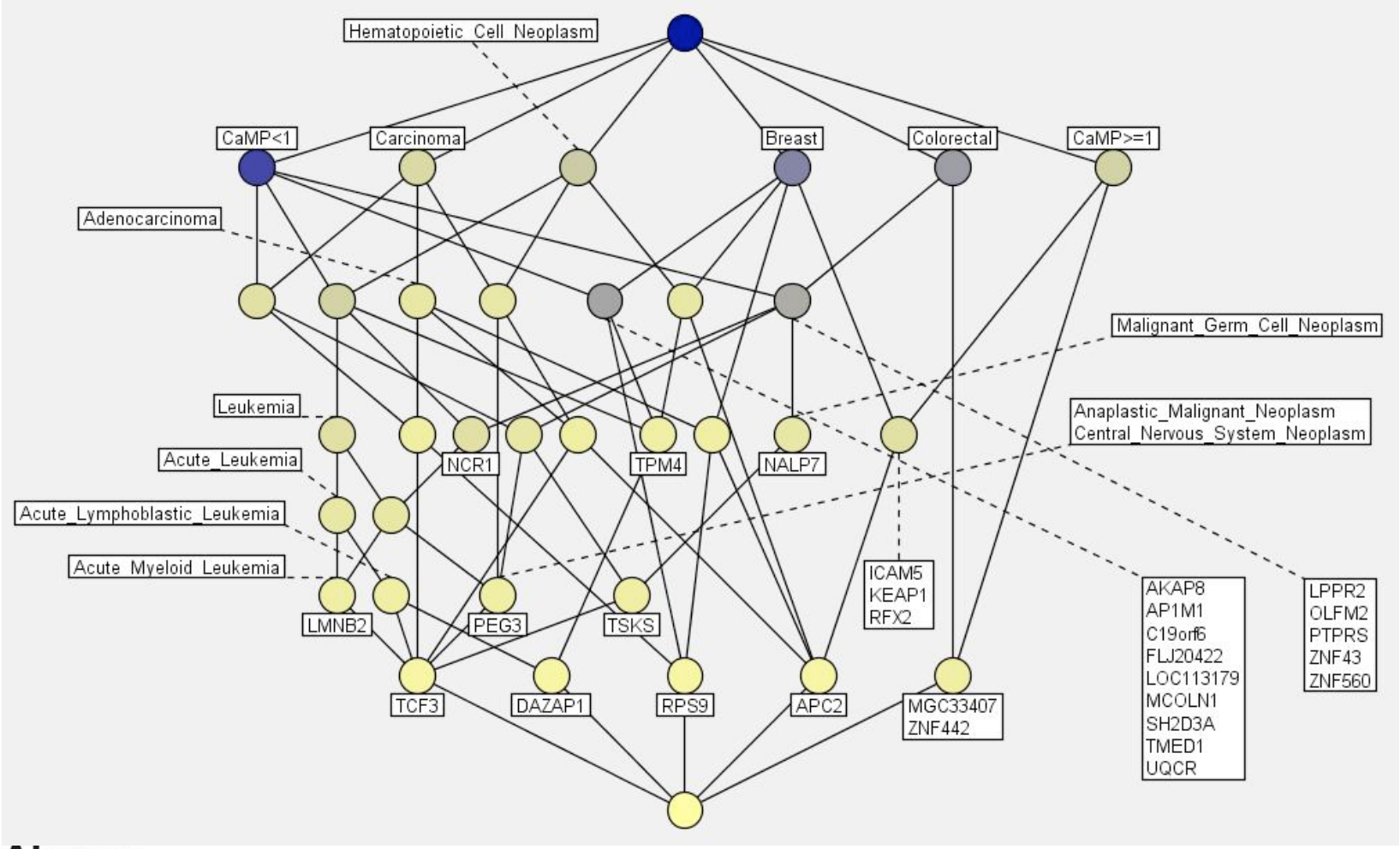


Attribute Metrics

- Lattice metrics between all pairwise attributes
- CaMP <(H)igh,(L)ow> maximally far apart: mutually exclusive
- <(C)olon,(B)reast> genes largely distinct
- Chromosomes far apart
- Pairs <H,17P>, <H,19P> very close: those chromosome locations may have many cancer-implicated genes

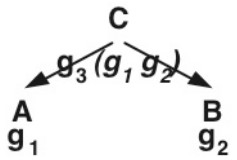


Adding National Cancer Institute Thesaurus (NCIT) terms

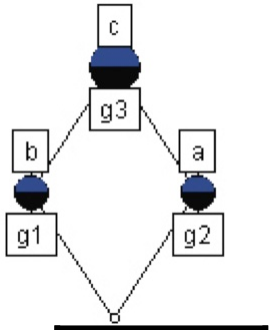


Cancer genes: next steps

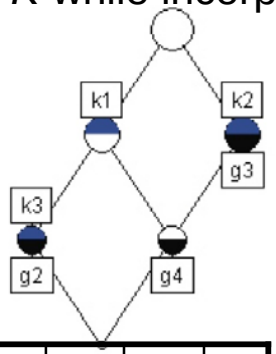
- Integrate more semantic information
  - Gene Ontology annotations
  - MeSH keyterms / extracted keyterms from associated literature



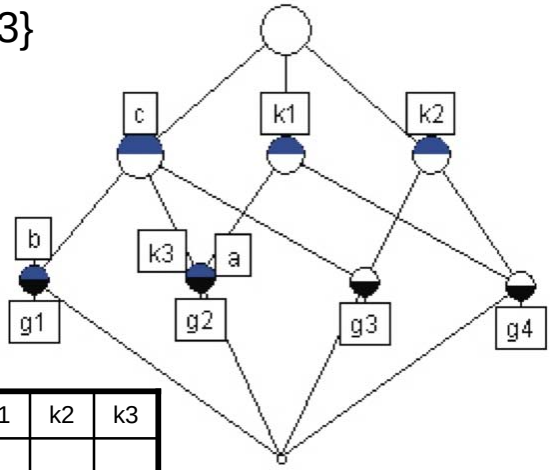
- Will lead to ontology integration and induction
  - {g1, g2, g3}: annotated into ontology *O*
  - {g2, g3, g4}: annotated to keywords *K* = {k1, k2, k3}
  - Induce order on *K* while incorporating order on *O*



	a	b	c
g1		√	√
g2	√		√
g3			√



	k1	k2	k3
g2	√		√
g3		√	
g4	√	√	

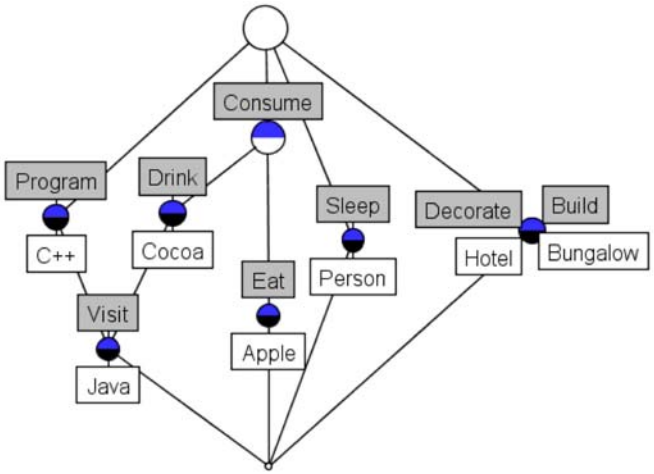


	a	b	c	k1	k2	k3
g1		√	√			
g2	√		√	√		√
g3			√		√	
g4				√	√	

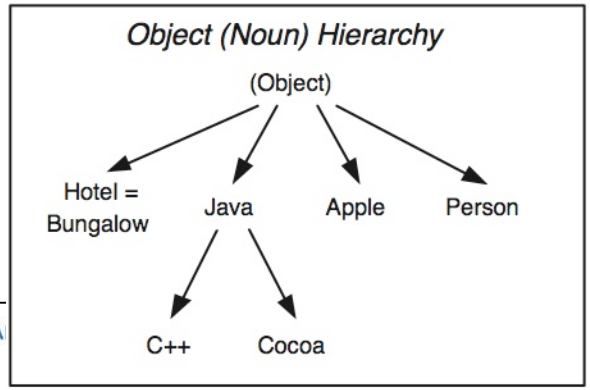
CA Joslyn, DDG Gessler, KM Verspoor (2007). “Knowledge Integration in Open Worlds: Utilizing the Mathematics of Hierarchical Structure”. In IEEE, *Int. Conf. on Semantic Computing.*

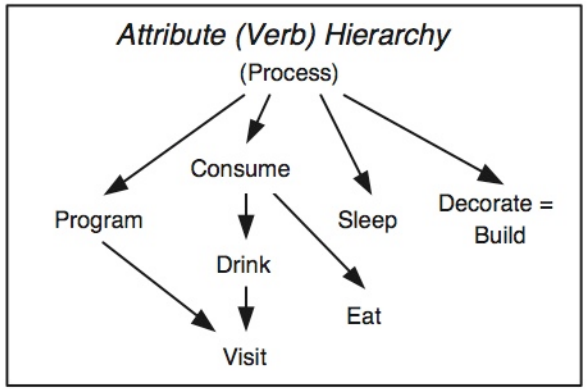
Formal concepts and text: new work

- Ontology induction from text
  - (Verb) Predicates as attributes
  - (Noun) Predicate arguments as objects
- Apply FCA to induce a dual order over and verbs
- Read hierarchy off lattice



	Program	Drink	Visit	Sleep	Build	Decorate	Consume	Eat
Java	X	X	X				X	
Cocoa		X					X	
C++	X							
Person				X				
Hotel					X	X		
Bungalow					X	X		
Apple							X	X





Los Alamos NATIONAL LABORATORY EST.1943
The World's Greatest Science Protecting A

Colleagues and Collaborators

- LANL
  - Cliff Joslyn
  - Judith Cohn
  - Sue Mniszewski
- National Center for Genome Resources
  - Damian Gessler
- Technische Universität Dresden
  - Stephan Schmidt
  - Bjoern Koester

With special thanks to William Baumgartner and Kevin Cohen at U. Colorado School of Medicine for sharing graphs from their ISMB 2007 paper, and to Betty Korber at LANL for information on the HIV database

Thanks for listening!