GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

19 October 2004

Lecture 14: Biology: Functions and Networks

- GeneOntology
- GoMiner
- Regulatory networks and metabolic pathways

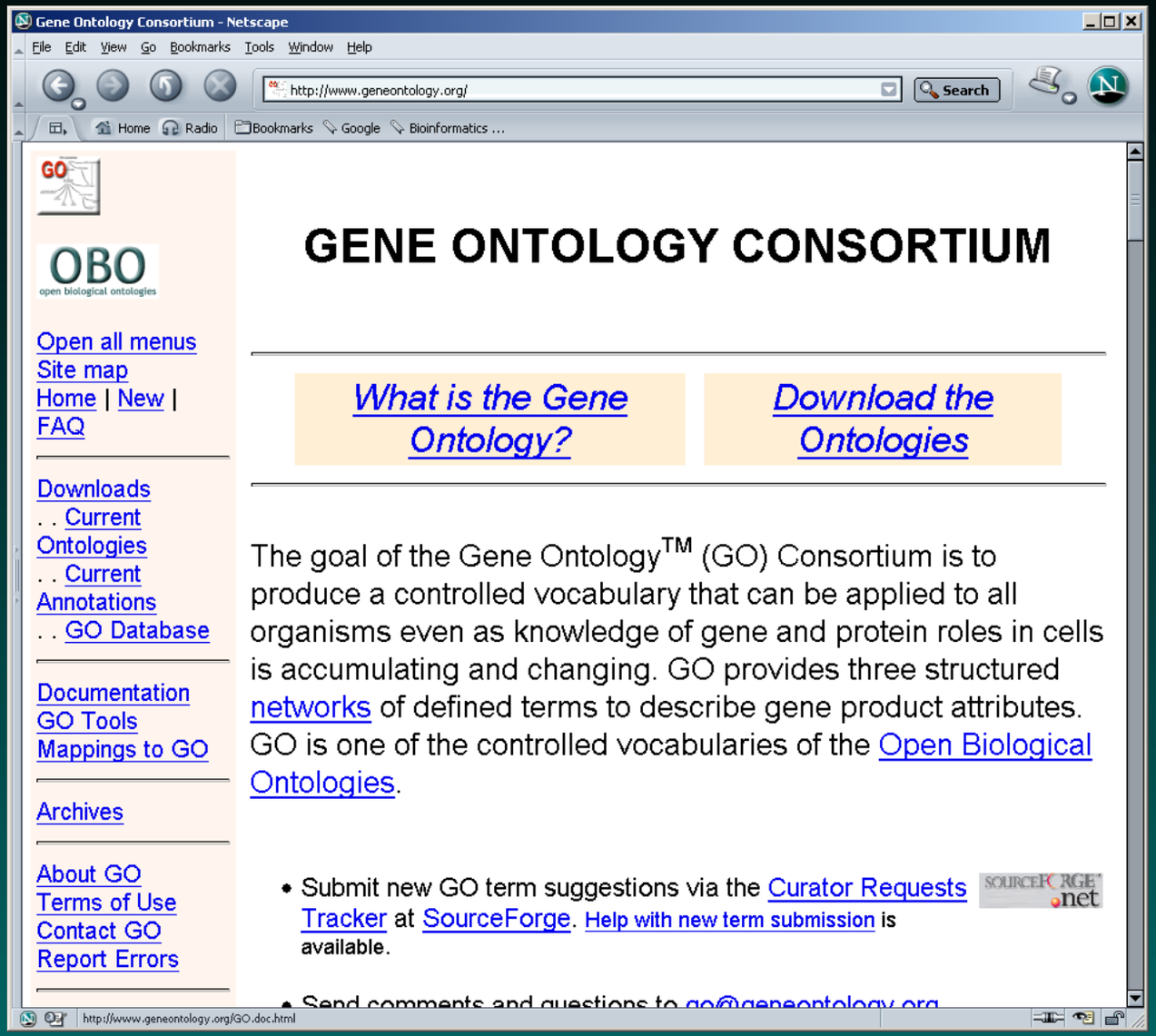
GeneOntology

GeneOntology (GO) uses controlled vocabularies to create a directed acyclic graph (DAG; a generalized tree) that describes the kinds of functions or properties that a gene might have.

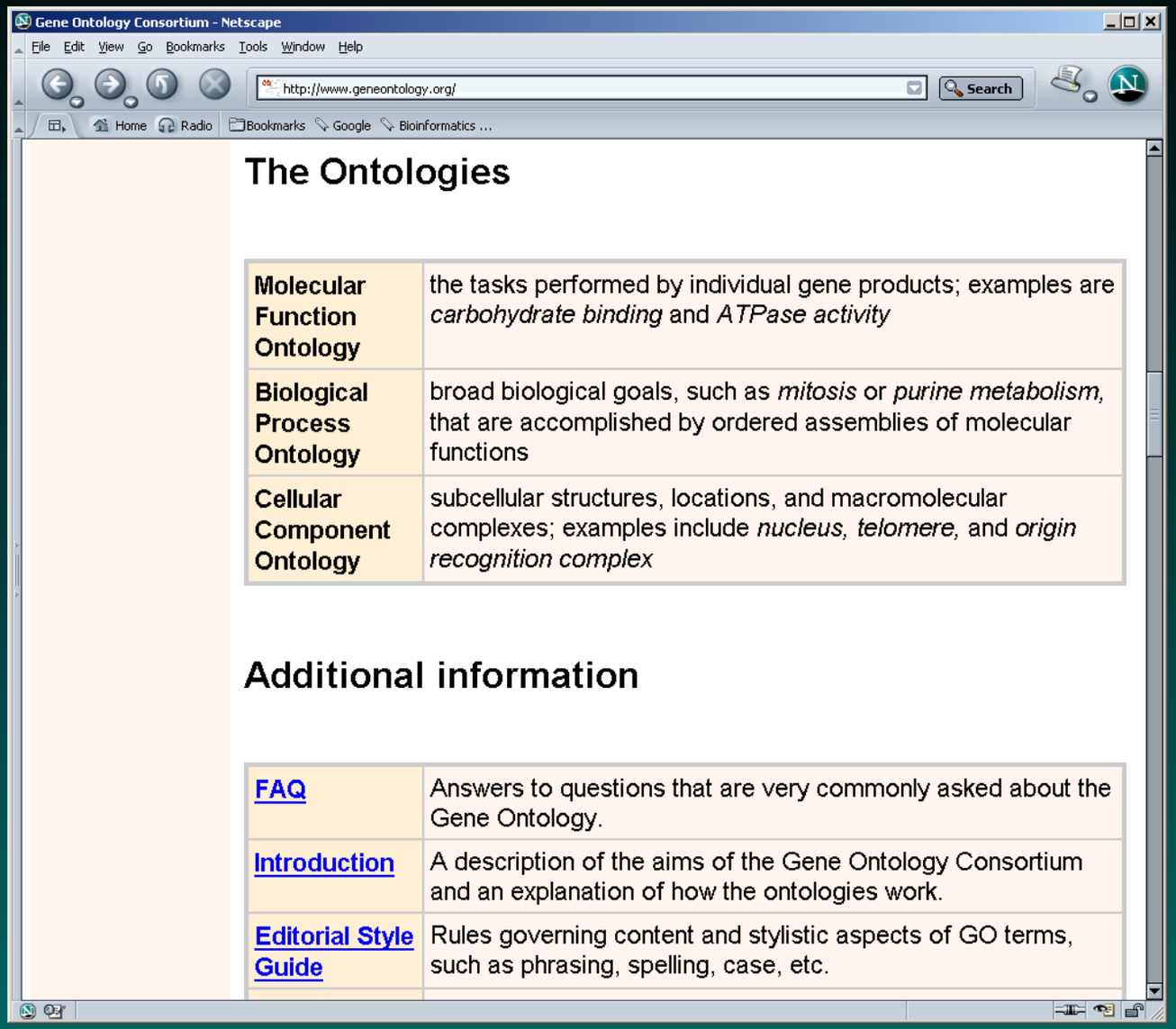
GeneOntology consists fundamentally of two pieces:

1. The DAG that describes functions and relations between them
2. Annotations that describe which genes actually have which functions

http://www.geneontology.org



GeneOntology: The top level

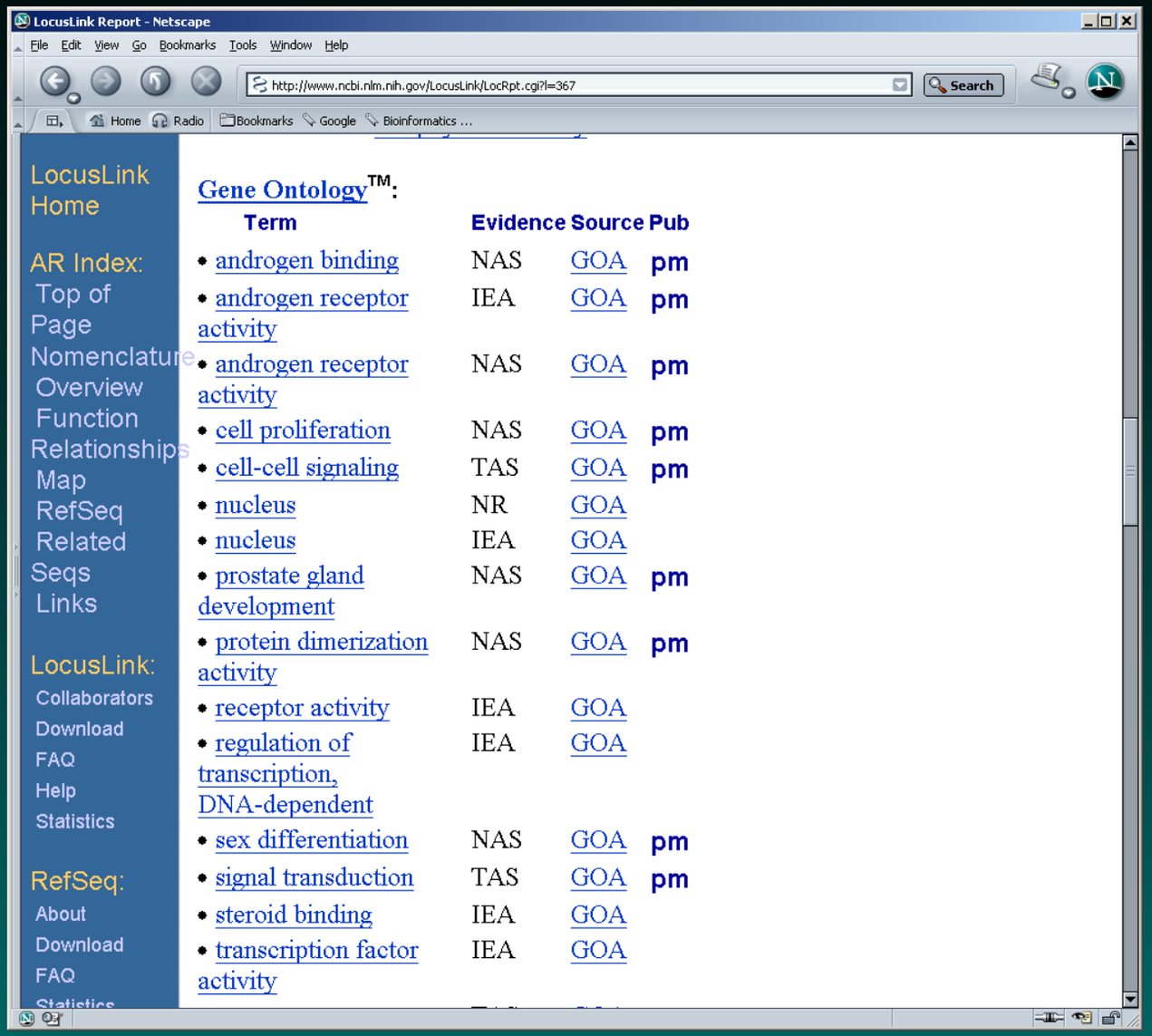


GeneOntology annotations in LocusLink

You can find the GeneOntology annotations for individual genes in LocusLink. For genes with known functions, the LocusLink page will conmtain a section titled “GeneOntology”, which contains a list of the known functions for that gene.

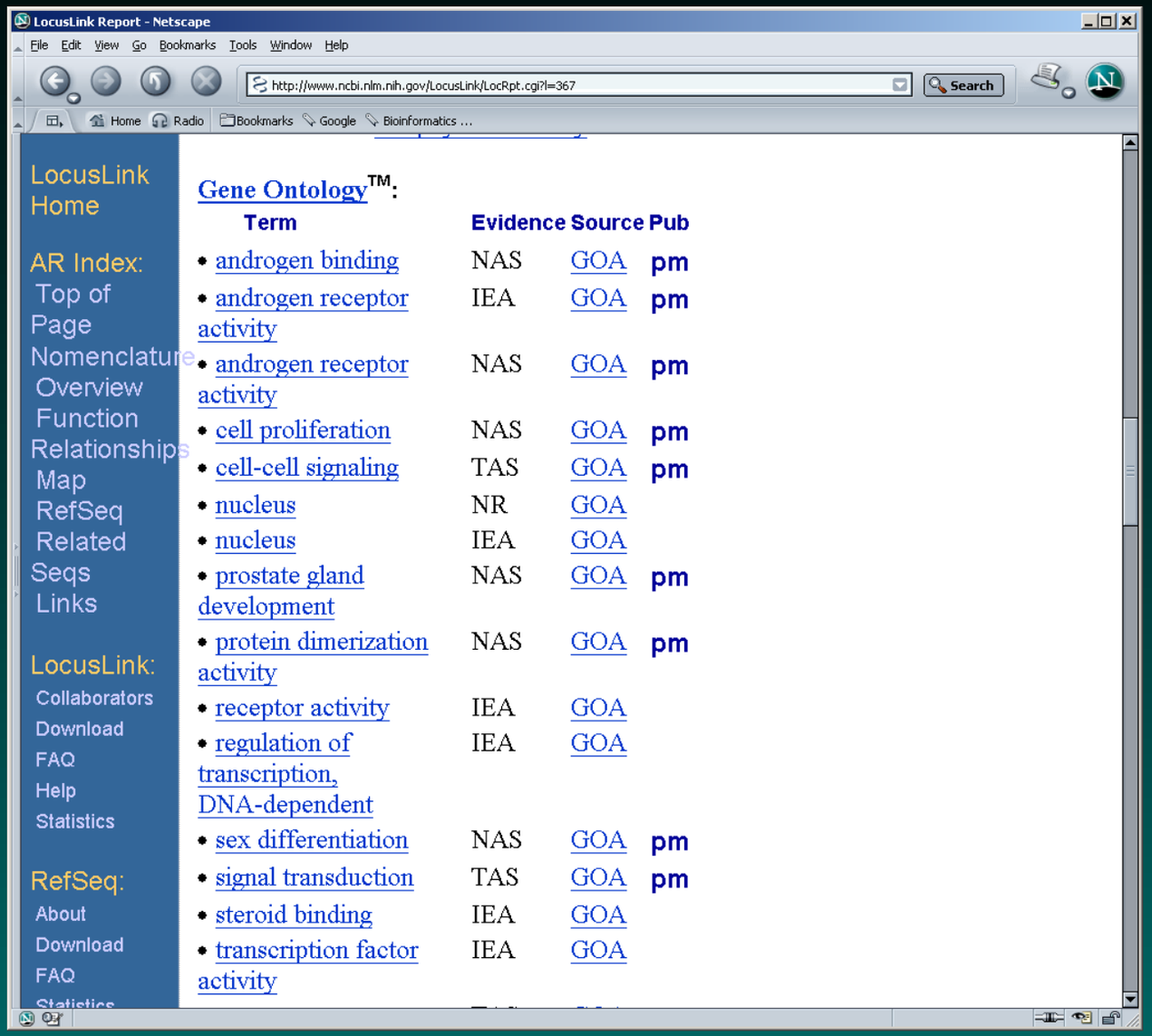
Every GO annotation asserts that a specific gene has a specific function. As part of the design of GO, each assertion is itself annotated to explain the kinds of evidence the assertion is based on, as well as the organization or individual that supplied the annotation.

GO annotations of the androgen receptor

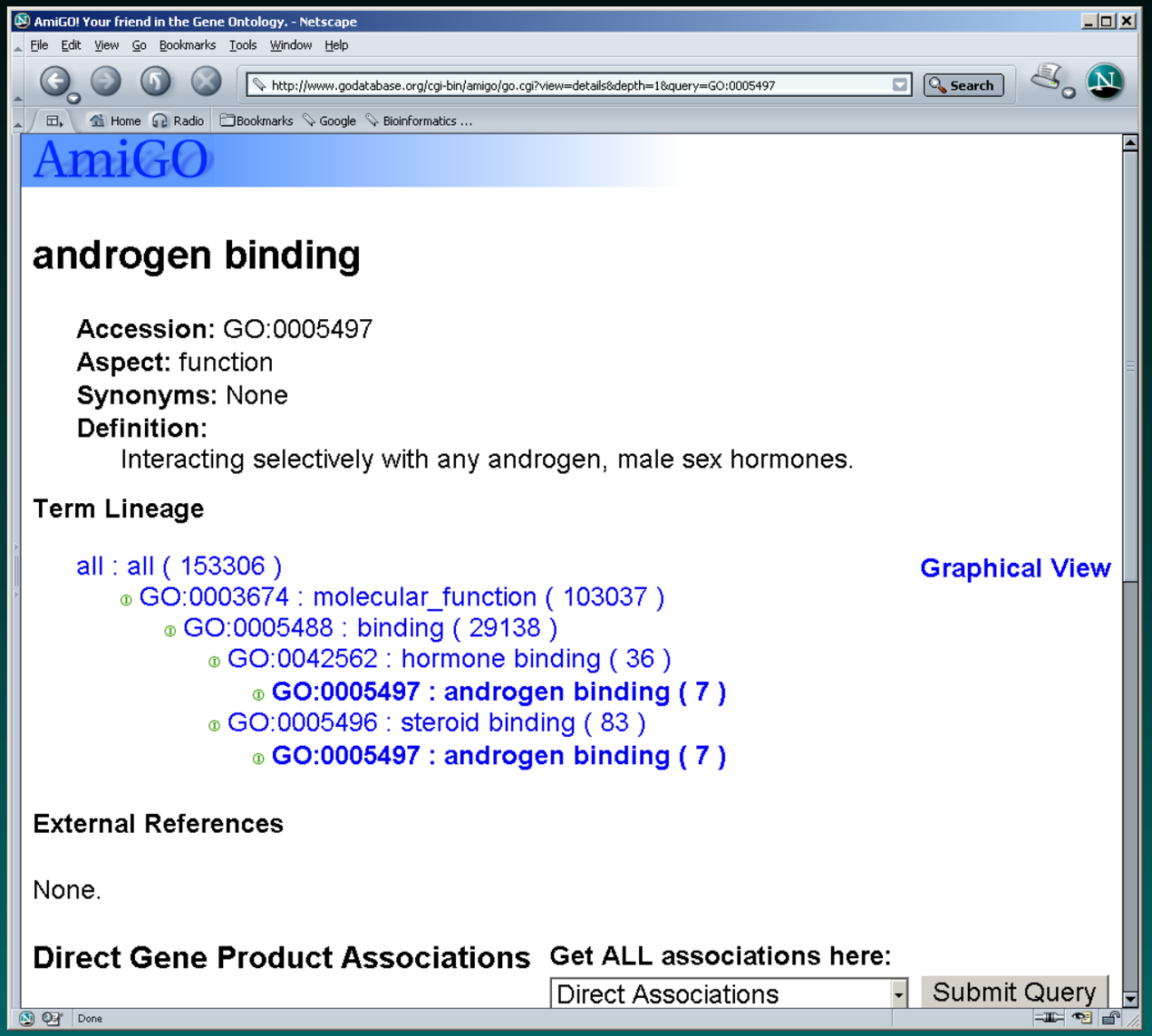


http://www.ebi.ac.uk/GOA/

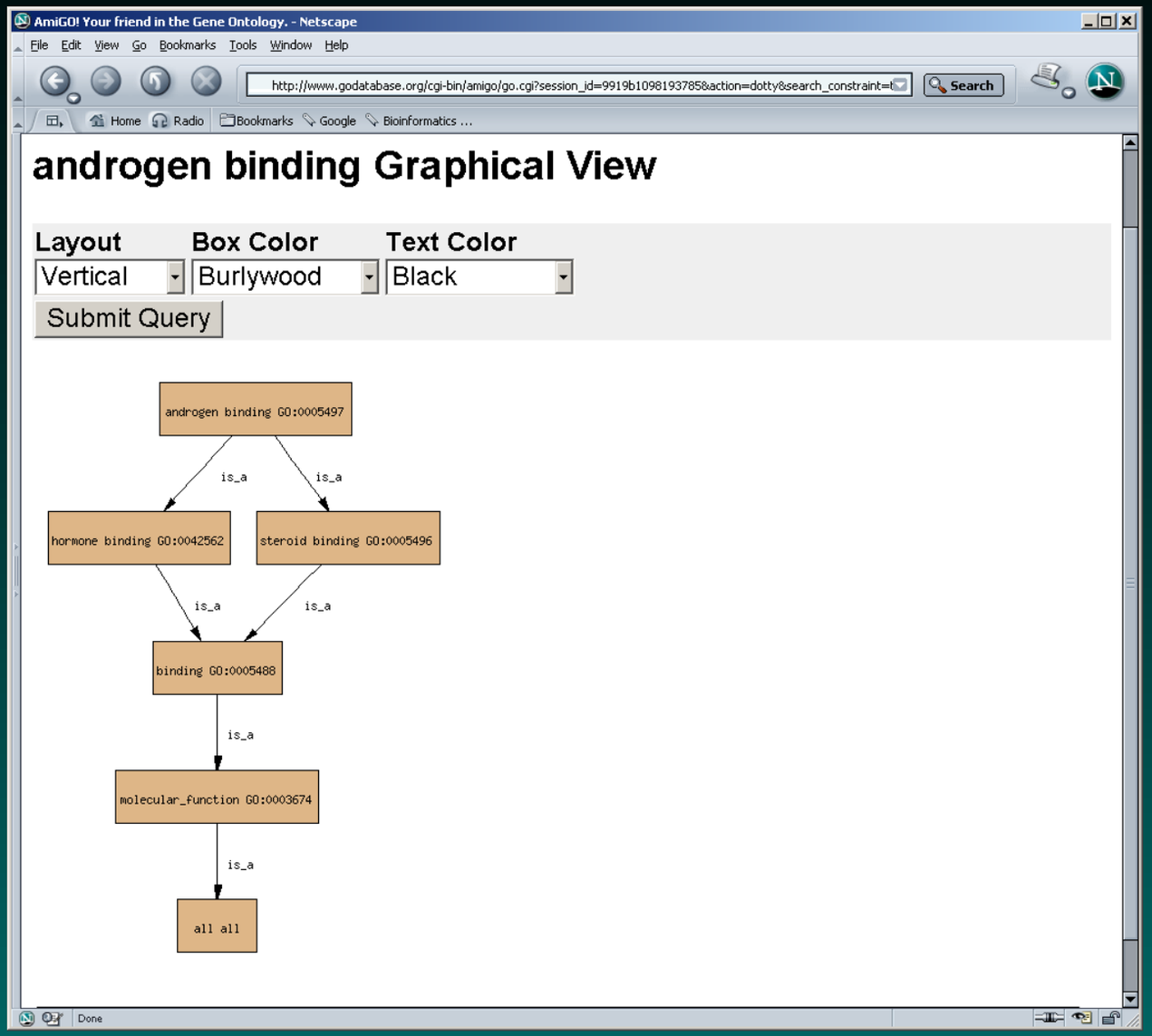
GO annotations of the androgen receptor



GO browsing



GO browsing



Edges are relationships

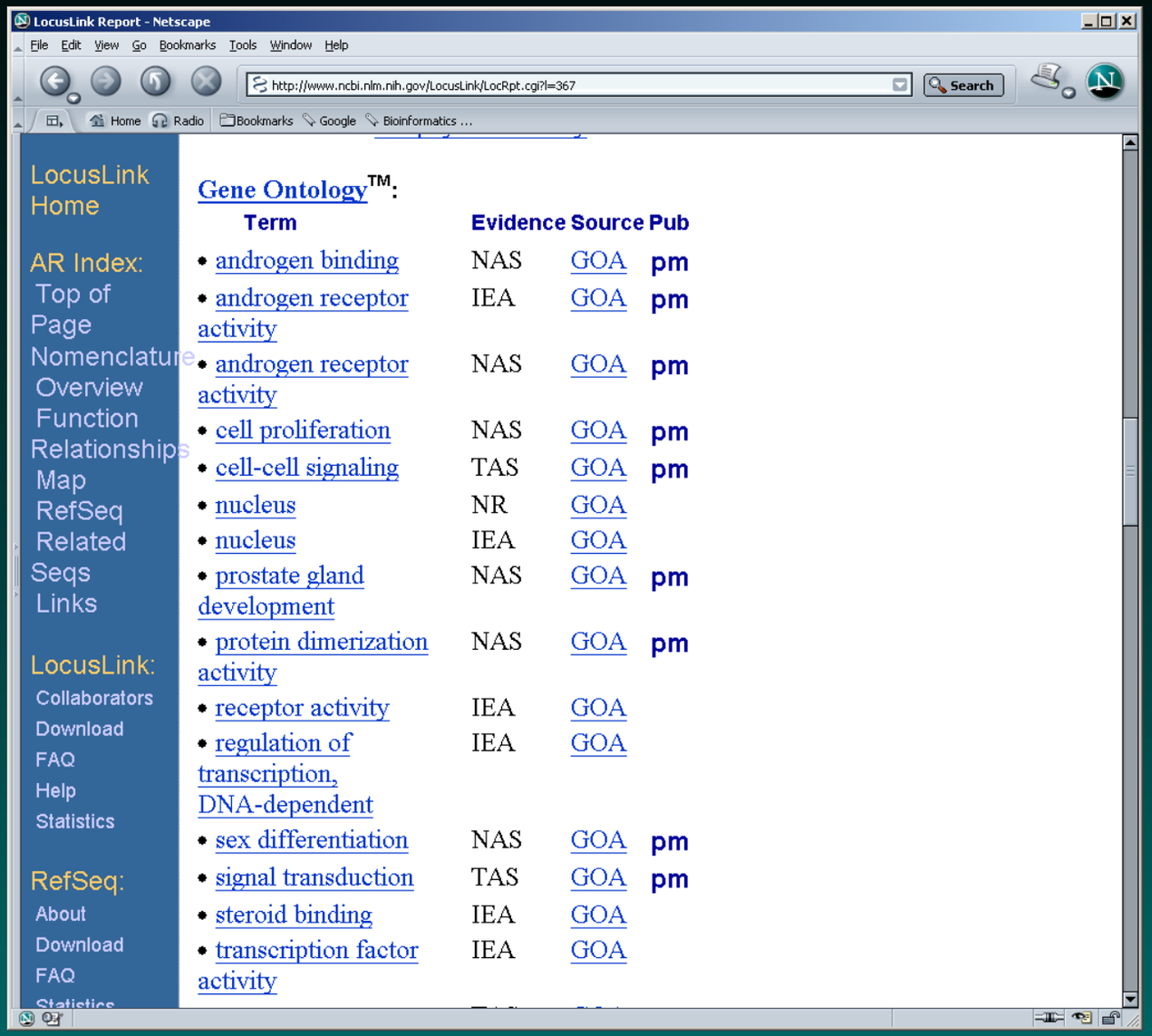
Edges in the DAG represent two kinds of relationships:

is_a : Used when the child node is a special case of the parent node. For example, `hormone binding` is_a kind of `binding`.

part_of : Used when the child node is a component of the parent node. For example, a `membrane` is part_of a `cell`

Genes may be annotated into different levels of the hierarchy, depending on how detailed the evidence is. In general, a gene not only has the function corresponding to the node with direct annotation, but also has every property at parent nodes up through the hierarchy.

GO annotations of the androgen receptor



GeneOntology: Evidence Codes

IDA : inferred from direct assay; indicates that the annotation is based on a paper describing an experiment that directly tested this function for this gene

TAS : traceable author statement; based on a review article or textbook that includes references to the original experiments

IMP : inferred from mutant phenotype; based on experiments involving mutations, knockouts, antisense, etc.

IPI : inferred from physical interation; based on assays (like co-immunoprecipitation) that demonstrate physical interactions between the gene in question and other gene products

IGI : inferred from genetic interaction; based on experiments (such as synthetic lethals, suppressors, functional complementation) that show a genetic interaction between the gene in question and another gene

ISS : inferred from sequence or structure similarity; based on BLAST results that have been reviewed for accuracy by a curator

IEP : inferred from expression pattern; based on Northerns, Westerns, or microarray experiments that reveal information about the timing or location of expression

NAS : non-traceable author statement; statements in papers (abstract, introduction, discussion) that a curator cannot trace to another publication

IEA : inferred from electronic annotation; based on sequence similarity searches or database records that have not been reviewed by a curator

IC : inferred by curator; even though no direct evidence is available, the property can reasonably be inferred by the curator. For example, it is reasonable to infer from direct evidence of “transcription factor activity” that the gene product is found in the nucleus

ND : no biological data available; only used for annotations to “unknown”

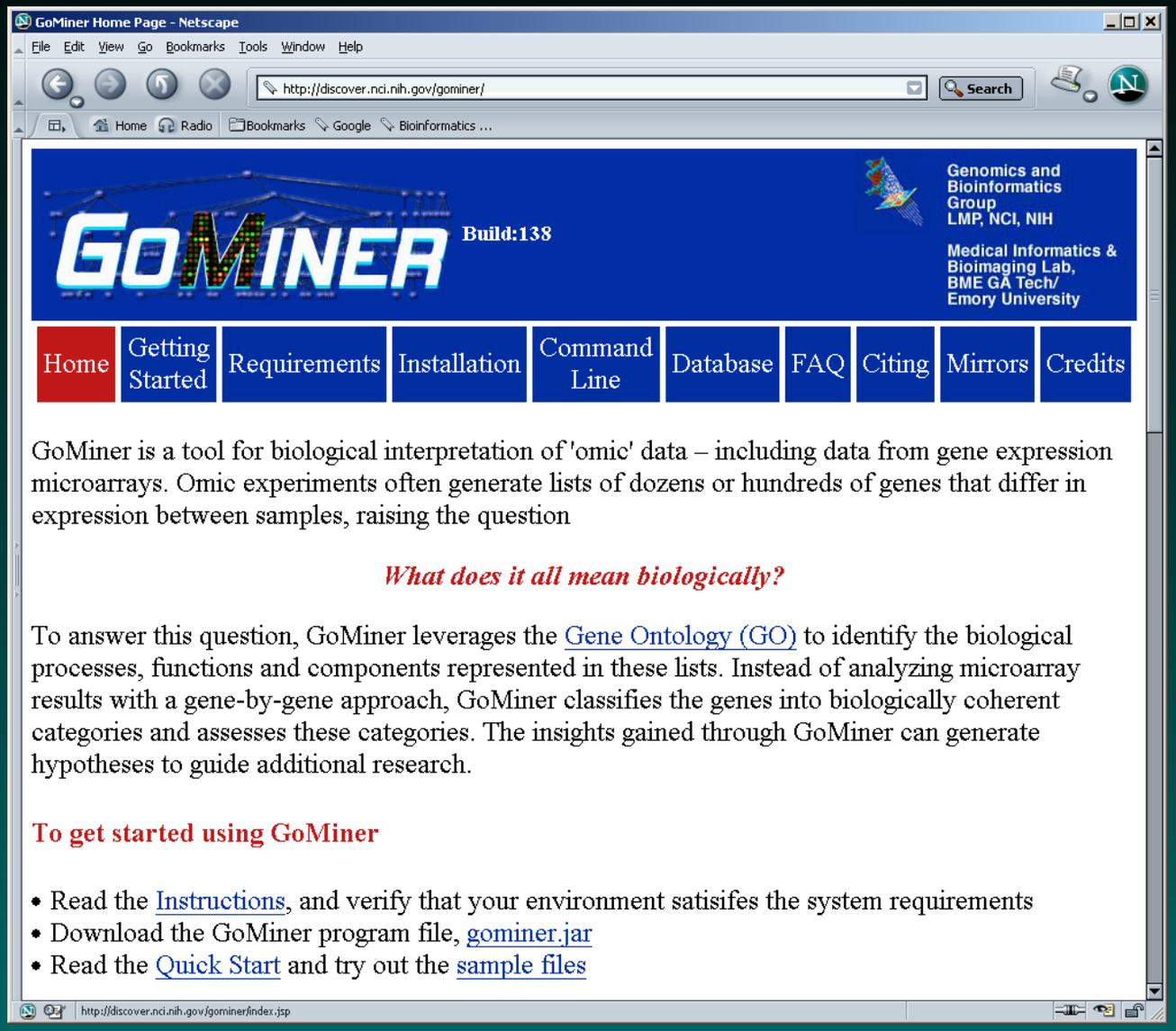
NR : not recorded; used only for annotations created before curators started adding evidence codes

Quality of evidence

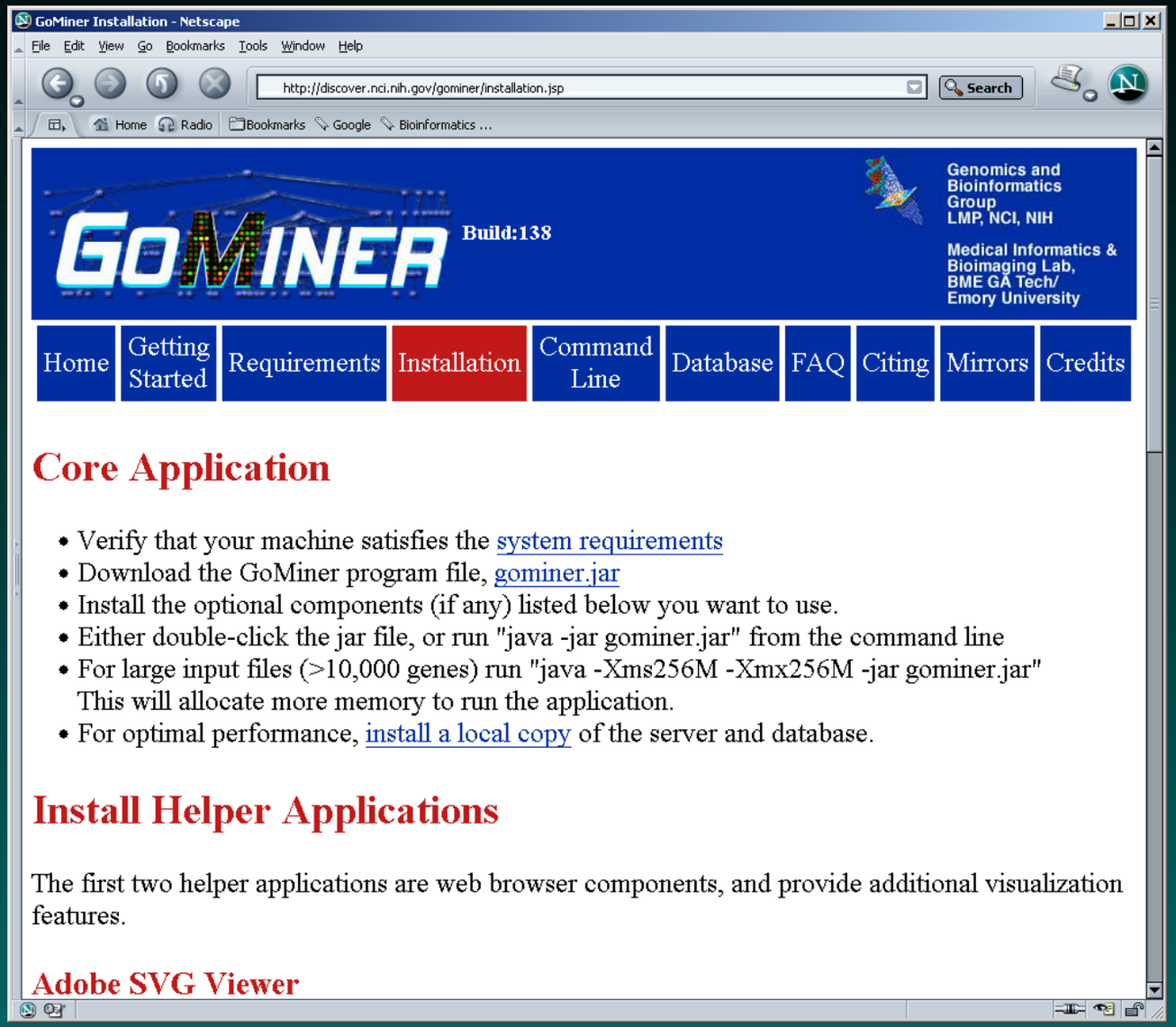
The evidence codes fall into a rough hierarchy indicating how strongly the annotation of function should be believed.

1. IDA, TAS
2. IMP, IPI, IGI
3. ISS, IEP
4. NAS
5. IEA
6. IC

GoMiner



http://discover.nci.nih.gov/gominer



GoMiner: Getting Started

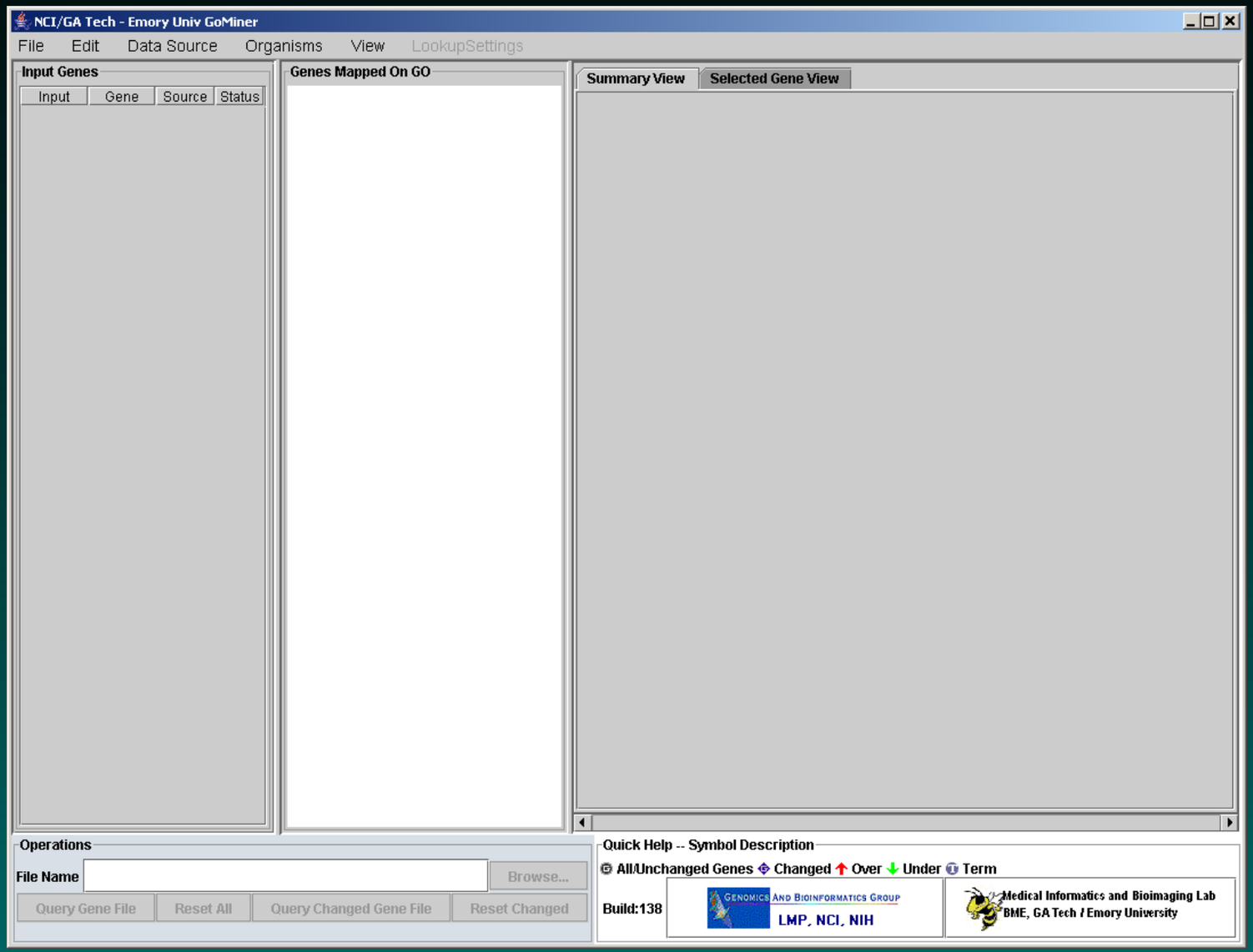
You need a machine with

- Java 1.3 or higher
- Windows 98 or higher, Mac OS X or higher, Solaris, Linux, or FreeBSD
- High-speed ointernet access

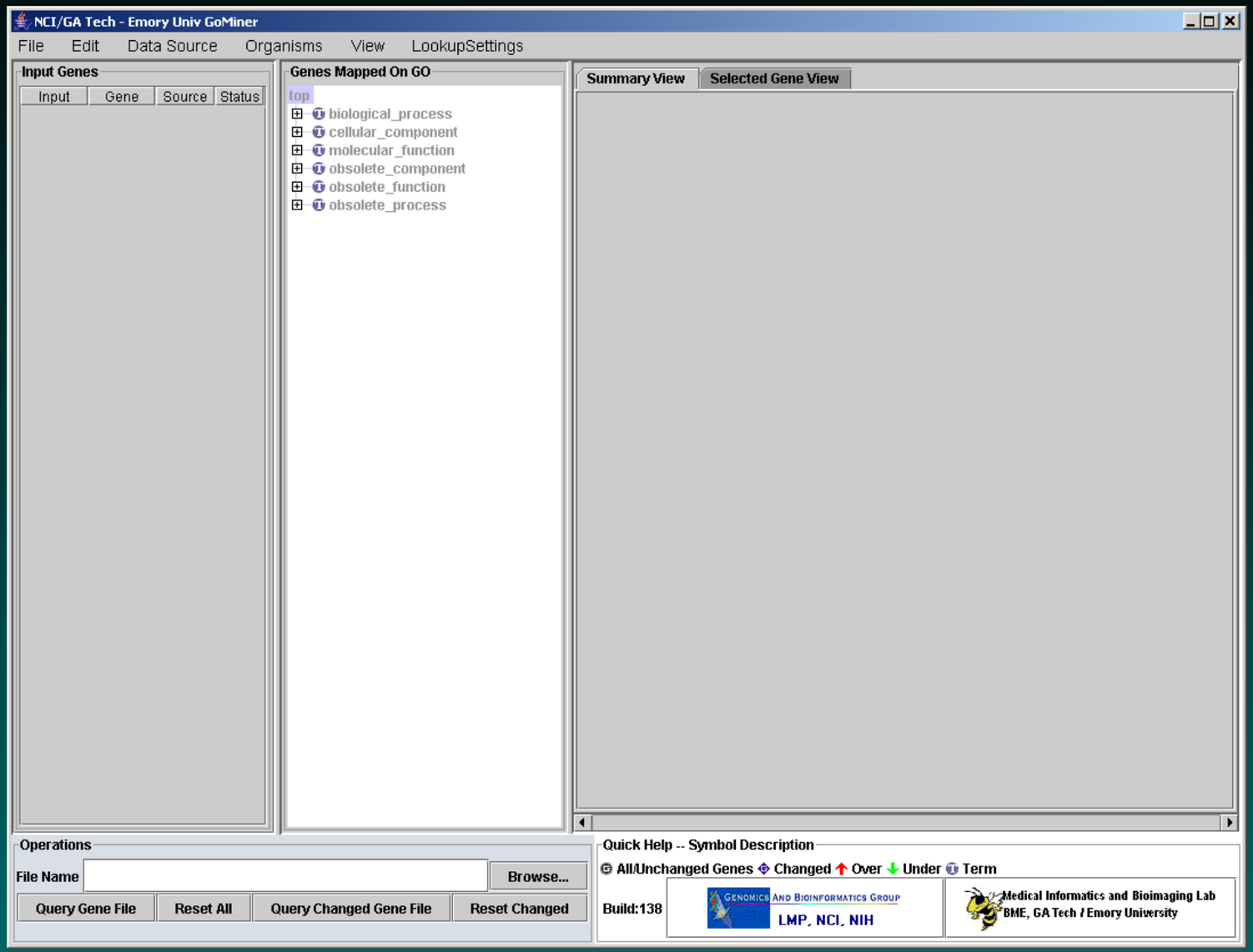
Download the GoMiner Java code, install it, and double-click on it to start the program.

Then go to "File" $->$ "Load GO Terms" and click "OK".

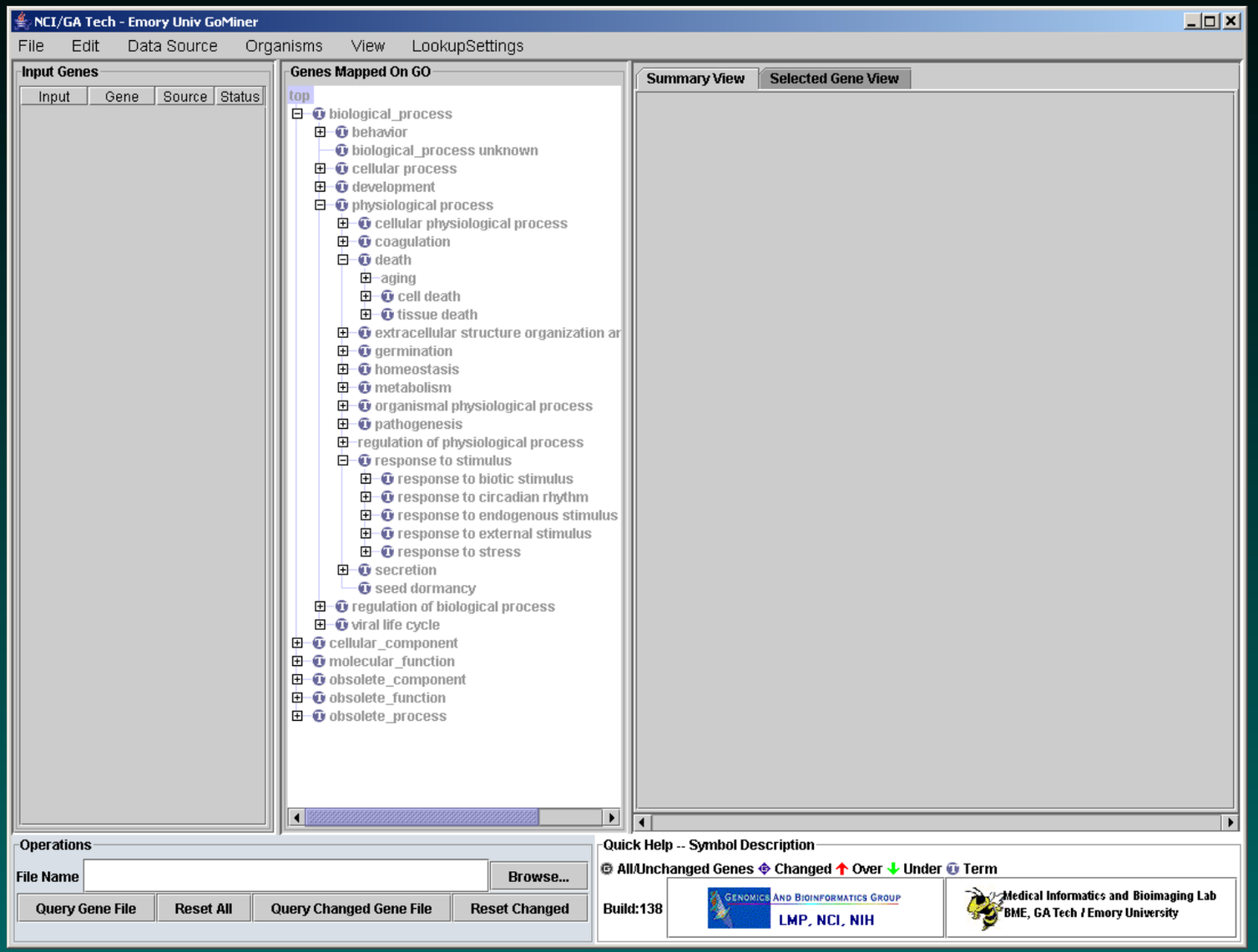
GoMiner Start



GoMiner: GO terms loaded



GoMiner as GO browser



Getting array data into GoMiner

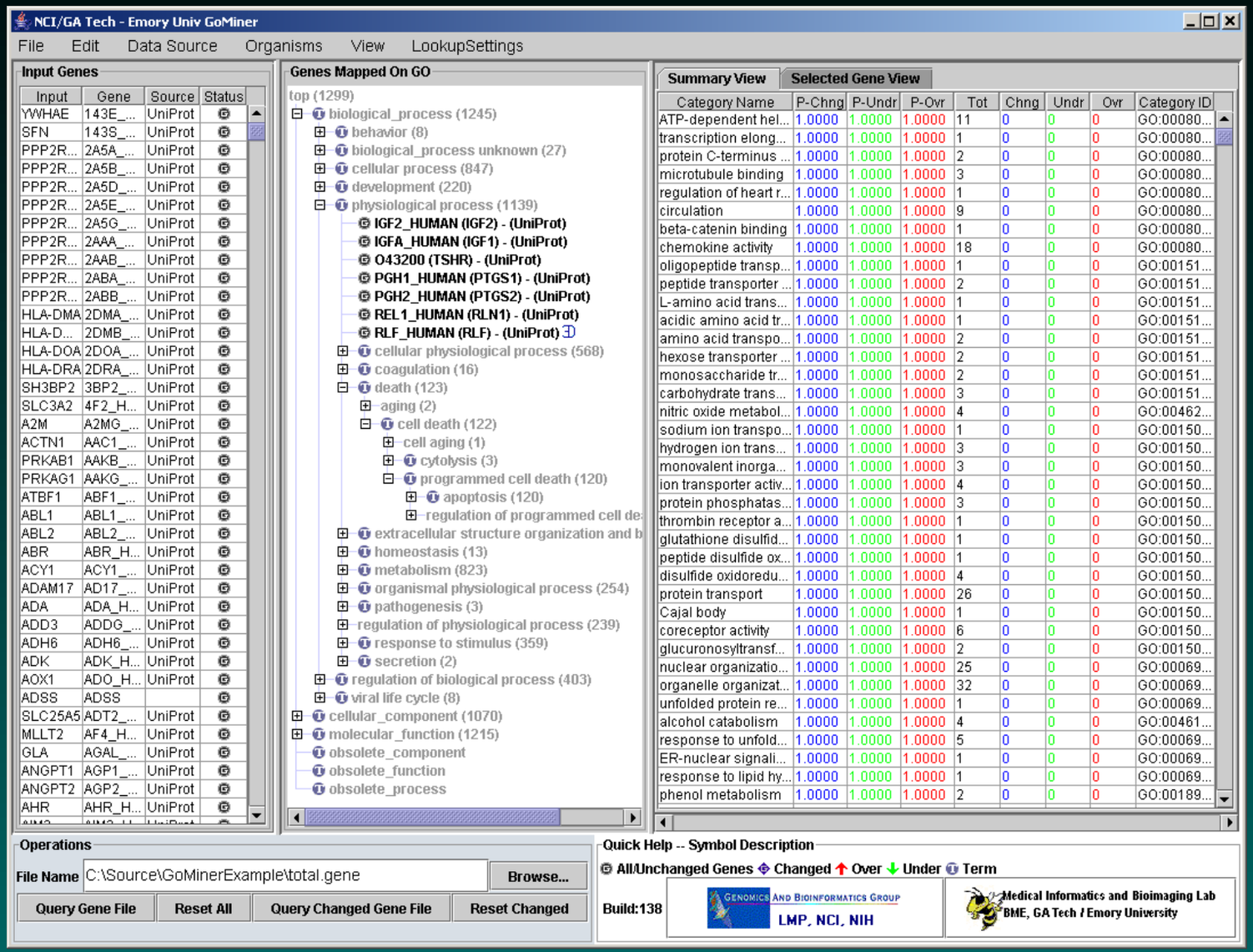
1. Go to “Data Source” and select “UniProt (Hs)” to restrict to human gene annotations

2. Need a file containing a list of all genes in the experiment, one HUGO symbol per line. Use the “Browse” button, and then click “Query Gene File” to load this information. This may take some time...

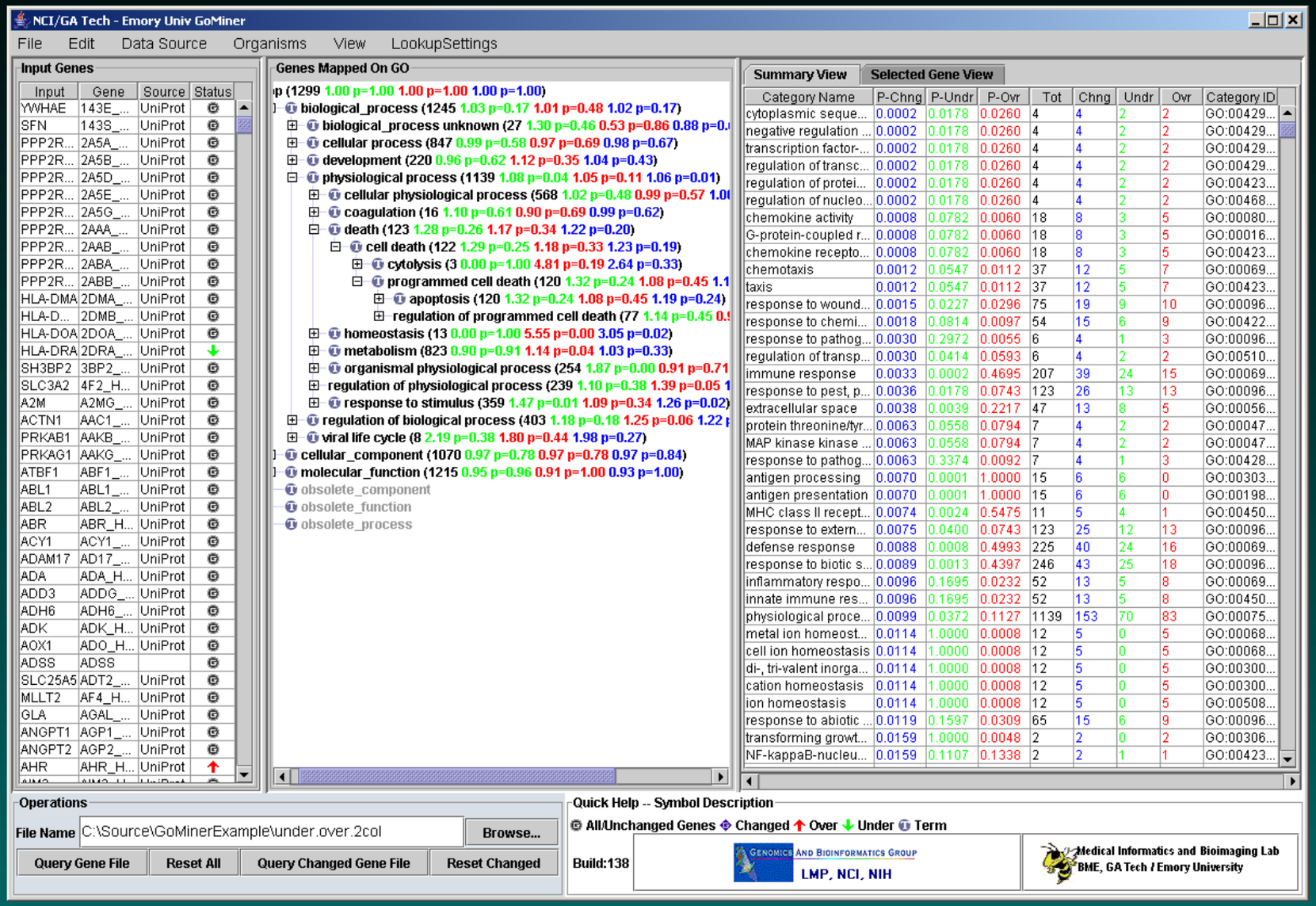
3. Need a file containing a list of genes that changed. Can be one HUGO symbol per line. Optionally, you can include a second column with 1 (overexpressed) or -1 (underexpressed). Use “Browse” and “Query Changed Gene File” to load this data.

Note: GeneLink or Source can convert from various gene ids to HUGO symbols.

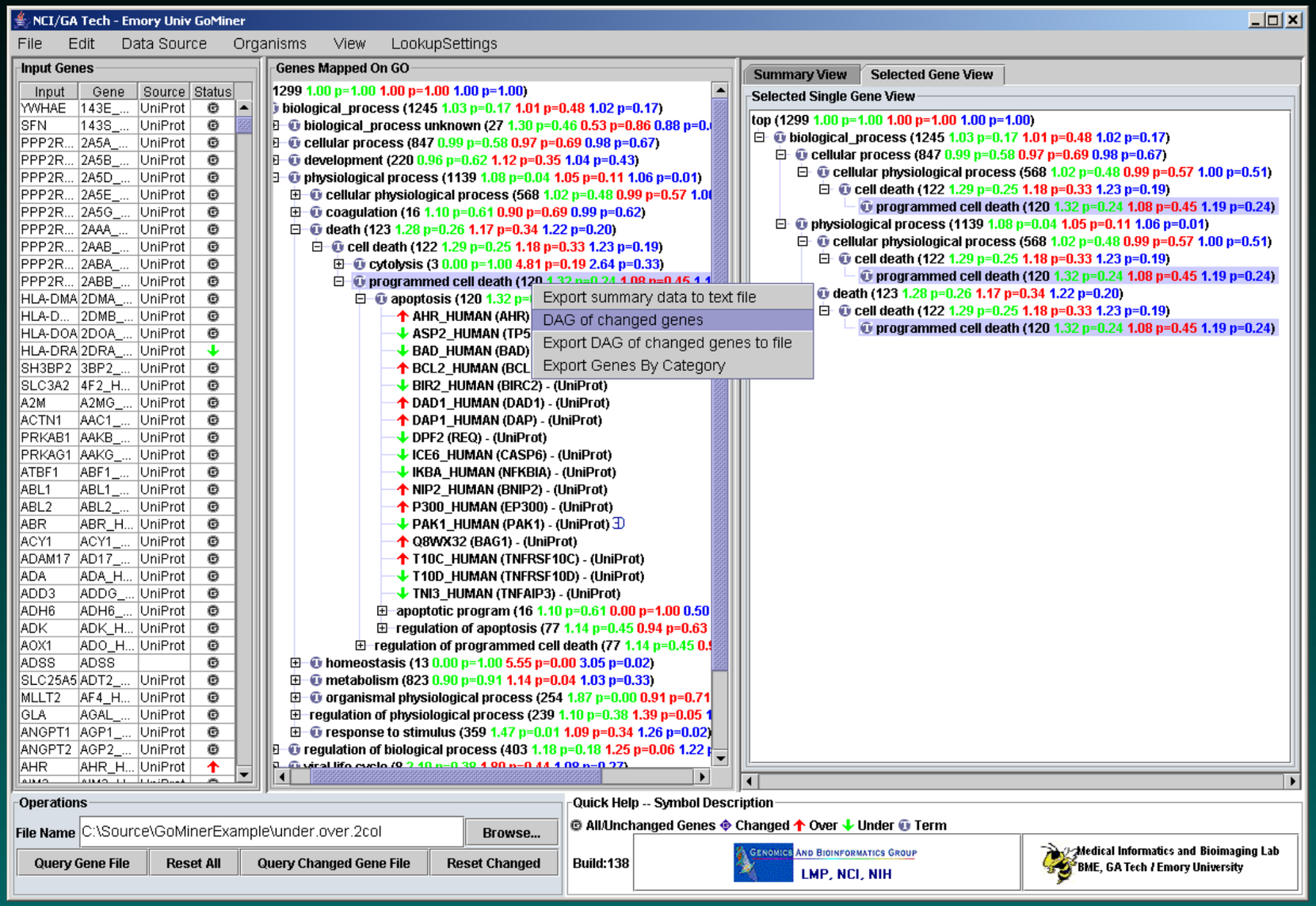
GoMiner with array gene list loaded



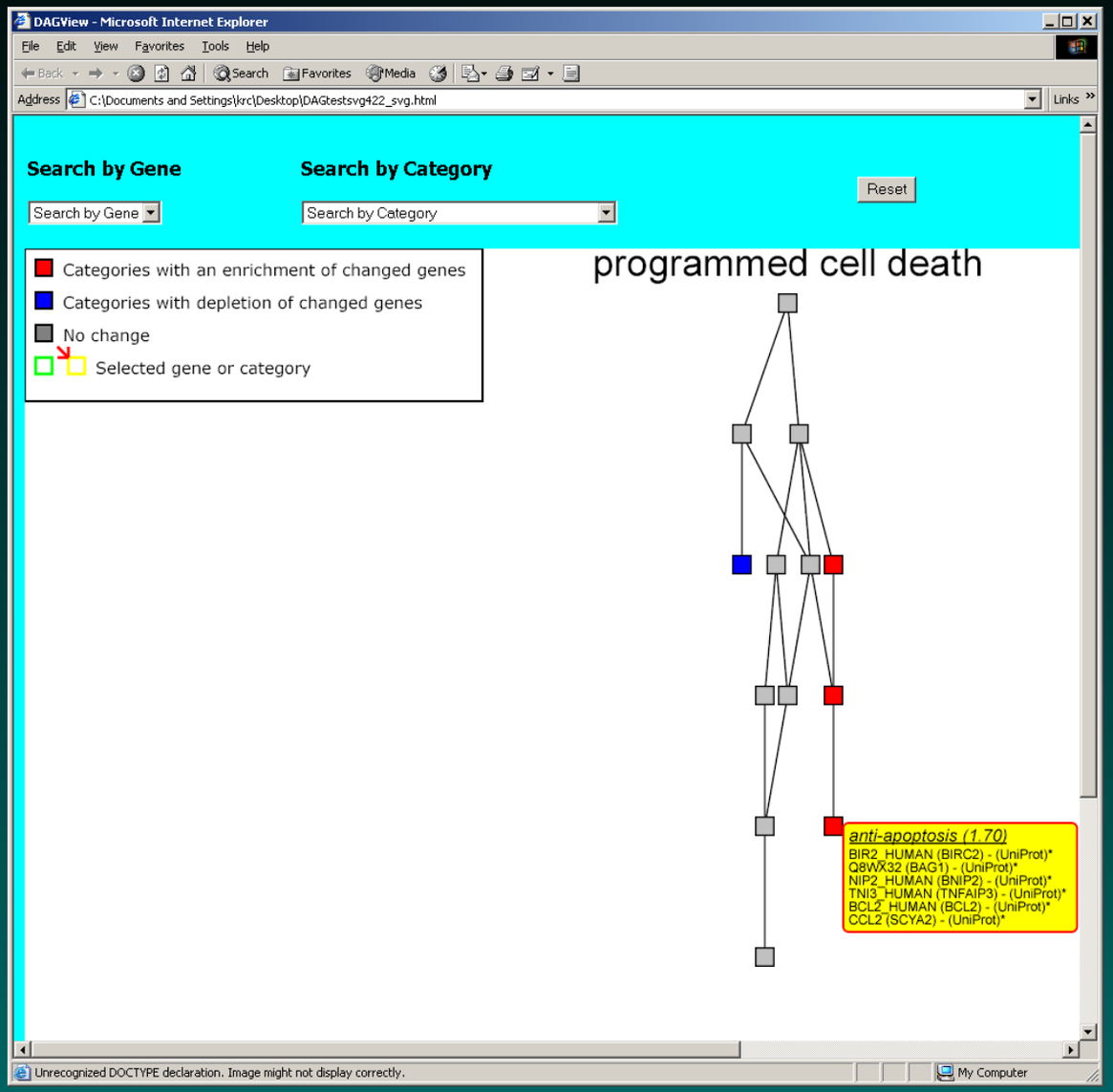
GoMiner with changed gene list loaded



GoMiner subgraphs



GoMiner subgraphs



Intepreting GoMiner results

Enrichment is computed as

$$\frac{\textit{changed genes in category} \,/\, \textit{total genes in category}}{\textit{changed genes on array} \,/\, \textit{all genes on array}}$$

Statistical evidence of enrichment is based on a Fisher exact test.

Example

With the sample files supplied from GoMiner, they have an array with 1399 genes, of which 177 change expression. A total of 37 genes on the array are annotated to the biological process category of `chemotaxis`, and 12 of those 37 genes changed expression. So, we have a 2×2 contingency table that looks like

chemotaxis:	yes	no	all
Changed	12	165	177
Unchanged	25	1197	1222
Total	37	1362	1399

The p-value arising from a Fisher exact test with these values is 0.001226, which is the value reported by GoMiner for the enrichment of the chemotaxis category.

Intepreting GoMiner results

The p-values from the Fisher test are not corrected for multiple testing, but they should be since one is potentially looking at all GO categories. The categories are not independent, so it is not clear how to correct for multiple testing.

If one filters the gene list from the array before testing differential expression (for example, by removing low expressing or low variance genes), should those genes be included in the “query gene file” for the experiment?

The Fisher exact test is not completely appropriate, since genes can have multiple overlapping annoptations into the GO DAG.

No existing test exploits the quality of evidence for the GO annotations.

Regulatory networks and metabolic pathways

The harder question to be asked about the results of a microarray experiment is: what pathways or networks are changed?

Distinction:

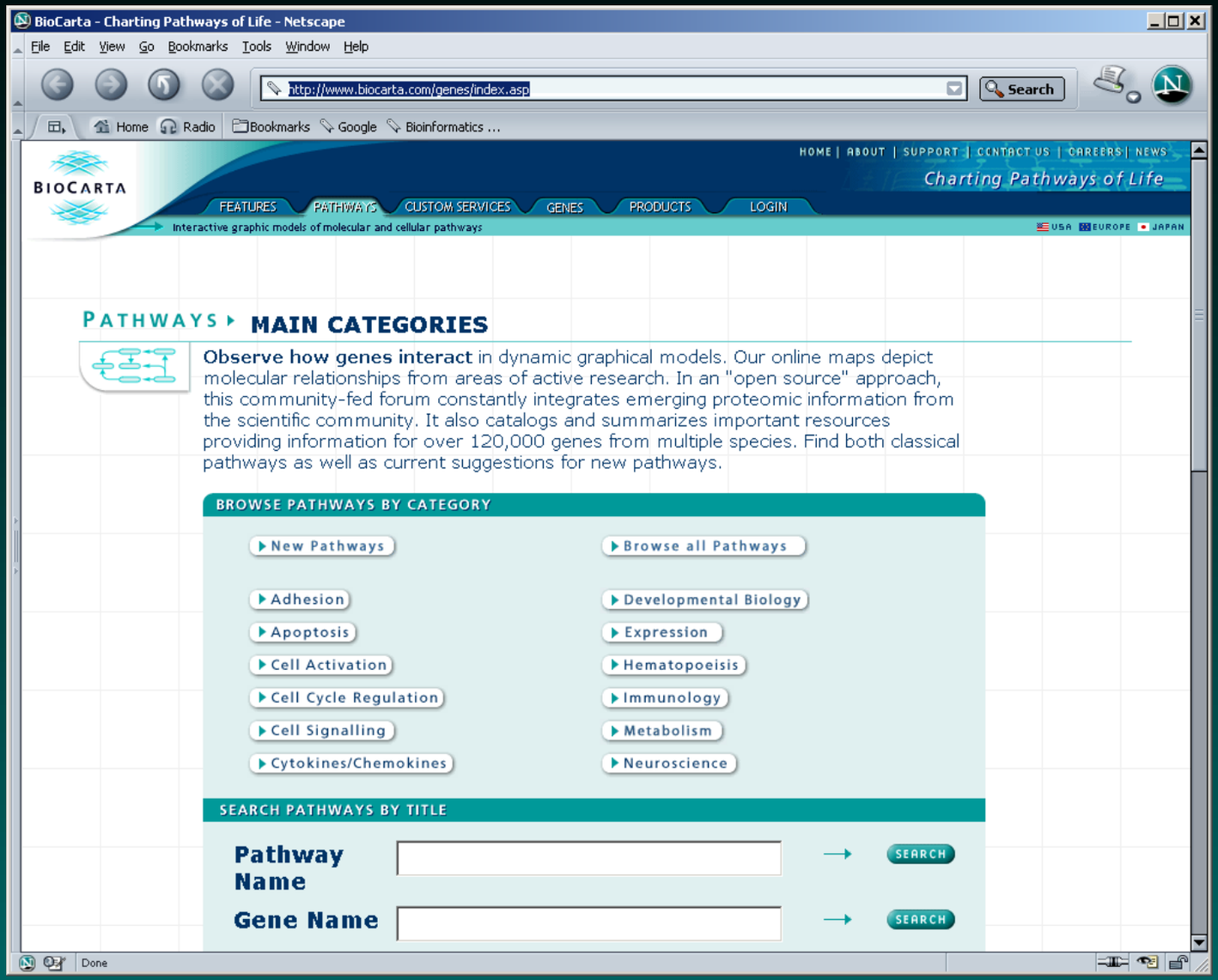
- Pathways typically refer to metabolic pathways (that process various small molecules) or signalling pathways that carry information from outside the cellinto the nucleus.
- Networks typically refer to interactons that control the expression level of various gene products.

Pathways are characterized by direct physical interactions betwen proteins. Networks are characterized by indirect interactions between genetic DNA or RNA or protein.

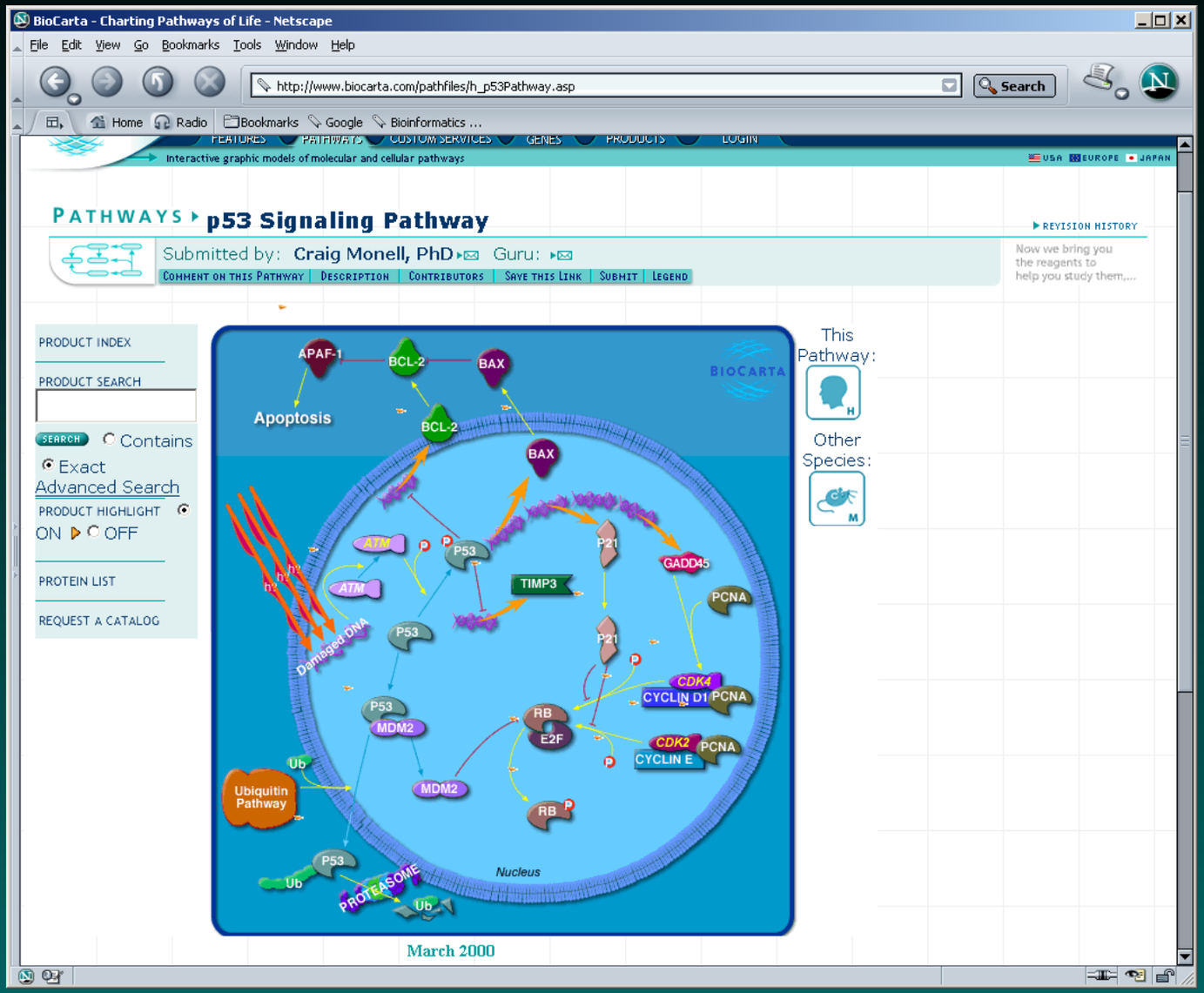
Pathway and netowrk resources

- KEGG (Kyoto Encyclopedia of Genes and Genomes) is a publically available tool, probably best used for metabolic pathways and secondarily for signalling pathways.
- PathArt (Jubilant Biosystems) is a commercial tool for investigating networks and pathways. Underlying network is based on curated interactions from the literature.
- Ingenuity Pathway Analysis (Ingenuity) is another curated commercial tool for investigating networks and pathways.

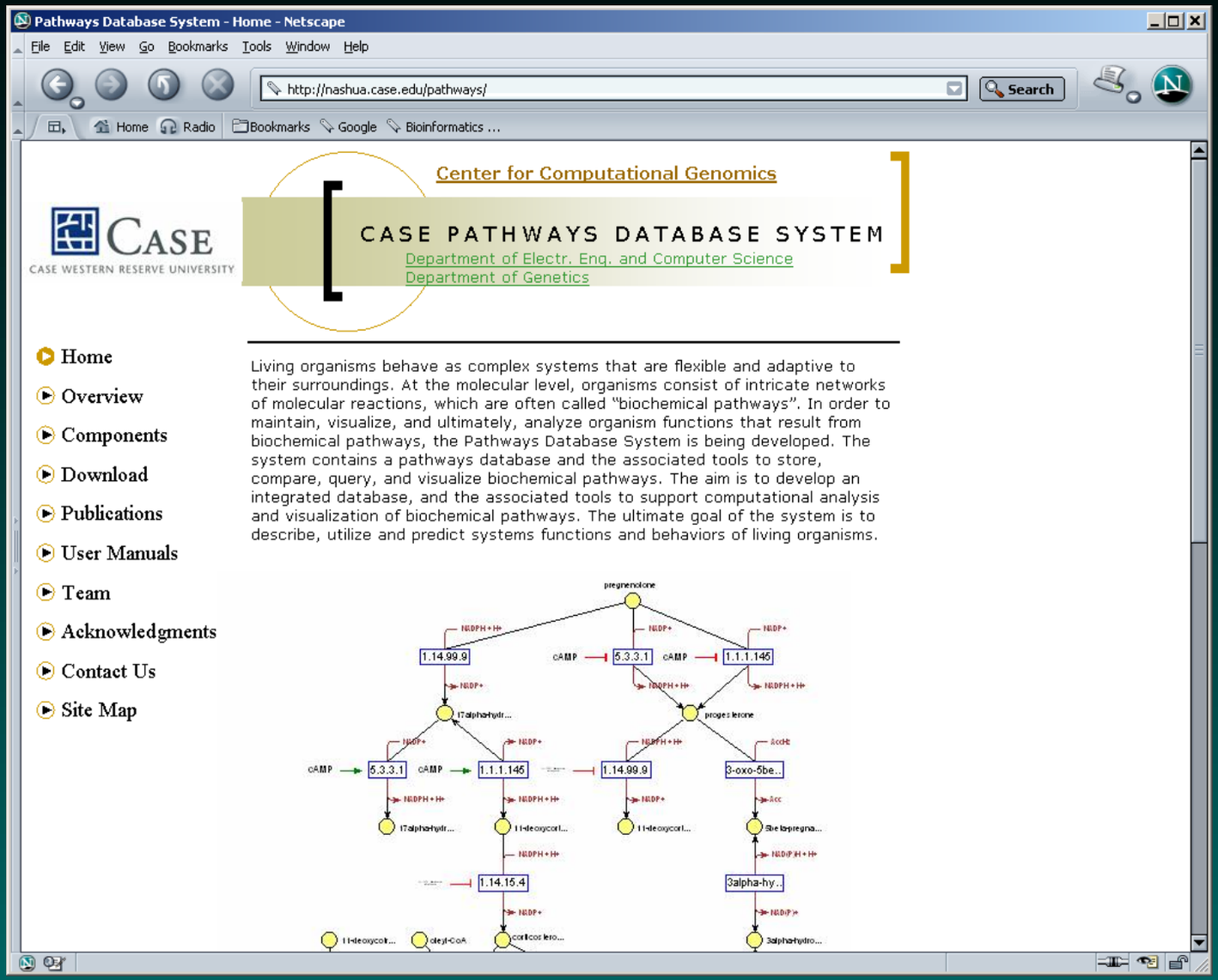
http://www.biocarta.com/genes/index.asp



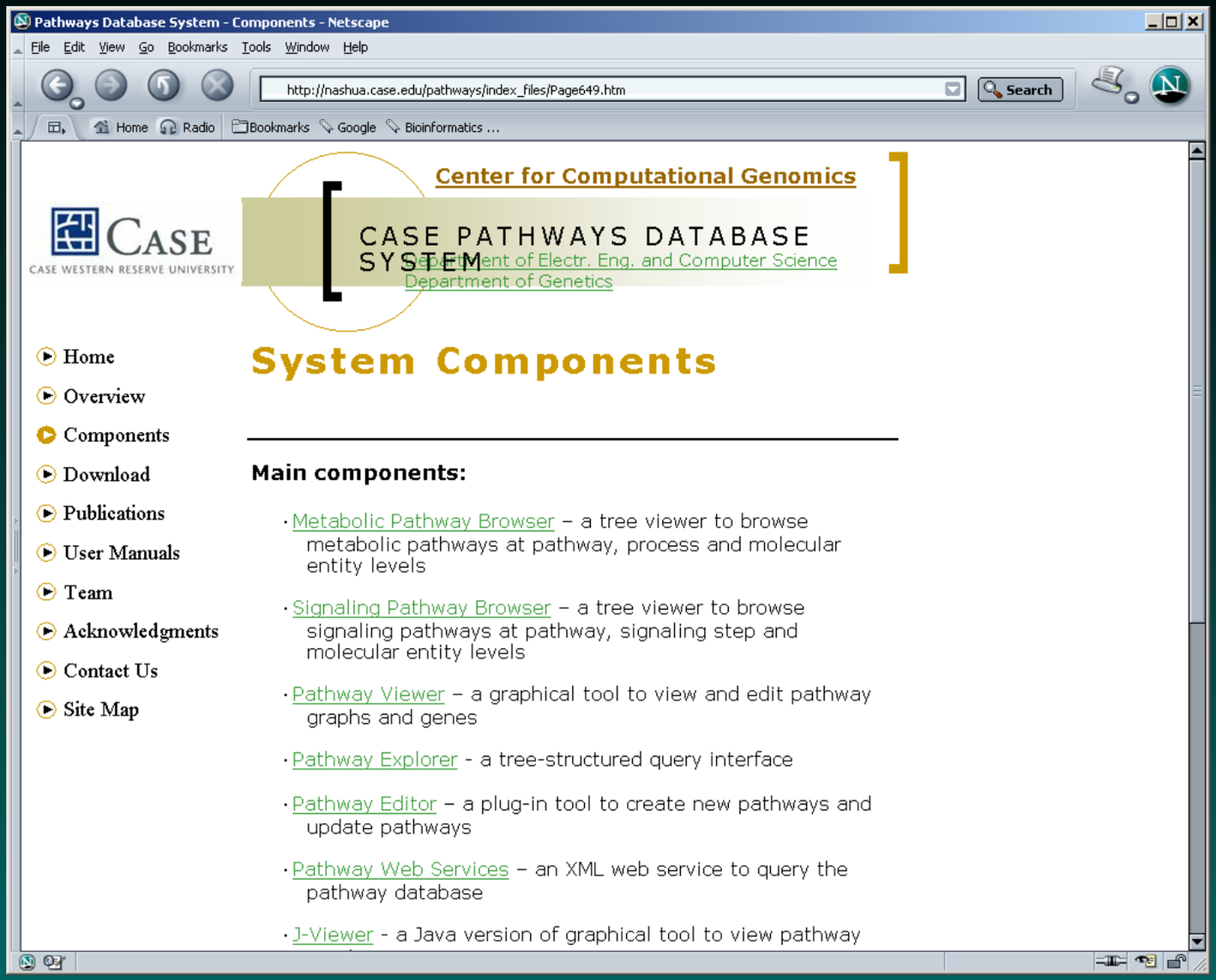
You can't compute on pictures...



http://nashua.case.edu/pathways



Case-Western Pathway Browser



Jubilant PathArt