Microarray annotation and biological information

Benedikt Brors

Dept. Intelligent Bioinformatics Systems

German Cancer Research Center

b.brors@dkfz.de

Why do we need microarray clone annotation?

- Often, the result of microarray data analysis is a list of genes.

- The list has to be summarized with respect to its biological meaning. For this, information about the genes and the related proteins has to be gathered.

- If the list is small (let’s say, 1–30), this is easily done by reading database information and/or the available literature.

- Sometimes, lists are longer (100s or even 1000s of genes). Automatic parsing and extracting of information is needed.

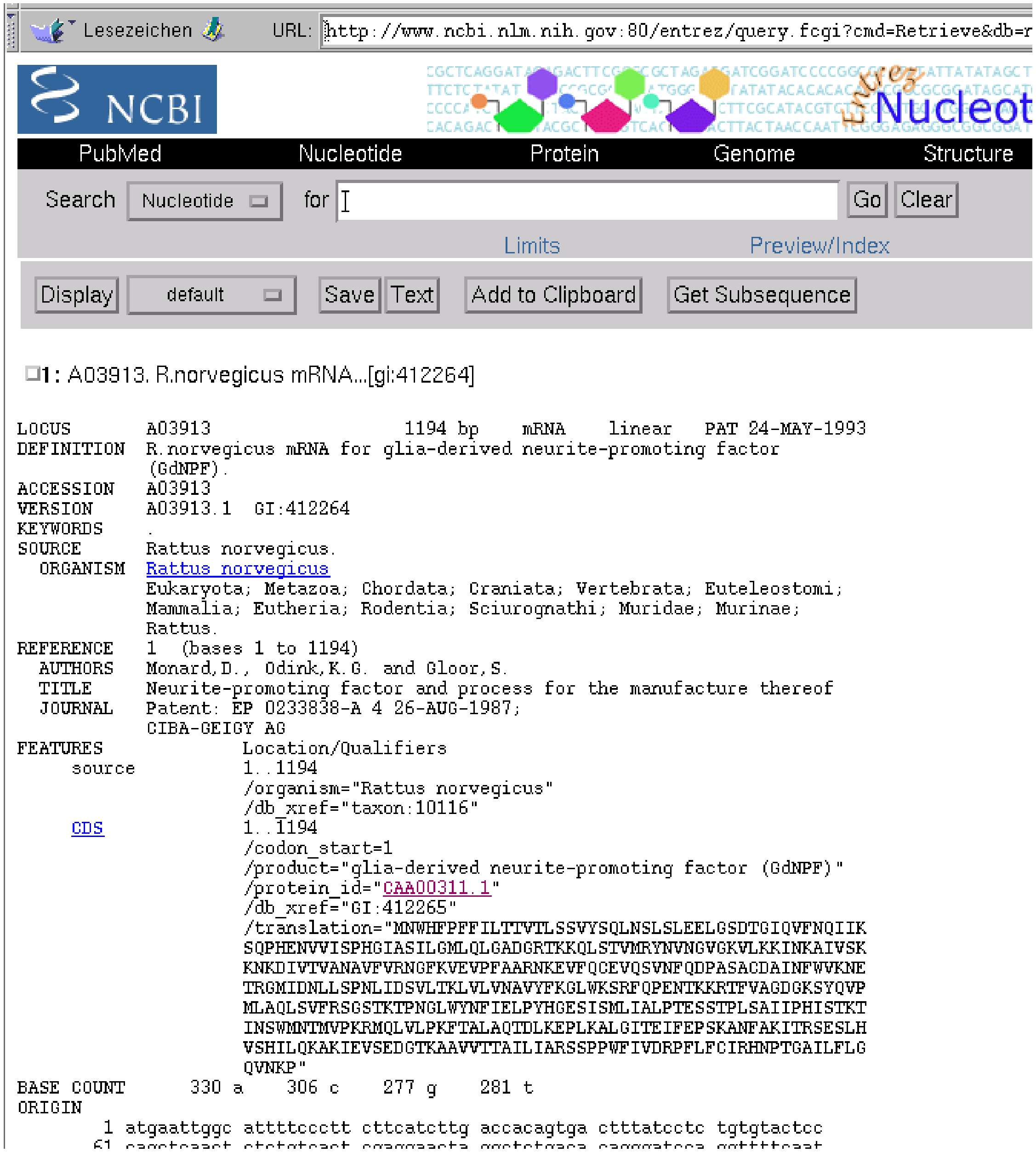
- To get complete information, you will need the help of an experienced computational biologist (aka ‘bioinformatician’). However, there is a lot that you can do on your own.

Primary databases

- Some information about genes and the encoded proteins is available already from sequence databases, e.g. database accession number, nucleotide and protein sequences, database cross references, and a sequence name that may or may not give a hint to the function. To find a sequence in another database, use sequence comparison tools like BLAST.

- There are large repositories for sequence data, the most prominent being EMBL, GenBank and DDBJ (these 3 are redundant). Because they are so large, nobody cares about the quality of the data. Everybody having internet access can deposit sequence information there. Errors introduced long time ago will stay there forever.

GenBank information from NCBI



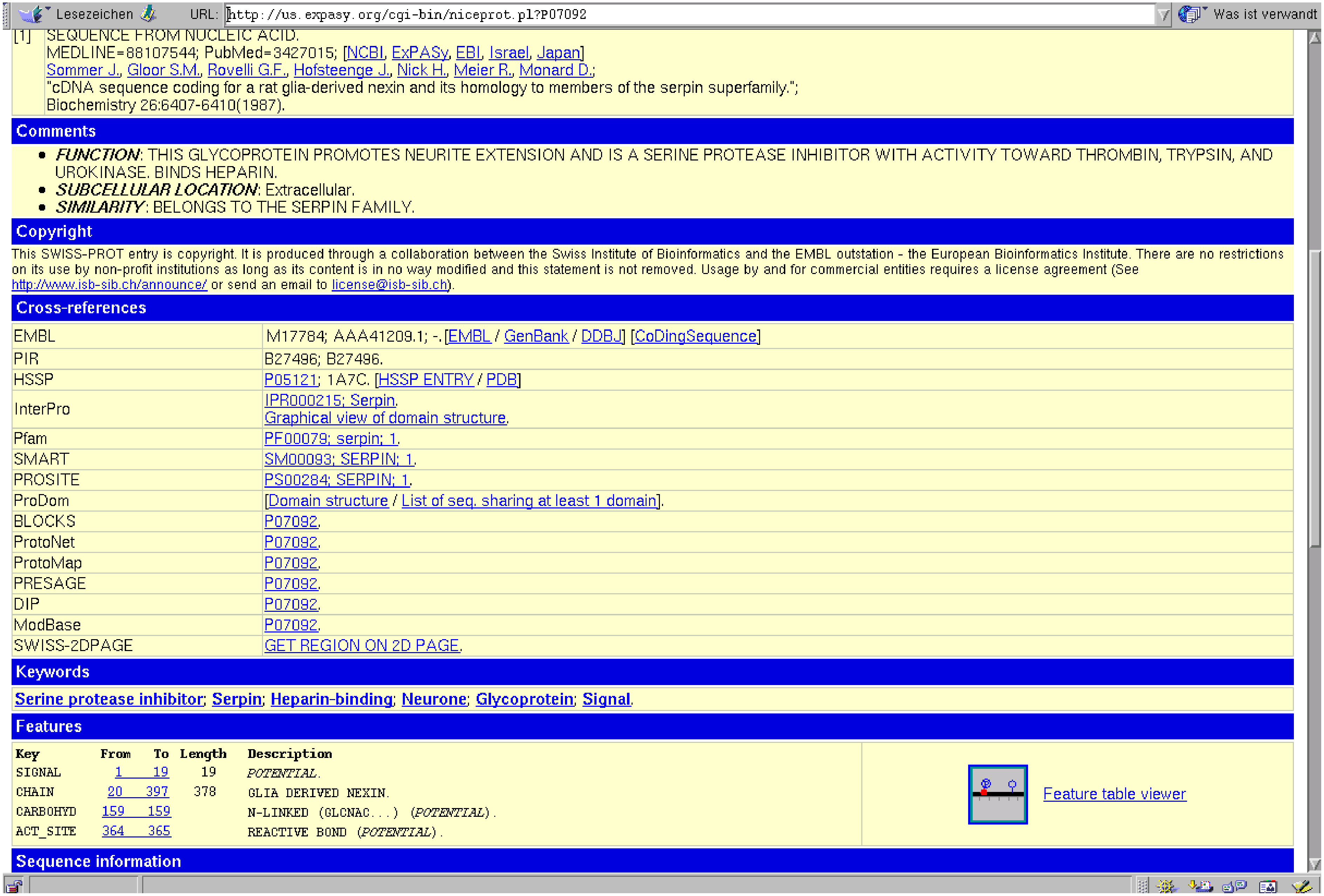
□1: A03913. R.norvegicus mRNA...[gi:412264]

```
LOCUS       A03913                  1194 bp    mRNA    linear   PAT 24-MAY-1993
DEFINITION  R.norvegicus mRNA for glia-derived neurite-promoting factor
            (GdNPF).
ACCESSION   A03913
VERSION     A03913.1  GI:412264
KEYWORDS    .
SOURCE      Rattus norvegicus.
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae;
            Rattus.
REFERENCE   1  (bases 1 to 1194)
  AUTHORS   Monard,D., Odink,K.G. and Gloor,S.
  TITLE     Neurite-promoting factor and process for the manufacture thereof
  JOURNAL   Patent: EP 0233838-A 4 26-AUG-1987;
            CIBA-GEIGY AG
FEATURES             Location/Qualifiers
     source          1..1194
                     /organism="Rattus norvegicus"
                     /db_xref="taxon:10116"
     CDS             1..1194
                     /codon_start=1
                     /product="glia-derived neurite-promoting factor (GdNPF)"
                     /protein_id="CAA00311.1"
                     /db_xref="GI:412265"
                     /translation="MNWHFPFFILTTVTLSSVYSQLNSLSLEELGSDTGIQVFNQIIK
                     SQPHENVVISPHGIASILGMLQLGADGRTKKQLSTVMRYNVNGVGKVLKKINKAIVSK
                     KNKDIVTVANAVFVRNGFKVEVPFAARNKEVFQCEVQSVNFQDPASACDAINFWVKNE
                     TRGMIDNLLSPNLIDSVLTKLVLVNAVYFKGLWKSRFQPENTKKRTFVAGDGKSYQVP
                     MLAQLSVFRSGSTKTPNGLWYNFIELPYHGESISMLIALPTESSTPLSAIIPHISTKT
                     INSWMNTMVPKRMQLVLPKFTALAQTDLKEPLKALGITEIFEPSKANFAKITRSESLH
                     VSHILQKAKIEVSEDGTKAAVVTTAILIARSSPPWFIVDRPFLFCIRHNPTGAILFLG
                     QVNKP"
BASE COUNT      330 a    306 c    277 g    281 t
ORIGIN
        1 atgaattggc attttccctt cttcatcttg accacagtga ctttatcctc tgtgtactcc
```

Curated databases

- In contrast, some databases are *curated*. That means that biologists will get the information first and compare them with literature before it goes into the database. Thus, the database is of high quality, but it takes some time until a newly discovered sequence is entered. Because information is only entered by curators, *annotation* can be unified. Rules can be put in place that say, e.g., that all enzymes cutting off phosphates are called *phosphatases*, not ‘phosphate hydrolases’. A very famous curated database is Amos Bairoch’s SWISSPROT (http://www.expasy.ch).

SwissProt entry



http://us.expasy.org/cgi-bin/niceprot.pl?P07092

[1] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=88107544; PubMed=3427015; [NCBI, ExPASy, EBI, Israel, Japan]
Sommer J., Gloor S.M., Rovelli G.F., Hofsteenge J., Nick H., Meier R., Monard D.;
"cDNA sequence coding for a rat glia-derived nexin and its homology to members of the serpin superfamily.";
Biochemistry 26:6407-6410(1987).

Comments

- ***FUNCTION***: THIS GLYCOPROTEIN PROMOTES NEURITE EXTENSION AND IS A SERINE PROTEASE INHIBITOR WITH ACTIVITY TOWARD THROMBIN, TRYPSIN, AND UROKINASE. BINDS HEPARIN.
- ***SUBCELLULAR LOCATION***: Extracellular.
- ***SIMILARITY***: BELONGS TO THE SERPIN FAMILY.

Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch).

Cross-references

EMBL	M17784; AAA41209.1; -.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	B27496; B27496.
HSSP	P05121; 1A7C. [HSSP ENTRY / PDB]
InterPro	IPR000215; Serpin. Graphical view of domain structure.
Pfam	PF00079; serpin; 1.
SMART	SM00093; SERPIN; 1.
PROSITE	PS00284; SERPIN; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
BLOCKS	P07092.
ProtoNet	P07092.
ProtoMap	P07092.
PRESAGE	P07092.
DIP	P07092.
ModBase	P07092.
SWISS-2DPAGE	GET REGION ON 2D PAGE.

Keywords

Serine protease inhibitor; **Serpin**; **Heparin-binding**; **Neurone**; **Glycoprotein**; **Signal**.

Features

Key	From	To	Length	Description
SIGNAL	1	19	19	*POTENTIAL*.
CHAIN	20	397	378	GLIA DERIVED NEXIN.
CARBOHYD	159	159		N-LINKED (GLCNAC...) (*POTENTIAL*).
ACT_SITE	364	365		REACTIVE BOND (*POTENTIAL*).

Feature table viewer

Sequence information

Other databases

- There are databases that connect sequence information with other data like literature references, three-dimensional (protein) structure, genomic localisation, or disease relatedness. Usually, they are indexed with primary database accession numbers. Often, they also have an interface to search with the sequence itself (mostly by BLAST).

- Some examples:

OMIM (Online Mendelian inheritance in man): Lists genes that are important in human disease (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM).

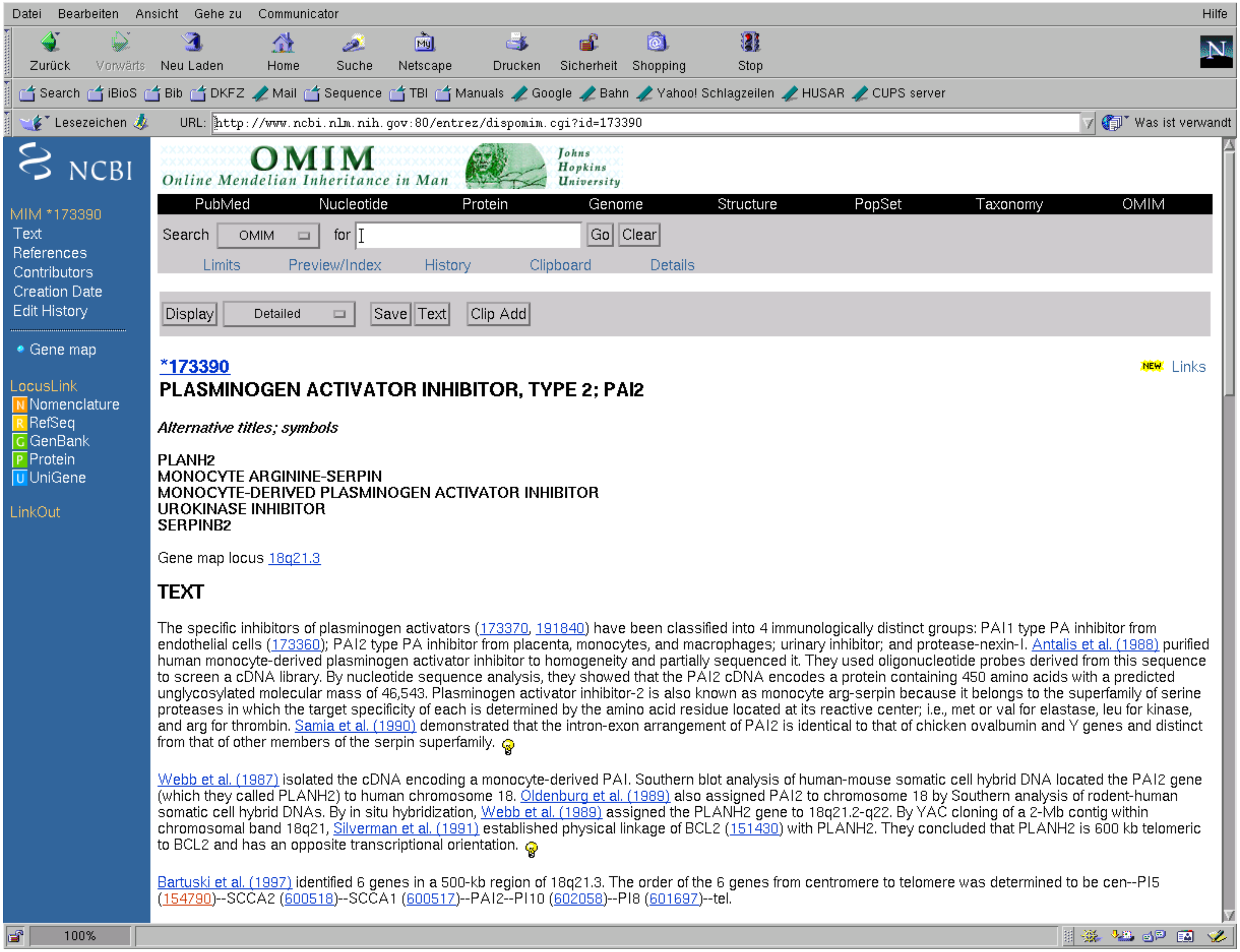
- Examples (continued):

Locus Link Links to a supposedly unique locus on the genome, some cross links; only available for some organisms. (http://www.ncbi.nlm.nih.gov/LocusLink/)

PFAM Gives information about domain structure and relations to other proteins containing these domains (http://www.sanger.ac.uk/Software/Pfam/).

Gene Cards Gives concise information for human genes, including links to other (non-primary) databases (http://bioinformatics.weizmann.ac.il/cards/, mirror in Heidelberg http://www.dkfz-heidelberg.de/GeneCards/).

Sample entry in OMIM



***173390**

PLASMINOGEN ACTIVATOR INHIBITOR, TYPE 2; PAI2

Alternative titles; symbols

PLANH2
MONOCYTE ARGININE-SERPIN
MONOCYTE-DERIVED PLASMINOGEN ACTIVATOR INHIBITOR
UROKINASE INHIBITOR
SERPINB2

Gene map locus 18q21.3

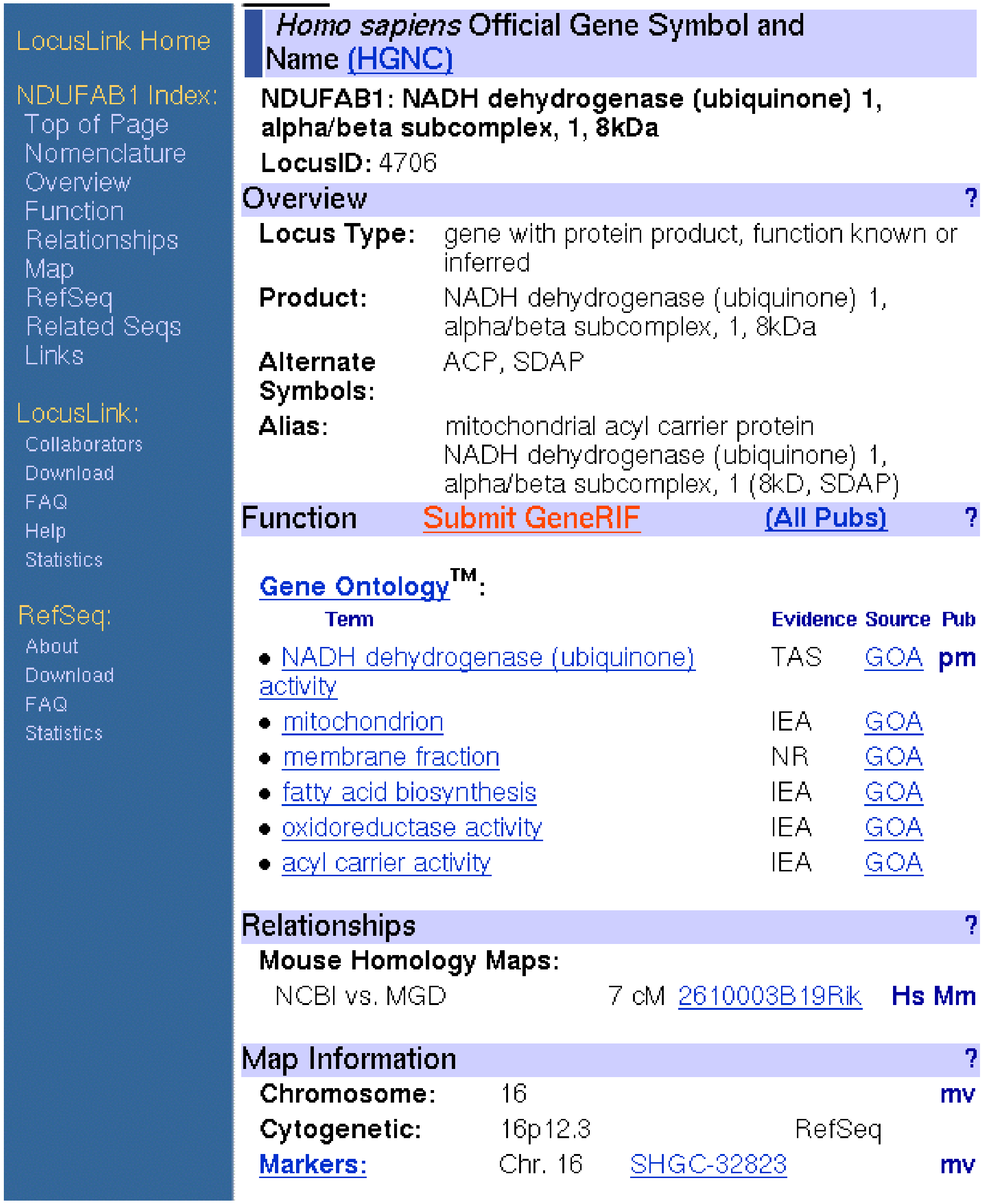
TEXT

The specific inhibitors of plasminogen activators (173370, 191840) have been classified into 4 immunologically distinct groups: PAI1 type PA inhibitor from endothelial cells (173360); PAI2 type PA inhibitor from placenta, monocytes, and macrophages; urinary inhibitor; and protease-nexin-I. Antalis et al. (1988) purified human monocyte-derived plasminogen activator inhibitor to homogeneity and partially sequenced it. They used oligonucleotide probes derived from this sequence to screen a cDNA library. By nucleotide sequence analysis, they showed that the PAI2 cDNA encodes a protein containing 450 amino acids with a predicted unglycosylated molecular mass of 46,543. Plasminogen activator inhibitor-2 is also known as monocyte arg-serpin because it belongs to the superfamily of serine proteases in which the target specificity of each is determined by the amino acid residue located at its reactive center; i.e., met or val for elastase, leu for kinase, and arg for thrombin. Samia et al. (1990) demonstrated that the intron-exon arrangement of PAI2 is identical to that of chicken ovalbumin and Y genes and distinct from that of other members of the serpin superfamily.

Webb et al. (1987) isolated the cDNA encoding a monocyte-derived PAI. Southern blot analysis of human-mouse somatic cell hybrid DNA located the PAI2 gene (which they called PLANH2) to human chromosome 18. Oldenburg et al. (1989) also assigned PAI2 to chromosome 18 by Southern analysis of rodent-human somatic cell hybrid DNAs. By in situ hybridization, Webb et al. (1989) assigned the PLANH2 gene to 18q21.2-q22. By YAC cloning of a 2-Mb contig within chromosomal band 18q21, Silverman et al. (1991) established physical linkage of BCL2 (151430) with PLANH2. They concluded that PLANH2 is 600 kb telomeric to BCL2 and has an opposite transcriptional orientation.

Bartuski et al. (1997) identified 6 genes in a 500-kb region of 18q21.3. The order of the 6 genes from centromere to telomere was determined to be cen--PI5 (154790)--SCCA2 (600518)--SCCA1 (600517)--PAI2--PI10 (602058)--PI8 (601697)--tel.

Sample entry in Locus Link



Homo sapiens Official Gene Symbol and Name (HGNC)

NDUFAB1: NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1, 8kDa

LocusID: 4706

Overview ?

Locus Type:	gene with protein product, function known or inferred
Product:	NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1, 8kDa
Alternate Symbols:	ACP, SDAP
Alias:	mitochondrial acyl carrier protein NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1 (8kD, SDAP)

Function Submit GeneRIF (All Pubs) ?

Gene Ontology^{TM}**:**

Term	Evidence	Source	Pub
● NADH dehydrogenase (ubiquinone) activity	TAS	GOA	pm
● mitochondrion	IEA	GOA	
● membrane fraction	NR	GOA	
● fatty acid biosynthesis	IEA	GOA	
● oxidoreductase activity	IEA	GOA	
● acyl carrier activity	IEA	GOA	

Relationships ?

Mouse Homology Maps:

NCBI vs. MGD	7 cM	2610003B19Rik	Hs Mm

Map Information ?

Chromosome:	16		mv
Cytogenetic:	16p12.3	RefSeq	
Markers:	Chr. 16	SHGC-32823	mv

The relation of clone information to genes and proteins

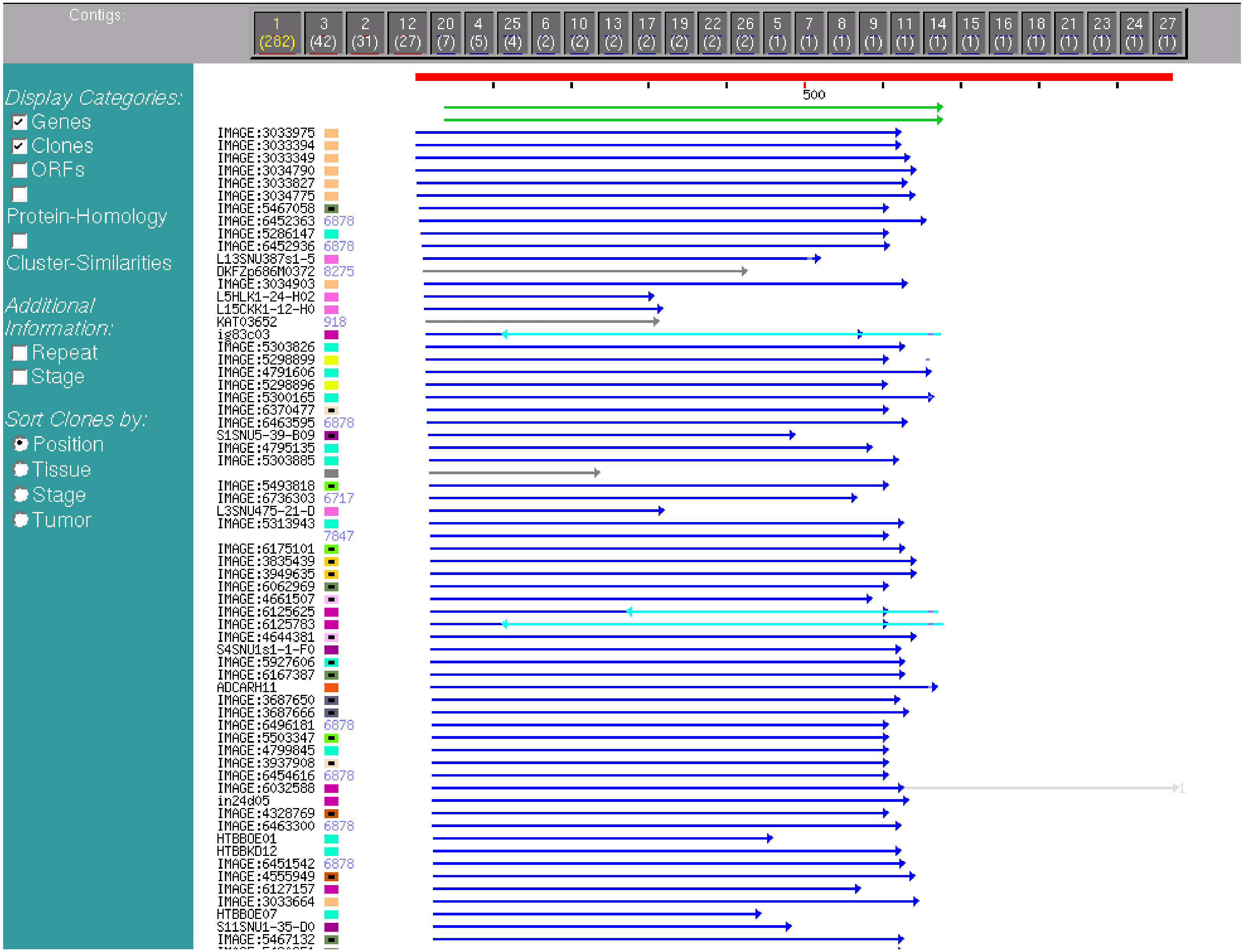
- Microarrays are produced using information on *expressed sequences* as EST clones, cDNAs, partial cDNAs etc.

- At the other end, functional information is generated (and available) for *proteins*. Hence, there is a need to map a clone sequence ID to a protein ID. This is non-trivial.

- First, there are usually hundreds of ESTs (and several cDNA sequences) that map to the same gene. The Database *Unigene* tries to resolve this clustering by sequence clustering. However, this is still imperfect, thus certain ESTs map to more than one cluster, and certain genes are split across several clusters.

- For some reason, Bioconductor does not use Unigene, but Locus Link to link features on an array to a gene or protein.

ESTs in a Unigene cluster



Contigs: 1 (282) 3 (42) 2 (31) 12 (27) 20 (7) 4 (5) 25 (4) 6 (2) 10 (2) 13 (2) 17 (2) 19 (2) 22 (2) 26 (2) 5 (1) 7 (1) 8 (1) 9 (1) 11 (1) 14 (1) 15 (1) 16 (1) 18 (1) 21 (1) 23 (1) 24 (1) 27 (1)

Display Categories:

☑ Genes
☑ Clones
□ ORFs
□ Protein-Homology
□ Cluster-Similarities

Additional Information:

□ Repeat
□ Stage

Sort Clones by:

● Position
○ Tissue
○ Stage
○ Tumor

```
IMAGE:3033975
IMAGE:3033394
IMAGE:3033349
IMAGE:3034790
IMAGE:3033827
IMAGE:3034775
IMAGE:5467058
IMAGE:6452363 6878
IMAGE:5286147
IMAGE:6452936 6878
L13SNU387s1-5
DKFZp686M0372 8275
IMAGE:3034903
L5HLK1-24-H02
L15CKK1-12-H0
KAT03652      918
ig83c03
IMAGE:5303826
IMAGE:5298899
IMAGE:4791606
IMAGE:5298896
IMAGE:5300165
IMAGE:6370477
IMAGE:6463595 6878
S1SNU5-39-B09
IMAGE:4795135
IMAGE:5303885

IMAGE:5493818
IMAGE:6736303 6717
L3SNU475-21-D
IMAGE:5313943
              7847
IMAGE:6175101
IMAGE:3835439
IMAGE:3949635
IMAGE:6062969
IMAGE:4661507
IMAGE:6125625
IMAGE:6125783
IMAGE:4644381
S4SNU1s1-1-F0
IMAGE:5927606
IMAGE:6167387
ADCARH11
IMAGE:3687650
IMAGE:3687666
IMAGE:6496181 6878
IMAGE:5503347
IMAGE:4799845
IMAGE:3937908
IMAGE:6454616 6878
IMAGE:6032588
in24d05
IMAGE:4328769
IMAGE:6463300 6878
HTBBOE01
HTBBKD12
IMAGE:6451542 6878
IMAGE:4555949
IMAGE:6127157
IMAGE:3033664
HTBBOE07
S11SNU1-35-D0
IMAGE:5467132
```

The Human Genome Sequence

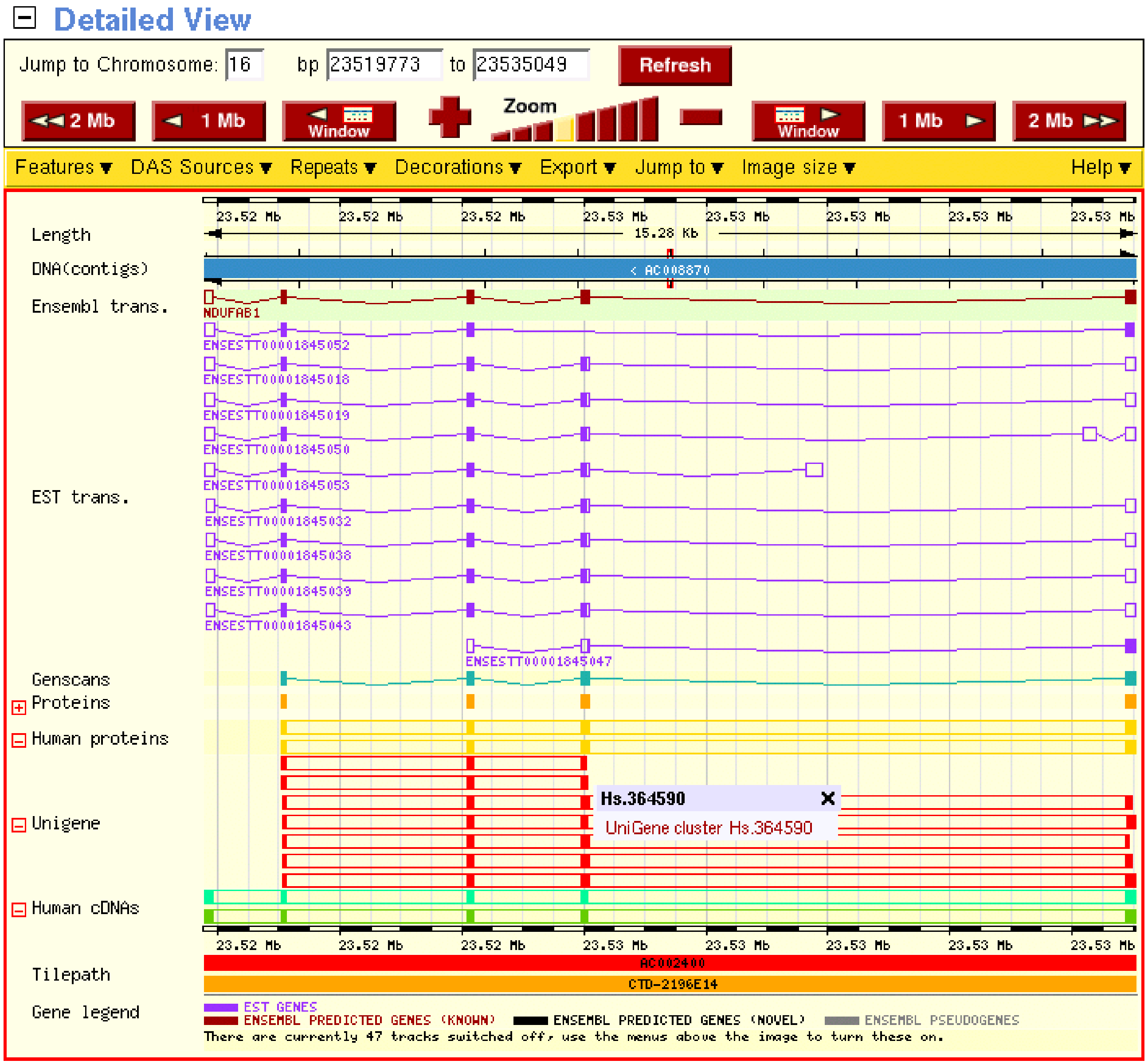
- With the completion of the human genome sequence, you'd think that such ambiguities can be resolved. In fact, that is not the case.

- Part of the problem is due to the fact that it is hard to predict gene structure (intron/exon) without knowing the entire mRNA sequence, wich happens for about two-thirds of all genes.

- Then, there are errors in the assembly (putting together the sequence snippets). A typical symptom is that a gene appears to map to multiple loci on the same chromosome, with very high sequence similarity. Then, the same sequence was probably introduced several times in the assembly.

- We will later discuss the consequences for microarray annotation.

Genomic mapping: ENSEMBL Browser



Detailed View

Jump to Chromosome: 16 bp 23519773 to 23535049

Hs.364590 ✕

UniGene cluster Hs.364590

🟪 EST GENES

🟥 ENSEMBL PREDICTED GENES (KNOWN) ■ ENSEMBL PREDICTED GENES (NOVEL) ▬ ENSEMBL PSEUDOGENES

There are currently 47 tracks switched off, use the menus above the image to turn these on.

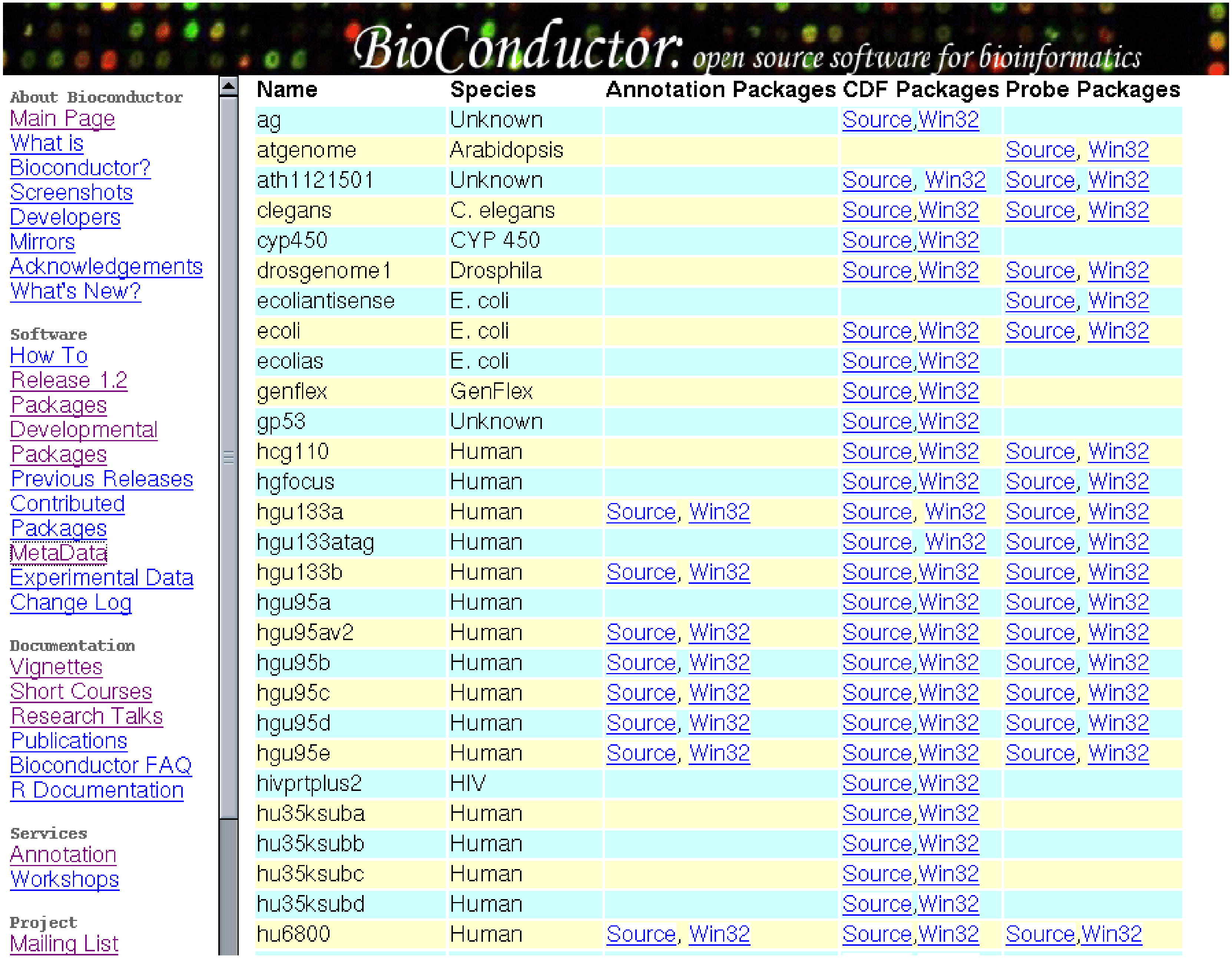
Where do we get all this information?

- Of course, all this can be looked up in a database for the “interesting” genes. For more than some dozens of genes, this is tedious.

- If you’re experienced in writing scripts and dealing with biological and relational databases, you can collect the information on your own and build your annotation database. This requires some experience with bioinformatics. You can even prepare your data to access them inside **R**, using the package `AnnBuilder`. It’s beyond the scope of this course to explain how this works.

- Fortunately, some data have been precompiled by kind people, and they are distributed via **Bioconductor**.

Data packages in Bioconductor



BioConductor: open source software for bioinformatics

About Bioconductor
- Main Page
- What is Bioconductor?
- Screenshots
- Developers
- Mirrors
- Acknowledgements
- What's New?

Software
- How To
- Release 1.2 Packages
- Developmental Packages
- Previous Releases
- Contributed Packages
- MetaData
- Experimental Data
- Change Log

Documentation
- Vignettes
- Short Courses
- Research Talks
- Publications
- Bioconductor FAQ
- R Documentation

Services
- Annotation
- Workshops

Project
- Mailing List

Name	Species	Annotation Packages	CDF Packages	Probe Packages
ag	Unknown		Source,Win32	
atgenome	Arabidopsis			Source, Win32
ath1121501	Unknown		Source, Win32	Source, Win32
celegans	C. elegans		Source,Win32	Source, Win32
cyp450	CYP 450		Source,Win32	
drosgenome1	Drosphila		Source,Win32	Source, Win32
ecoliantisense	E. coli			Source, Win32
ecoli	E. coli		Source,Win32	Source, Win32
ecolias	E. coli		Source,Win32	
genflex	GenFlex		Source,Win32	
gp53	Unknown		Source,Win32	
hcg110	Human		Source,Win32	Source, Win32
hgfocus	Human		Source,Win32	Source, Win32
hgu133a	Human	Source, Win32	Source, Win32	Source, Win32
hgu133atag	Human		Source, Win32	Source, Win32
hgu133b	Human	Source, Win32	Source,Win32	Source, Win32
hgu95a	Human		Source,Win32	Source, Win32
hgu95av2	Human	Source, Win32	Source,Win32	Source, Win32
hgu95b	Human	Source, Win32	Source,Win32	Source, Win32
hgu95c	Human	Source, Win32	Source,Win32	Source, Win32
hgu95d	Human	Source, Win32	Source,Win32	Source, Win32
hgu95e	Human	Source, Win32	Source,Win32	Source, Win32
hivprtplus2	HIV		Source,Win32	
hu35ksuba	Human		Source,Win32	
hu35ksubb	Human		Source,Win32	
hu35ksubc	Human		Source,Win32	
hu35ksubd	Human		Source,Win32	
hu6800	Human	Source, Win32	Source,Win32	Source,Win32

Bioconductor metadata packages

- These packages contain one-to-one and one-to-many mappings for frequently used chips, especially Affymetrix arrays.

- Information available includes gene names, gene symbol, database accession numbers, Gene Ontology function description, enzmye classification number (EC), relations to PubMed abstracts, and others.

- The data use the framework of the `annotate` package, so I will briefly explain how it works.

Environments in R

- To quickly find information on one subject in a long list, a data structure called *hash table* is frequently used in computer science.

- A hash table is a list of key/value pairs, where the key is used to find the corresponding value. To go the other way round, you have to use pattern matching, which is much slower.

- In R, hash tables are implemented as *environments*. For the moment, we do not care about the philosophy behind it and simply treat it as another word for hash table.

Setting up environments

To set up a new environment:

```
symbol.hash = new.env(hash=TRUE)
```

To create a key/value pair:

```
assign("1234_at", "EphA3", env=symbol.hash)
```

To list all keys of an environment:

```
ls(env=symbol.hash)
```

To get the value for a certain key:

```
get("1234_at", env=symbol.hash)
```

The annotate package

- That's all standard R. The annotate package gives one further function, `multiget`, which retrieves more than one entry at a time, and definitions for special data, e.g. PubMed abstracts, or chromosomal location objects.

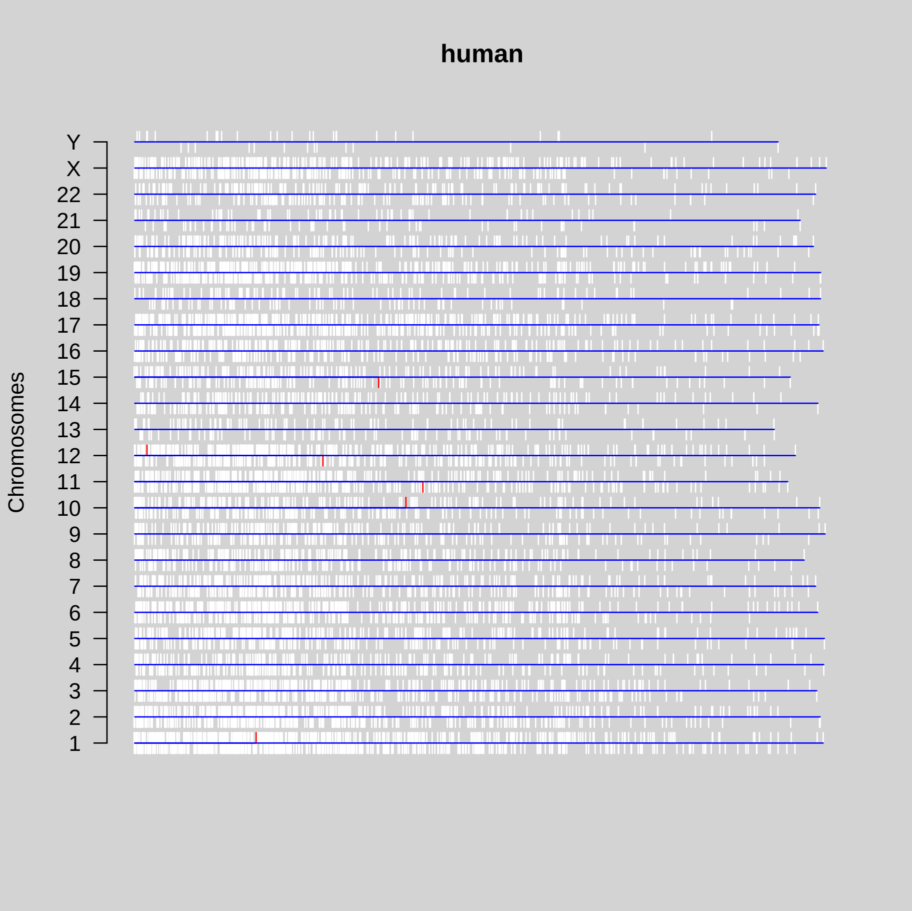
- ChromLoc objects are quite useful if you want to associate gene expression with certain positions on a chromosome, e.g. if aberration occurs in your samples.

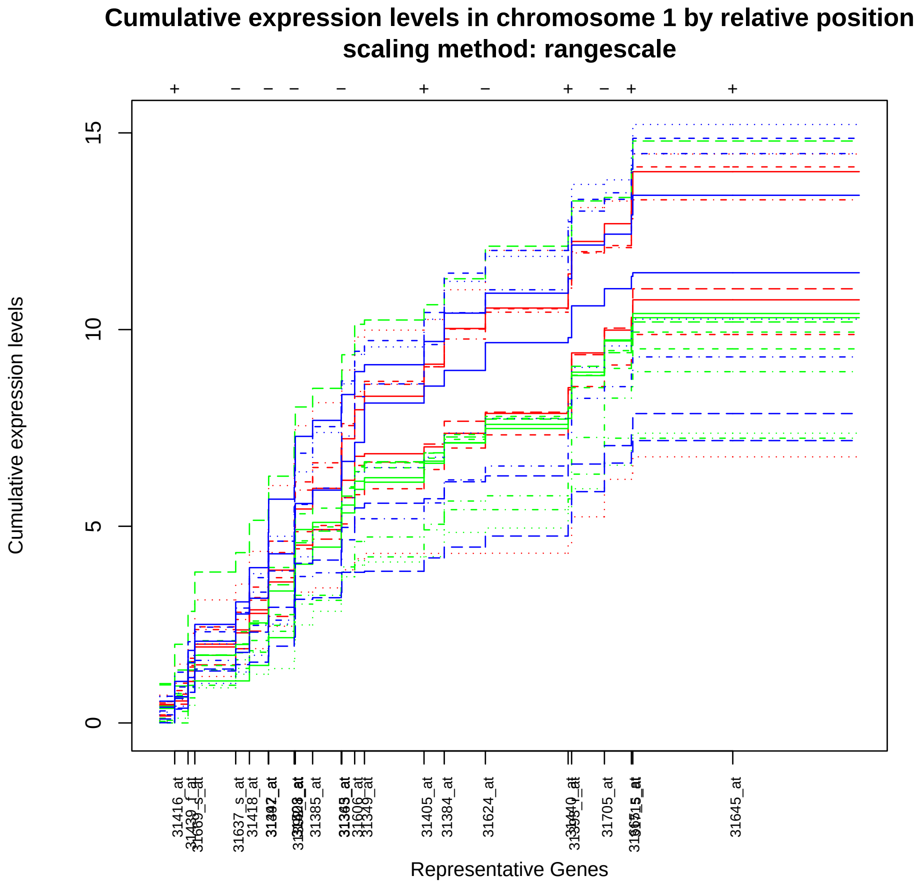
- You can construct a ChromLoc object on your own (⟶ Vignette), or use the function `buildChromLocation`. For chip HGU95a_v2:

```
library(hgu95av2)
cl.95a = buildChromLocation("hgu95av2")
```

Plots for ChromLocation objects

- Plotting methods are available via library `geneplotter`





How to get annotation for a set of genes

- Suppose you have found some interesting genes. The index in the matrix is in `index.int`. To get the gene names:

```
gnam.int = geneNames(exprset)[index.int]
```

- To find the description:

```
multiget(gnam.int, env=hgu95av2GENENAME)
```

- To get EC Numbers (relating to KEGG pathways):

```
multiget(gnam.int, env=hgu95av2ENZYME)
```

Some caveats

- Because of the non-unique matching of sequences to the genome, array features are sometimes annotated with more than one position:

```
a = ls(env=hgu95av2CHRLOC)
table(sapply(multiget(a, env=hgu95av2CHRLOC),
   length))
```

```
    1      2      3      4      5      6      7      8
11793    647    127     29     13     10      2      1
```

- For the 800 or so sequences with more than one location, only the first one is used, although there is no warning. It should be desirable to resolve the ambiguities by hand, but nobody has done yet.

- There are even 14 probe sets on HGU95A_v2 that map to 2 chromosomes; however, these are located on some special extrachromosomal segment and annotated with “X” and “Y”.

Pattern matching

- To find something in character vectors or character lists, some pattern matching is required.

- If you have real full names, use `match`, e.g.

```
match("1234_at", rownames(exprs(exprset)) )
```

- This will give you the index of `‘‘1234_at’’`. It works also with more than one gene:

```
match(gnam.int, rownames(exprs(exprset)) )
```

will give all indeces for genes in `gnam.int`.

- If you want to use regular expression matching, use `grep`.

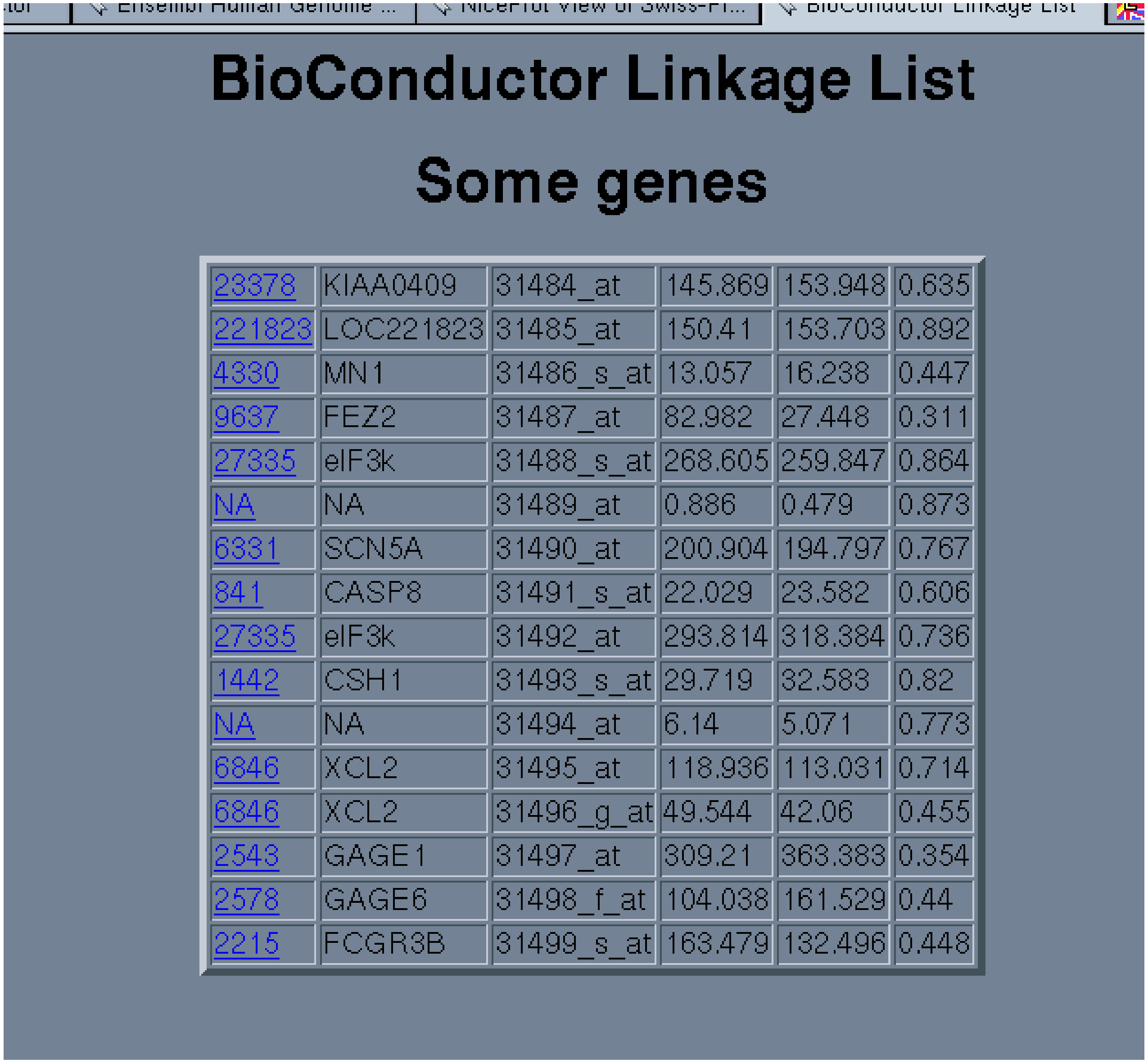
Export of annotation to HTML

- `annotate` is able to export tables of gene annotations to HTML, which is much nicer to browse than text tables

- Suppose, from a t-test you have for some genes `igenes`: mean of genes in class 1, `igenes.gp1`, mean in class 2, `igenes.gp2`, and P-value `igenes.pval`. To construct pretty HTML output:

```
igenes.ll = multiget(igenes, env=hgu95av2LOCUSID)
igenes.sym = multiget(igenes, env=hgu95av2SYMBOL)
ll.htmlpage(igenes.ll, "HOWTO.igenes", "Some genes",
  list(igenes,sym, igenes, round(igenes.gp1,3),
  round(igenes.gp2,3),round(igenes.pval,3)))
```

The result



BioConductor Linkage List

Some genes

23378	KIAA0409	31484_at	145.869	153.948	0.635
221823	LOC221823	31485_at	150.41	153.703	0.892
4330	MN1	31486_s_at	13.057	16.238	0.447
9637	FEZ2	31487_at	82.982	27.448	0.311
27335	eIF3k	31488_s_at	268.605	259.847	0.864
NA	NA	31489_at	0.886	0.479	0.873
6331	SCN5A	31490_at	200.904	194.797	0.767
841	CASP8	31491_s_at	22.029	23.582	0.606
27335	eIF3k	31492_at	293.814	318.384	0.736
1442	CSH1	31493_s_at	29.719	32.583	0.82
NA	NA	31494_at	6.14	5.071	0.773
6846	XCL2	31495_at	118.936	113.031	0.714
6846	XCL2	31496_g_at	49.544	42.06	0.455
2543	GAGE1	31497_at	309.21	363.383	0.354
2578	GAGE6	31498_f_at	104.038	161.529	0.44
2215	FCGR3B	31499_s_at	163.479	132.496	0.448

Function annotation

- Probably, the must important thing you want to know is what the genes or their products are concerned with, i.e. their **function**.

- Function annotation is difficult: Different people use different words for the same function, or may mean different things by the same word. The context in which a gene was found (e.g. “TGFβ-induced gene”) may not be particularly associated with its function.

- Inference of function from sequence alone is error-prone and sometimes unreliable. The best function annotation systems (GO, SwissProt) use human beings who read the literature before assigning a function to a gene.

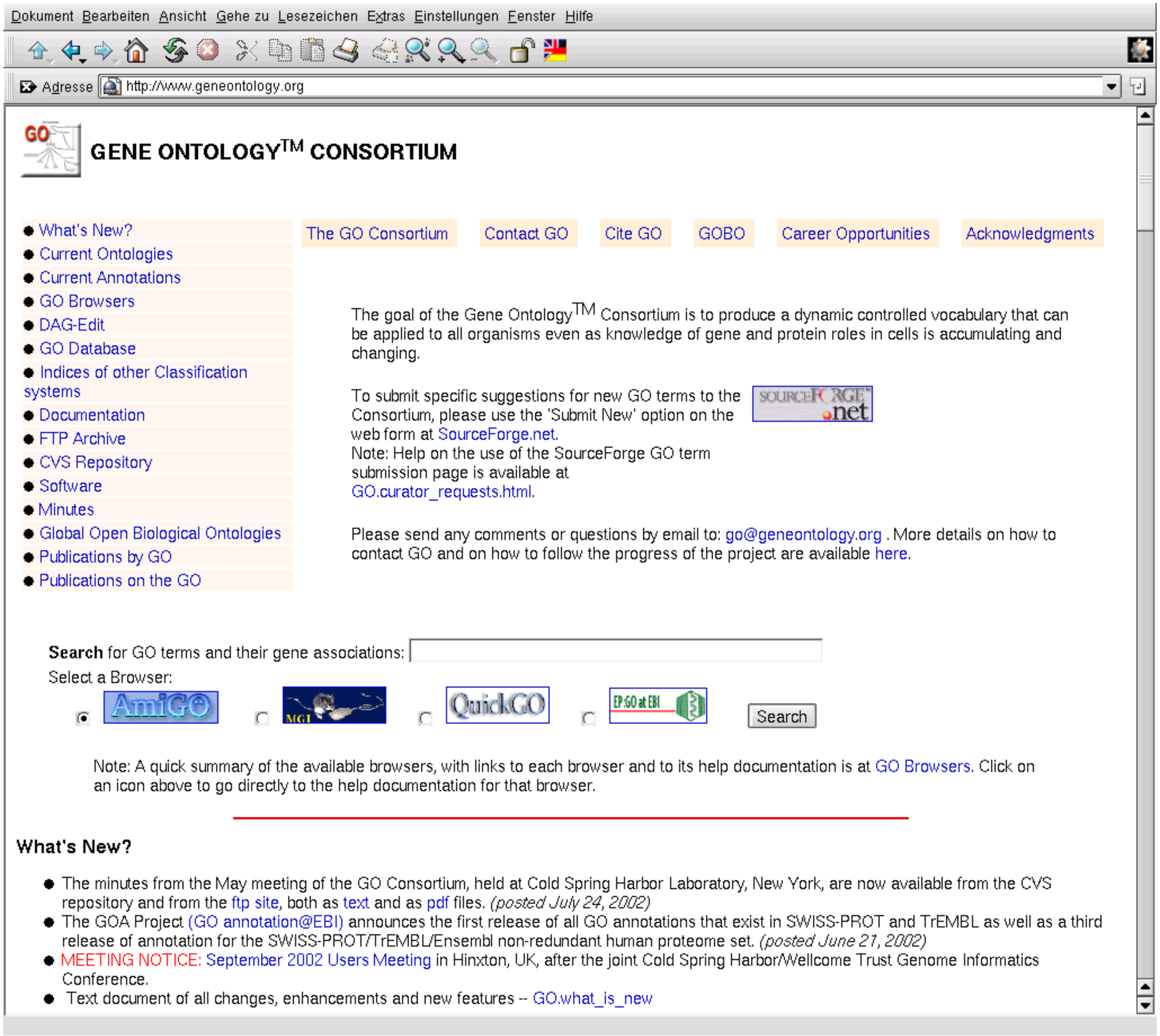
The Gene Ontology system

- To overcome some of the problems, an annotation system has been created: Gene Ontology (http://www.geneontology.org). Ontology means here the art (or science) of giving everything its correct name.

- It represents a unified, consistent system, i.e. terms occur only once, and there is a dictionary of allowed words.

- Furthermore, terms are related to each other: the hierarchy goes from very general terms to very detailed ones.

The Gene Ontology site


Dokument Bearbeiten Ansicht Gehe zu Lesezeichen Extras Einstellungen Fenster Hilfe

Adresse http://www.geneontology.org

GENE ONTOLOGY™ CONSORTIUM

- What's New?
- Current Ontologies
- Current Annotations
- GO Browsers
- DAG-Edit
- GO Database
- Indices of other Classification systems
- Documentation
- FTP Archive
- CVS Repository
- Software
- Minutes
- Global Open Biological Ontologies
- Publications by GO
- Publications on the GO

The GO Consortium | Contact GO | Cite GO | GOBO | Career Opportunities | Acknowledgments

The goal of the Gene Ontology™ Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.

To submit specific suggestions for new GO terms to the Consortium, please use the 'Submit New' option on the web form at SourceForge.net.
Note: Help on the use of the SourceForge GO term submission page is available at GO.curator_requests.html.

SOURCEFORGE.net

Please send any comments or questions by email to: go@geneontology.org . More details on how to contact GO and on how to follow the progress of the project are available here.

Search for GO terms and their gene associations:

Select a Browser:

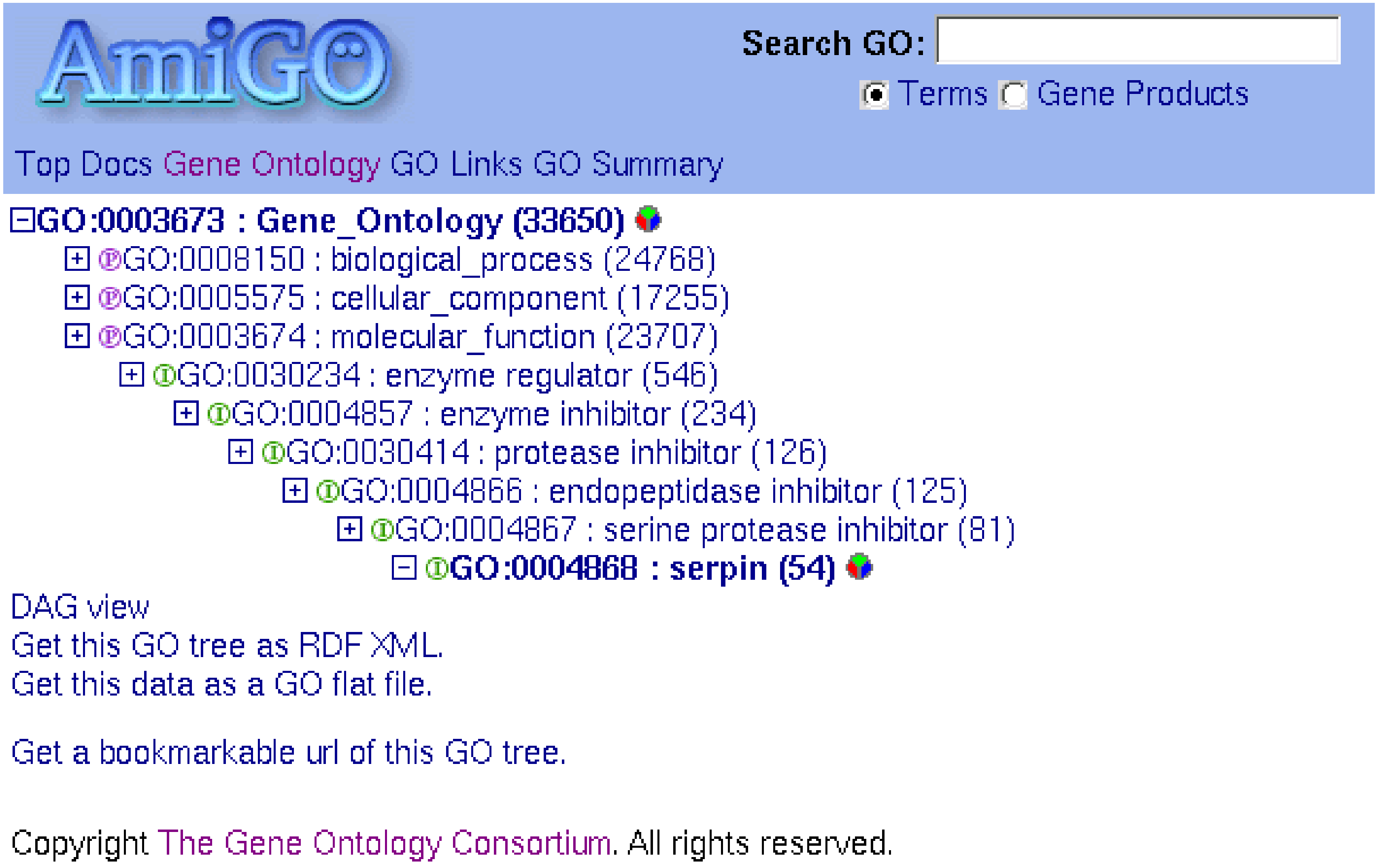
● AmiGO ○ MGI ○ QuickGO ○ EP:GO at EBI Search

Note: A quick summary of the available browsers, with links to each browser and to its help documentation is at GO Browsers. Click on an icon above to go directly to the help documentation for that browser.

What's New?

- The minutes from the May meeting of the GO Consortium, held at Cold Spring Harbor Laboratory, New York, are now available from the CVS repository and from the ftp site, both as text and as pdf files. *(posted July 24, 2002)*
- The GOA Project (GO annotation@EBI) announces the first release of all GO annotations that exist in SWISS-PROT and TrEMBL as well as a third release of annotation for the SWISS-PROT/TrEMBL/Ensembl non-redundant human proteome set. *(posted June 21, 2002)*
- MEETING NOTICE: September 2002 Users Meeting in Hinxton, UK, after the joint Cold Spring Harbor/Wellcome Trust Genome Informatics Conference.
- Text document of all changes, enhancements and new features -- GO.what_is_new

The Gene Ontology hierarchy



Search GO: □

● Terms ○ Gene Products

Top Docs Gene Ontology GO Links GO Summary

- ⊟**GO:0003673 : Gene_Ontology (33650)**
  - ⊞ ⓅGO:0008150 : biological_process (24768)
  - ⊞ ⓅGO:0005575 : cellular_component (17255)
  - ⊞ ⓅGO:0003674 : molecular_function (23707)
    - ⊞ ⒾGO:0030234 : enzyme regulator (546)
      - ⊞ ⒾGO:0004857 : enzyme inhibitor (234)
        - ⊞ ⒾGO:0030414 : protease inhibitor (126)
          - ⊞ ⒾGO:0004866 : endopeptidase inhibitor (125)
            - ⊞ ⒾGO:0004867 : serine protease inhibitor (81)
              - ⊟ Ⓘ**GO:0004868 : serpin (54)**

DAG view
Get this GO tree as RDF XML.
Get this data as a GO flat file.

Get a bookmarkable url of this GO tree.

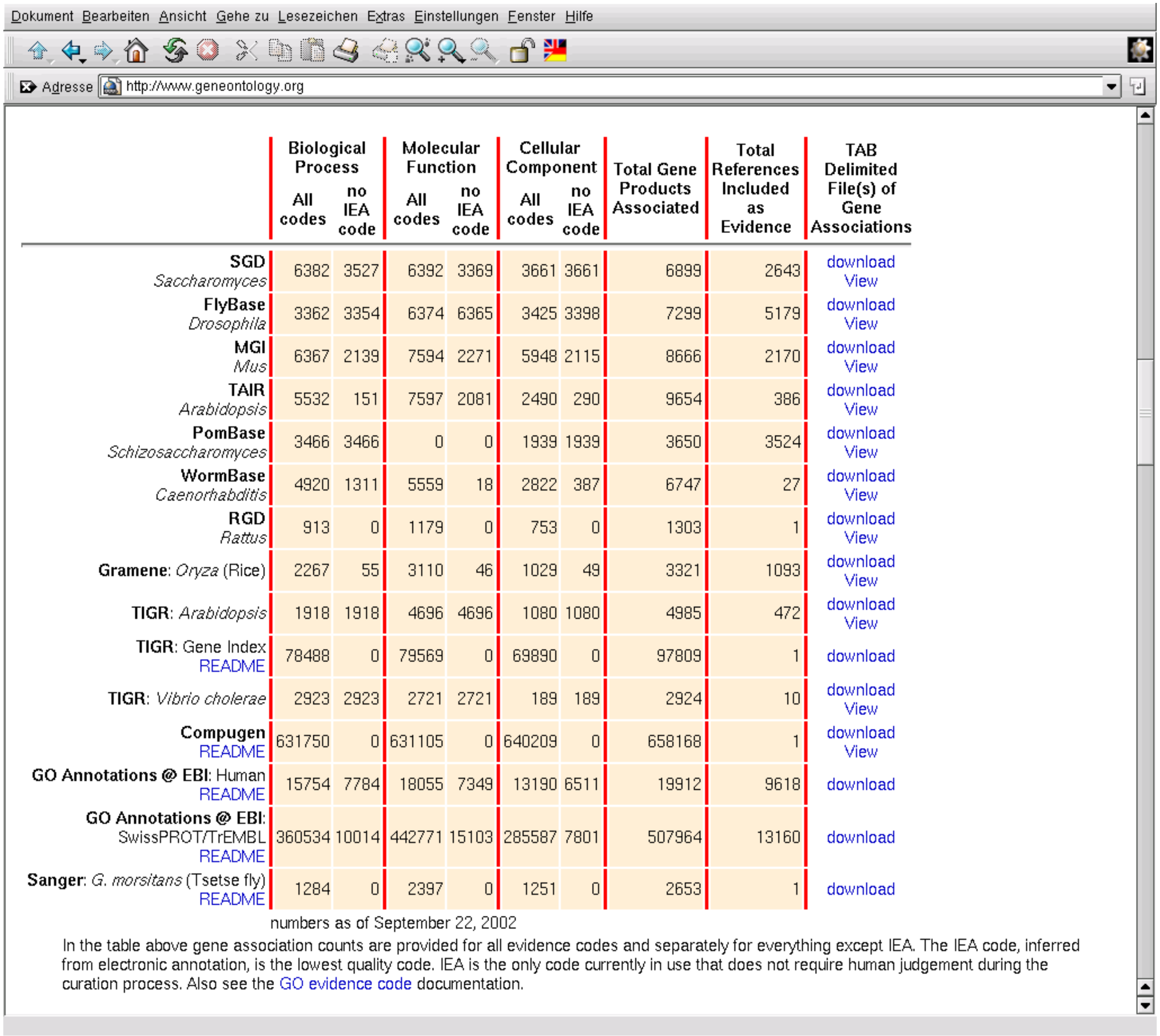
Copyright The Gene Ontology Consortium. All rights reserved.

Actual annotation

- Gene Ontology by itself is only a system for annotating genes and proteins. It does not relate database entries to a special annotation value.

- Luckily, research communities for several model organisms have agreed on entering Gene Ontology information into the databases. As this is done ‘by hand’, GO annotation for most organisms is far from complete.

Available Gene Ontology information


Dokument Bearbeiten Ansicht Gehe zu Lesezeichen Extras Einstellungen Fenster Hilfe
Adresse
http://www.geneontology.org

	Biological Process		Molecular Function		Cellular Component		Total Gene Products Associated	Total References Included as Evidence	TAB Delimited File(s) of Gene Associations
	All codes	no IEA code	All codes	no IEA code	All codes	no IEA code			
SGD *Saccharomyces*	6382	3527	6392	3369	3661	3661	6899	2643	download View
FlyBase *Drosophila*	3362	3354	6374	6365	3425	3398	7299	5179	download View
MGI *Mus*	6367	2139	7594	2271	5948	2115	8666	2170	download View
TAIR *Arabidopsis*	5532	151	7597	2081	2490	290	9654	386	download View
PomBase *Schizosaccharomyces*	3466	3466	0	0	1939	1939	3650	3524	download View
WormBase *Caenorhabditis*	4920	1311	5559	18	2822	387	6747	27	download View
RGD *Rattus*	913	0	1179	0	753	0	1303	1	download View
Gramene: *Oryza* (Rice)	2267	55	3110	46	1029	49	3321	1093	download View
TIGR: *Arabidopsis*	1918	1918	4696	4696	1080	1080	4985	472	download View
TIGR: Gene Index README	78488	0	79569	0	69890	0	97809	1	download
TIGR: *Vibrio cholerae*	2923	2923	2721	2721	189	189	2924	10	download View
Compugen README	631750	0	631105	0	640209	0	658168	1	download View
GO Annotations @ EBI: Human README	15754	7784	18055	7349	13190	6511	19912	9618	download
GO Annotations @ EBI: SwissPROT/TrEMBL README	360534	10014	442771	15103	285587	7801	507964	13160	download
Sanger: *G. morsitans* (Tsetse fly) README	1284	0	2397	0	1251	0	2653	1	download

numbers as of September 22, 2002

In the table above gene association counts are provided for all evidence codes and separately for everything except IEA. The IEA code, inferred from electronic annotation, is the lowest quality code. IEA is the only code currently in use that does not require human judgement during the curation process. Also see the GO evidence code documentation.

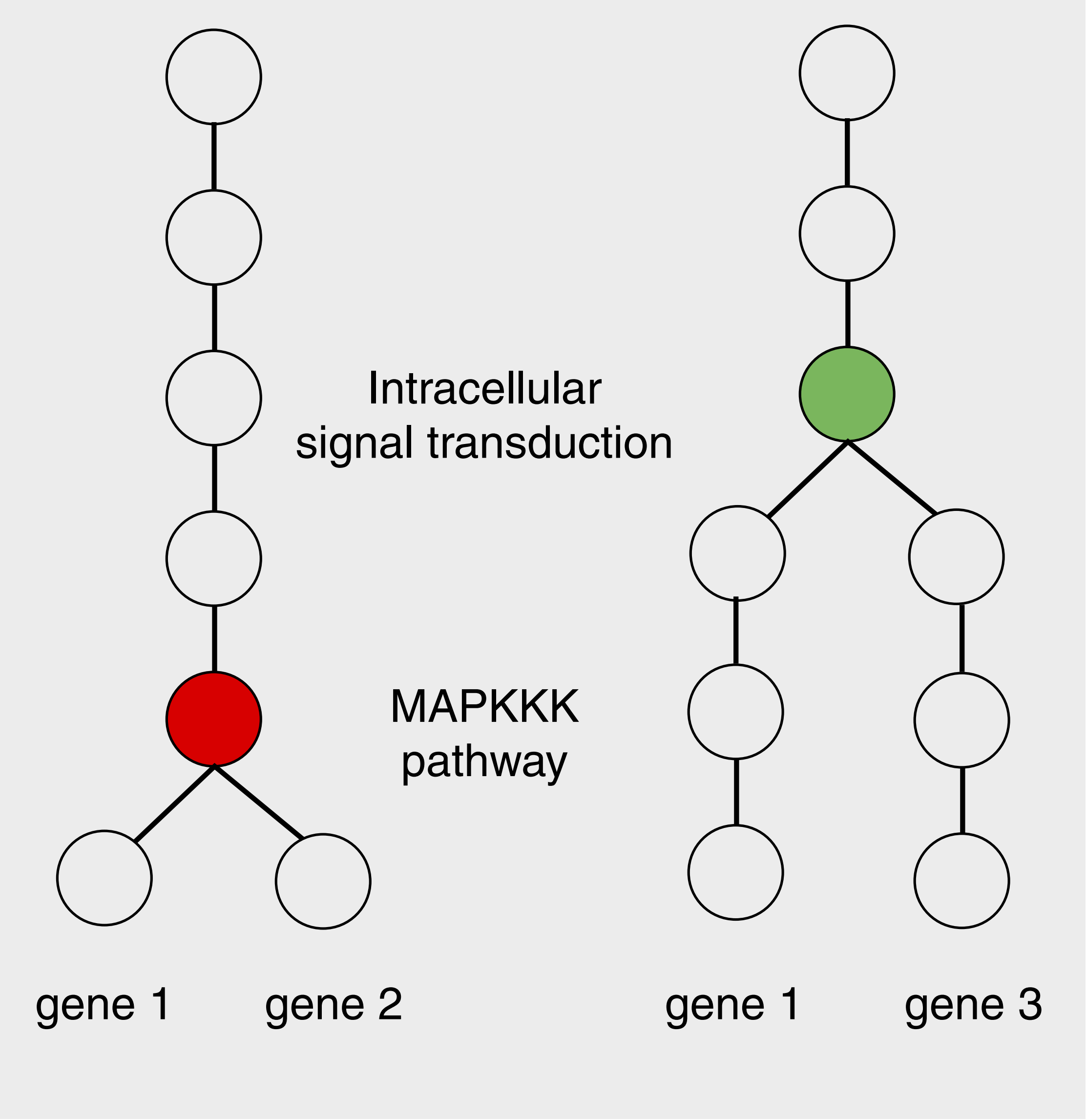
Dealing with GO annotations

- Since the annotation system is hierarchical, i.e. for each term there is a hierarchical list of more general terms, we can compare functions of genes on every level we wish.

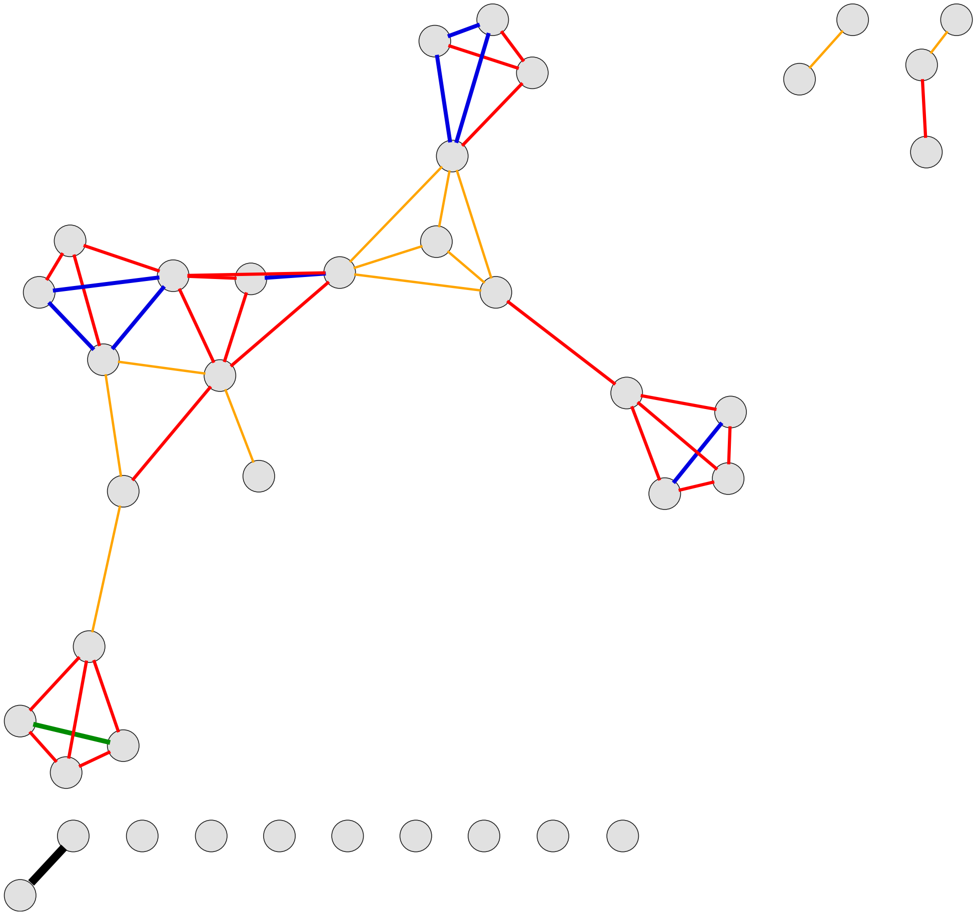
- Technically, this amounts to the problem of finding the least common parent node between to genes of interest.

- This can be used to find clusters of functionally related genes in a list that comes out of some other analysis.

Comparing GO-annotated genes



GO functional clusters as a graph



Graphs as analysis tools

- Graphs are quite useful for bioinformatic analysis, and have a long-standing history in sequence analysis.

- Recently, some functionality has been built into R to deal with graphs (`graph, Rgraphviz, RBGL`). Certainly, the most useful capability is to visualize graphs via `Rgraphviz`. The R package is an interface to the external program `graphviz` (from AT&T). Big graphs should be visualized by means of `ggobi`, however.

- Some other immediate use is to construct PubMed co-citation graphs for genes of interest. Functions for this exist. However, for many other applications the meaning of graphs or graph-theoretic algorithms is not clear, so a lot of work remains to be done.