Sequence annotation and biological information

Benedikt Brors

Dept. Intelligent Bioinformatics Systems

German Cancer Research Center

Why do we need sequence annotation?

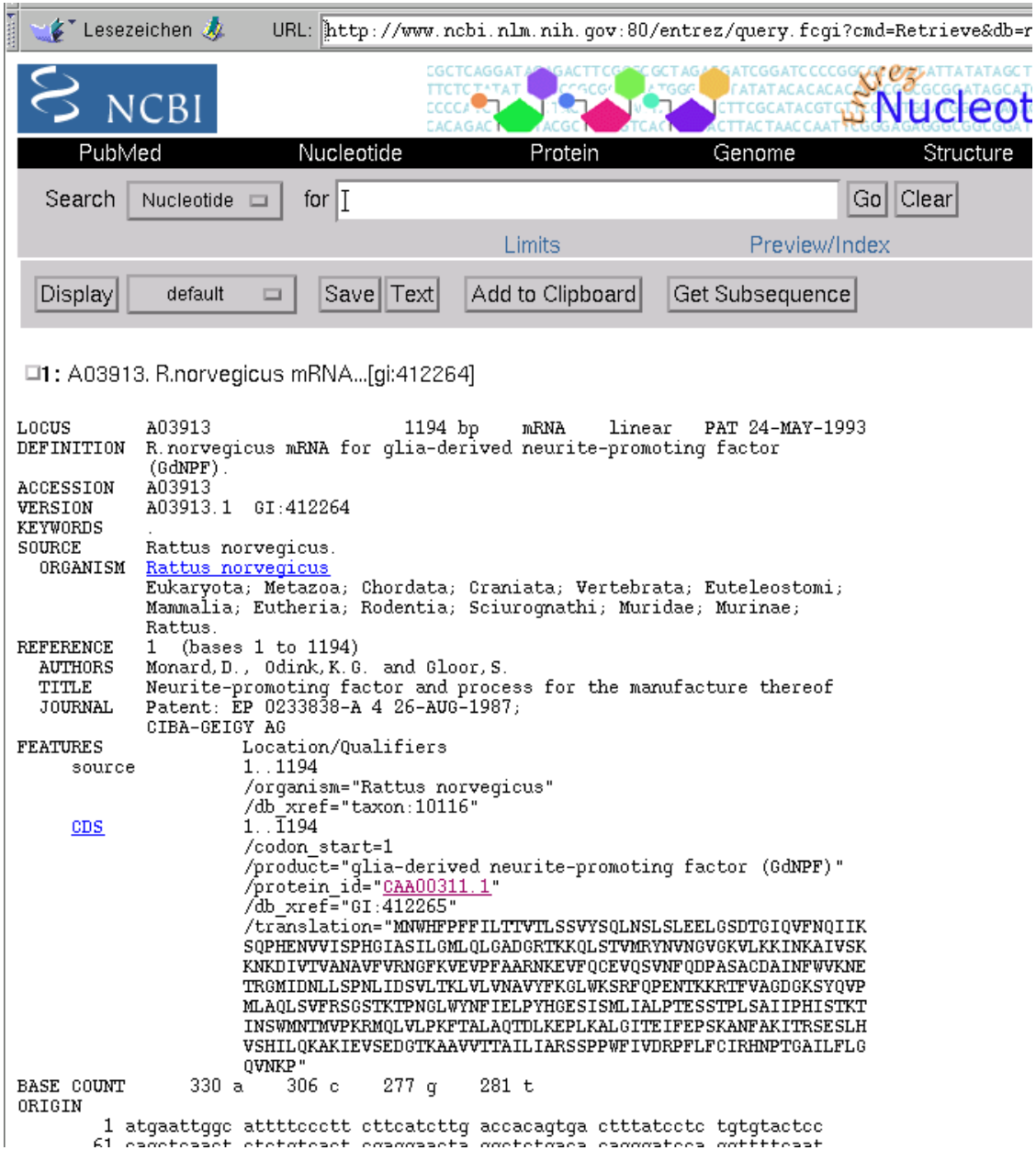
- Often, the result of microarray data analysis is a list of genes.
- The list has to be summarized with respect to its biological meaning. For this, information about the genes and the related proteins has to be gathered.
- If the list is small (let's say, 1–30), this is easily done by reading database 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



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

Lesezeichen URL: http://us.expasy.org/cgi-bin/niceprot.pl?P07092 Was ist verwandt

[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

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

dkfz

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:

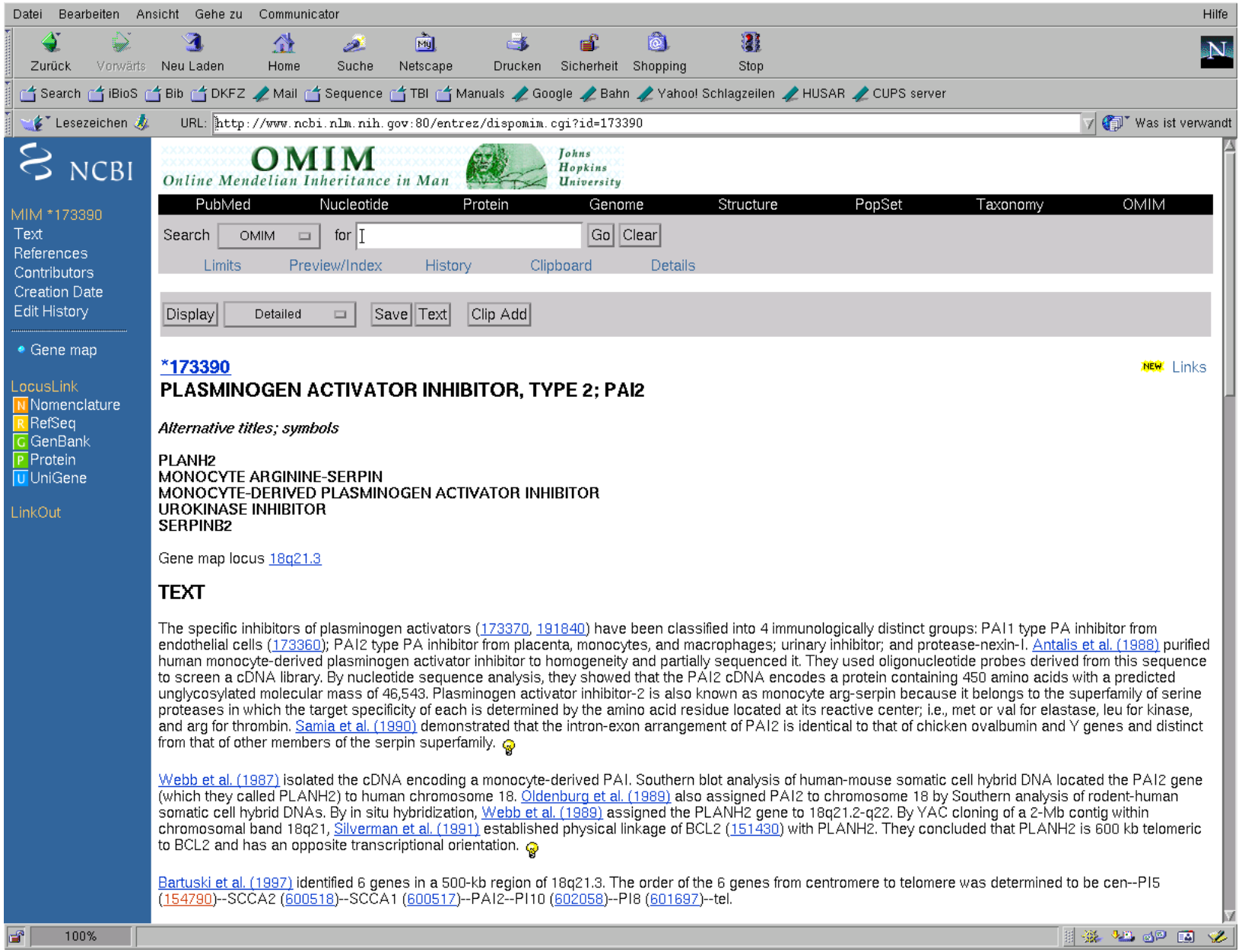
  OMIM (Online Mendelian inheritance in man): Lists genes that are important in human disease (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM).

- Examples (continued):

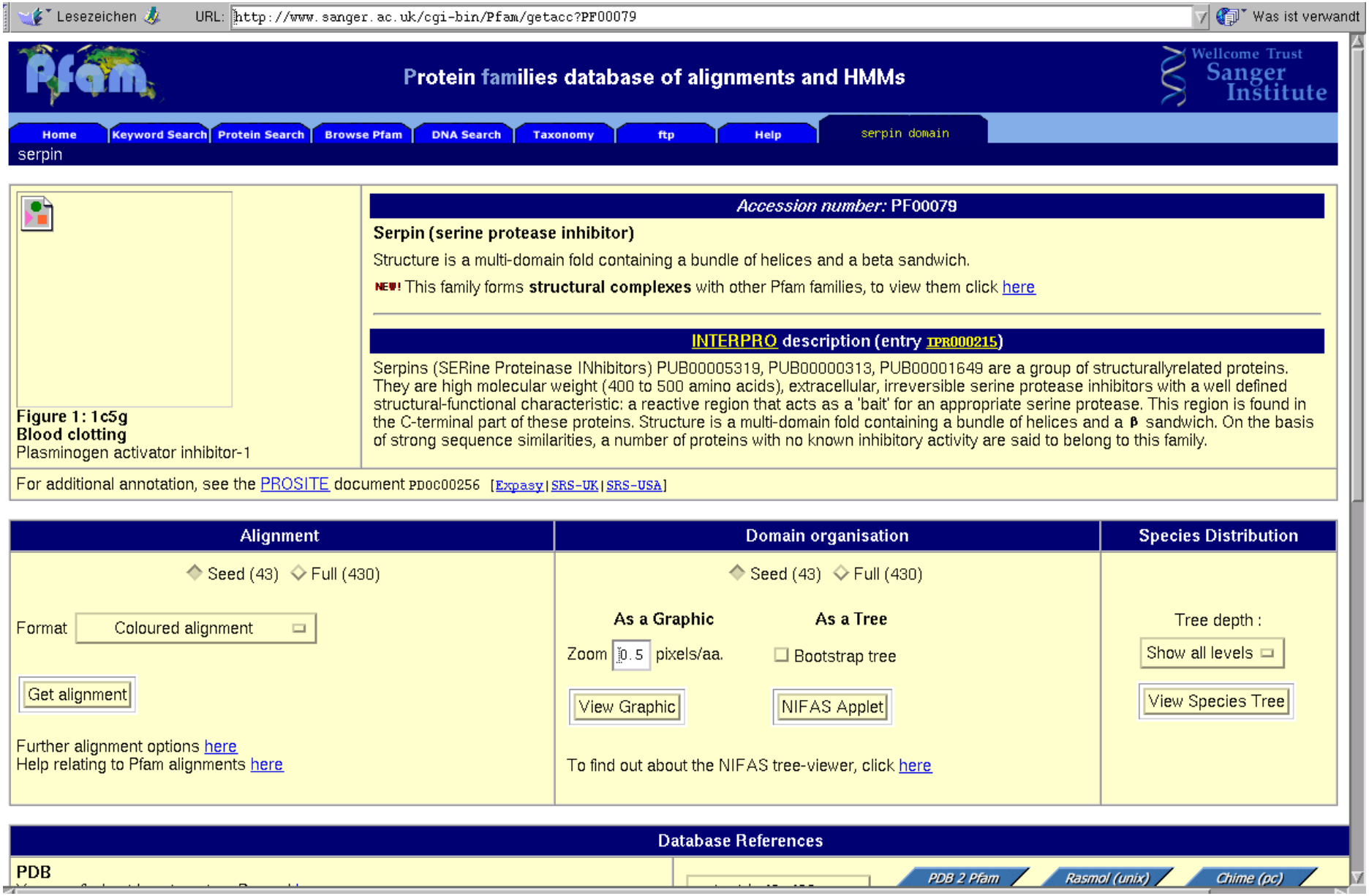
PFAM Gives information about domain structure and relations to other proteins containing these domains (http://www.sanger.ac.uk/Software/Pfam/).

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



Sample entry in PFAM



How to retrieve large-scale data

- Although most or all of the databases mentioned are ‘real’ databases, you won’t have direct access to them. There is no interface to get *only* the information you want by using a database query language like SQL.

- It is also annoying that the information is presented on different HTML pages. You usually don’t want to go through 100 pages and extract the two or three words you need by cut-and-paste.

- Fortunately, a number of languages have an interface to HTML retrieval. By far the easiest to learn is PERL (http://www.perl.org). You can write a small script that retrieves the web page, parses it and extracts only the relevant information. If you like, you can parse 12,000 web pages on a weekend (speed depends on your internet connection, though).

A sample perl script

File Edit Mule Apps Options Buffers Tools Perl Help

Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug News

```
#!/usr/bin/perl

use LWP::Simple ;

$server="http://iubio.bio.indiana.edu:8089/.bin/moquery?" ;
$accno = $ARGV[0];

$url = $server.$accno ;
#print $url;

$r .= get($url) ;

$r =~ s/<[^>]+>//g;

$r =~ /Molecular function\n(.+?)\n\n.+Biological process\n(.+?)\n\n/s; #\n\n/s ;
$molfunc = $1;

#$r =~ /Biological Process\n(.+?)\n\n/s;
$bioproc = $2;

#print $1;

#print $r;
print "EMBL-Acc:$accno\n\nMolecular function:\n$molfunc\n\nBiological Process:\n$bioproc\n";



#end ;
```

ISO8-----XEmacs: gethttp.pl (CPerl Font)----L1--C0--All----------

Fontifying gethttp.pl... done.

The HTML page ...

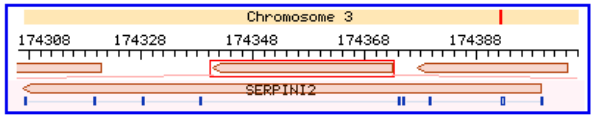
Lesezeichen URL: http://iubio.bio.indiana.edu:8089/.bin/fbidq.html?HUgn0005276 Was ist verwandt

euGenes .. Fish .. Fly .. Human .. Mouse .. Mosquito .. Weed .. Worm .. Yeast 

Help .. Preferences

euGenes Report

Human* Gene *SERPINI2

Symbol SERPINI2	**Full name** serine (or cysteine) proteinase inhibitor, clade I (neuroserpin), member 2
Molecular function	serpin, tumor suppressor, Inhibitor or repressor
Biological process	protease inhibitor 14, cell motility, Cell migration/motility
Cellular component	--
Protein domains	--
Chromosome Map location	3 3q26.1-q26.2 
Ref. sequence Ref. protein	REFSEQ:NM_006217 REFPROT:NP_006208
Similar genes	*Fruitfly* Spn6 FBgn0028983 (31%); Spn2 FBgn0028987 (29%); Spn4 FBgn0028985 (29%); Spn43Aa FBgn0024294 (28%); nec FBgn0002930 (27%) *Human* SERPINI1 HUgn0005274 (39%) *Mosquito* SPI1B AGgn0013344 (30%) *Mouse* Serpini1 MGgn0011165 (39%) *Weed* At1g47710 ATgn0005921 (28%); At2g25240 ATgn0009085 (26%) *Worm* srp-6 CEgn0003919 (31%); srp-7 CEgn0008892 (31%); srp-2 CEgn0004098 (30%) *Zfish* hsp47 ZFgn0000260 (26%) *rat* -- pir ||S49162 (75%)

Interesting information

Output of Perl script

```
bbrors@durga:bin > ./gethttp.pl NM_006217
EMBL-Acc:NM_006217

Molecular function:
serpin,
tumor suppressor,
Inhibitor or repressor

Biological Process:
protease inhibitor 14,
cell motility,
Cell migration/motility
bbrors@durga:bin >
```

Problems with *text mining*

- Computers are dumb: they cannot extract the sense of written words. Thus, a number of problems can arise:
  - ⋆ Context: terms can be negated, or cited as a relation; the meaning may be reversed, and so on. Luckily, databases are structured and this problem does rarely occur.
  - ⋆ Synonyms: two terms referring to the same thing
  - ⋆ Ambiguity: one term referring to different things
  - ⋆ typos
  - ⋆ some more ...

- There is a more fundamental problem: annotation can be detailed, or rather general (like ‘phosphatase’ or ‘inositol-1,4,5-trisphosphonate 5-phosphatase’). This reflects the different depths of knowledge about a gene product. However, when summarizing information, this may lead to bias in the summary statistics (like ‘phosphatase’ occurring 20 times, while specialized terms may occur only once).

- Furthermore, the composition of a microarray may be biased as well. If a high number of the probes (probe sets) on it are annotated with ‘apoptosis’, it is no surprise if a large proportion of a gene list also bears this annotation.

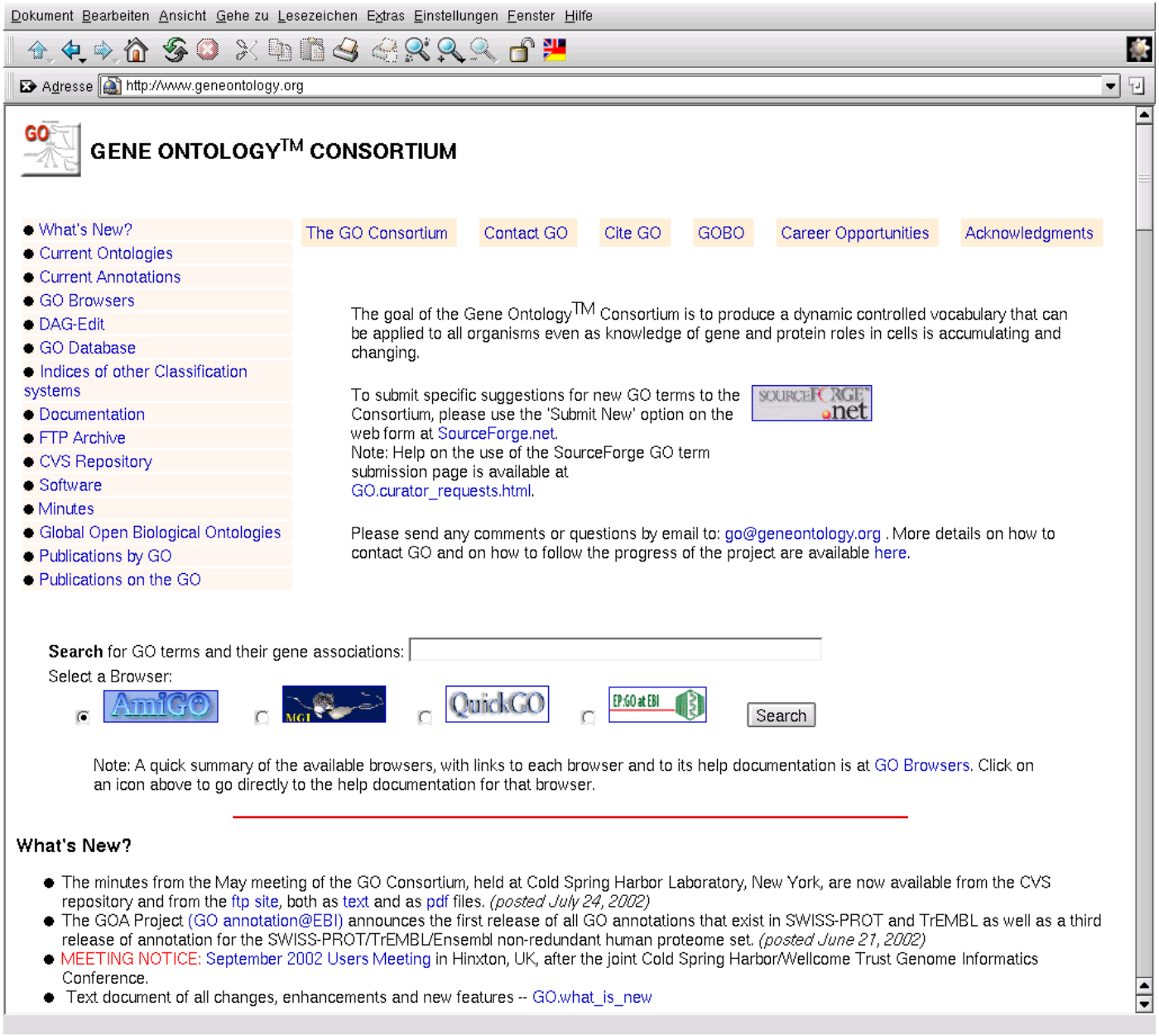
The Gene Ontology system

- To overcome some of these 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

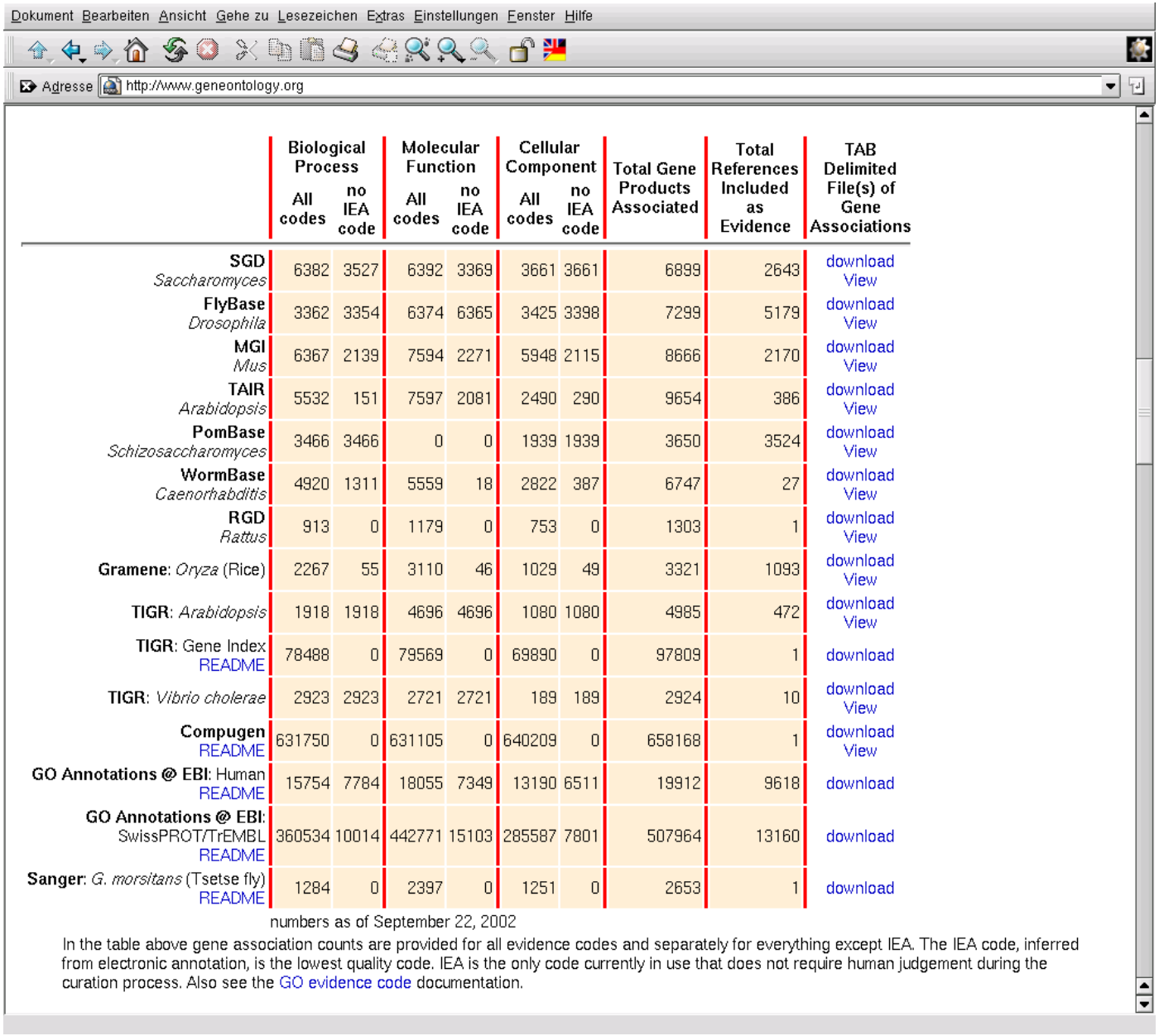


Actual annotation

- 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

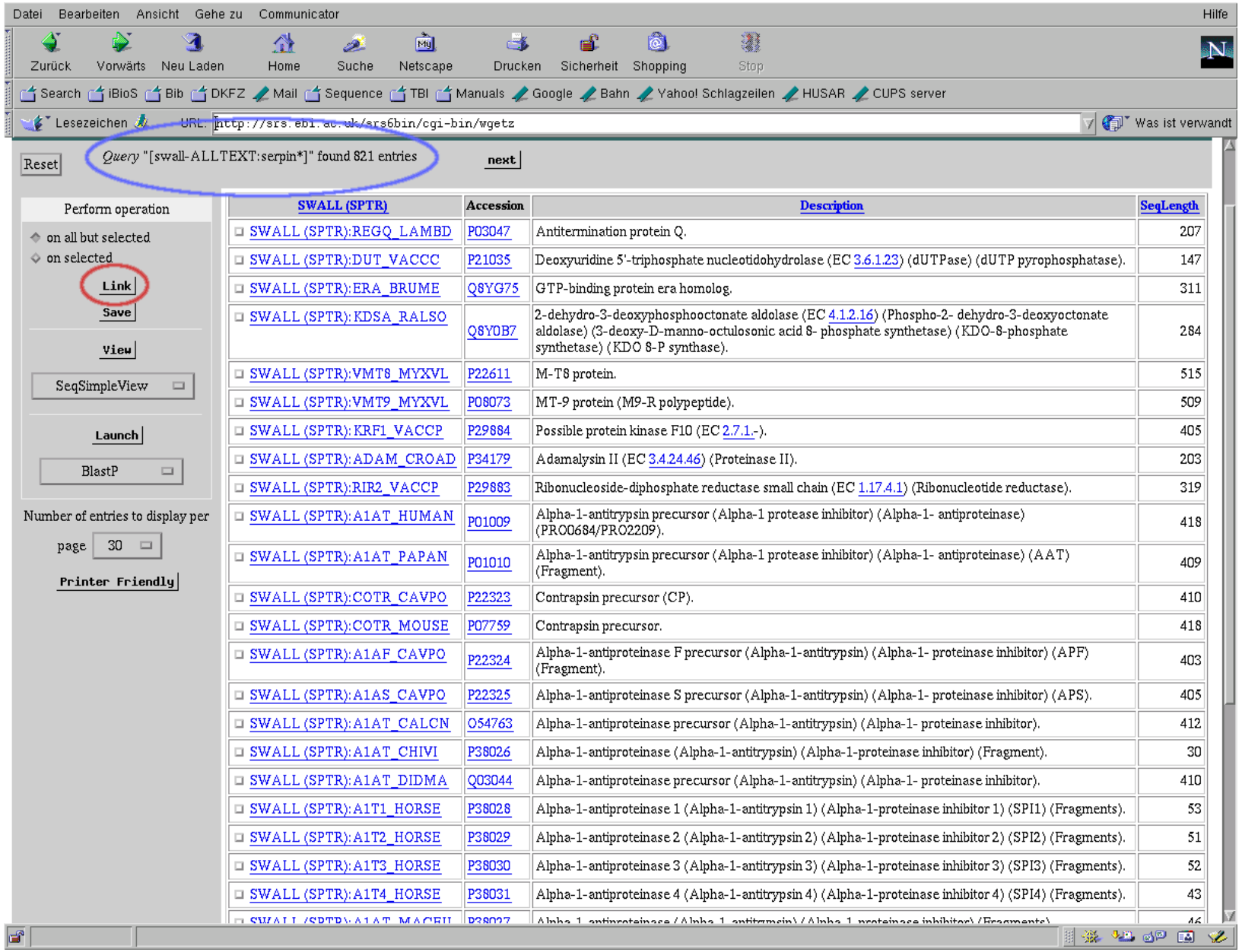
numbers as of September 22, 2002

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.

A tutorial: how to get GO information from EBI

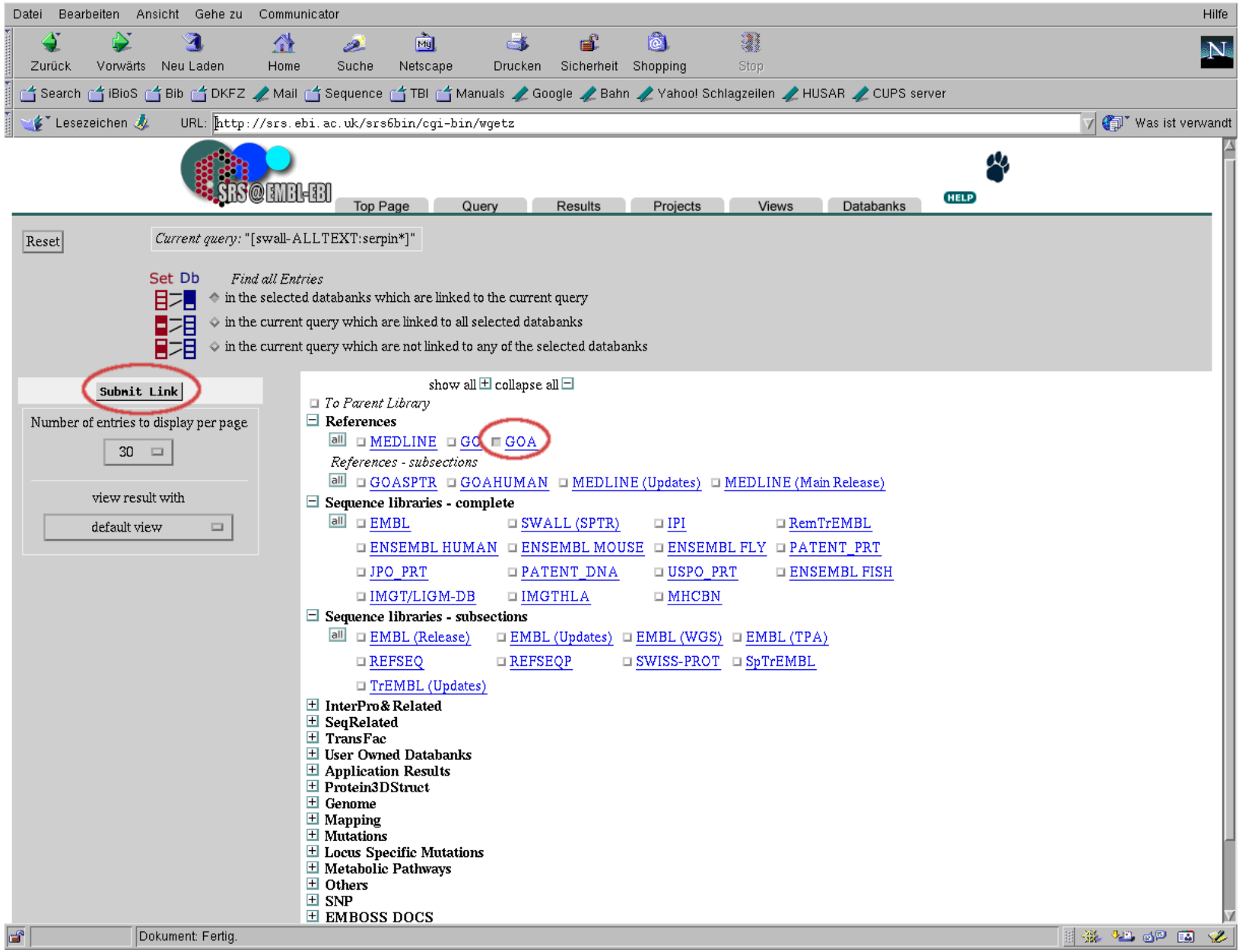
- The European Bioinformatics Institute (EBI) has started to annotate human genes (and some more) with GO terms.
- The database is called **GOA**. It is available via SRS (Sequence Retrieval System, http://srs.ebi.ac.uk).
- The following slides show you how it works.

SRS tutorial 1

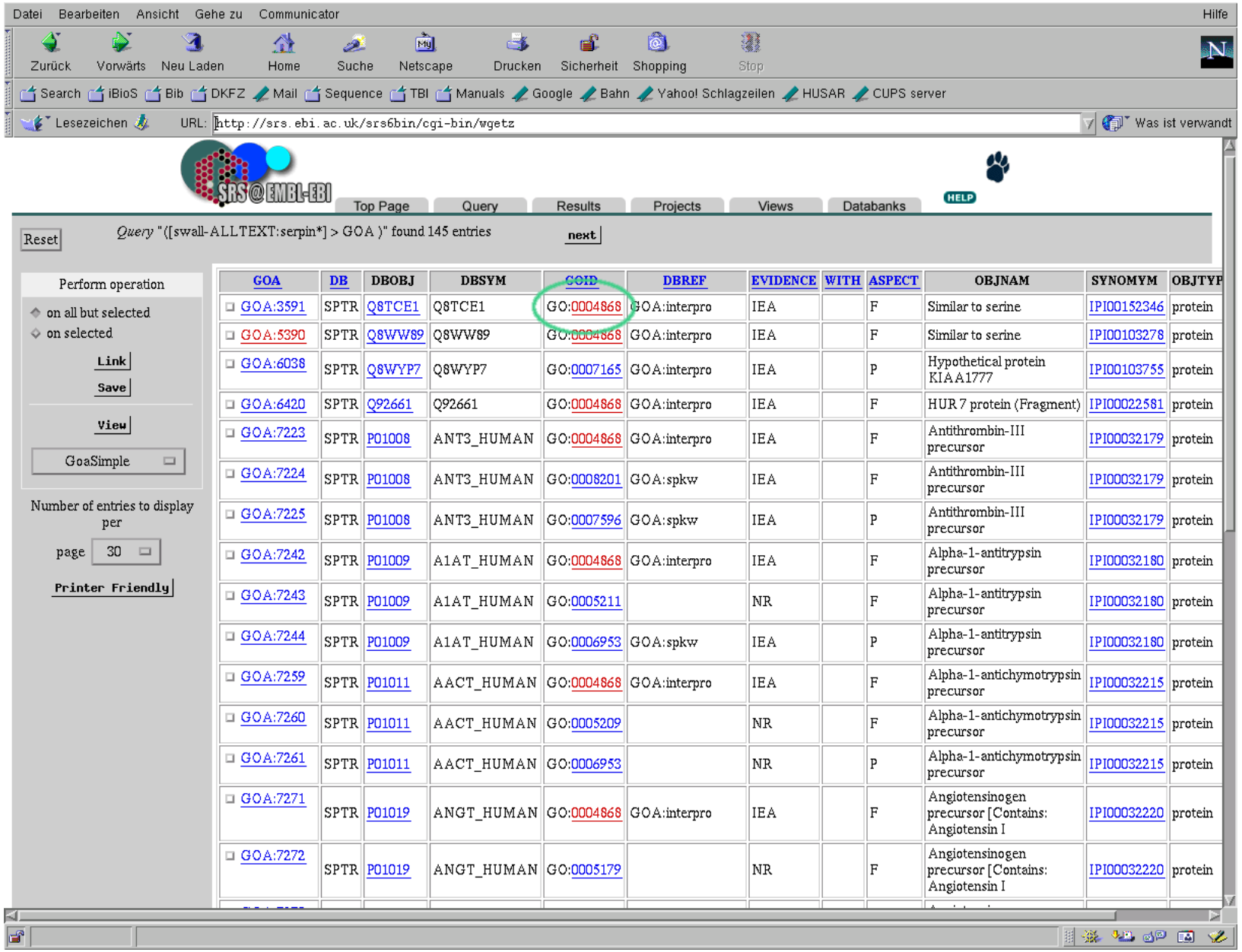


SWALL (SPTR)	Accession	Description	SeqLength
SWALL (SPTR):REGQ_LAMBD	P03047	Antitermination protein Q.	207
SWALL (SPTR):DUT_VACCC	P21035	Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase).	147
SWALL (SPTR):ERA_BRUME	Q8YG75	GTP-binding protein era homolog.	311
SWALL (SPTR):KDSA_RALSO	Q8Y0B7	2-dehydro-3-deoxyphosphooctonate aldolase (EC 4.1.2.16) (Phospho-2- dehydro-3-deoxyoctonate aldolase) (3-deoxy-D-manno-octulosonic acid 8- phosphate synthetase) (KDO-8-phosphate synthetase) (KDO 8-P synthase).	284
SWALL (SPTR):VMT8_MYXVL	P22611	M-T8 protein.	515
SWALL (SPTR):VMT9_MYXVL	P08073	MT-9 protein (M9-R polypeptide).	509
SWALL (SPTR):KRF1_VACCP	P29884	Possible protein kinase F10 (EC 2.7.1.-).	405
SWALL (SPTR):ADAM_CROAD	P34179	Adamalysin II (EC 3.4.24.46) (Proteinase II).	203
SWALL (SPTR):RIR2_VACCP	P29883	Ribonucleoside-diphosphate reductase small chain (EC 1.17.4.1) (Ribonucleotide reductase).	319
SWALL (SPTR):A1AT_HUMAN	P01009	Alpha-1-antitrypsin precursor (Alpha-1 protease inhibitor) (Alpha-1- antiproteinase) (PRO0684/PRO2209).	418
SWALL (SPTR):A1AT_PAPAN	P01010	Alpha-1-antitrypsin precursor (Alpha-1 protease inhibitor) (Alpha-1- antiproteinase) (AAT) (Fragment).	409
SWALL (SPTR):COTR_CAVPO	P22323	Contrapsin precursor (CP).	410
SWALL (SPTR):COTR_MOUSE	P07759	Contrapsin precursor.	418
SWALL (SPTR):A1AF_CAVPO	P22324	Alpha-1-antiproteinase F precursor (Alpha-1-antitrypsin) (Alpha-1- proteinase inhibitor) (APF) (Fragment).	403
SWALL (SPTR):A1AS_CAVPO	P22325	Alpha-1-antiproteinase S precursor (Alpha-1-antitrypsin) (Alpha-1- proteinase inhibitor) (APS).	405
SWALL (SPTR):A1AT_CALCN	O54763	Alpha-1-antiproteinase precursor (Alpha-1-antitrypsin) (Alpha-1- proteinase inhibitor).	412
SWALL (SPTR):A1AT_CHIVI	P38026	Alpha-1-antiproteinase (Alpha-1-antitrypsin) (Alpha-1-proteinase inhibitor) (Fragment).	30
SWALL (SPTR):A1AT_DIDMA	Q03044	Alpha-1-antiproteinase precursor (Alpha-1-antitrypsin) (Alpha-1- proteinase inhibitor).	410
SWALL (SPTR):A1T1_HORSE	P38028	Alpha-1-antiproteinase 1 (Alpha-1-antitrypsin 1) (Alpha-1-proteinase inhibitor 1) (SPI1) (Fragments).	53
SWALL (SPTR):A1T2_HORSE	P38029	Alpha-1-antiproteinase 2 (Alpha-1-antitrypsin 2) (Alpha-1-proteinase inhibitor 2) (SPI2) (Fragments).	51
SWALL (SPTR):A1T3_HORSE	P38030	Alpha-1-antiproteinase 3 (Alpha-1-antitrypsin 3) (Alpha-1-proteinase inhibitor 3) (SPI3) (Fragments).	52
SWALL (SPTR):A1T4_HORSE	P38031	Alpha-1-antiproteinase 4 (Alpha-1-antitrypsin 4) (Alpha-1-proteinase inhibitor 4) (SPI4) (Fragments).	43
SWALL (SPTR):A1AT_MACEU	P38027	Alpha-1-antiproteinase (Alpha-1-antitrypsin) (Alpha-1-proteinase inhibitor) (Fragments).	46

SRS tutorial 2



SRS tutorial 3



GOA	DB	DBOBJ	DBSYM	GOID	DBREF	EVIDENCE	WITH	ASPECT	OBJNAM	SYNOMYM	OBJTYP
GOA:3591	SPTR	Q8TCE1	Q8TCE1	GO:0004868	GOA:interpro	IEA		F	Similar to serine	IPI00152346	protein
GOA:5390	SPTR	Q8WW89	Q8WW89	GO:0004868	GOA:interpro	IEA		F	Similar to serine	IPI00103278	protein
GOA:6038	SPTR	Q8WYP7	Q8WYP7	GO:0007165	GOA:interpro	IEA		P	Hypothetical protein KIAA1777	IPI00103755	protein
GOA:6420	SPTR	Q92661	Q92661	GO:0004868	GOA:interpro	IEA		F	HUR 7 protein (Fragment)	IPI00022581	protein
GOA:7223	SPTR	P01008	ANT3_HUMAN	GO:0004868	GOA:interpro	IEA		F	Antithrombin-III precursor	IPI00032179	protein
GOA:7224	SPTR	P01008	ANT3_HUMAN	GO:0008201	GOA:spkw	IEA		F	Antithrombin-III precursor	IPI00032179	protein
GOA:7225	SPTR	P01008	ANT3_HUMAN	GO:0007596	GOA:spkw	IEA		P	Antithrombin-III precursor	IPI00032179	protein
GOA:7242	SPTR	P01009	A1AT_HUMAN	GO:0004868	GOA:interpro	IEA		F	Alpha-1-antitrypsin precursor	IPI00032180	protein
GOA:7243	SPTR	P01009	A1AT_HUMAN	GO:0005211		NR		F	Alpha-1-antitrypsin precursor	IPI00032180	protein
GOA:7244	SPTR	P01009	A1AT_HUMAN	GO:0006953	GOA:spkw	IEA		P	Alpha-1-antitrypsin precursor	IPI00032180	protein
GOA:7259	SPTR	P01011	AACT_HUMAN	GO:0004868	GOA:interpro	IEA		F	Alpha-1-antichymotrypsin precursor	IPI00032215	protein
GOA:7260	SPTR	P01011	AACT_HUMAN	GO:0005209		NR		F	Alpha-1-antichymotrypsin precursor	IPI00032215	protein
GOA:7261	SPTR	P01011	AACT_HUMAN	GO:0006953		NR		P	Alpha-1-antichymotrypsin precursor	IPI00032215	protein
GOA:7271	SPTR	P01019	ANGT_HUMAN	GO:0004868	GOA:interpro	IEA		F	Angiotensinogen precursor [Contains: Angiotensin I	IPI00032220	protein
GOA:7272	SPTR	P01019	ANGT_HUMAN	GO:0005179		NR		F	Angiotensinogen precursor [Contains: Angiotensin I	IPI00032220	protein

dkfz

The Gene Ontology hierarchy