Genome Annotation

Bioinformatics 301
David Wishart
david.wishart@ualberta.ca

Notes at: http://wishartlab.com

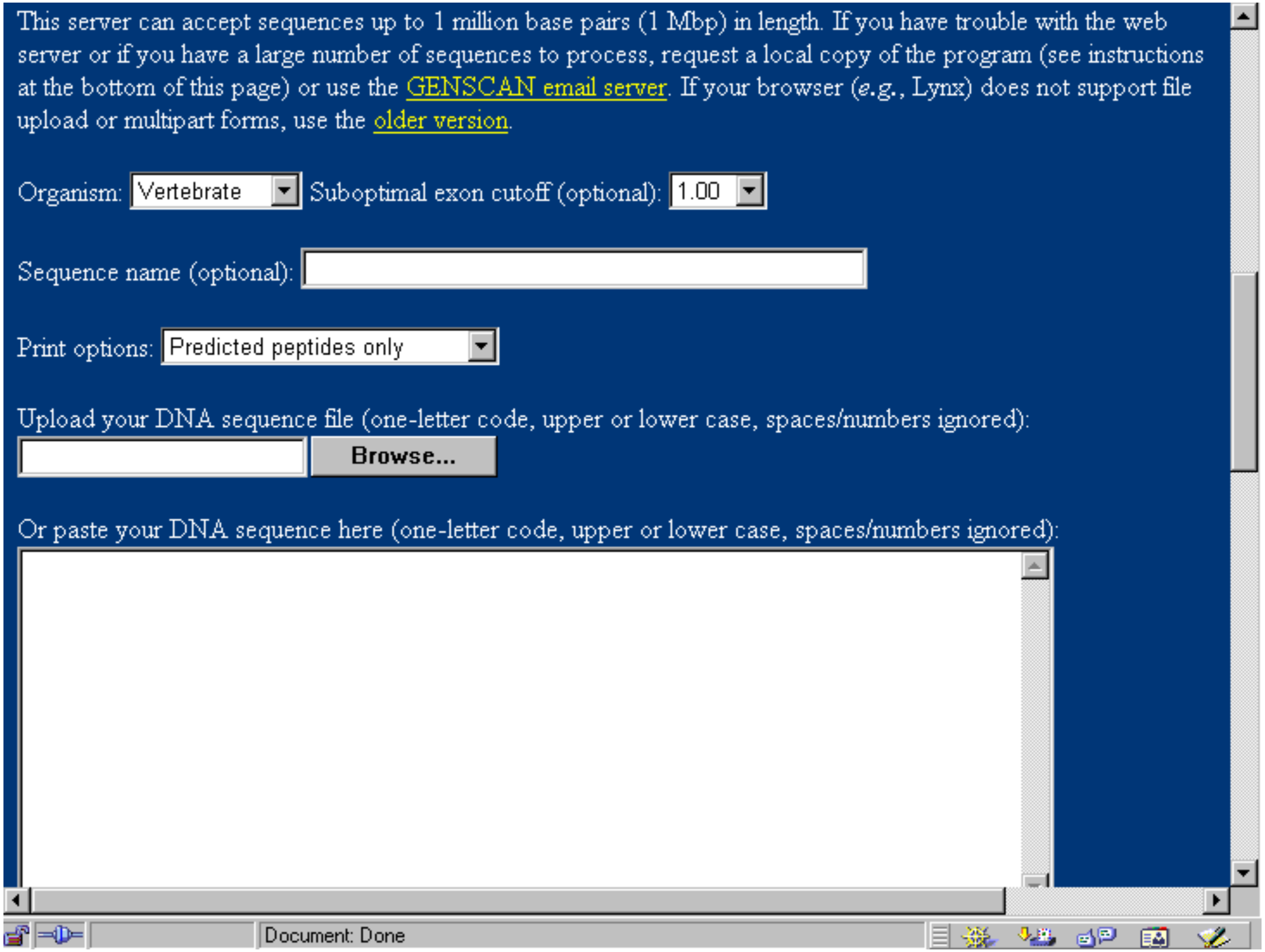
Objectives*

- To demonstrate the growing importance of gene and genome annotation in biology and the role bioinformatics plays
- To make students aware of new trends in gene and genome annotation (i.e. “deep” annotation)
- To make students aware of the methods, algorithms and tools used for gene and genome annotation

Genome Sequence

```
>P12345 Yeast chromosome1
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
```

Predict Genes



This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate ▾ Suboptimal exon cutoff (optional): 1.00 ▾

Sequence name (optional):

Print options: Predicted peptides only ▾

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Browse...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

The Result...

```
>P12346 Sequence 1
ATGTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGAT

>P12347 Sequence 2
ATGAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATT

>P12348 Sequence 3
ATGTTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACA...
```

Is This Annotated?

```
>P12346 Sequence 1
ATGTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGAT

>P12347 Sequence 2
ATGAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATT

>P12348 Sequence 3
ATGTTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACA...
```

How About This?

```
>P12346 Sequence 1
MEKGQASRTDHNMCLKPGAAERTPESTSPASDAAGG
IPQNLKGFYQALNNWLKDSQLKPPPSSGTREWAALK
LPNTHIALD

>P12347 Sequence 2
MKPQRTLNASELVISLIVESINTHISHOUSEPLEAS
EWILLITALLCEASE

>P12348 Sequence 3
MQWERTGHFDALKPQWERTYHEREISANTHERS...
```

Gene Annotation*

- ***Annotation*** **– to identify and describe all the physico-chemical, functional and structural properties of a gene including its DNA sequence, protein sequence, sequence corrections, name(s), position, function(s), abundance, location, mass, pI, absorptivity, solubility, active sites, binding sites, reactions, substrates, homologues, 2^{o} structure, 3D structure, domains, pathways, interacting partners**

Gene Annotation

||

Protein Annotation

Protein/Gene vs. Proteome/ Genome Annotation

- Gene/Protein annotation is concerned with one or a small number (<50) genes or proteins from one or several types of organisms

- Genome/Proteome annotation is concerned with entire proteomes (>2000 proteins) from a specific organism (or for all organisms) *- need for speed*

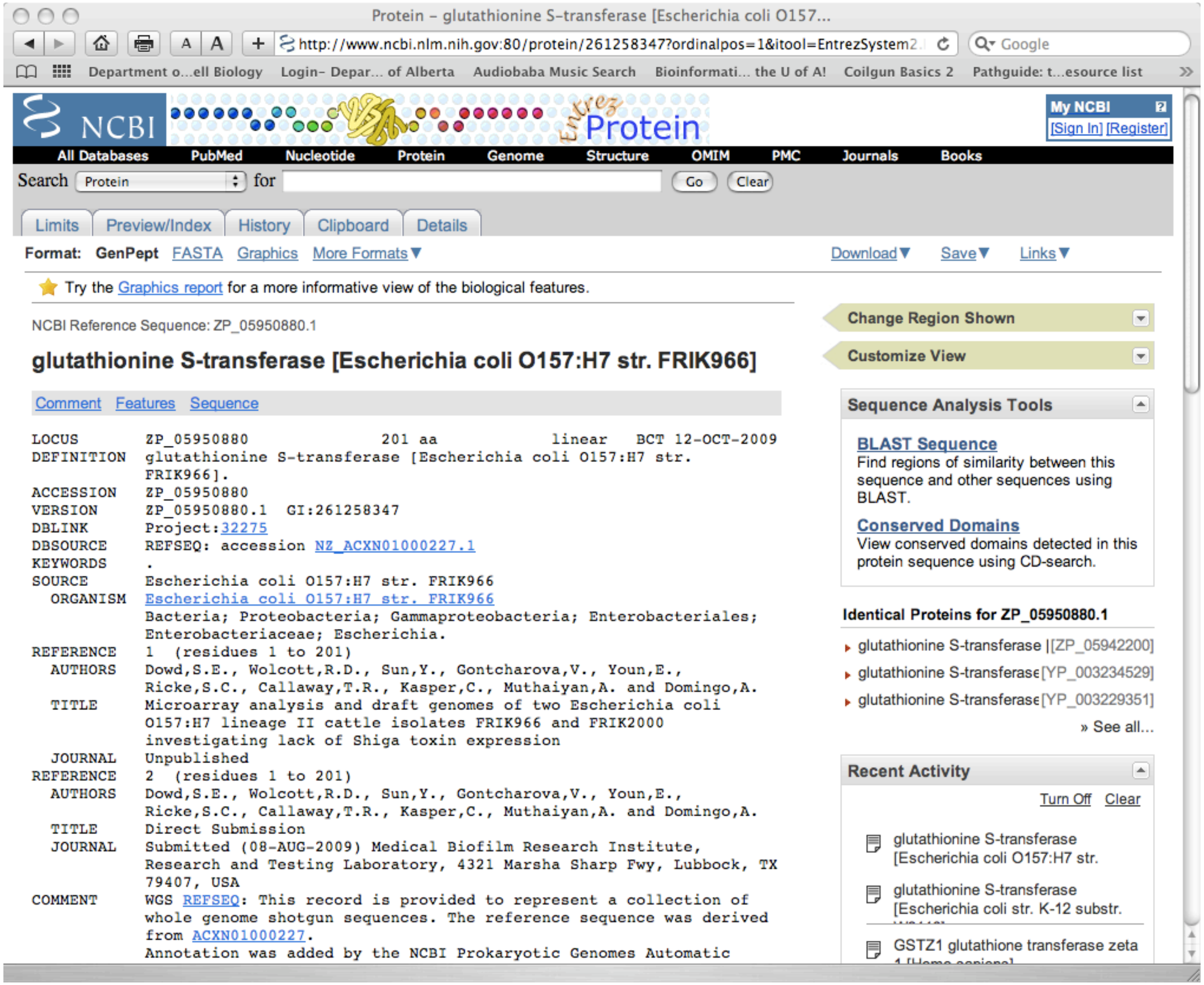
Different Levels of Annotation*

- Sparse – typical of archival databanks like GenBank, usually just includes name, depositor, accession number, dates, ID #
- Moderate – typical of many curated protein sequence databanks (UniProt or TrEMBL)
- Detailed – not typical (occasionally found in organism-specific databases)

Different Levels of Database Annotation*

- GenBank (large # of sequences, minimal annotation)
- TrEMBL (large # of sequences, slightly better [computer] annotation)
- UniProtKB (small # of sequences, even better [hand] annotation)
- Organsim-specific DB (very small # of sequences, best annotation)

GenBank Annotation (GST)



Try the Graphics report for a more informative view of the biological features.

NCBI Reference Sequence: ZP_05950880.1

glutathionine S-transferase [Escherichia coli O157:H7 str. FRIK966]

```
LOCUS       ZP_05950880             201 aa            linear   BCT 12-OCT-2009
DEFINITION  glutathionine S-transferase [Escherichia coli O157:H7 str.
            FRIK966].
ACCESSION   ZP_05950880
VERSION     ZP_05950880.1  GI:261258347
DBLINK      Project:32275
DBSOURCE    REFSEQ: accession NZ_ACXN01000227.1
KEYWORDS    .
SOURCE      Escherichia coli O157:H7 str. FRIK966
  ORGANISM  Escherichia coli O157:H7 str. FRIK966
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (residues 1 to 201)
  AUTHORS   Dowd,S.E., Wolcott,R.D., Sun,Y., Gontcharova,V., Youn,E.,
            Ricke,S.C., Callaway,T.R., Kasper,C., Muthaiyan,A. and Domingo,A.
  TITLE     Microarray analysis and draft genomes of two Escherichia coli
            O157:H7 lineage II cattle isolates FRIK966 and FRIK2000
            investigating lack of Shiga toxin expression
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 201)
  AUTHORS   Dowd,S.E., Wolcott,R.D., Sun,Y., Gontcharova,V., Youn,E.,
            Ricke,S.C., Callaway,T.R., Kasper,C., Muthaiyan,A. and Domingo,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-AUG-2009) Medical Biofilm Research Institute,
            Research and Testing Laboratory, 4321 Marsha Sharp Fwy, Lubbock, TX
            79407, USA
COMMENT     WGS REFSEQ: This record is provided to represent a collection of
            whole genome shotgun sequences. The reference sequence was derived
            from ACXN01000227.
            Annotation was added by the NCBI Prokaryotic Genomes Automatic
```

Sequence Analysis Tools

BLAST Sequence
Find regions of similarity between this sequence and other sequences using BLAST.

Conserved Domains
View conserved domains detected in this protein sequence using CD-search.

Identical Proteins for ZP_05950880.1

- glutathionine S-transferase [ZP_05942200]
- glutathionine S-transferase [YP_003234529]
- glutathionine S-transferase [YP_003229351]

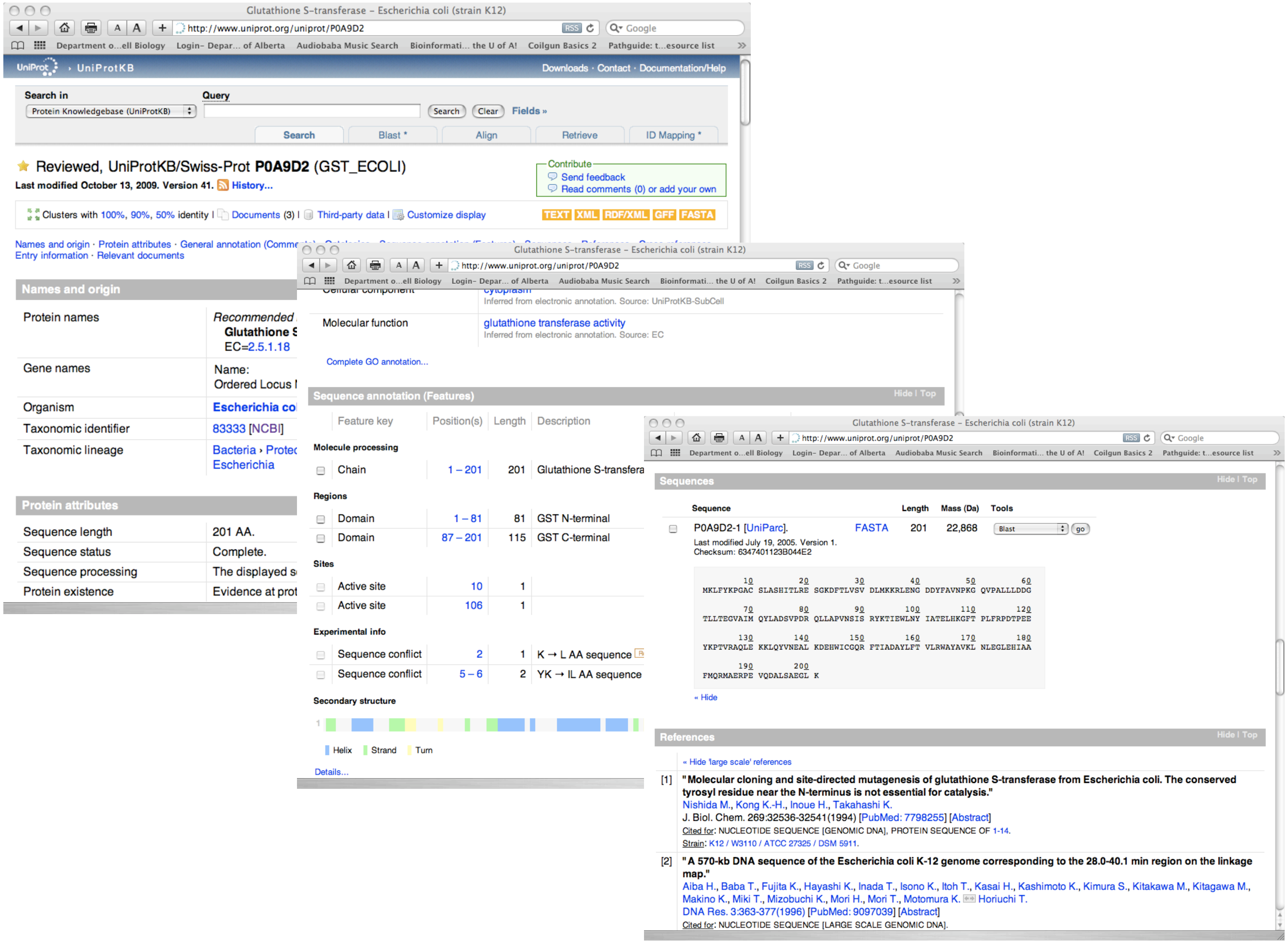
» See all...

Recent Activity

Turn Off Clear

- glutathionine S-transferase [Escherichia coli O157:H7 str.
- glutathionine S-transferase [Escherichia coli str. K-12 substr.
- GSTZ1 glutathione transferase zeta

UniProtKB Annotation (GST)


Glutathione S–transferase – Escherichia coli (strain K12)
http://www.uniprot.org/uniprot/P0A9D2
RSS
Google
Department o...ell Biology
Login– Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list
UniProt
UniProtKB
Downloads · Contact · Documentation/Help
Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Clear
Fields »
Search
Blast *
Align
Retrieve
ID Mapping *
Contribute
Send feedback
Read comments (0) or add your own
TEXT
XML
RDF/XML
GFF
FASTA
Hide | Top
Blast
go

★ Reviewed, UniProtKB/Swiss-Prot **P0A9D2** (GST_ECOLI)

Last modified October 13, 2009. Version 41. History...

Clusters with 100%, 90%, 50% identity | Documents (3) | Third-party data | Customize display

Names and origin · Protein attributes · General annotation (Comme
Entry information · Relevant documents

Names and origin

Protein names	*Recommended* **Glutathione S** EC=2.5.1.18
Gene names	Name: Ordered Locus N
Organism	**Escherichia col**
Taxonomic identifier	83333 [NCBI]
Taxonomic lineage	Bacteria › Protec Escherichia

Protein attributes

Sequence length	201 AA.
Sequence status	Complete.
Sequence processing	The displayed s
Protein existence	Evidence at prot

Cellular component	cytoplasm Inferred from electronic annotation. Source: UniProtKB-SubCell
Molecular function	glutathione transferase activity Inferred from electronic annotation. Source: EC

Complete GO annotation...

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description
Molecule processing				
□	Chain	1 – 201	201	Glutathione S-transfera
Regions				
□	Domain	1 – 81	81	GST N-terminal
□	Domain	87 – 201	115	GST C-terminal
Sites				
□	Active site	10	1	
□	Active site	106	1	
Experimental info				
□	Sequence conflict	2	1	K → L AA sequence
□	Sequence conflict	5 – 6	2	YK → IL AA sequence

Secondary structure

1

🟦 Helix 🟩 Strand 🟨 Turn

Details...

Sequences

	Sequence		Length	Mass (Da)	Tools
□	P0A9D2-1 [UniParc]. Last modified July 19, 2005. Version 1. Checksum: 6347401123B044E2	FASTA	201	22,868	Blast go

```
        10         20         30         40         50         60
MKLFYKPGAC SLASHITLRE SGKDFTLVSV DLMKKRLENG DDYFAVNPKG QVPALLLDDG

        70         80         90        100        110        120
TLLTEGVAIM QYLADSVPDR QLLAPVNSIS RYKTIEWLNY IATELHKGFT PLFRPDTPEE

       130        140        150        160        170        180
YKPTVRAQLE KKLQYVNEAL KDEHWICGQR FTIADAYLFT VLRWAYAVKL NLEGLEHIAA

       190        200
FMQRMAERPE VQDALSAEGL K
```

« Hide

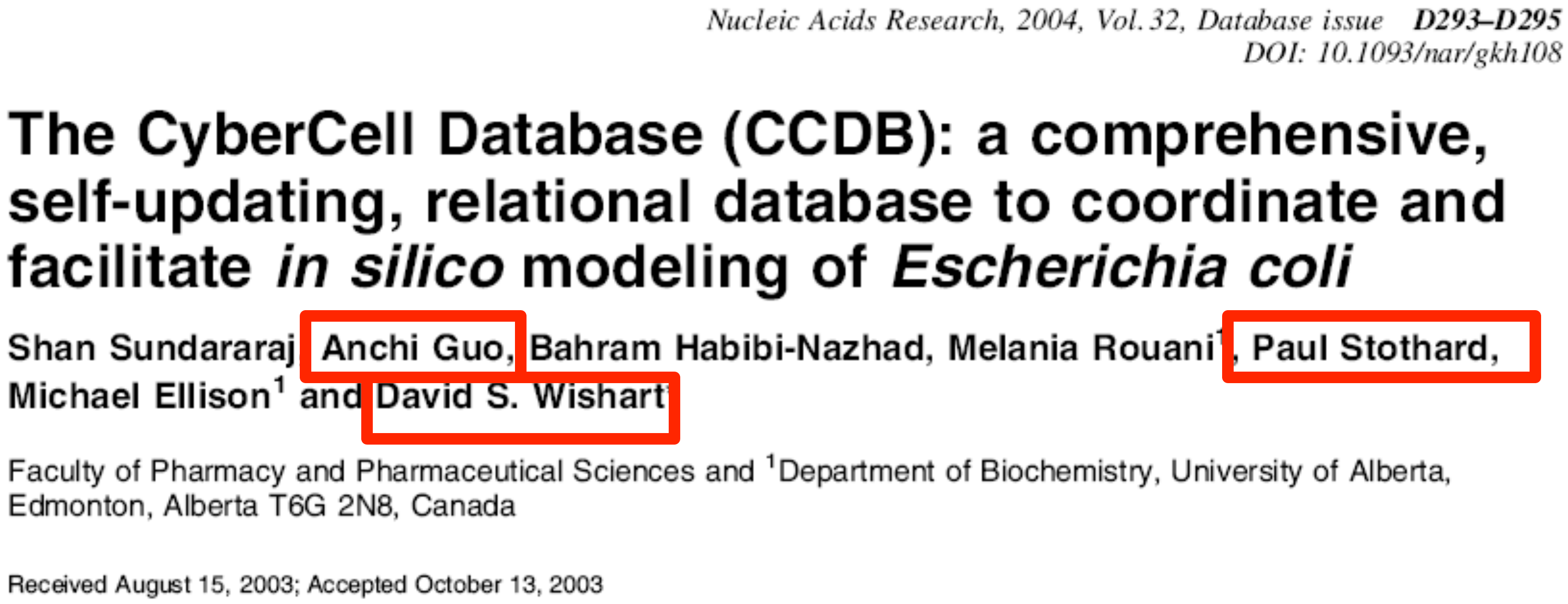
References

« Hide 'large scale' references

[1] **"Molecular cloning and site-directed mutagenesis of glutathione S-transferase from Escherichia coli. The conserved tyrosyl residue near the N-terminus is not essential for catalysis."**
Nishida M., Kong K.-H., Inoue H., Takahashi K.
J. Biol. Chem. 269:32536-32541(1994) [PubMed: 7798255] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 1-14.
Strain: K12 / W3110 / ATCC 27325 / DSM 5911.

[2] **"A 570-kb DNA sequence of the Escherichia coli K-12 genome corresponding to the 28.0-40.1 min region on the linkage map."**
Aiba H., Baba T., Fujita K., Hayashi K., Inada T., Isono K., Itoh T., Kasai H., Kashimoto K., Kimura S., Kitakawa M., Kitagawa M., Makino K., Miki T., Mizobuchi K., Mori H., Mori T., Motomura K. ⇔ Horiuchi T.
DNA Res. 3:363-377(1996) [PubMed: 9097039] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].

The CCDB*



Nucleic Acids Research, 2004, Vol. 32, Database issue **D293–D295**
DOI: 10.1093/nar/gkh108

The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*

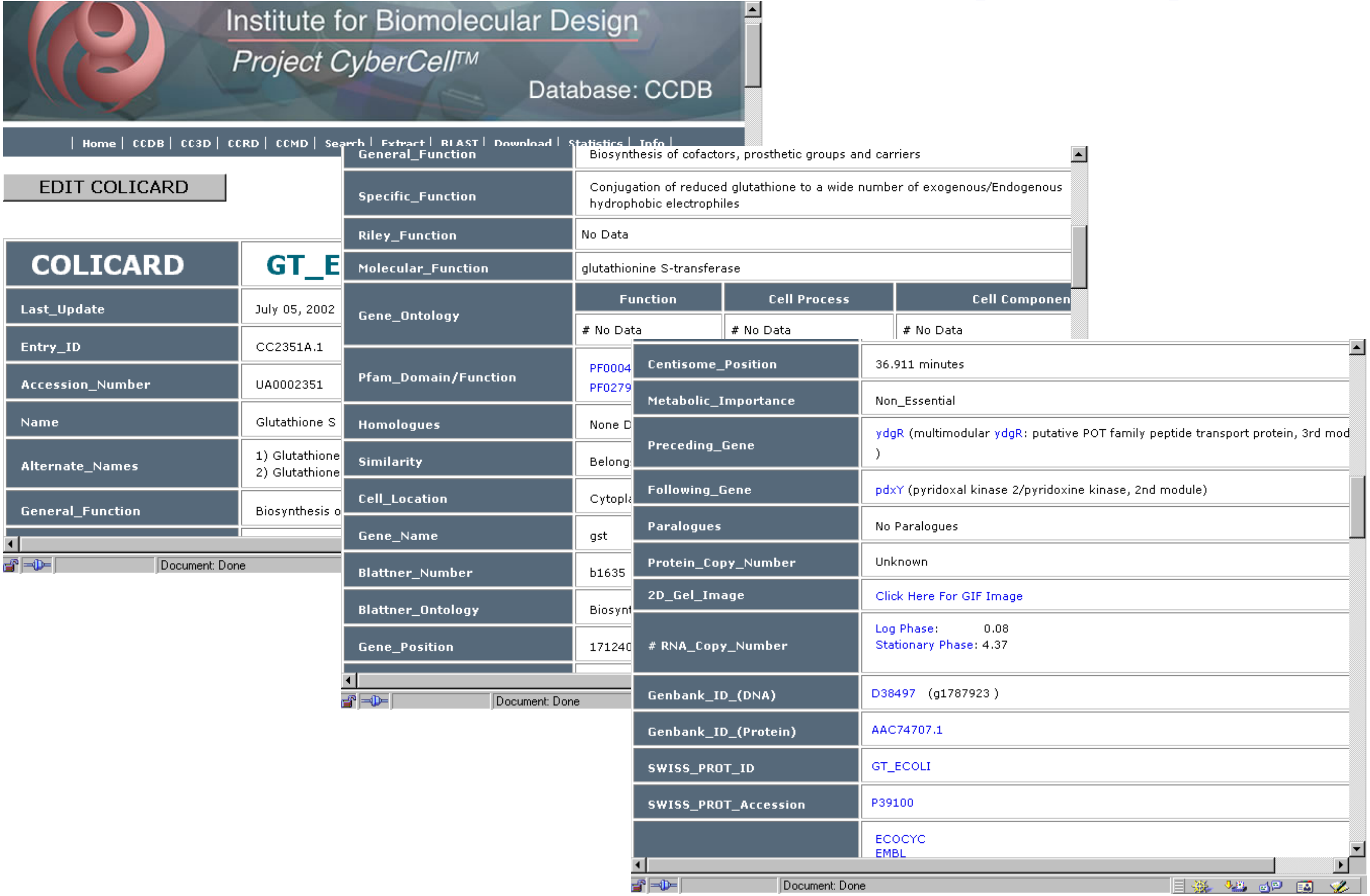
Shan Sundararaj, Anchi Guo, Bahram Habibi-Nazhad, Melania Rouani[1], Paul Stothard, Michael Ellison[1] and David S. Wishart

Faculty of Pharmacy and Pharmaceutical Sciences and [1]Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2N8, Canada

Received August 15, 2003; Accepted October 13, 2003

http://ccdb.wishartlab.com/CCDB/

CCDB Annotation (GST)

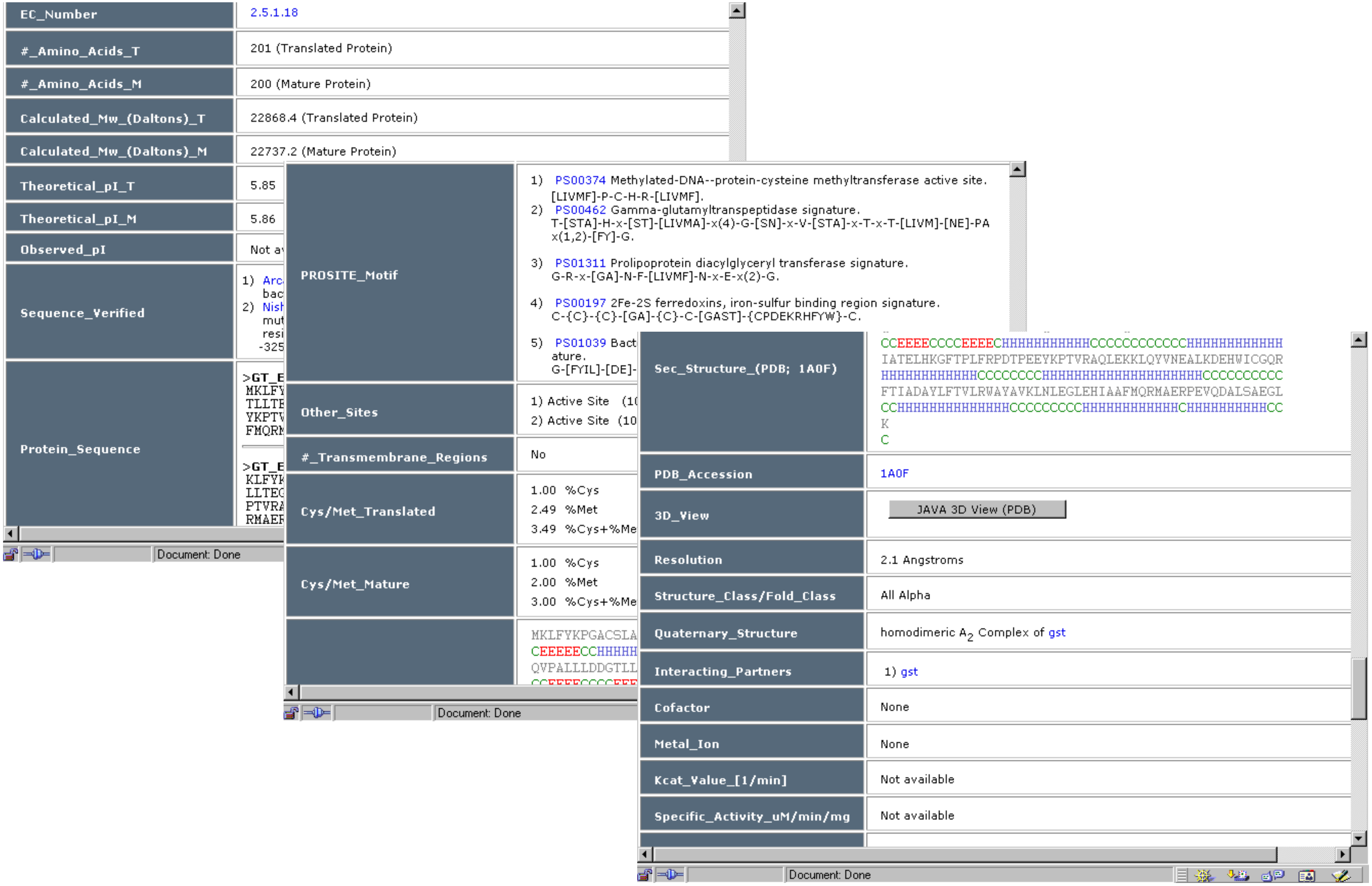

Institute for Biomolecular Design
Project CyberCell™
Database: CCDB
| Home | CCDB | CC3D | CCRD | CCMD | Search | Extract | BLAST | Download | Statistics | Info |
EDIT COLICARD
Document: Done

COLICARD	GT_E
Last_Update	July 05, 2002
Entry_ID	CC2351A.1
Accession_Number	UA0002351
Name	Glutathione S
Alternate_Names	1) Glutathione 2) Glutathione
General_Function	Biosynthesis o

General_Function	Biosynthesis of cofactors, prosthetic groups and carriers		
Specific_Function	Conjugation of reduced glutathione to a wide number of exogenous/Endogenous hydrophobic electrophiles		
Riley_Function	No Data		
Molecular_Function	glutathionine S-transferase		
Gene_Ontology	Function	Cell Process	Cell Componen
	# No Data	# No Data	# No Data
Pfam_Domain/Function	PF0004 PF0279		
Homologues	None D		
Similarity	Belong		
Cell_Location	Cytopl		
Gene_Name	gst		
Blattner_Number	b1635		
Blattner_Ontology	Biosynt		
Gene_Position	171240		

Centisome_Position	36.911 minutes
Metabolic_Importance	Non_Essential
Preceding_Gene	ydgR (multimodular ydgR: putative POT family peptide transport protein, 3rd mod)
Following_Gene	pdxY (pyridoxal kinase 2/pyridoxine kinase, 2nd module)
Paralogues	No Paralogues
Protein_Copy_Number	Unknown
2D_Gel_Image	Click Here For GIF Image
# RNA_Copy_Number	Log Phase: 0.08 Stationary Phase: 4.37
Genbank_ID_(DNA)	D38497 (g1787923)
Genbank_ID_(Protein)	AAC74707.1
SWISS_PROT_ID	GT_ECOLI
SWISS_PROT_Accession	P39100
	ECOCYC EMBL

CCDB Annotation



EC_Number	2.5.1.18
#_Amino_Acids_T	201 (Translated Protein)
#_Amino_Acids_M	200 (Mature Protein)
Calculated_Mw_(Daltons)_T	22868.4 (Translated Protein)
Calculated_Mw_(Daltons)_M	22737.2 (Mature Protein)
Theoretical_pI_T	5.85
Theoretical_pI_M	5.86
Observed_pI	Not a
Sequence_Verified	1) Arc bac 2) Nish mut resi -325
Protein_Sequence	>GT_E MKLFY TLLTE YKPTV FMQRM >GT_E KLFYK LLTEG PTVRA RMAER

PROSITE_Motif	1) PS00374 Methylated-DNA--protein-cysteine methyltransferase active site. [LIVMF]-P-C-H-R-[LIVMF]. 2) PS00462 Gamma-glutamyltranspeptidase signature. T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-PA x(1,2)-[FY]-G. 3) PS01311 Prolipoprotein diacylglyceryl transferase signature. G-R-x-[GA]-N-F-[LIVMF]-N-x-E-x(2)-G. 4) PS00197 2Fe-2S ferredoxins, iron-sulfur binding region signature. C-{C}-{C}-[GA]-{C}-C-[GAST]-{CPDEKRHFYW}-C. 5) PS01039 Bact ature. G-[FYIL]-[DE]-
Other_Sites	1) Active Site (1(2) Active Site (10
#_Transmembrane_Regions	No
Cys/Met_Translated	1.00 %Cys 2.49 %Met 3.49 %Cys+%Me
Cys/Met_Mature	1.00 %Cys 2.00 %Met 3.00 %Cys+%Me
	MKLFYKPGACSLA CEEEEECCHHHHH QVPALLLDDGTLL CCEEEECCCCEEE

Sec_Structure_(PDB; 1A0F)	CCEEEECCCCEEEECHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHH IATELHKGFTPLFRPDTPEEYKPTVRAQLEKKLQYVNEALKDEHWICGQR HHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCCCCCCCC FTIADAYLFTVLRWAYAVKLNLEGLEHIAAFMQRMAERPEVQDALSAEGL CCHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHCHHHHHHHHHHCC K C
PDB_Accession	1A0F
3D_View	JAVA 3D View (PDB)
Resolution	2.1 Angstroms
Structure_Class/Fold_Class	All Alpha
Quaternary_Structure	homodimeric A_2 Complex of gst
Interacting_Partners	1) gst
Cofactor	None
Metal_Ion	None
Kcat_Value_[1/min]	Not available
Specific_Activity_uM/min/mg	Not available

CCDB Contents*

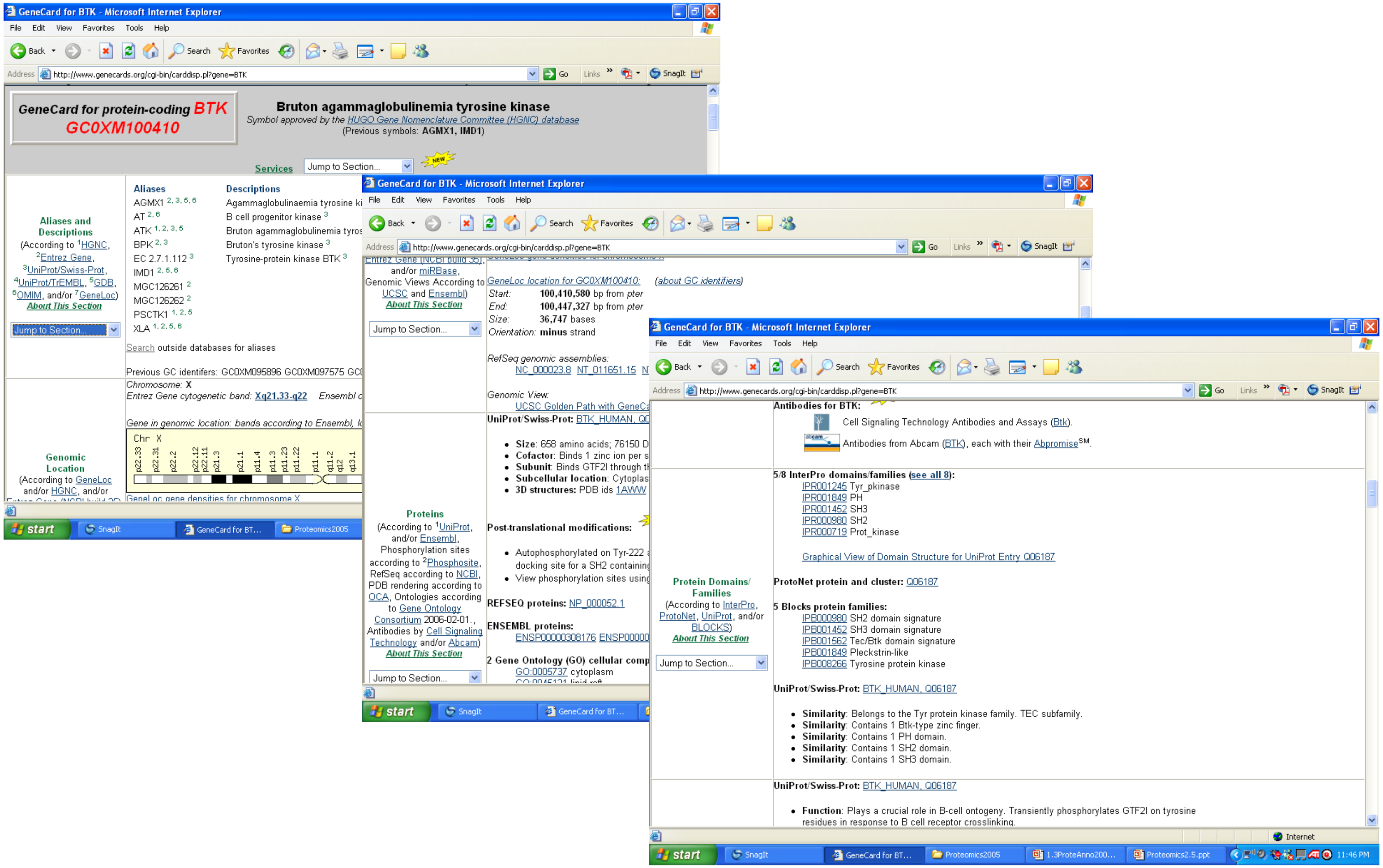
- Functional info (predicted or known)
- Sequence information (sites, modifications, pI, MW, cleavage)
- Location information (in chromosome & cell)
- Interacting partners (known & predicted)
- Structure (2^o, 3^o, 4^o, predicted)
- Enzymatic rate and binding constants
- Abundance, copy number, concentration
- Links to other sites & viewing tools
- Integrated version of all major Db’s
- *70+ fields for each entry*

GeneCards Content

- Aliases
- Databases
- Disorders
- Domains
- Drugs/Cmpds
- Expression
- Function
- Location
- Orthologs/Paralogs
- Pathways and Interactions
- References
- Proteins/MAbs
- SNPs
- Transcripts
- Gene Maps

http://www.genecards.org/index.shtml

GeneCards Annotation


GeneCard for BTK - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.genecards.org/cgi-bin/carddisp.pl?gene=BTK
Go Links SnagIt
start SnagIt GeneCard for BT... Proteomics2005 1.3ProteAnno200... Proteomics2.5.ppt
Internet
11:46 PM

GeneCard for protein-coding BTK
GC0XM100410

Bruton agammaglobulinemia tyrosine kinase
Symbol approved by the HUGO Gene Nomenclature Committee (HGNC) database
(Previous symbols: **AGMX1, IMD1**)

Services Jump to Section...

Aliases and Descriptions
(According to [1]HGNC, [2]Entrez Gene, [3]UniProt/Swiss-Prot, [4]UniProt/TrEMBL, [5]GDB, [6]OMIM, and/or [7]GeneLoc)
About This Section

Jump to Section...

Aliases	Descriptions
AGMX1 [2, 3, 5, 6]	Agammaglobulinaemia tyrosine ki
AT [2, 6]	B cell progenitor kinase [3]
ATK [1, 2, 3, 5]	Bruton agammaglobulinemia tyros
BPK [2, 3]	Bruton's tyrosine kinase [3]
EC 2.7.1.112 [3]	Tyrosine-protein kinase BTK [3]
IMD1 [2, 5, 6]	
MGC126261 [2]	
MGC126262 [2]	
PSCTK1 [1, 2, 5]	
XLA [1, 2, 5, 6]	

Search outside databases for aliases

Previous GC identifers: GC0XM095896 GC0XM097575 GC

Genomic Location
(According to GeneLoc and/or HGNC, and/or Entrez Gene (NCBI build 35), and/or miRBase, Genomic Views According to UCSC and Ensembl)
About This Section

Jump to Section...

Chromosome: **X**
Entrez Gene cytogenetic band: **Xq21.33-q22** *Ensembl c*

Gene in genomic location: bands according to Ensembl, l

Chr X
p22.33 p22.31 p22.2 p22.12 p22.11 p21.3 p21.1 p11.4 p11.3 p11.23 p11.22 p11.1 q11.2 q12 q13.1

GeneLoc gene densities for chromosome X

GeneLoc location for GC0XM100410: *(about GC identifiers)*

Start:	**100,410,580** bp from *pter*
End:	**100,447,327** bp from *pter*
Size:	**36,747** bases
Orientation:	**minus** strand

RefSeq genomic assemblies:
NC_000023.8 NT_011651.15 N

Genomic View:
UCSC Golden Path with GeneCa

Proteins
(According to [1]UniProt, and/or Ensembl, Phosphorylation sites according to [2]Phosphosite, RefSeq according to NCBI, PDB rendering according to OCA, Ontologies according to Gene Ontology Consortium 2006-02-01., Antibodies by Cell Signaling Technology and/or Abcam)
About This Section

Jump to Section...

UniProt/Swiss-Prot: BTK_HUMAN, Q0

- **Size**: 658 amino acids; 76150 D
- **Cofactor**: Binds 1 zinc ion per s
- **Subunit**: Binds GTF2I through th
- **Subcellular location**: Cytoplas
- **3D structures:** PDB ids 1AWW

Post-translational modifications:

- Autophosphorylated on Tyr-222 a docking site for a SH2 containin
- View phosphorylation sites using

REFSEQ proteins: NP_000052.1

ENSEMBL proteins:
ENSP00000308176 ENSP00000

2 Gene Ontology (GO) cellular comp
GO:0005737 cytoplasm
GO:0045121 lipid raft

Antibodies for BTK:
Cell Signaling Technology Antibodies and Assays (Btk).
Antibodies from Abcam (BTK), each with their Abpromise^{SM}.

Protein Domains/ Families
(According to InterPro, ProtoNet, UniProt, and/or BLOCKS)
About This Section

Jump to Section...

5/8 InterPro domains/families (see all 8):
IPR001245 Tyr_pkinase
IPR001849 PH
IPR001452 SH3
IPR000980 SH2
IPR000719 Prot_kinase

Graphical View of Domain Structure for UniProt Entry Q06187

ProtoNet protein and cluster: Q06187

5 Blocks protein families:
IPB000980 SH2 domain signature
IPB001452 SH3 domain signature
IPB001562 Tec/Btk domain signature
IPB001849 Pleckstrin-like
IPB008266 Tyrosine protein kinase

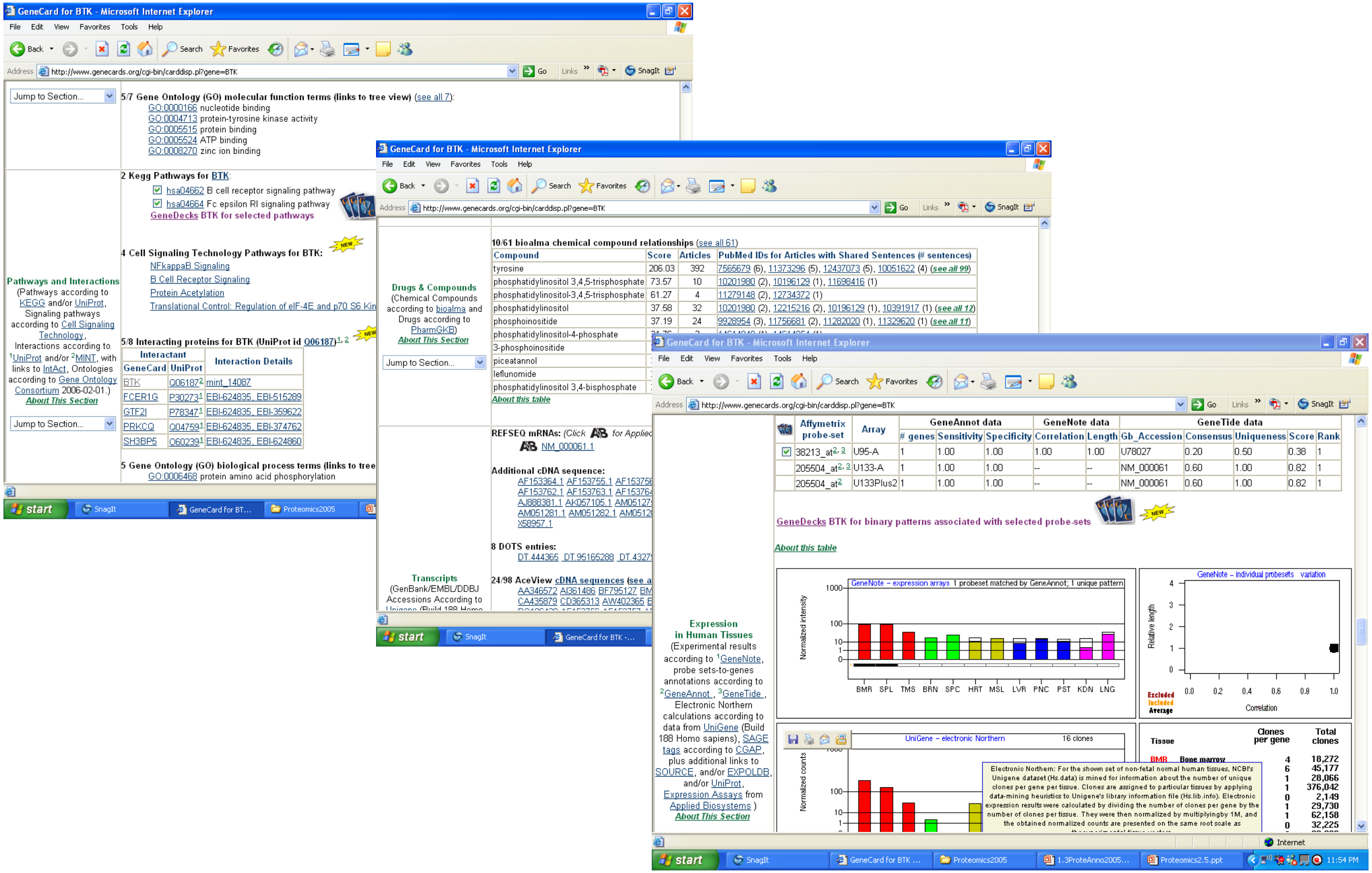
UniProt/Swiss-Prot: BTK_HUMAN, Q06187

- **Similarity**: Belongs to the Tyr protein kinase family. TEC subfamily.
- **Similarity**: Contains 1 Btk-type zinc finger.
- **Similarity**: Contains 1 PH domain.
- **Similarity**: Contains 1 SH2 domain.
- **Similarity**: Contains 1 SH3 domain.

UniProt/Swiss-Prot: BTK_HUMAN, Q06187

- **Function**: Plays a crucial role in B-cell ontogeny. Transiently phosphorylates GTF2I on tyrosine residues in response to B cell receptor crosslinking.

GeneCards Annotation



GeneCard for BTK - Microsoft Internet Explorer

Address: http://www.genecards.org/cgi-bin/carddisp.pl?gene=BTK

5/7 Gene Ontology (GO) molecular function terms (links to tree view) (see all 7):
GO:0000166 nucleotide binding
GO:0004713 protein-tyrosine kinase activity
GO:0005515 protein binding
GO:0005524 ATP binding
GO:0008270 zinc ion binding

Pathways and Interactions
(Pathways according to KEGG and/or UniProt, Signaling pathways according to Cell Signaling Technology, Interactions according to [1]UniProt and/or [2]MINT, with links to IntAct, Ontologies according to Gene Ontology Consortium 2006-02-01.)
About This Section

2 Kegg Pathways for BTK:
☑ hsa04662 B cell receptor signaling pathway
☑ hsa04664 Fc epsilon RI signaling pathway
GeneDecks BTK for selected pathways

4 Cell Signaling Technology Pathways for BTK:
NFkappaB Signaling
B Cell Receptor Signaling
Protein Acetylation
Translational Control: Regulation of eIF-4E and p70 S6 Kin

5/8 Interacting proteins for BTK (UniProt id Q06187)[1, 2]

Interactant		Interaction Details
GeneCard	UniProt	
BTK	Q06187[2]	mint_14087
FCER1G	P30273[1]	EBI-624835, EBI-515289
GTF2I	P78347[1]	EBI-624835, EBI-359622
PRKCQ	Q04759[1]	EBI-624835, EBI-374762
SH3BP5	O60239[1]	EBI-624835, EBI-624860

5 Gene Ontology (GO) biological process terms (links to tree
GO:0006468 protein amino acid phosphorylation

GeneCard for BTK - Microsoft Internet Explorer

Address: http://www.genecards.org/cgi-bin/carddisp.pl?gene=BTK

Drugs & Compounds
(Chemical Compounds according to bioalma and Drugs according to PharmGKB)
About This Section

10/61 bioalma chemical compound relationships (see all 61)

Compound	Score	Articles	PubMed IDs for Articles with Shared Sentences (# sentences)
tyrosine	206.03	392	7565679 (6), 11373296 (5), 12437073 (5), 10051622 (4) (*see all 99*)
phosphatidylinositol 3,4,5-trisphosphate	73.57	10	10201980 (2), 10196129 (1), 11698416 (1)
phosphatidylinositol-3,4,5-trisphosphate	61.27	4	11279148 (2), 12734372 (1)
phosphatidylinositol	37.58	32	10201980 (2), 12215216 (2), 10196129 (1), 10391917 (1) (*see all 12*)
phosphoinositide	37.19	24	9928954 (3), 11756681 (2), 11282020 (1), 11329620 (1) (*see all 11*)
phosphatidylinositol-4-phosphate			
3-phosphoinositide			
piceatannol			
leflunomide			
phosphatidylinositol 3,4-bisphosphate			

About this table

Transcripts
(GenBank/EMBL/DDBJ Accessions According to Unigene (Build 188 Homo

REFSEQ mRNAs: *(Click AB for Applied*
AB NM_000061.1

Additional cDNA sequence:
AF153364.1 AF153755.1 AF15375
AF153762.1 AF153763.1 AF15376
AJ888381.1 AK057105.1 AM05127
AM051281.1 AM051282.1 AM0512
X58957.1

8 DOTS entries:
DT.444365 DT.95165288 DT.4327

24/98 AceView cDNA sequences (see a
AA346572 AI361486 BF795127 BM
CA435879 CD365313 AW402365 E

GeneCard for BTK - Microsoft Internet Explorer

Address: http://www.genecards.org/cgi-bin/carddisp.pl?gene=BTK

	Affymetrix probe-set	Array	GeneAnnot data			GeneNote data		GeneTide data				
			# genes	Sensitivity	Specificity	Correlation	Length	Gb_Accession	Consensus	Uniqueness	Score	Rank
☑	38213_at[2, 3]	U95-A	1	1.00	1.00	1.00	1.00	U78027	0.20	0.50	0.38	1
	205504_at[2, 3]	U133-A	1	1.00	1.00	--	--	NM_000061	0.60	1.00	0.82	1
	205504_at[2]	U133Plus2	1	1.00	1.00	--	--	NM_000061	0.60	1.00	0.82	1

GeneDecks BTK for binary patterns associated with selected probe-sets

About this table

Expression in Human Tissues
(Experimental results according to [1]GeneNote, probe sets-to-genes annotations according to [2]GeneAnnot , [3]GeneTide , Electronic Northern calculations according to data from UniGene (Build 188 Homo sapiens), SAGE tags according to CGAP, plus additional links to SOURCE, and/or EXPOLDB, and/or UniProt, Expression Assays from Applied Biosystems)
About This Section

GeneNote – expression arrays 1 probeset matched by GeneAnnot; 1 unique pattern
Normalized intensity
1000
100
10
1
0
BMR SPL TMS BRN SPC HRT MSL LVR PNC PST KDN LNG
GeneNote – individual probesets variation
Relative length
4
3
2
1
0
0.0 0.2 0.4 0.6 0.8 1.0
Correlation
Excluded
Included
Average
UniGene – electronic Northern
16 clones
Normalized counts
100
10
1

Electronic Northern: For the shown set of non-fetal normal human tissues, NCBI's Unigene dataset (Hs.data) is mined for information about the number of unique clones per gene per tissue. Clones are assigned to particular tissues by applying data-mining heuristics to Unigene's library information file (Hs.lib.info). Electronic expression results were calculated by dividing the number of clones per gene by the number of clones per tissue. They were then normalized by multiplyingby 1M, and the obtained normalized counts are presented on the same root scale as

Tissue		Clones per gene	Total clones
BMR	Bone marrow	4	18,272
		6	45,177
		1	28,066
		1	376,042
		0	2,149
		1	29,730
		1	62,158
		0	32,225

Ultimate Goal...

- To achieve the same level of protein/ proteome annotation as found in CCDB or GeneCards for all genes/proteins -- ***automatically***

How?

Annotation Methods*

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools

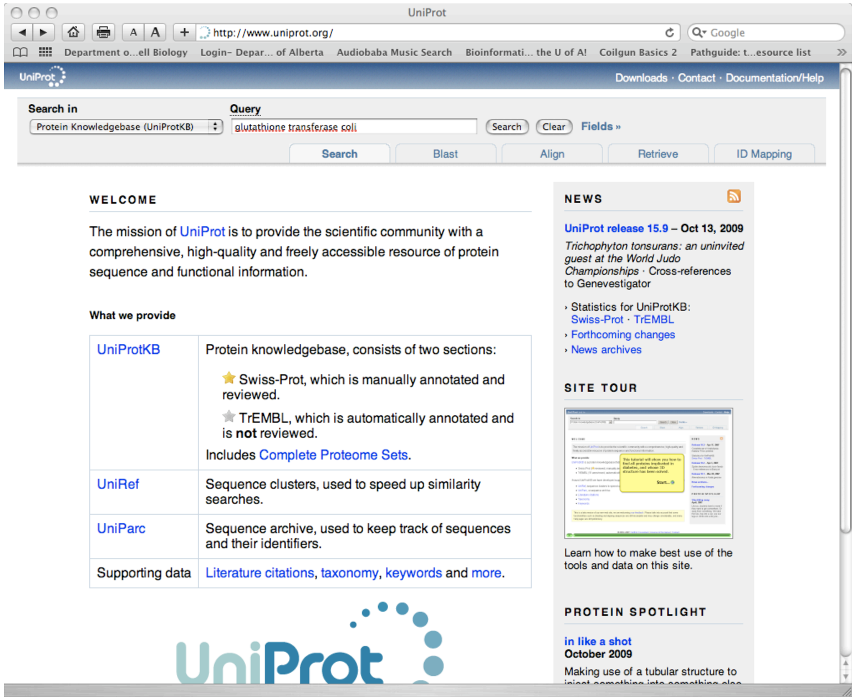
Annotation by Homology*

- Statistically significant sequence matches identified by BLAST searches against GenBank (nr), UniProt, DDBJ, PDB, InterPro, KEGG, Brenda, STRING
- Properties or annotation inferred by name, keywords, features, comments

Databases Are Key

Sequence Databases*

- GenBank
  - www.ncbi.nlm.nih.gov/
- UniProt/trEMBL
  - http://www.uniprot.org/
- DDBJ
  - http://www.ddbj.nig.ac.jp


UniProt
http://www.uniprot.org/
Google
Department o...ell Biology
Login– Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list
UniProt
Downloads · Contact · Documentation/Help
Search
Clear
Fields »
Search
Blast
Align
Retrieve
ID Mapping
UniProt

Search in: Protein Knowledgebase (UniProtKB)

Query: glutathione transferase coli

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes Complete Proteome Sets.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.

NEWS

UniProt release 15.9 – Oct 13, 2009

Trichophyton tonsurans: an uninvited guest at the World Judo Championships · Cross-references to Genevestigator

- Statistics for UniProtKB: Swiss-Prot · TrEMBL
- Forthcoming changes
- News archives

SITE TOUR

Learn how to make best use of the tools and data on this site.

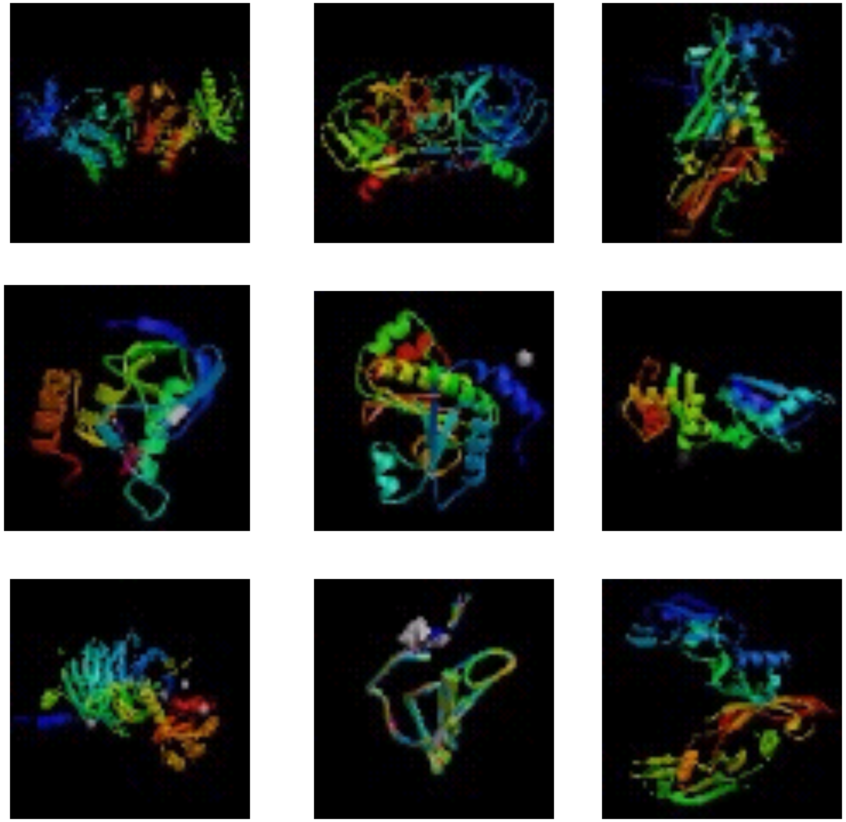
PROTEIN SPOTLIGHT

in like a shot
October 2009

Making use of a tubular structure to

Structure Databases*

- RCSB-PDB
  - http://www.rcsb.org/pdb/
- PDBe
  - http://www.ebi.ac.uk/pdbe/
- CATH
  - http://www.cathdb.info/
- SCOP
  - http://scop.mrc-lmb.cam.ac.uk/scop/

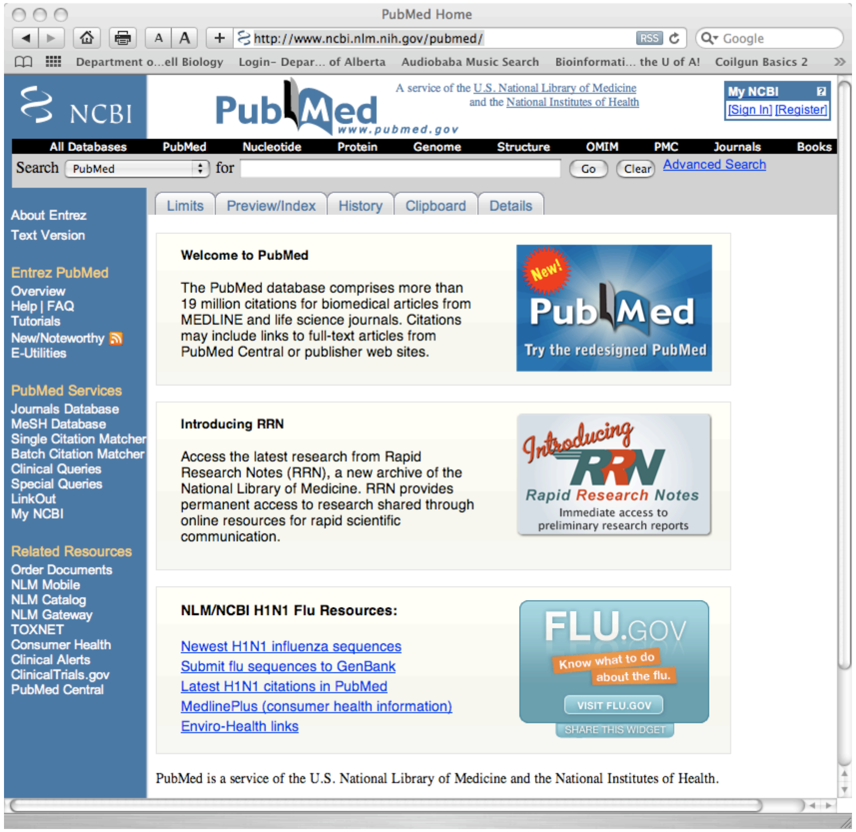


Interaction Databases*

- STRING
  - http://string.embl.de/
- DIP
  - http://dip.doe-mbi.ucla.edu/
- PIM
  - http://www.ebi.ac.uk/intact/main.xhtml
- MINT
  - http://mint.bio.uniroma2.it/mint/Welcome.do

Bibliographic Databases

- PubMed Medline
  - http://www.ncbi.nlm.nih.gov/PubMed/
- Google Scholar
  - http://scholar.google.ca/
- Your Local eLibrary
  - www.XXXX.ca
- Current Contents
  - http://science.thomsonreuters.com/



A service of the U.S. National Library of Medicine and the National Institutes of Health

Welcome to PubMed

The PubMed database comprises more than 19 million citations for biomedical articles from MEDLINE and life science journals. Citations may include links to full-text articles from PubMed Central or publisher web sites.

Introducing RRN

Access the latest research from Rapid Research Notes (RRN), a new archive of the National Library of Medicine. RRN provides permanent access to research shared through online resources for rapid scientific communication.

NLM/NCBI H1N1 Flu Resources:

Newest H1N1 influenza sequences
Submit flu sequences to GenBank
Latest H1N1 citations in PubMed
MedlinePlus (consumer health information)
Enviro-Health links

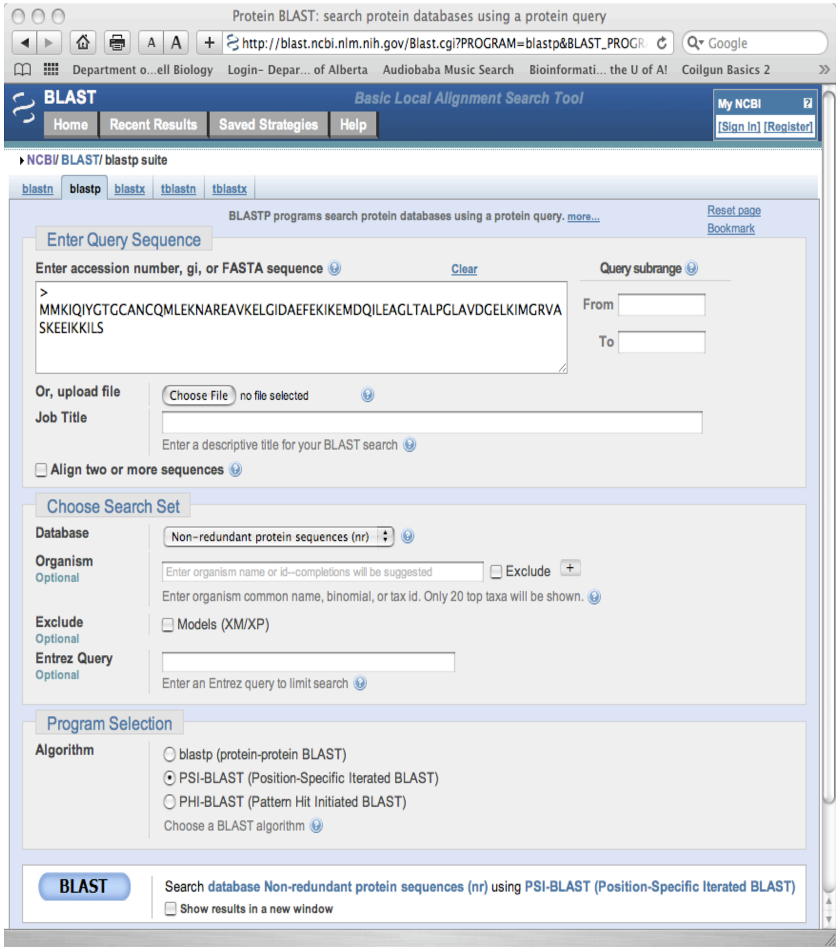
PubMed is a service of the U.S. National Library of Medicine and the National Institutes of Health.

Annotation by Homology
An Example

- **76 residue protein from *Methanobacter thermoautotrophicum* (newly sequenced)**
- **What does it do?**

- **MMKIQIYGTGCANCQMLEKNAREAVKELGIDAE**
  FEKIKEMDQILEAGLTALPGLAVDGELKIMGRV
  ASKEEIKKILS

PSI BLAST



Protein BLAST: search protein databases using a protein query
http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGR
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence — Clear

```
>
MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGELKIMGRVA
SKEEIKKILS
```

Query subrange — From: — To:

Or, upload file: Choose File — no file selected

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP)

Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

Algorithm:
- ○ blastp (protein-protein BLAST)
- ● PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

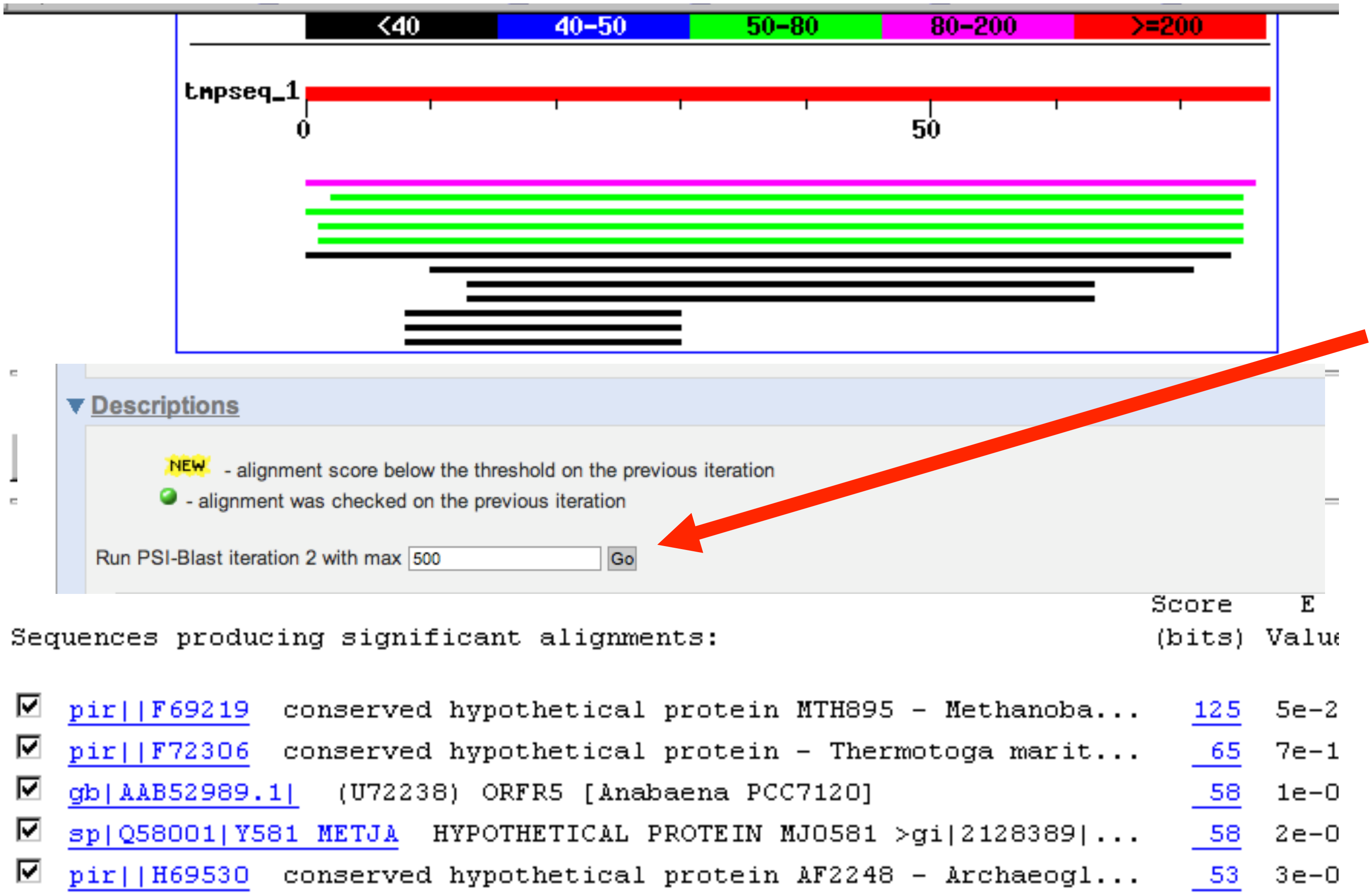
Choose a BLAST algorithm

BLAST — Search database Non-redundant protein sequences (nr) using PSI-BLAST (Position-Specific Iterated BLAST)

□ Show results in a new window

- PSI-BLAST – position specific iterative BLAST
- Derives a position-specific scoring matrix (PSSM) from the multiple sequence alignment of sequences detected above a given score threshold using protein BLAST
- This PSSM is used to further search the database for new matches, and is updated for subsequent iterations with these newly detected sequences
- PSI-BLAST provides a means of detecting distant relationships between proteins

PSI-BLAST



▼ Descriptions

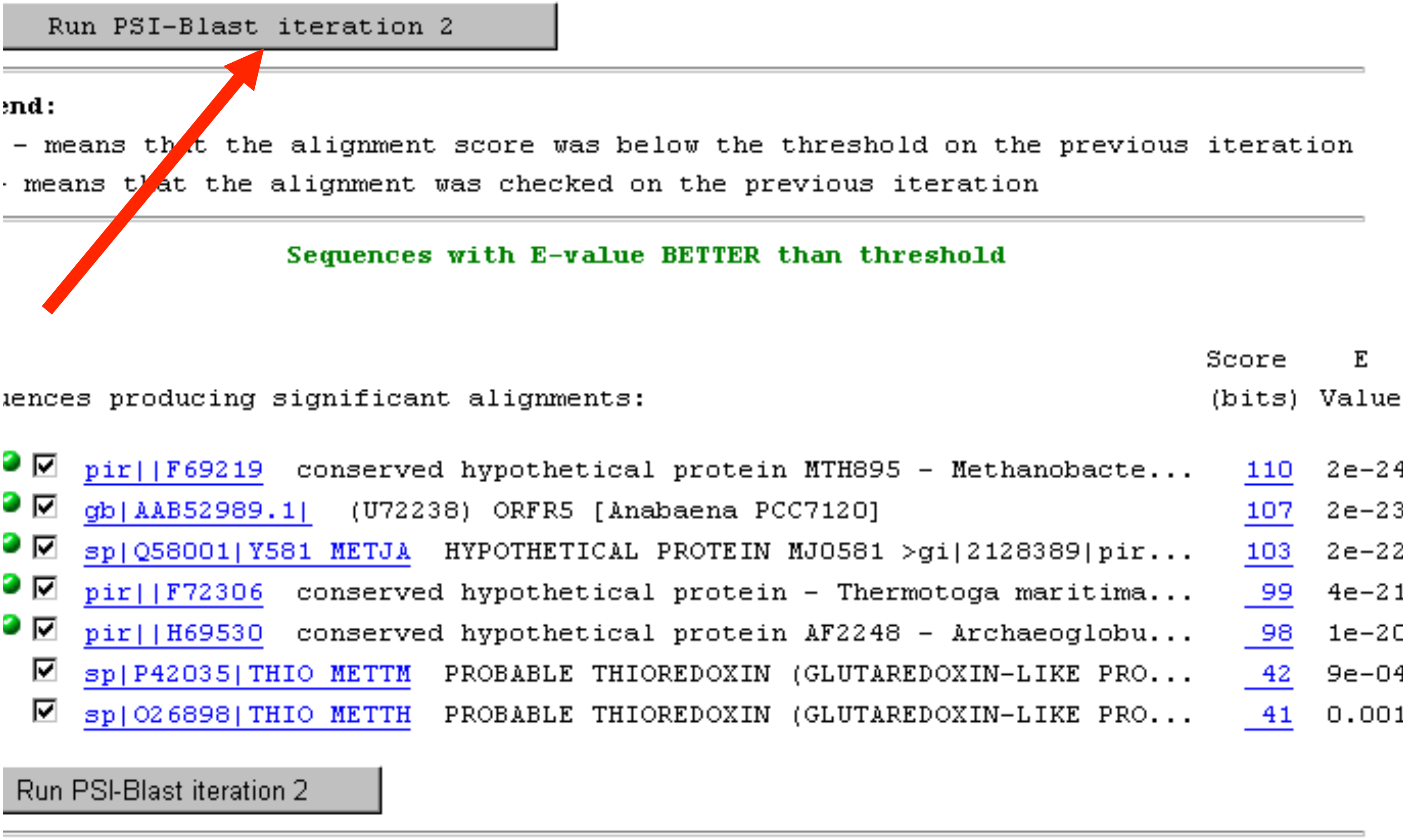
NEW - alignment score below the threshold on the previous iteration

● - alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max 500 Go

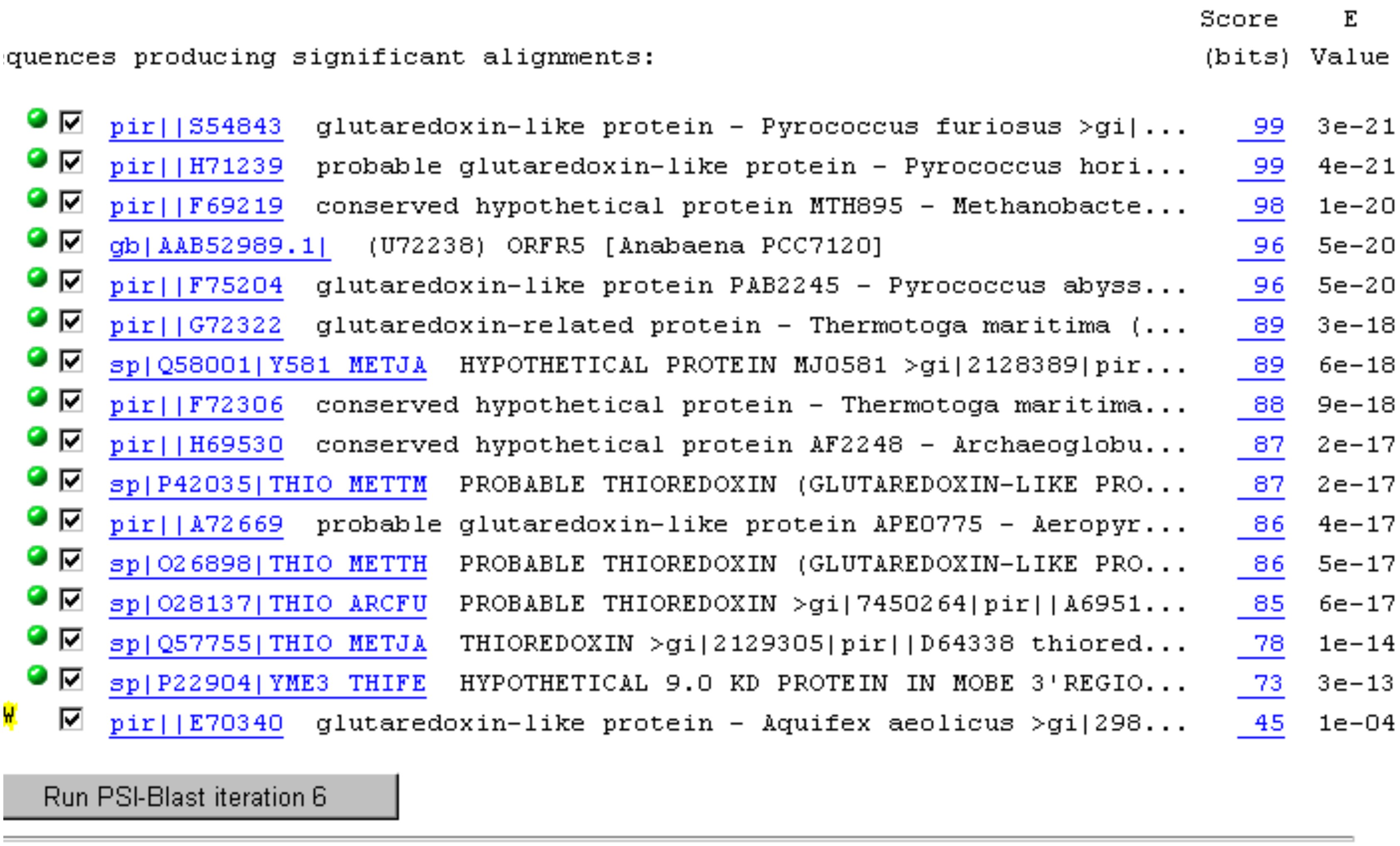
```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Valu

☑ pir||F69219  conserved hypothetical protein MTH895 - Methanoba...   125   5e-2
☑ pir||F72306  conserved hypothetical protein - Thermotoga marit...    65   7e-1
☑ gb|AAB52989.1|  (U72238) ORFR5 [Anabaena PCC7120]                    58   1e-0
☑ sp|Q58001|Y581_METJA  HYPOTHETICAL PROTEIN MJ0581 >gi|2128389|...   58   2e-0
☑ pir||H69530  conserved hypothetical protein AF2248 - Archaeogl...    53   3e-0
```

PSI-BLAST*



```
Run PSI-Blast iteration 2

end:
 - means that the alignment score was below the threshold on the previous iteration
- means that the alignment was checked on the previous iteration

                 Sequences with E-value BETTER than threshold

                                                                           Score     E
uences producing significant alignments:                                   (bits)  Value

● ☑ pir||F69219  conserved hypothetical protein MTH895 - Methanobacte...    110    2e-24
● ☑ gb|AAB52989.1|  (U72238) ORFR5 [Anabaena PCC7120]                        107    2e-23
● ☑ sp|Q58001|Y581_METJA  HYPOTHETICAL PROTEIN MJ0581 >gi|2128389|pir...    103    2e-22
● ☑ pir||F72306  conserved hypothetical protein - Thermotoga maritima...     99    4e-21
● ☑ pir||H69530  conserved hypothetical protein AF2248 - Archaeoglobu...     98    1e-20
  ☑ sp|P42035|THIO_METTM  PROBABLE THIOREDOXIN (GLUTAREDOXIN-LIKE PRO...     42    9e-04
  ☑ sp|O26898|THIO_METTH  PROBABLE THIOREDOXIN (GLUTAREDOXIN-LIKE PRO...     41    0.001

Run PSI-Blast iteration 2
```

PSI-BLAST*



```
                                                                     Score    E
quences producing significant alignments:                           (bits) Value

● ☑ pir||S54843   glutaredoxin-like protein - Pyrococcus furiosus >gi|...   99   3e-21
● ☑ pir||H71239   probable glutaredoxin-like protein - Pyrococcus hori...   99   4e-21
● ☑ pir||F69219   conserved hypothetical protein MTH895 - Methanobacte...   98   1e-20
● ☑ gb|AAB52989.1|   (U72238) ORFR5 [Anabaena PCC7120]                      96   5e-20
● ☑ pir||F75204   glutaredoxin-like protein PAB2245 - Pyrococcus abyss...   96   5e-20
● ☑ pir||G72322   glutaredoxin-related protein - Thermotoga maritima (...   89   3e-18
● ☑ sp|Q58001|Y581_METJA   HYPOTHETICAL PROTEIN MJ0581 >gi|2128389|pir...   89   6e-18
● ☑ pir||F72306   conserved hypothetical protein - Thermotoga maritima...   88   9e-18
● ☑ pir||H69530   conserved hypothetical protein AF2248 - Archaeoglobu...   87   2e-17
● ☑ sp|P42035|THIO_METTM   PROBABLE THIOREDOXIN (GLUTAREDOXIN-LIKE PRO...   87   2e-17
● ☑ pir||A72669   probable glutaredoxin-like protein APE0775 - Aeropyr...   86   4e-17
● ☑ sp|O26898|THIO_METTH   PROBABLE THIOREDOXIN (GLUTAREDOXIN-LIKE PRO...   86   5e-17
● ☑ sp|O28137|THIO_ARCFU   PROBABLE THIOREDOXIN >gi|7450264|pir||A6951...   85   6e-17
● ☑ sp|Q57755|THIO_METJA   THIOREDOXIN >gi|2129305|pir||D64338 thiored...   78   1e-14
● ☑ sp|P22904|YME3_THIFE   HYPOTHETICAL 9.0 KD PROTEIN IN MOBE 3'REGIO...   73   3e-13
W ☑ pir||E70340   glutaredoxin-like protein - Aquifex aeolicus >gi|298...   45   1e-04
```

Conclusions

- Protein is a thioredoxin or glutaredoxin (function, family)
- Protein has thioredoxin fold (2º and 3D structure)
- Active site is from residues 11-14 (active site location)
- Protein is soluble, cytoplasmic (cellular location)

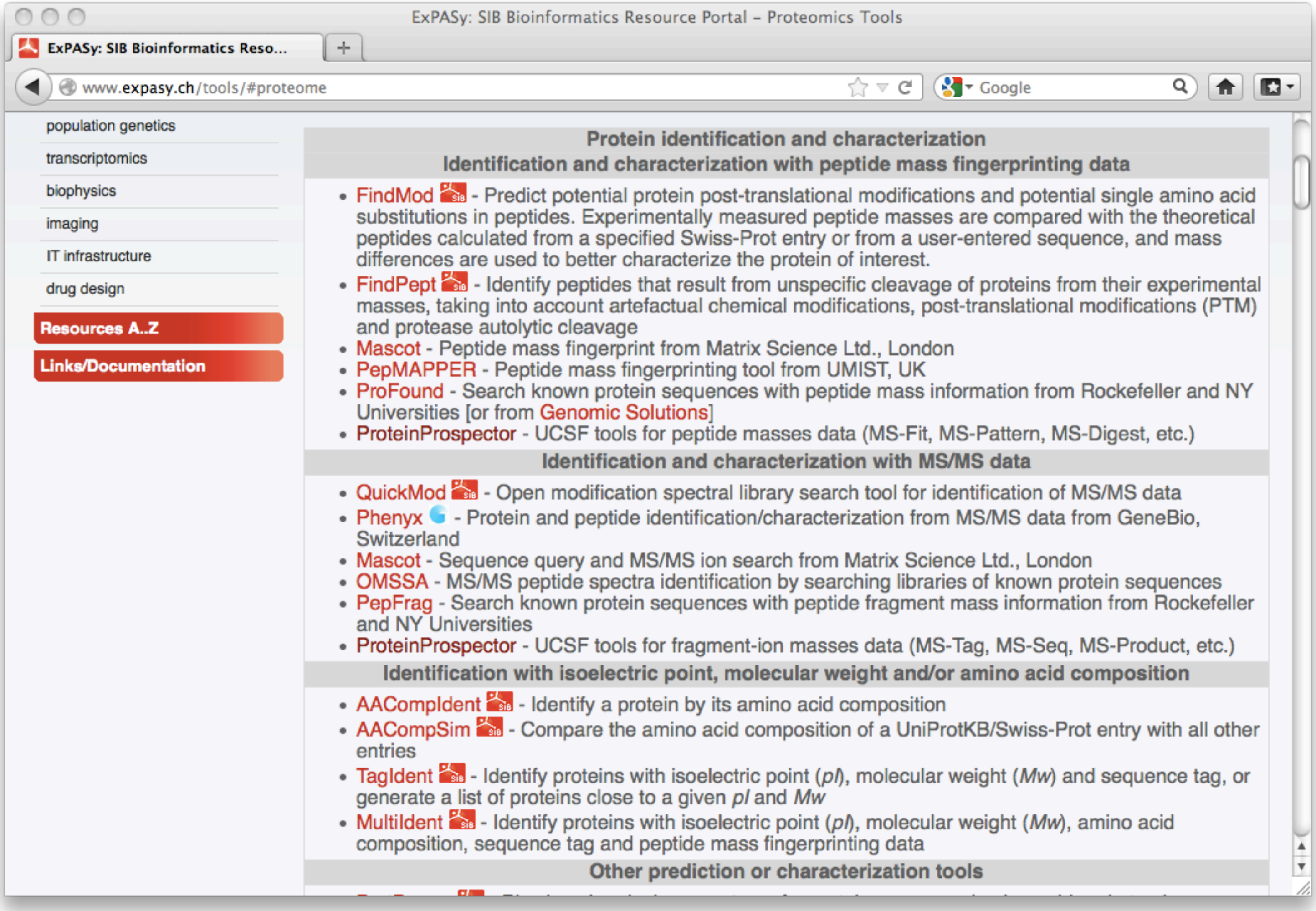
Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools

Annotation by Composition*

- Molecular Weight
- Isoelectric Point
- UV Absorptivity
- Hydrophobicity

Where To Go


ExPASy: SIB Bioinformatics Resource Portal – Proteomics Tools
ExPASy: SIB Bioinformatics Reso...
www.expasy.ch/tools/#proteome
Google
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
drug design
Resources A..Z
Links/Documentation

Protein identification and characterization

Identification and characterization with peptide mass fingerprinting data

- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- Mascot - Peptide mass fingerprint from Matrix Science Ltd., London
- PepMAPPER - Peptide mass fingerprinting tool from UMIST, UK
- ProFound - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from Genomic Solutions]
- ProteinProspector - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)

Identification and characterization with MS/MS data

- QuickMod - Open modification spectral library search tool for identification of MS/MS data
- Phenyx - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities
- ProteinProspector - UCSF tools for fragment-ion masses data (MS-Tag, MS-Seq, MS-Product, etc.)

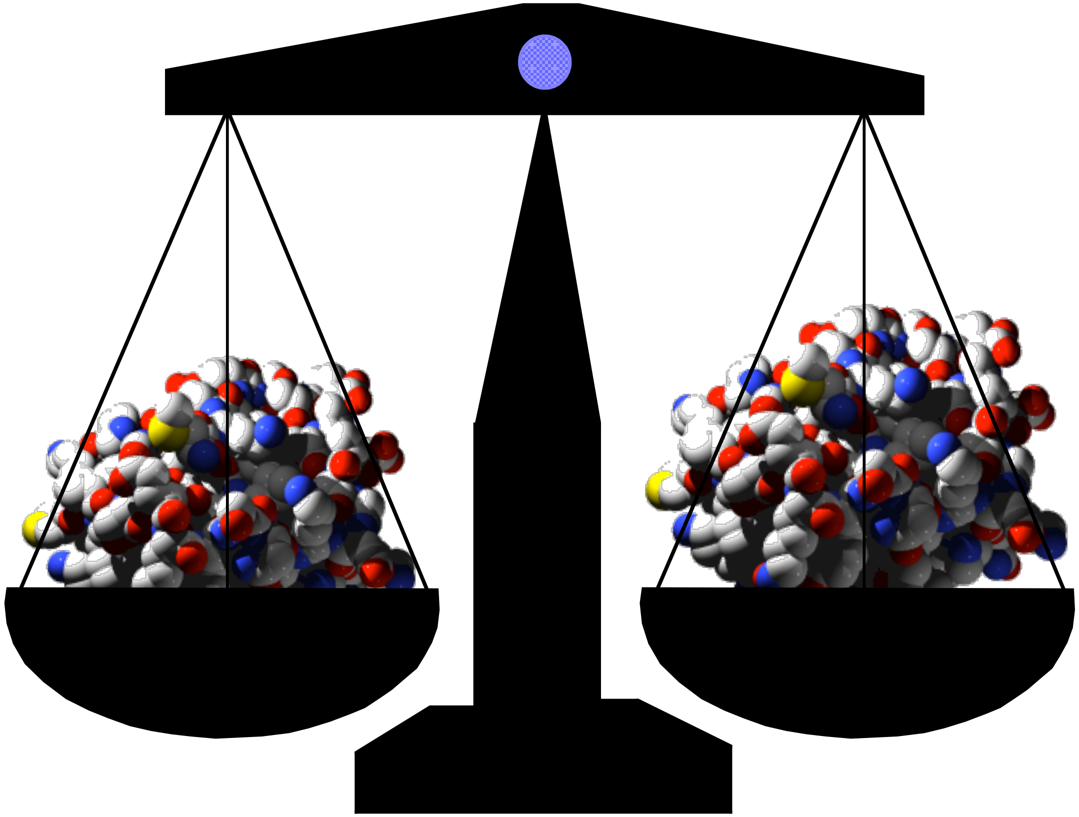
Identification with isoelectric point, molecular weight and/or amino acid composition

- AACompIdent - Identify a protein by its amino acid composition
- AACompSim - Compare the amino acid composition of a UniProtKB/Swiss-Prot entry with all other entries
- TagIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*) and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- MultiIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*), amino acid composition, sequence tag and peptide mass fingerprinting data

Other prediction or characterization tools

http://www.expasy.ch/tools/#proteome

Molecular Weight



Molecular Weight*

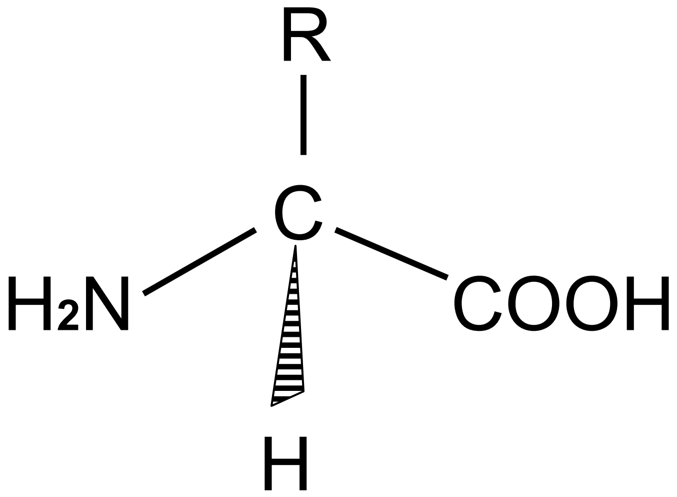
- Useful for SDS PAGE and 2D gel analysis
- Useful for deciding on SEC matrix
- Useful for deciding on MWC for dialysis
- Essential in synthetic peptide analysis
- Essential in peptide sequencing (classical or mass-spectrometry based)
- Essential in proteomics and high throughput protein characterization

Molecular Weight*

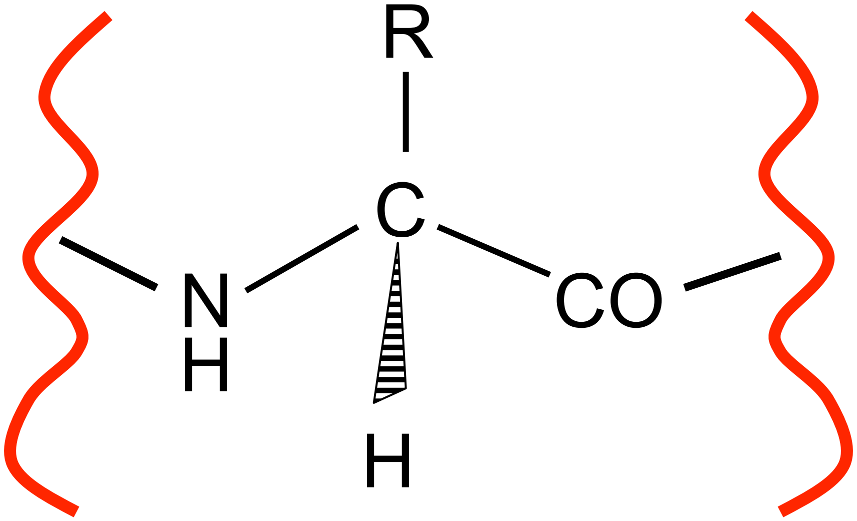
- Crude MW calculation: MW = 110 X Numres
- Exact MW calculation: $MW = \Sigma nAA_i \text{ x } MW_i$
- Remember to add 1 water (18.01 amu) after adding all res.
- Corrections for CHO, PO4, Acetyl, CONH2

Amino Acid Residue Weights			
Residue	Weight	Residue	Weight
A	71.08	M	131.21
C	103.14	N	114.11
D	115.09	P	97.12
E	129.12	Q	128.14
F	147.18	R	156.2
G	57.06	S	87.08
H	137.15	T	101.11
I	113.17	V	99.14
K	128.18	W	186.21
L	113.17	Y	163.18

Amino Acid versus Residue

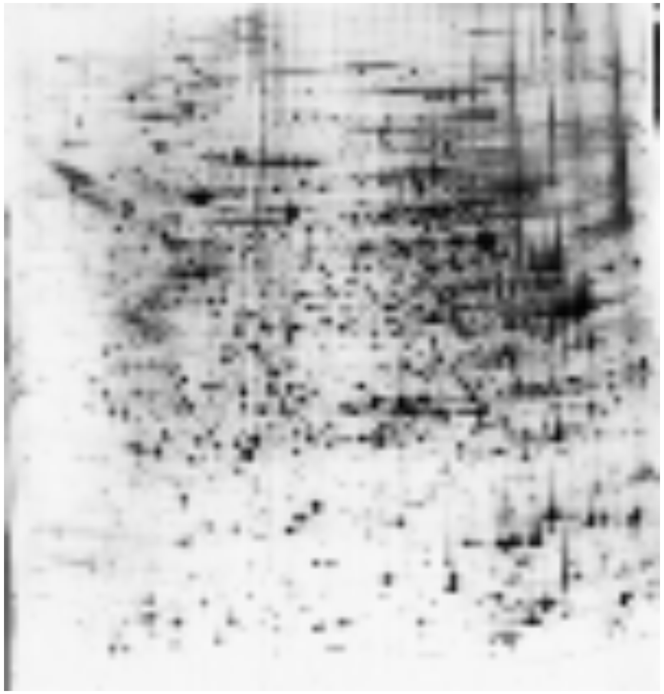


Amino Acid

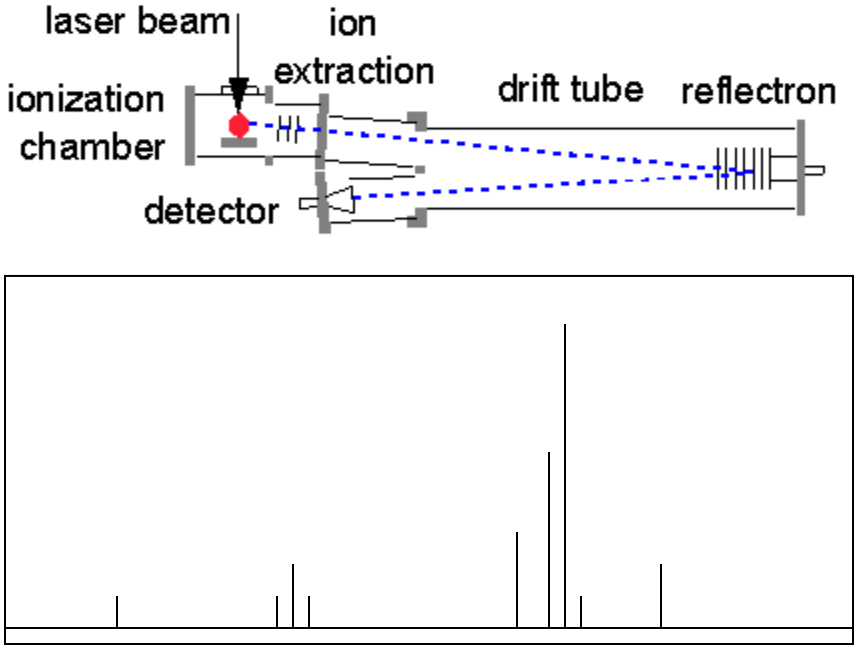


Residue

Molecular Weight & Proteomics



2-D Gel

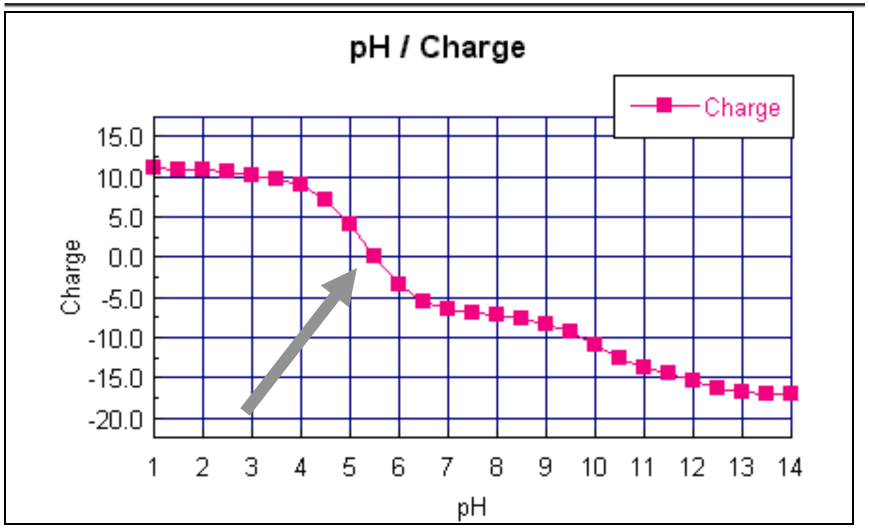


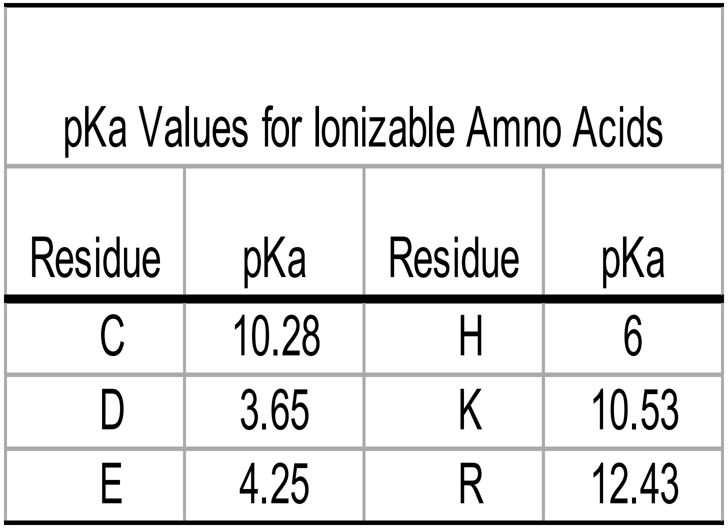
QTOF Mass Spectrometry

Isoelectric Point*

- **The pH at which a protein has a net charge=0**
- $$Q = \Sigma Ni/(1 + 10^{pH-pKi})$$

This is a transcendental equation



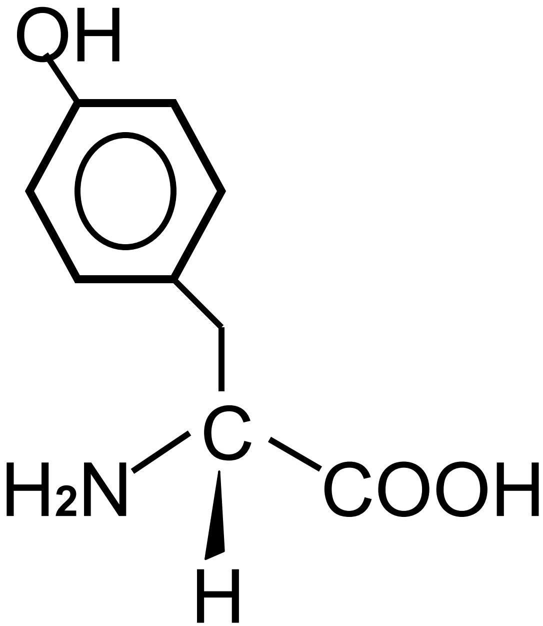


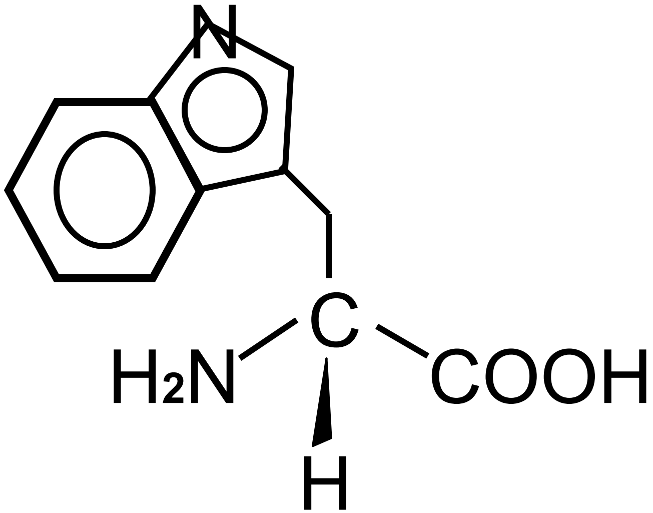
pKa Values for Ionizable Amno Acids

Residue	pKa	Residue	pKa
C	10.28	H	6
D	3.65	K	10.53
E	4.25	R	12.43

UV Absorptivity*

- **OD_{280} = (5690 x #W + 1280 x #Y)/MW x Conc.**
- **Conc. = OD_{280} x MW/(5690 X #W + 1280 x #Y)**





Very useful for measuring protein concentration

Hydrophobicity*

- **Average Hphob calculation: H_{ave} = (ΣnAA_i x $Hphob_i$)/N**
- **Indicates Solubility, stability, location**
- **If H_{ave} < 1 the protein is soluble**
- **If H_{ave} > 1 it is likely a membrane protein**

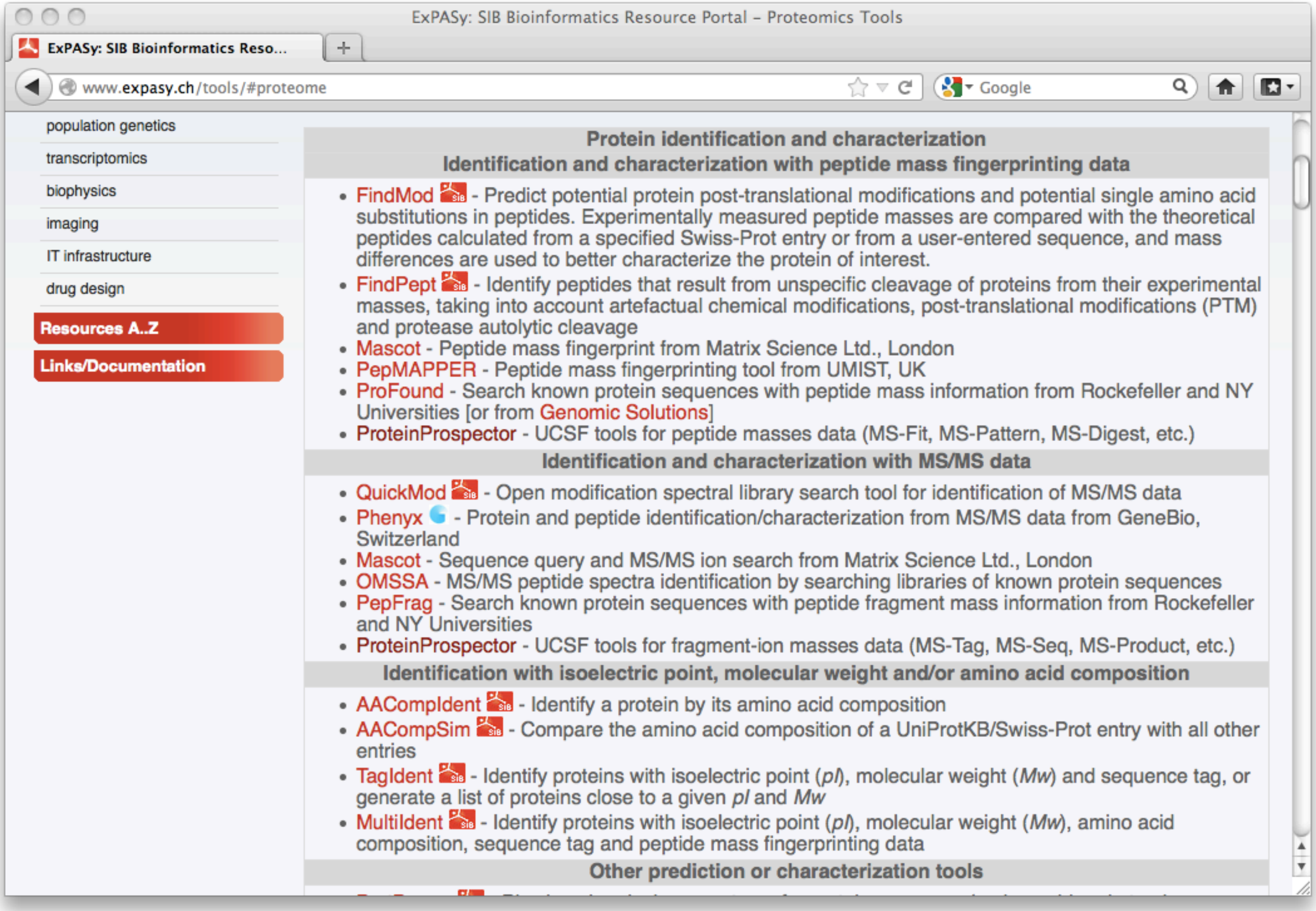
Kyte / Doolittle Hyrophobicity Scale

Residue	Hphob	Residue	Hphob
A	1.8	M	1.9
C	2.5	N	-3.5
D	-3.5	P	-1.6
E	-3.5	Q	-3.5
F	2.8	R	-4.5
G	-0.4	S	-0.8
H	-3.2	T	-0.7
I	4.5	V	4.2
K	-3.9	W	-0.9
L	3.8	Y	-1.3

Annotation Methods

- Annotation by homology (BLAST)
  - requires a large, well annotated database of protein sequences
- Annotation by sequence composition
  - simple statistical/mathematical methods
- Annotation by sequence features, profiles or motifs
  - requires sophisticated sequence analysis tools

Where To Go



ExPASy: SIB Bioinformatics Resource Portal – Proteomics Tools
ExPASy: SIB Bioinformatics Reso...
www.expasy.ch/tools/#proteome
Google
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
drug design
Resources A..Z
Links/Documentation

Protein identification and characterization

Identification and characterization with peptide mass fingerprinting data

- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- Mascot - Peptide mass fingerprint from Matrix Science Ltd., London
- PepMAPPER - Peptide mass fingerprinting tool from UMIST, UK
- ProFound - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from Genomic Solutions]
- ProteinProspector - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)

Identification and characterization with MS/MS data

- QuickMod - Open modification spectral library search tool for identification of MS/MS data
- Phenyx - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities
- ProteinProspector - UCSF tools for fragment-ion masses data (MS-Tag, MS-Seq, MS-Product, etc.)

Identification with isoelectric point, molecular weight and/or amino acid composition

- AACompIdent - Identify a protein by its amino acid composition
- AACompSim - Compare the amino acid composition of a UniProtKB/Swiss-Prot entry with all other entries
- TagIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*) and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- MultiIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*), amino acid composition, sequence tag and peptide mass fingerprinting data

Other prediction or characterization tools

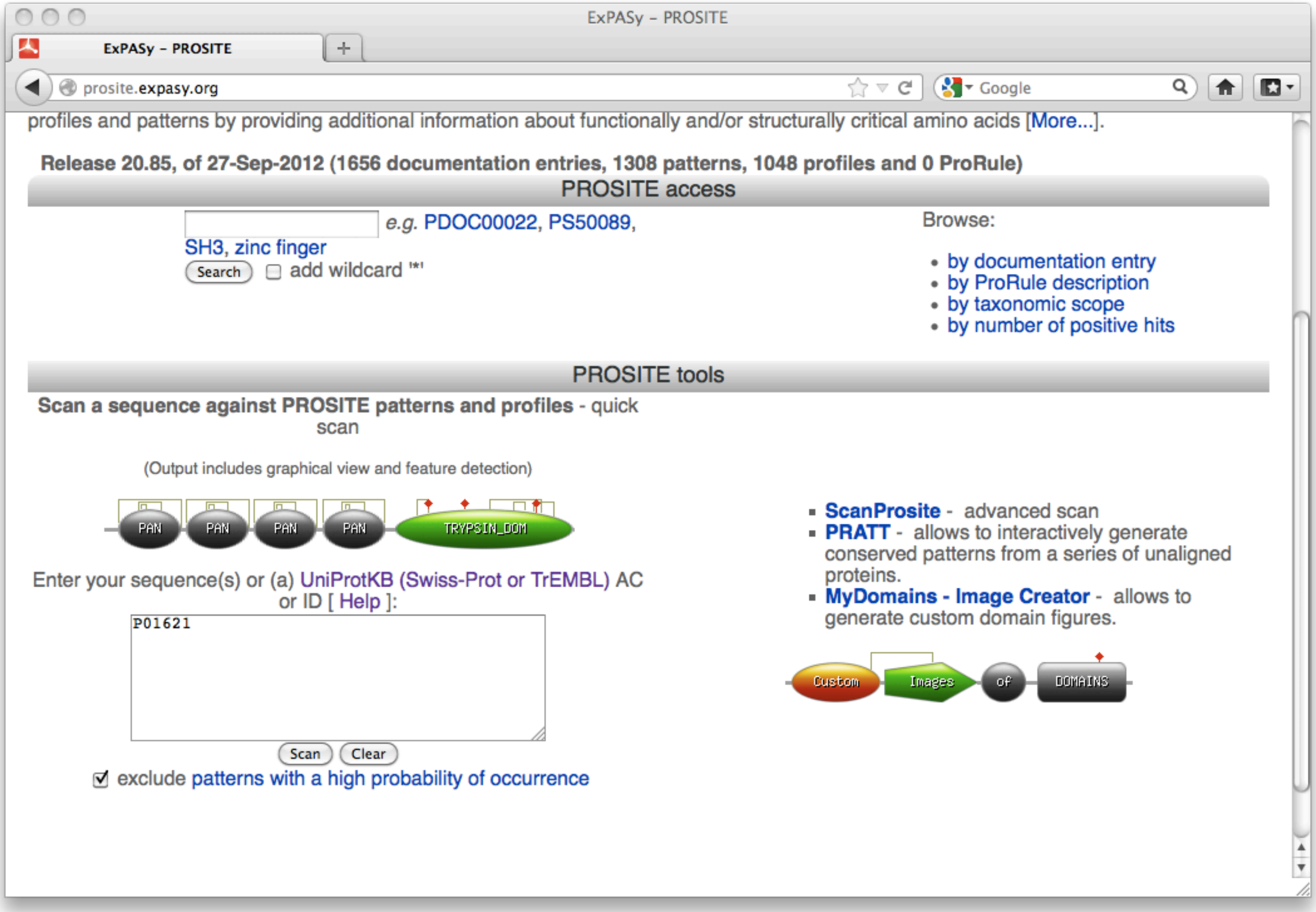
http://www.expasy.ch/tools/#proteome

Sequence Feature Databases

- PROSITE - http://www.expasy.ch/prosite/
- InterPro - http://www.ebi.ac.uk/interpro/
- PPT-DB - http://www.pptdb.ca/

To use these databases just submit your PROTEIN sequence to the database and download the output. They provide domain information, predicted disulfides, functional sites, active sites, secondary structure – IF THERE IS A MATCH

Using Prosite



profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 20.85, of 27-Sep-2012 (1656 documentation entries, 1308 patterns, 1048 profiles and 0 ProRule)

PROSITE access

e.g. PDOC00022, PS50089, SH3, zinc finger

Search □ add wildcard '*'

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence(s) or (a) UniProtKB (Swiss-Prot or TrEMBL) AC or ID [Help]:

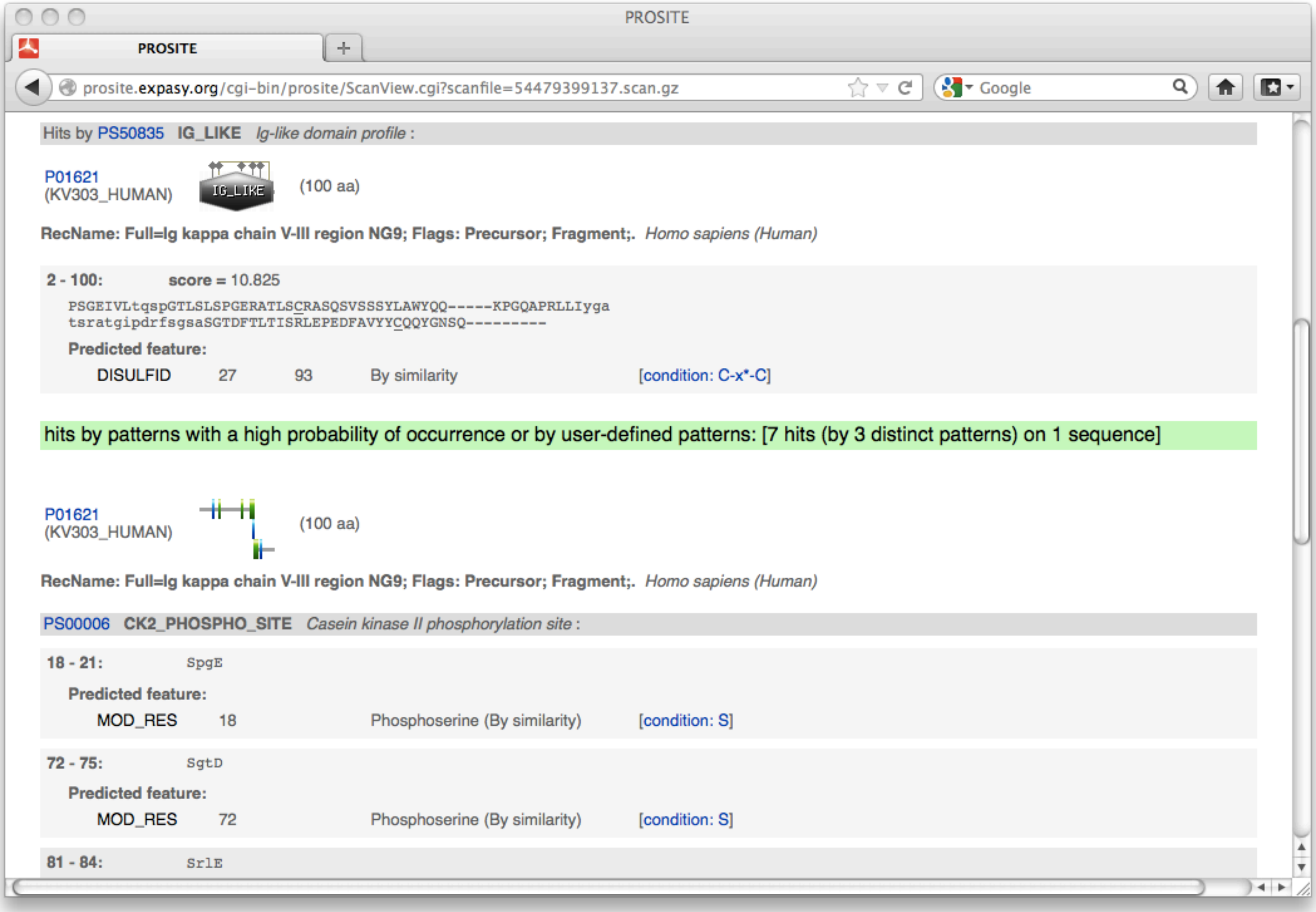
P01621

Scan Clear

☑ exclude patterns with a high probability of occurrence

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** - allows to generate custom domain figures.

Prosite Output



PROSITE

PROSITE

prosite.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=54479399137.scan.gz

Google

Hits by PS50835 **IG_LIKE** *Ig-like domain profile :*

P01621
(KV303_HUMAN) IG_LIKE (100 aa)

RecName: Full=Ig kappa chain V-III region NG9; Flags: Precursor; Fragment;. *Homo sapiens (Human)*

2 - 100: **score =** 10.825

```
PSGEIVLtqspGTLSLSPGERATLSCRASQSVSSSYLAWYQQ-----KPGQAPRLLIyga
tsratgipdrfsgsaSGTDFTLTISRLEPEDFAVYYCQQYGNSQ---------
```

Predicted feature:

DISULFID	27	93	By similarity	[condition: C-x*-C]

hits by patterns with a high probability of occurrence or by user-defined patterns: [7 hits (by 3 distinct patterns) on 1 sequence]

P01621
(KV303_HUMAN) (100 aa)

RecName: Full=Ig kappa chain V-III region NG9; Flags: Precursor; Fragment;. *Homo sapiens (Human)*

PS00006 **CK2_PHOSPHO_SITE** *Casein kinase II phosphorylation site :*

18 - 21: `SpgE`

Predicted feature:

MOD_RES	18	Phosphoserine (By similarity)	[condition: S]

72 - 75: `SgtD`

Predicted feature:

MOD_RES	72	Phosphoserine (By similarity)	[condition: S]

81 - 84: `SrlE`

What if your Sequence doesn’t match to Something in the Database?

- Don’t worry
- You can use prediction programs and freely available web servers that use machine learning, neural networks, HMMs and other cool bioinformatic tricks to predict some of the same things that your database matching tools try to identify

What Can Be Predicted?*

- O-Glycosylation Sites
- Phosphorylation Sites
- Protease Cut Sites
- Nuclear Targeting Sites
- Mitochondrial Targ Sites
- Chloroplast Targ Sites
- Signal Sequences
- Signal Sequence Cleav.
- Peroxisome Targ Sites
- ER Targeting Sites
- Transmembrane Sites
- Tyrosine Sulfation Sites
- GPInositol Anchor Sites
- PEST sites
- Coil-Coil Sites
- T-Cell/MHC Epitopes
- Protein Lifetime
- A whole lot more….

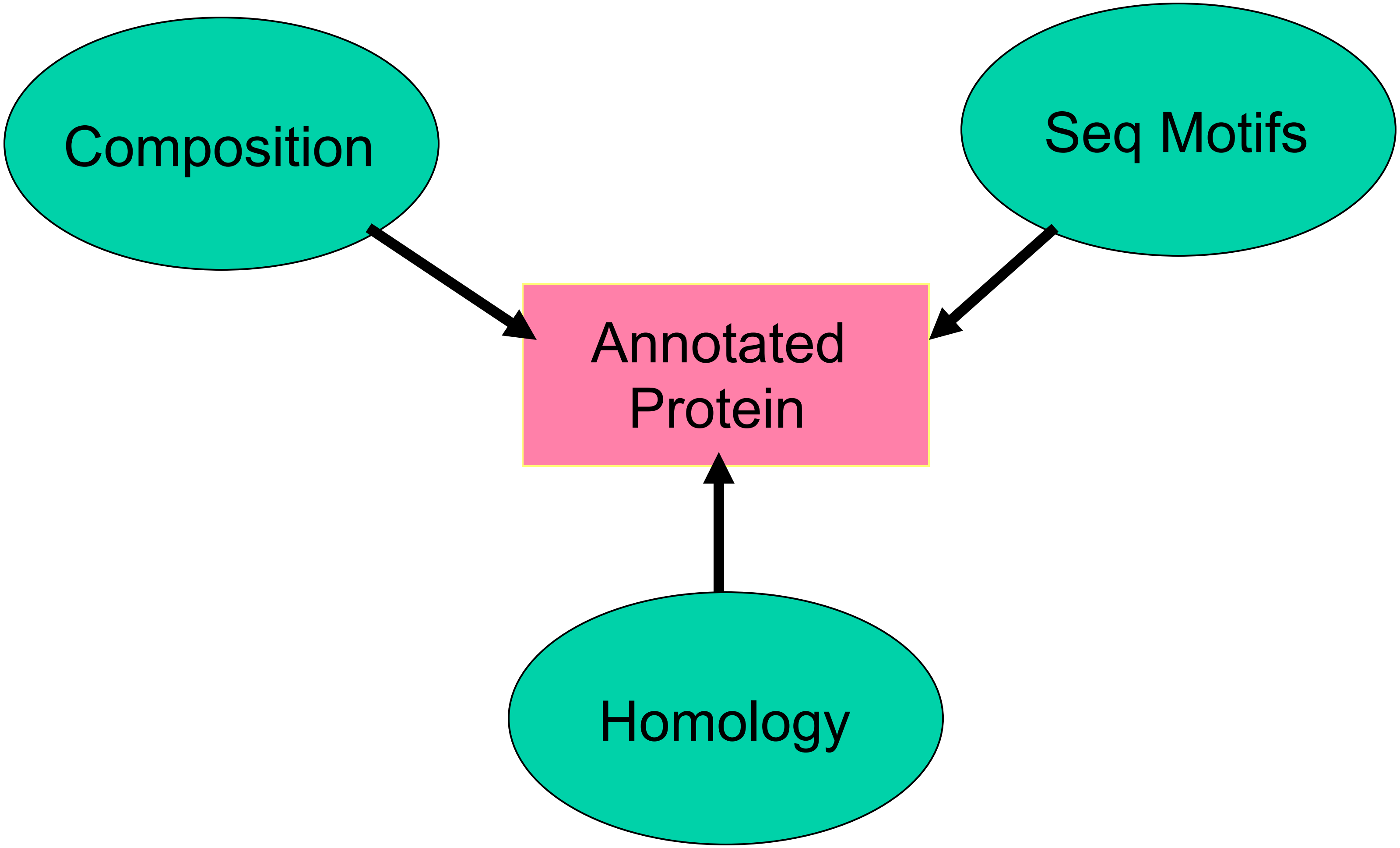
Cutting Edge Sequence Feature Servers*

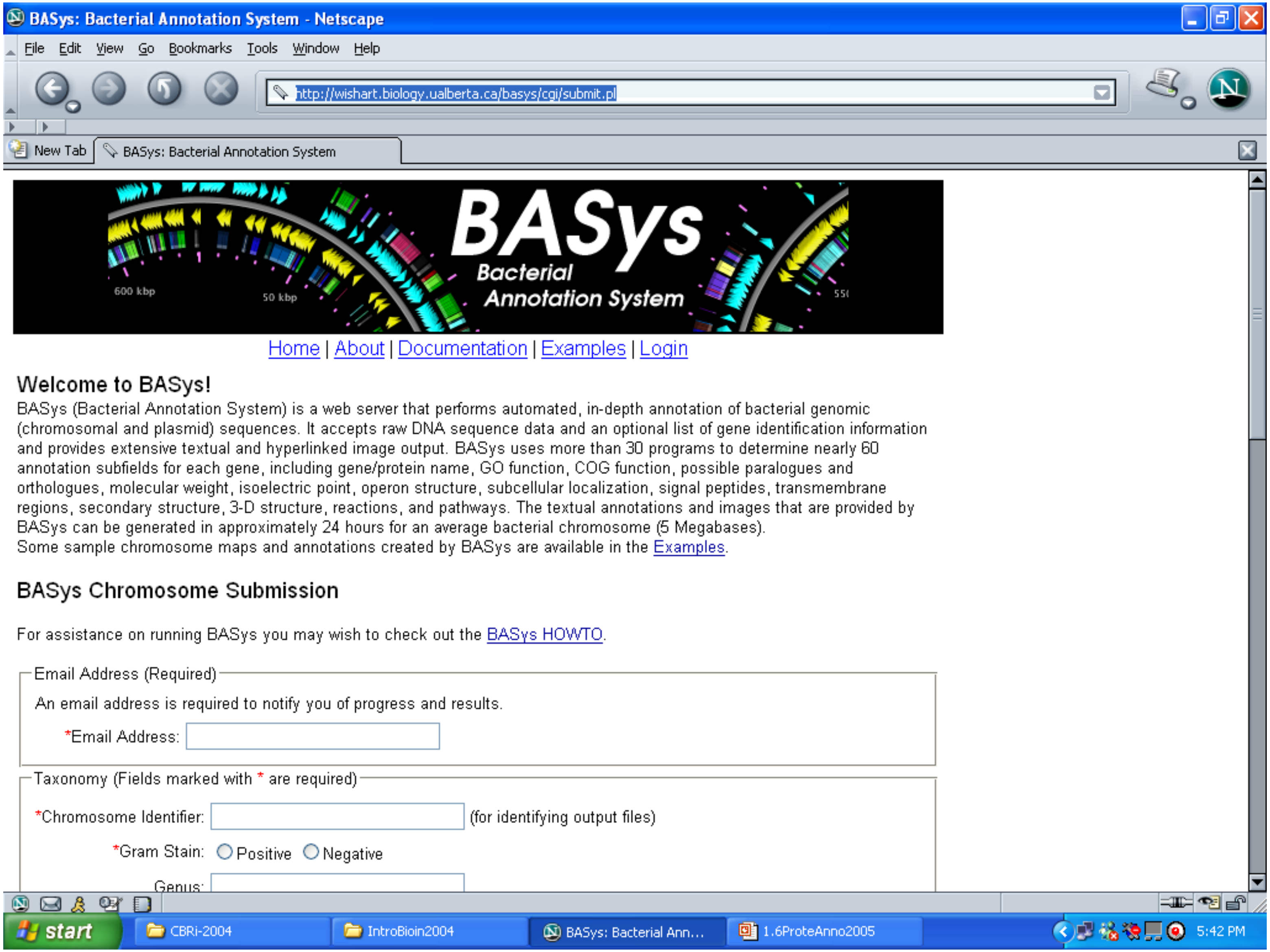
- Membrane Helix Prediction
  - http://www.cbs.dtu.dk/services/TMHMM-2.0/
- T-Cell Epitope Prediction
  - http://www.syfpeithi.de/home.htm
- O-Glycosylation Prediction
  - http://www.cbs.dtu.dk/services/NetOGlyc/
- Phosphorylation Prediction
  - http://www.cbs.dtu.dk/services/NetPhos/
- Protein Localization Prediction
  - http://psort.ims.u-tokyo.ac.jp/

2º Structure Prediction*

- PredictProtein-PHD (72%)
  - http://www.predictprotein.org
- Jpred (73-75%)
  - http://www.compbio.dundee.ac.uk/~www-jpred/
- PSIpred (77%)
  - http://bioinf.cs.ucl.ac.uk/psipred/
- Proteus2 (78-90%)
  - http://www.proteus2.ca/proteus2/

Putting It All Together




BASys: Bacterial Annotation System - Netscape
File Edit View Go Bookmarks Tools Window Help
http://wishart.biology.ualberta.ca/basys/cgi/submit.pl
New Tab
BASys: Bacterial Annotation System
BASys
Bacterial
Annotation System
600 kbp
50 kbp
Home | About | Documentation | Examples | Login
start
CBRi-2004
IntroBioin2004
BASys: Bacterial Ann...
1.6ProteAnno2005
5:42 PM

Welcome to BASys!

BASys (Bacterial Annotation System) is a web server that performs automated, in-depth annotation of bacterial genomic (chromosomal and plasmid) sequences. It accepts raw DNA sequence data and an optional list of gene identification information and provides extensive textual and hyperlinked image output. BASys uses more than 30 programs to determine nearly 60 annotation subfields for each gene, including gene/protein name, GO function, COG function, possible paralogues and orthologues, molecular weight, isoelectric point, operon structure, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3-D structure, reactions, and pathways. The textual annotations and images that are provided by BASys can be generated in approximately 24 hours for an average bacterial chromosome (5 Megabases).
Some sample chromosome maps and annotations created by BASys are available in the Examples.

BASys Chromosome Submission

For assistance on running BASys you may wish to check out the BASys HOWTO.

Email Address (Required)

An email address is required to notify you of progress and results.

*Email Address:

Taxonomy (Fields marked with * are required)

*Chromosome Identifier: (for identifying output files)

*Gram Stain: ○ Positive ○ Negative

Genus:

http://basys.ca/basys/cgi/submit.pl

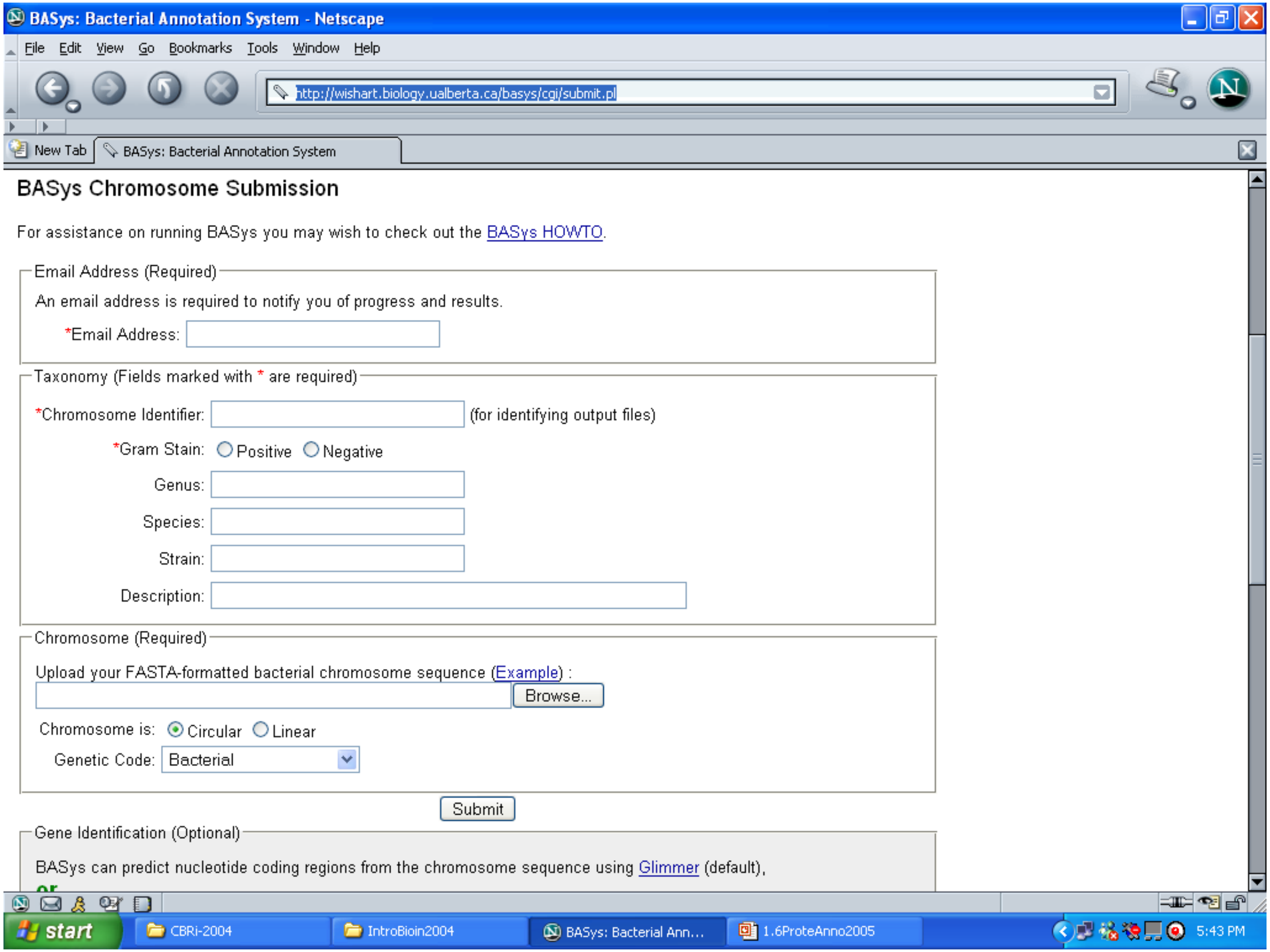
BASys

- BASys (Bacterial Annotation System) is a web server that performs automated, in-depth annotation of bacterial genomic sequences
- It accepts raw DNA sequence data and an optional list of gene identification information and provides extensive textual and hyperlinked image output

BASys

- BASys uses more than 30 programs to determine nearly 60 annotation subfields for each gene, including:
- Gene/protein name, GO function, COG function, possible paralogues and orthologues, molecular weight, isoelectric point, operon structure, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3-D structure and reactions

Submitting to BASys


BASys: Bacterial Annotation System - Netscape
File Edit View Go Bookmarks Tools Window Help
http://wishart.biology.ualberta.ca/basys/cgi/submit.pl
New Tab
BASys: Bacterial Annotation System
start
CBRi-2004
IntroBioin2004
BASys: Bacterial Ann...
1.6ProteAnno2005
5:43 PM

BASys Chromosome Submission

For assistance on running BASys you may wish to check out the BASys HOWTO.

Email Address (Required)

An email address is required to notify you of progress and results.

*Email Address:

Taxonomy (Fields marked with * are required)

*Chromosome Identifier: (for identifying output files)

*Gram Stain: ○ Positive ○ Negative

Genus:

Species:

Strain:

Description:

Chromosome (Required)

Upload your FASTA-formatted bacterial chromosome sequence (Example) :

Browse...

Chromosome is: ● Circular ○ Linear

Genetic Code: Bacterial

Submit

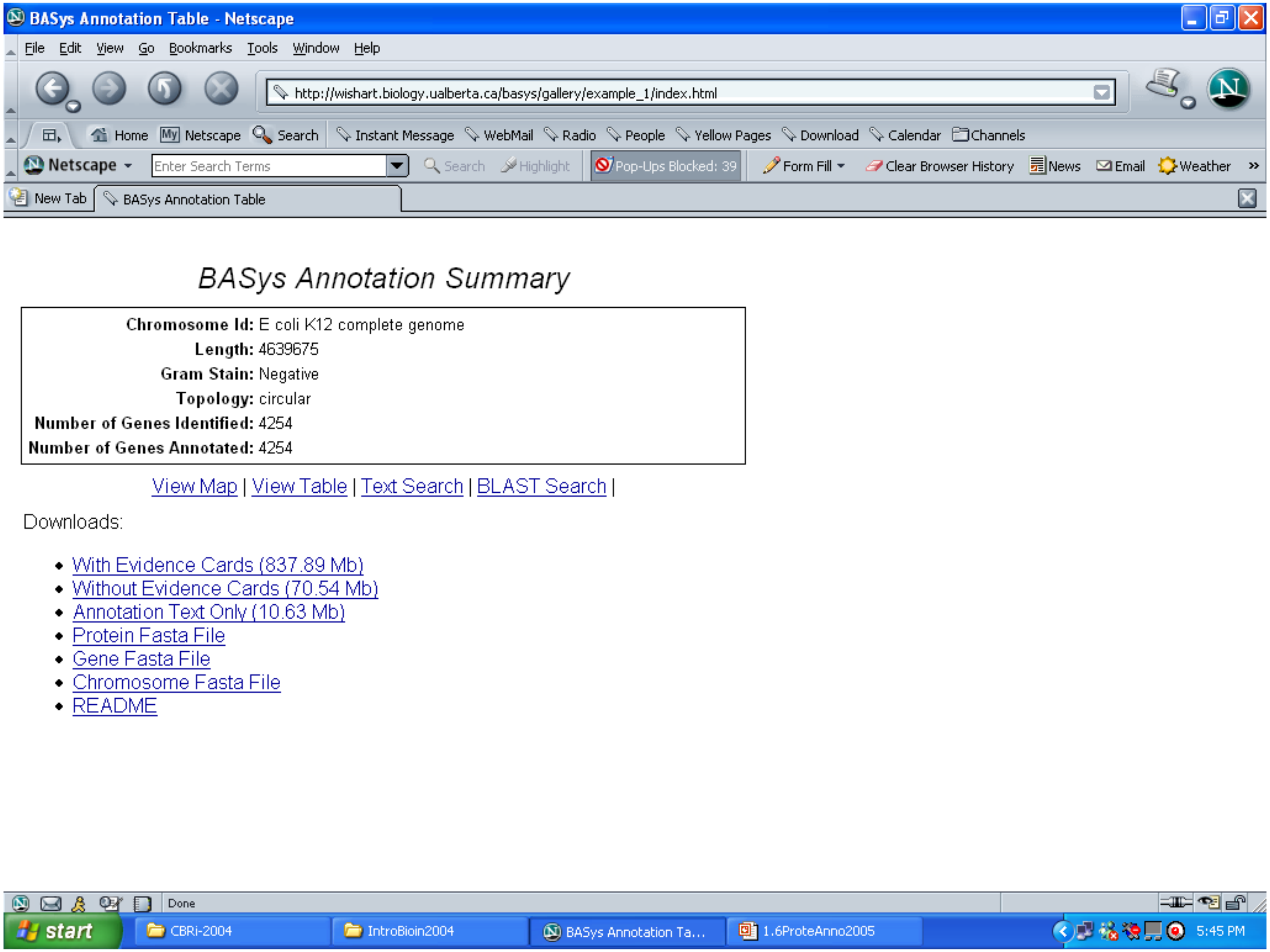
Gene Identification (Optional)

BASys can predict nucleotide coding regions from the chromosome sequence using Glimmer (default),

or

Wait…

BASys Output



BASys Annotation Table - Netscape
File Edit View Go Bookmarks Tools Window Help
http://wishart.biology.ualberta.ca/basys/gallery/example_1/index.html
Home Netscape Search Instant Message WebMail Radio People Yellow Pages Download Calendar Channels
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 39 Form Fill Clear Browser History News Email Weather
New Tab BASys Annotation Table
Done
start CBRi-2004 IntroBioin2004 BASys Annotation Ta... 1.6ProteAnno2005 5:45 PM

BASys Annotation Summary

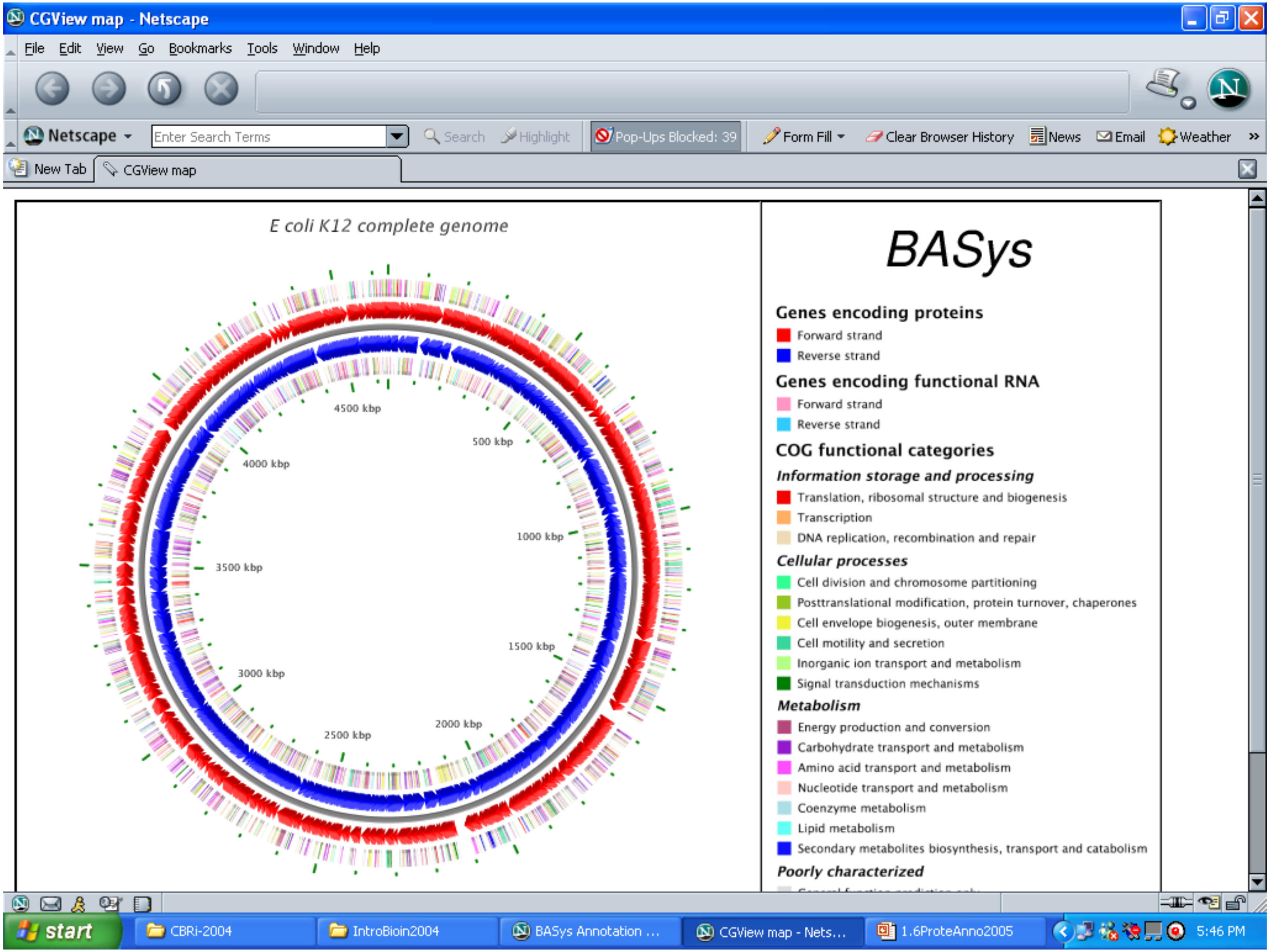
Chromosome Id:	E coli K12 complete genome
Length:	4639675
Gram Stain:	Negative
Topology:	circular
Number of Genes Identified:	4254
Number of Genes Annotated:	4254

View Map | View Table | Text Search | BLAST Search |

Downloads:

- With Evidence Cards (837.89 Mb)
- Without Evidence Cards (70.54 Mb)
- Annotation Text Only (10.63 Mb)
- Protein Fasta File
- Gene Fasta File
- Chromosome Fasta File
- README

BASys Output (Map)



BASys

Genes encoding proteins

- 🟥 Forward strand
- 🟦 Reverse strand

Genes encoding functional RNA

- 🟪 Forward strand
- 🟦 Reverse strand

COG functional categories

Information storage and processing

- 🟥 Translation, ribosomal structure and biogenesis
- 🟧 Transcription
- 🟫 DNA replication, recombination and repair

Cellular processes

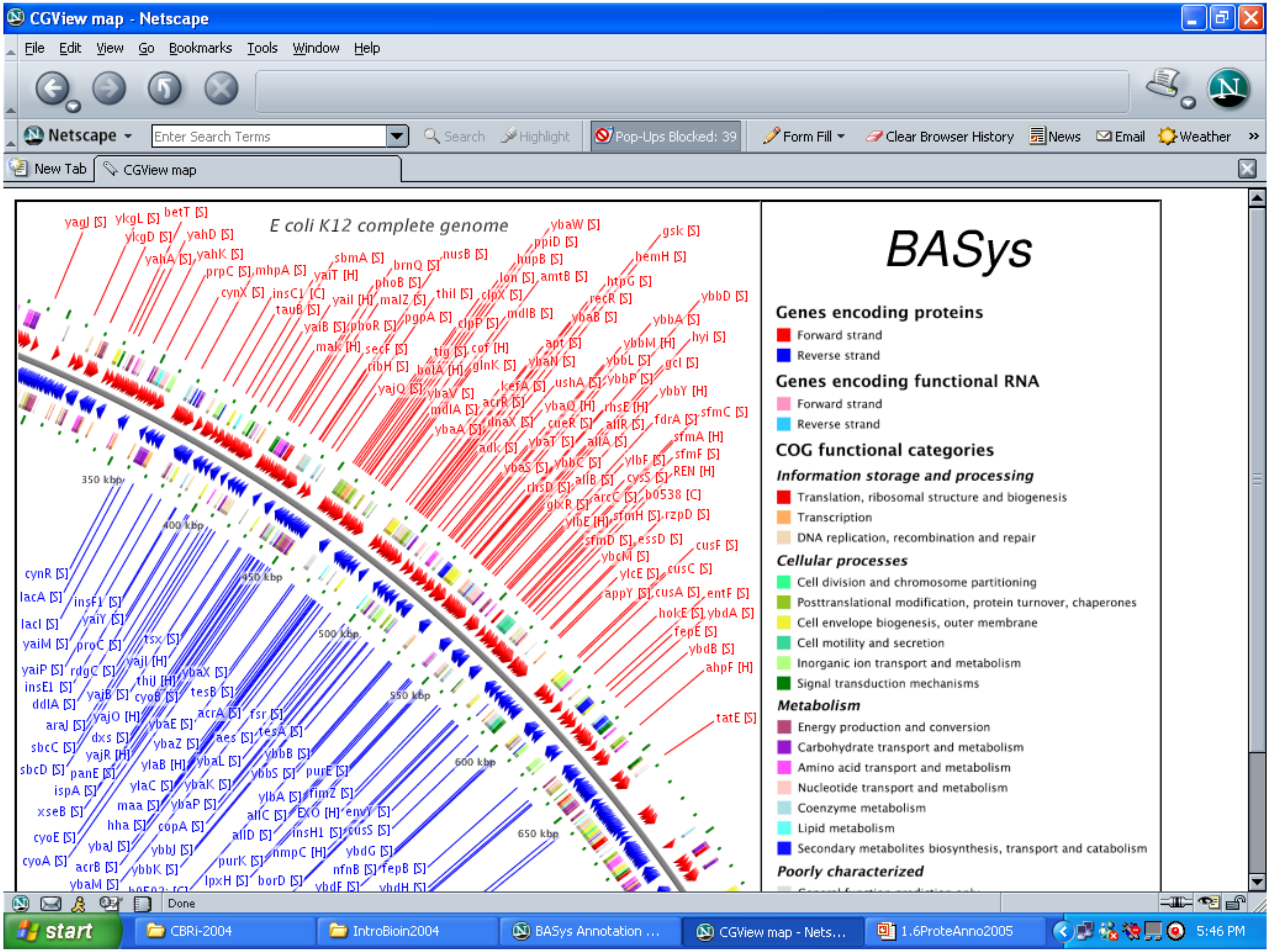
- 🟩 Cell division and chromosome partitioning
- 🟩 Posttranslational modification, protein turnover, chaperones
- 🟨 Cell envelope biogenesis, outer membrane
- 🟩 Cell motility and secretion
- 🟩 Inorganic ion transport and metabolism
- 🟩 Signal transduction mechanisms

Metabolism

- 🟪 Energy production and conversion
- 🟪 Carbohydrate transport and metabolism
- 🟪 Amino acid transport and metabolism
- 🟥 Nucleotide transport and metabolism
- 🟦 Coenzyme metabolism
- 🟦 Lipid metabolism
- 🟦 Secondary metabolites biosynthesis, transport and catabolism

Poorly characterized

BASys Output (Map)



BASys

Genes encoding proteins

- 🟥 Forward strand
- 🟦 Reverse strand

Genes encoding functional RNA

- 🟪 Forward strand
- 🟦 Reverse strand

COG functional categories

Information storage and processing

- 🟥 Translation, ribosomal structure and biogenesis
- 🟧 Transcription
- 🟫 DNA replication, recombination and repair

Cellular processes

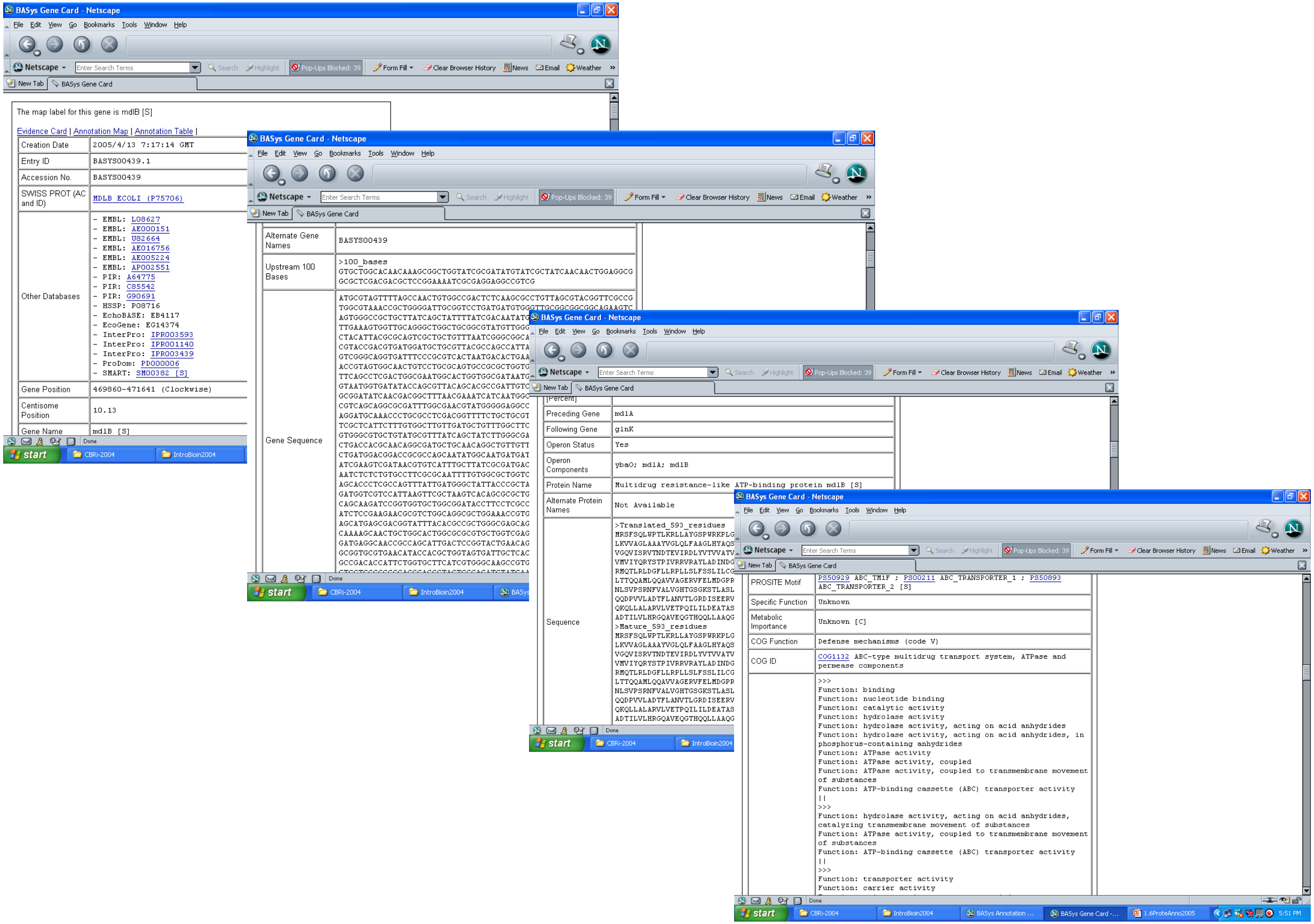
- 🟩 Cell division and chromosome partitioning
- 🟩 Posttranslational modification, protein turnover, chaperones
- 🟨 Cell envelope biogenesis, outer membrane
- 🟩 Cell motility and secretion
- 🟩 Inorganic ion transport and metabolism
- 🟩 Signal transduction mechanisms

Metabolism

- 🟪 Energy production and conversion
- 🟪 Carbohydrate transport and metabolism
- 🟪 Amino acid transport and metabolism
- 🟥 Nucleotide transport and metabolism
- 🟦 Coenzyme metabolism
- 🟦 Lipid metabolism
- 🟦 Secondary metabolites biosynthesis, transport and catabolism

Poorly characterized

BASys Output (Gene Link)



The map label for this gene is mdlB [S]

Evidence Card | Annotation Map | Annotation Table |

Creation Date	2005/4/13 7:17:14 GMT
Entry ID	BASYS00439.1
Accession No.	BASYS00439
SWISS PROT (AC and ID)	MDLB_ECOLI (P75706)
Other Databases	- EMBL: L08627 - EMBL: AE000151 - EMBL: U82664 - EMBL: AE016756 - EMBL: AE005224 - EMBL: AP002551 - PIR: A64775 - PIR: C85542 - PIR: G90691 - HSSP: P08716 - EchoBASE: EB4117 - EcoGene: EG14374 - InterPro: IPR003593 - InterPro: IPR001140 - InterPro: IPR003439 - ProDom: PD000006 - SMART: SM00382 [S]
Gene Position	469860-471641 (Clockwise)
Centisome Position	10.13
Gene Name	mdlB [S]

Alternate Gene Names	BASYS00439
Upstream 100 Bases	>100_bases GTGCTGGCACAACAAAGCGGCTGGTATCGCGATATGTATCGCTATCAACAACTGGAGGCG GCGCTCGACGACGCTCCGGAAAATCGCGAGGAGGCCGTCG
Gene Sequence	ATGCGTAGTTTTAGCCAACTGTGGCCGACTCTCAAGCGCCTGTTAGCGTACGGTTCGCCG TGGCGTAAACCGCTGGGGATTGCGGTCCTGATGATGTGGGTTGCGGCGGCGGCAGAAGTC AGTGGGCCGCTGCTTATCAGCTATTTTATCGACAATATG TTGAAAGTGGTTGCAGGGCTGGCTGCGGCGTATGTTGGG CTACATTACGCGCAGTCGCTGCTGTTTAATCGGGCGGCA CGTACCGACGTGATGGATGCTGCGTTACGCCAGCCATTA GTCGGGCAGGTGATTTCCCGCGTCACTAATGACACTGAA ACCGTAGTGGCAACTGTCCTGCGCAGTGCCGCGCTGGTG TTCAGCCTCGACTGGCGAATGGCACTGGTGGCGATAATG GTAATGGTGATATACCAGCGTTACAGCACGCCGATTGTC GCGGATATCAACGACGGCTTTAACGAAATCATCAATGGC CGTCAGCAGGCGCGATTTGGCGAACGTATGGGGGAGGCC AGGATGCAAACCCTGCGCCTCGACGGTTTTCTGCTGCGT TCGCTCATTCTTTGTGGCTTGTTGATGCTGTTTGGCTTC GTGGGCGTGCTGTATGCGTTTATCAGCTATCTTGGGCGA CTGACCACGCAACAGGCGATGCTGCAACAGGCTGTTGTT CTGATGGACGGACCGCGCCAGCAATATGGCAATGATGAT ATCGAAGTCGATAACGTGTCATTTGCTTATCGCGATGAC AATCTCTCTGTGCCTTCGCGCAATTTTGTGGCGCTGGTC AGCACCCTCGCCAGTTTATTGATGGGCTATTACCCGCTA GATGGTCGTCCATTAAGTTCGCTAAGTCACAGCGCGCTG CAGCAAGATCCGGTGGTGCTGGCGGATACCTTCCTCGCC ATCTCCGAAGAACGCGTCTGGCAGGCGCTGGAAACCGTG AGCATGAGCGACGGTATTTACACGCCGCTGGGCGAGCAG CAAAAGCAACTGCTGGCACTGGCGCGCGTGCTGGTCGAG GATGAGGCAACCGCCAGCATTGACTCCGGTACTGAACAG GCGGTGCGTGAACATACCACGCTGGTAGTGATTGCTCAC GCCGACACCATTCTGGTGCTTCATCGTGGCAAGCCGTG

[Percent]	
Preceding Gene	mdlA
Following Gene	glnK
Operon Status	Yes
Operon Components	ybaO; mdlA; mdlB
Protein Name	Multidrug resistance-like ATP-binding protein mdlB [S]
Alternate Protein Names	Not Available
Sequence	>Translated_593_residues MRSFSQLWPTLKRLLAYGSPWRKPLG LKVVAGLAAAYVGLQLFAAGLHYAQS VGQVISRVTNDTEVIRDLYVTVVATV VMVIYQRYSTPIVRRVRAYLADINDG RMQTLRLDGFLLRPLLSLFSSLILCG LTTQQAMLQQAVVAGERVFELMDGPR NLSVPSRNFVALVGHTGSGKSTLASL QQDPVVLADTFLANVTLGRDISEERV QKQLLALARVLVETPQILILDEATAS ADTILVLHRGQAVEQGTHQQLLAAQG >Mature_593_residues MRSFSQLWPTLKRLLAYGSPWRKPLG LKVVAGLAAAYVGLQLFAAGLHYAQS VGQVISRVTNDTEVIRDLYVTVVATV VMVIYQRYSTPIVRRVRAYLADINDG RMQTLRLDGFLLRPLLSLFSSLILCG LTTQQAMLQQAVVAGERVFELMDGPR NLSVPSRNFVALVGHTGSGKSTLASL QQDPVVLADTFLANVTLGRDISEERV QKQLLALARVLVETPQILILDEATAS ADTILVLHRGQAVEQGTHQQLLAAQG

PROSITE Motif	PS50929 ABC_TM1F ; PS00211 ABC_TRANSPORTER_1 ; PS50893 ABC_TRANSPORTER_2 [S]
Specific Function	Unknown
Metabolic Importance	Unknown [C]
COG Function	Defense mechanisms (code V)
COG ID	COG1132 ABC-type multidrug transport system, ATPase and permease components
	>>> Function: binding Function: nucleotide binding Function: catalytic activity Function: hydrolase activity Function: hydrolase activity, acting on acid anhydrides Function: hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides Function: ATPase activity Function: ATPase activity, coupled Function: ATPase activity, coupled to transmembrane movement of substances Function: ATP-binding cassette (ABC) transporter activity || >>> Function: hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances Function: ATPase activity, coupled to transmembrane movement of substances Function: ATP-binding cassette (ABC) transporter activity || >>> Function: transporter activity Function: carrier activity

Conclusion

- Genome annotation is the same as proteome annotation – required after any gene sequencing and gene ID effort
- Can be done either manually or automatically
- Need for high throughput, automated “pipelines” to keep up with the volume of genome sequence data
- Area of active research and development with about ½ of all bioinformaticians working on some aspect of this process