Bioinformatics and Functional Genomics

Course Overview, Introduction of Bioinformatics, Biology Background

Biol4230 Thurs, Jan 18, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:

- Overview of the course
- Introduction to Bioinformatics – questions, algorithms, resources, data types
- Introduction to Genome Biology – DNA, RNA, and protein (molecule types, sizes, and abundance), gene structure, protein structure
- Preparation for tomorrow's Unix Lecture/Lab

What should you do to reinforce the lecture material?

- Luscombe et al. (2001) "What is Bioinformatics? An introduction and overview" Methods Inf Med. 40:346-58. PMID: 11552348
- Pevsner, Ch. 1, 2
- Recombinant DNA, Ch. 1,2
- Basic Biology:
  - what is the DNA alphabet? the protein alphabet?
  - how does genome organization change?
  - what is an "exon"? an "intron"? which sequences make mRNA?
  - what is an initiation codon (how many are there)? a termination codon (how many)?
- Visit the NCBI website (www.ncbi.nlm.nih.gov), and look up the plant protein alpha amylase in rice. (`alpha amylase AND rice[organism]`)
  - How many proteins are there? How many in RefSeq? What is the longest? The shortest? How many genes?
  - Pick a single rice alpha amylase (one longer than 400 aa) at the NCBI and check its domains (how many?), and gene structure (how many exons?, how many code for protein?).
- Look for rice alpha-amylase proteins at Uniprot (www.uniprot.org).
  - How many alpha-amylases are in SwissProt? In Trembl?
  - Can you find a long (>400 aa) rice alpha amylase in RefSeq that is not found in SwissProt? Can you find it in Trembl?
  - What information is available at the NCBI that is not available at Uniprot?
  - At Uniprot but not NCBI?

Bioinformatics and Functional Genomics – Overview

- Homology, Similarity searching, evolutionary tree reconstruction
  - BLAST and FASTA, scoring matrices, tree-building methods
- Unix at the command line, Python scripting
  - unix commands, directories and files, using an editor
  - writing/debugging Python scripts
- Gene expression analysis (RNAseq)
  - "NextGen" sequence analysis (cleaning, alignment, mapping)
  - 'R' and 'BioConductor'
- Identifying regulatory motifs

Why study/teach bioinformatics?

- The human genome project: 1991 – 2001 knowledge/assumptions before 2001
  - human genome size known (3 billion bp, haploid, 23 chromosomes)
  - E coli (4 million bp, had about 4,000 genes)
  - human gene estimates from 30,000 – 300,000 genes, with most estimates > 100,000
  - ~ 50% of genome was "single copy", 5 – 10% transcribed in most tissues (greater in brain)
- human genome, post 2001
  - correct genome size
  - 15,000 – 20,000 genes (smaller than plants)
  - <2% of genome transcribed into proteins
  - most individuals have 100 – 500 non-functional (truncated) protein coding genes
- Bioinformatics illustrates the shortcomings of "big data" approaches. The enormous increase in data "volume" seems to raise more questions than provide answers.

How to determine what's "true"?

Bioinformatics and Functional Genomics – What will you learn?

- Similarity searching, from the command line, and using scripts
- Multiple sequence alignment and phyogeny reconstruction
- Large-scale sequence mapping, and genome sequence manipulation
- (Regulatory) Motif finding
- Biological Pathway analysis

What are the algorithmic and biological reasons for errors and inconsistencies?
What can we trust?

What is Bioinformatics?

- Data organization
  - sequence/structure/expression/variation databases (resources)
  - Nucleic Acid Res. database, web server, issue
- Development of algorithms/statistics/tools
  - FASTA, BLAST, CLUSTAL, MUSCLE, PHYLIP, BALIPHY, MAC, TOPHAT, CUFFLINKS, BIOCONDUCTOR, DAVID
- Application and evaluation of analysis methods to understand biological processes
  - what does an unknown protein do (activity)?
  - what genes are up/down-regulated in cancer?
  - what mutations increase/reduce heart disease?

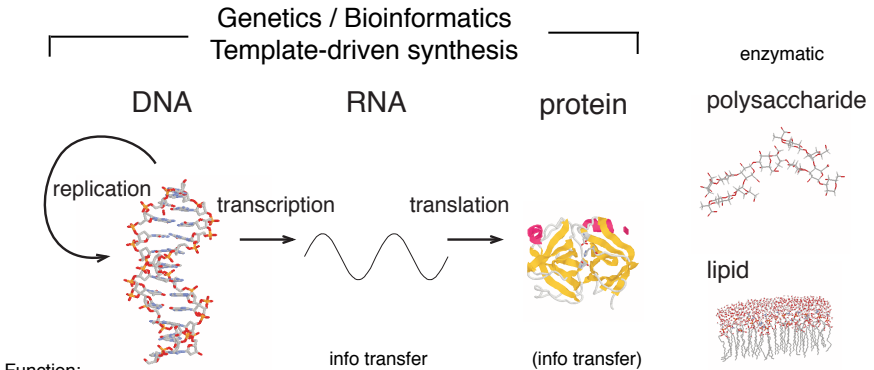
Luscombe et al. (2001) PMID: 11552348

What is Bioinformatics?

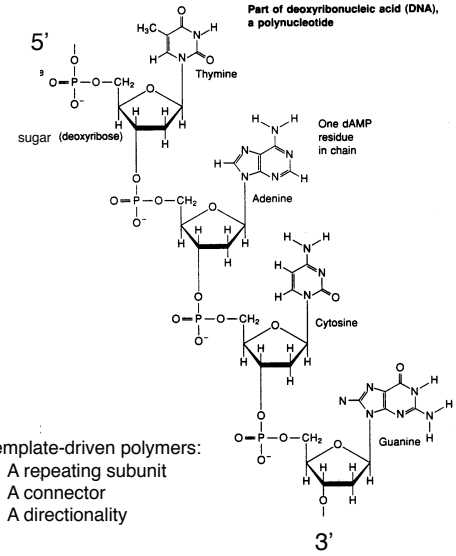
Bioinformatics explores differences (changes) in DNA, RNA, and protein sequence and abundance.

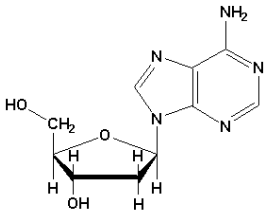
- genetic information – molecules made from a genetic template
- changes in DNA: variation
  - mutation (single site) or copy number variation (gene or multigene regions)
  - no changes in abundance, all cells have (almost) the same DNA content
- changes in RNA: structure and abundance
  - different cells express (make RNA for) different genes
  - different RNAs can be made from the same gene
- changes in protein:
  - abundance (dependent on RNA abundance, but other factors) – partially genetic
  - post-translational modification (non-template changes)
  - interactions and binding partners

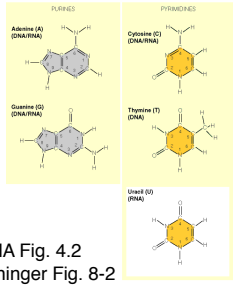
The Central Dogma of Molecular Biology
Molecules for Information transfer, storage, and function



Polymers and Monomers - DNA



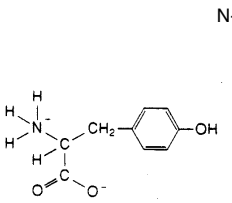




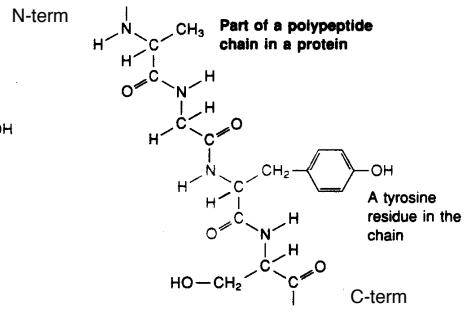
Template-driven polymers:
1. A repeating subunit
2. A connector
3. A directionality

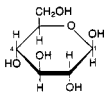
MvHA Fig. 4.2
Lehninger Fig. 8-2

Monomers and Polymers - proteins



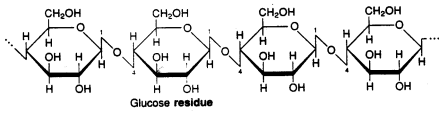
tyrosine monomer





β-D-glucose

carbohydrates



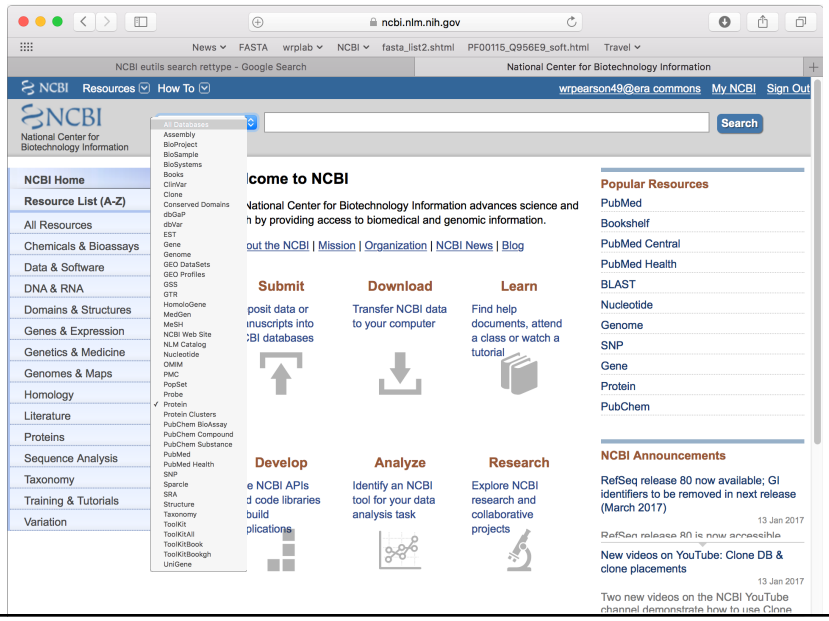
Basic biology you should know

- Central dogma DNA transcribed into RNA translated into protein
- Prokaryotes – bacteria, archaea
  - no nuclei or mitochondria
  - small genomes (1,000 – 5,000 genes), >90% of genome is protein coding
  - RNA transcript = mRNA (unspliced)
- Eukaryotes – higher organisms (yeast, plants, people)
  - nuclei, mitochondria, chloroplasts (plants)
  - small (yeast) to large (plants, metazoa) genomes
  - large genomes have similar numbers of genes (10,000 – 20,000), but < 5% of genome codes for protein
  - RNA transcripts can be spliced into mRNA
- proteins – (20 amino acids)
  - average size ~400 amino acids, range from 10 – 40,000 amino acids
  - are directional (start at N-terminus, initiation codon, AUG, end at C-terminus, stop codon, UAA, UAG, UGA)
  - fold into distinct 3-D structures, characterized by alpha-helices, beta-sheets
- mRNA – (4 nucleotides, 61 codons for amino acids + 3 termination)
  - average size ~2000 nucleotides (1200 nt code for protein, remainder short 5'-untranslated, long 3-untranslated), end with poly-A (added after transcription)
  - in prokaryotes, same as transcript
  - in eukaryotes, built from exons (separated by introns) from a much longer transcript
  - RNAs differ in abundance (>1000-fold) in different tissues

Protein Sequence and Structure Databases

1. NCBI/Entrez – Most comprehensive, linked to PubMed.
   - Best known: GenBank / GenPept , but probably least useful.
   - Most annotated: RefSeq
   - Best links to human disease: Entrez/Gene and OMIM.
2. Uniprot – Most information about proteins
   - Functional information (functional sites)
   - Links to other databases (InterPro for domains)
3. 1,500+ Biological/disease/genetic/variation databases
   - Nucleic Acids Research database issue nar.oxfordjournals.org/content/45/D1/D1.abstract

www.ncbi.nlm.nih.gov



13

NCBI Databases and Services

- GenBank primary sequence database
- Free public access to biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Software and databases for download

fasta.bioch.virginia.edu/biol4230

14

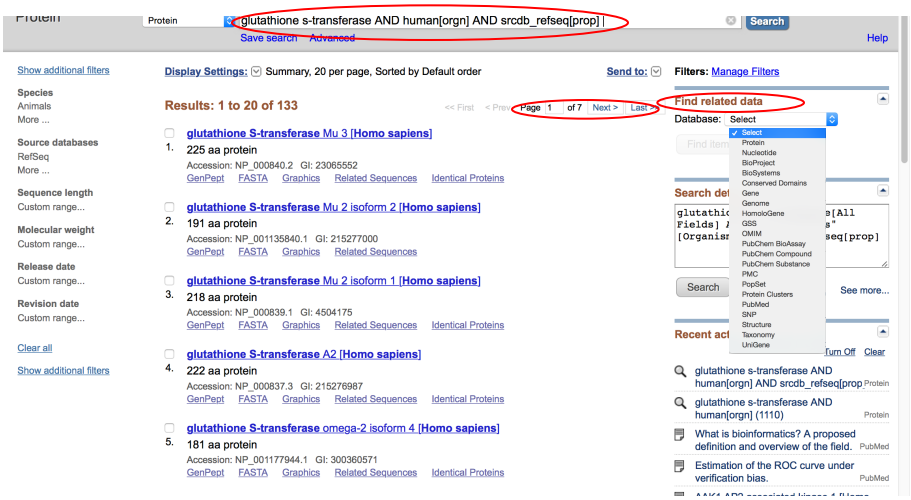
Types of Databases

Why do we care – database "dimensions"

1. Completeness (what is included, left out?)
2. Correctness (who corrects errors?)

- Primary Databases (avoid)
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: GenBank, SNP, GEO
- Derivative Databases (use)
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: NCBI Protein, Refseq, TPA, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain

Finding protein sequences with Entrez/Proteins



ncbi.nlm.nih.gov

Apple News FASTA wrplab NCBI Travel

glutathione S-transferase Mu 1 isoform 1 [Homo sapiens] - Protein - NCBI

Protein | Search | Advanced | Help

Display Settings: GenPept | Send to:

Change region shown

Customize view

glutathione S-transferase Mu 1 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP_000552.2

Identical Proteins FASTA Graphics

Pathways for the GSTM1 gene
- Chemical carcinogenesis
- Aflatoxin B1 metabolism
- Estrogen metabolism

See all...

Reference sequence information
- RefSeq genomic sequence
  See the genomic reference sequence for the GSTM1 gene (NG_009246.1).
- RefSeq mRNA
  See reference mRNA sequence for the GSTM1 gene (NM_000561.3).
- RefSeq protein isoforms
  See 4 reference sequence protein isoforms for the GSTM1 gene.

More about the GSTM1 gene
Cytosolic and membrane-bound forms of glutathione S-transferase are encoded by two distinct supergene families. At present, eight distinct c...
Also Known As: GST1, GSTM1-1, GSTM1a-1...

Homologs of the GSTM1 gene
The GSTM1 gene is conserved in chimpanzee, Rhesus monkey, cow, mouse, rat, and frog.

LinkOut to external resources
- NP_000552 [Domain Mapping of Disease Mut...]
- Ensembl [Ensembl]
- A selection of literature about the proteins [GoPubMed Proteins]
- Protein expression data [Model Organism Protein Expres...]
- Transcript/Protein Information [PANTHER Classification System]
- PSI Structural Biology Knowledgebase [PSI Structural Biology Knowle...]
- antibody review [ExactAntigen/Labome]
- biochemicals [ExactAntigen/Labome]
- protein and peptide [ExactAntigen/Labome]
- antibody [ExactAntigen/Labome]
- cDNA clone [ExactAntigen/Labome]
- siRNA and shRNA [ExactAntigen/Labome]
- others [ExactAntigen/Labome]
- Evolutionary Trace of Functional Site [Evolutionary Trace of Functio...]

Related information
- BLink
- Related Sequences
- Identical Proteins
- BioAssay by Target (Identical Proteins, List)
- BioAssay by Target (Identical Proteins, Summary)
- BioProject
- BioSystems
- CCDS
- CDD Search Results
- Conserved Domains (Concise)
- Conserved Domains (Full)
- Domain Relatives
- Encoding mRNA
- Full text in PMC
- Gene
- Gene Genotype
- GeneView in dbSNP
- Genome
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Protein (UniProtKB)
- PubMed

Analyze this sequence
- Run BLAST
- Identify Conserved Domains
- Highlight Sequence Features
- Find in this Sequence

Protein 3D Structure
no image yet
Structure Of Human Glutathione S-Transferase M1a-1a Complexed With
PDB: 2F3M
Source: Homo sapiens
Method: X-Ray Diffraction
Resolution: 2.7 Å

See all 5 structures...

Articles about the GSTM1 gene
- [Analysis of a GSTM1 gene deletion in the context of the [Mol Gen Mikrobiol Virusol. 2014]
- Glutathione S-transferase M1 null genotype meta-analysis on gastric c [Diagn Pathol. 2014]
- Genetic variants in the glutathione S-transferase ([Pharmacogenet Genomics. 2014]

See all...

fasta.bioch.vir

17

Entrez Gene: genetic/genomic information

Display Settings: Full Report | Send to: | Hide sidebar >>

Filters activated: Current only. Clear all

GSTM1 glutathione S-transferase mu 1 [*Homo sapiens* (human)]

Gene ID: 2944, updated on 4-Jan-2015

Summary

Official Symbol	GSTM1 provided by HGNC
Official Full Name	glutathione S-transferase mu 1 provided by HGNC
Primary source	HGNC:HGNC:4632
See related	Ensembl:ENSG00000134184; HPRD:00707; MIM:138350; Vega:OTTHUMG00000011635
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	MU; H-B; GST1; GTH4; GTM1; MU-1; GSTM1-1; GSTM1a-1a; GSTM1b-1b
Summary	Cytosolic and membrane-bound forms of glutathione S-transferase are encoded by two distinct supergene families. At present, eight distinct classes of the soluble cytoplasmic mammalian glutathione S-transferases have been identified: alpha, kappa, mu, omega, pi, sigma, theta and zeta. This gene encodes a glutathione S-transferase that belongs to the mu class. The mu class of enzymes functions in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione. The genes encoding the mu class of enzymes are organized in a gene cluster on chromosome 1p13.3 and are known to be highly polymorphic. These genetic variations can change an individual's susceptibility to carcinogens and toxins as well as affect the toxicity and efficacy of certain drugs. Null mutations of this class mu gene have been linked with an increase in a number of cancers, likely due to an increased susceptibility to environmental toxins and carcinogens. Multiple protein isoforms are encoded by transcript variants of this gene. [provided by RefSeq, Jul 2008]

Genomic context

Location: 1p13.3 — See GSTM1 in Epigenomics, MapViewer

Exon count: 8

Annotation release	Status	Assembly	Chr	Location

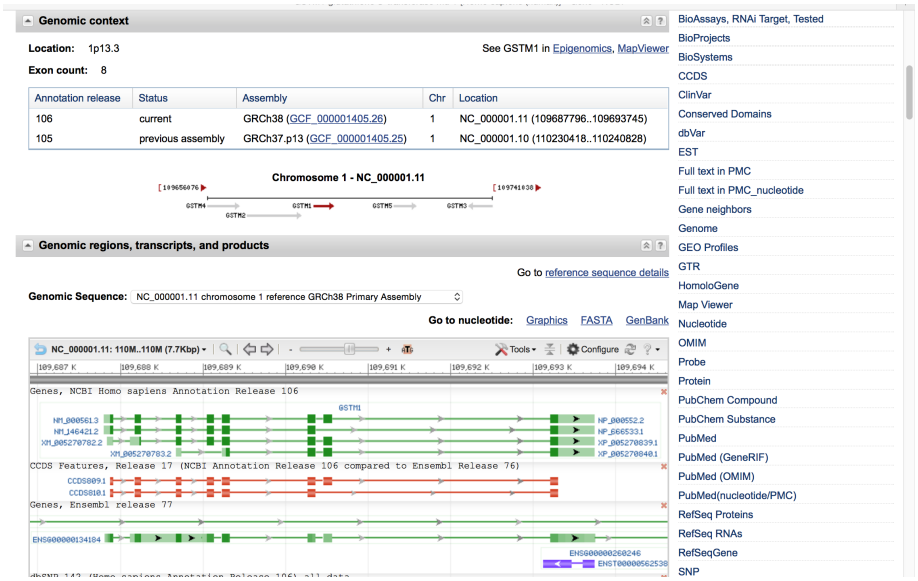
Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Clone Names, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Related information
- Order cDNA clone
- 3D structures
- BioAssay
- BioAssay by Target (List)
- BioAssay by Target (Summary)
- BioAssay, by Gene target
- BioAssays, RNAi Target, Tested
- BioProjects
- BioSystems
- CCDS
- ClinVar

fasta.bioch.virginia.edu/biol4230

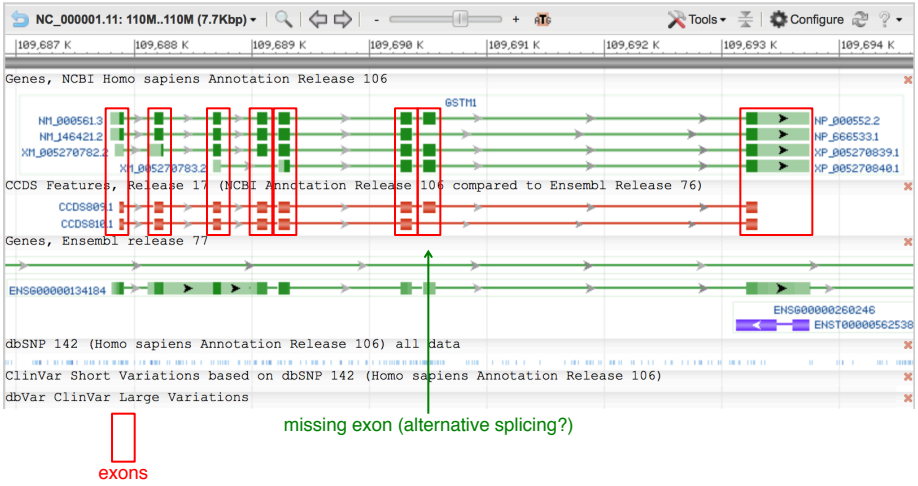
18

Entrez Gene: genetic/genomic information



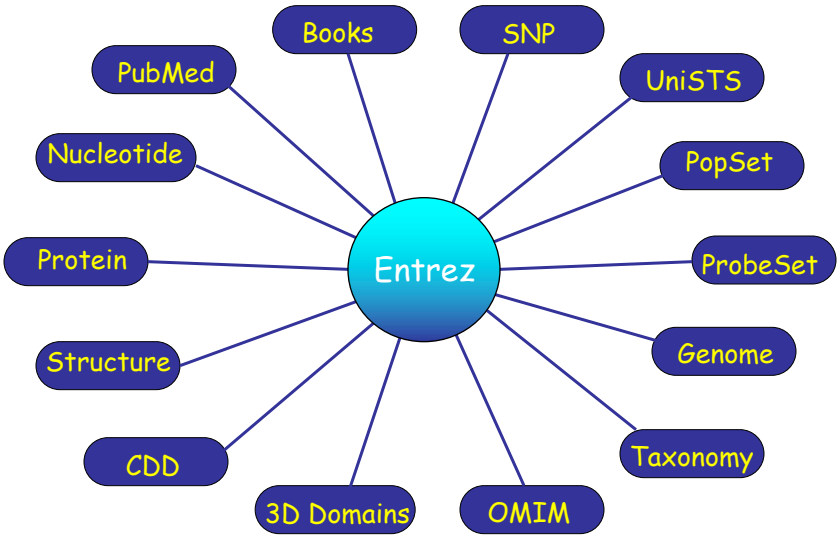
fasta.bioch.virginia.edu/biol4230 19

Entrez Gene: Genomic/transcript structure



fasta.bioch.virginia.edu/biol4230 20

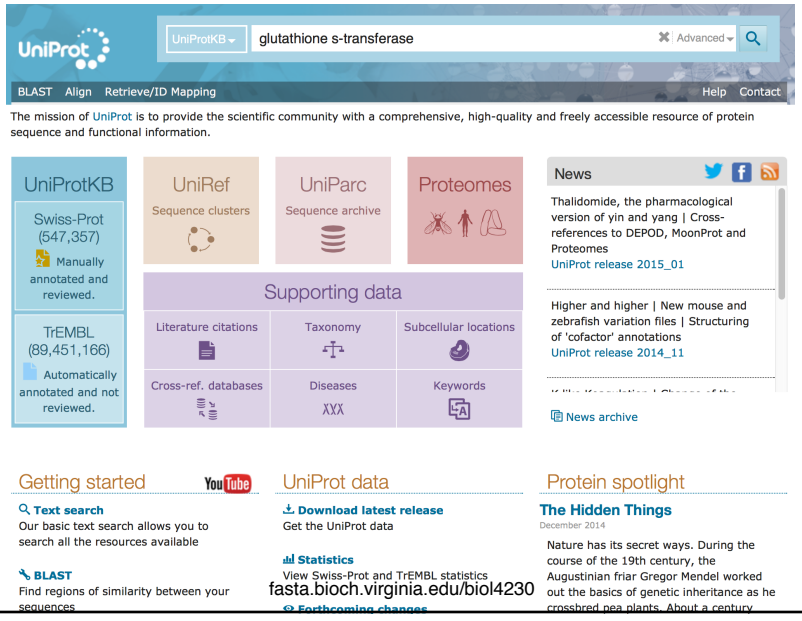
The (ever) Expanding Entrez System



fasta.bioch.virginia.edu/biol4230

21

Uniprot/SwissProt (uniprot.org) Comprehensive (inclusive) Database links



fasta.bioch.virginia.edu/biol4230

22

UniProtKB | glutathione "s transferase" | Advanced

BLAST Align Retrieve/ID Mapping Help Contact

Show help for UniProtKB

Basket

Results

High quality (reviewed) annotations

Filter by

Reviewed (349) Swiss-Prot

Unreviewed (64,350) TrEMBL

TREMBL/ unreviewed

Popular organisms

Human (104)
Rice (93)
A. thaliana (76)
Mouse (66)
Rat (57)

Other organisms

Go

Search terms

Filter "glutathione" as:
gene name (1)

View by

Columns | BLAST | Align | Download | Add to basket | 1 to 25 of 64,699 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
Q26387	Q26387_HELPZ	**Glutathione S-transferase**	**glutathione S-transferase: GST**	Heligmosomoides polygyrus (Parasitic roundworm)	216
P13745	GSTA1_MOUSE	**Glutathione S-transferase A1**	**Gsta1**, Gsta, Gstya	Mus musculus (Mouse)	223
P00502	GSTA1_RAT	**Glutathione S-transferase alpha-1**	**Gsta1**	Rattus norvegicus (Rat)	222
P30713	GSTT2_RAT	**Glutathione S-transferase theta-2**	**Gstt2**	Rattus norvegicus (Rat)	244
P30115	GSTA3_MOUSE	**Glutathione S-transferase A3**	**Gsta3**, Gstyc	Mus musculus (Mouse)	221
P78417	GSTO1_HUMAN	**Glutathione S-transferase omega-1**	**GSTO1**, GSTTLP28	Homo sapiens (Human)	241
P04905	GSTM1_RAT	**Glutathione S-transferase Mu 1**	**Gstm1**	Rattus norvegicus (Rat)	218
P08263	GSTA1_HUMAN	**Glutathione S-**	**GSTA1**	Homo sapiens	222

fasta.bioch.virginia.edu/biol4230 23

Basket

P09488 - GSTM1_HUMAN

Protein: **Glutathione S-transferase Mu 1**
Gene: **GSTM1**
Organism: *Homo sapiens (Human)*
Status: Reviewed - - Experimental evidence at protein level

Display None

FUNCTION
NAMES & TAXONOMY
SUBCELL. LOCATION
PATHOL./BIOTECH
PTM / PROCESSING
EXPRESSION
INTERACTION
STRUCTURE
FAMILY & DOMAINS
SEQUENCES (2)
CROSS-REFERENCES
PUBLICATIONS
ENTRY INFORMATION
MISCELLANEOUS
SIMILAR PROTEINS

BLAST | Align | Format | Add to basket | History | Feedback | Help video
Show feature tables

Function

Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles. 1 Publication

Catalytic activity
RX + glutathione = HX + R-S-glutathione. 1 Publication

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Binding site	116 – 116	1	Substrate			

GO - Molecular function
- enzyme binding Source: BHF-UCL
- glutathione transferase activity Source: BHF-UCL
- glutathione binding Source: BHF-UCL
- protein homodimerization activity Source: BHF-UCL

GO - Biological process
- cellular detoxification of nitrogen compound Source: BHF-UCL
- glutathione derivative biosynthetic process Source: Reactome
- glutathione metabolic process Source: BHF-UCL
- small molecule metabolic process Source: Reactome
- nitrobenzene metabolic process Source: BHF-UCL
- xenobiotic catabolic process Source: BHF-UCL

fasta.bioch.virginia.edu/biol4230 24

Structure

Secondary structure

1 — 218

Legend: Helix Turn Beta strand

Show more details

3D structure databases

Select the link destinations:
- PDBe
- RCSB PDB
- PDBj

Entry	Metho
1GTU	X-ray
1XW6	X-ray
1XWK	X-ray
1YJ6	X-ray
2F3M	X-ray

ProteinModelPortal	P09488.
SMR	P09488. Position:
ModBase	Search...
MobiDB	Search...

Miscellaneous databases

EvolutionaryTrace	P09488.

Family and domain databases

Gene3D	1.20.1050.10. 1 hit. 3.40.30.10. 1 hit.
InterPro	IPR010987. Glutathione-S-Trfase_C-like. IPR004045. Glutathione_S-Trfase_N. IPR004046. GST_C. IPR003081. GST_mu. IPR012336. Thioredoxin-like_fold. [Graphical view]
Pfam	PF00043. GST_C. 1 hit. PF02798. GST_N. 1 hit. [Graphical view]
PRINTS	PR01267. GSTRNSFRASEM.
SUPFAM	SSF47616. SSF47616. 1 hit. SSF52833. SSF52833. 1 hit.
PROSITE	PS50405. GST_CTER. 1 hit. PS50404. GST_NTER. 1 hit. [Graphical view]

Glutathione S-transferase GSTM1

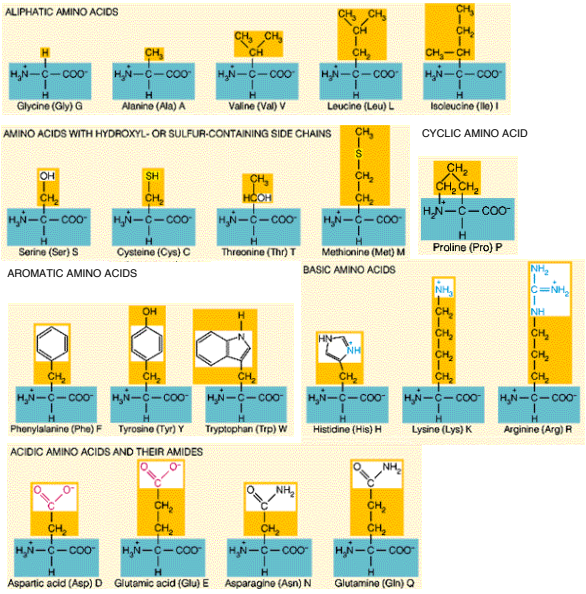
```
>sp|P09488|GSTM1_HUMAN Glutathione S-transferase Mu 1 GN=GSTM1
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNL
PYLIDGAHKITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEF
EKLKPKYLEELPEKLKLYSEFLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPN
LKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK
```

Sequence in "FASTA" format

Structure and properties of Amino-acids

Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamine	Gln	Q	Serine	Ser	S
Glutamic acid	Glu	E	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V
Asp/Asn	Asx	B	Glu/Gln	Glx	Z

Figure 5.3: The amino acids found in proteins.



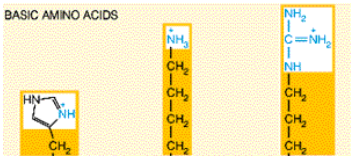
Some amino acids are more common than others:

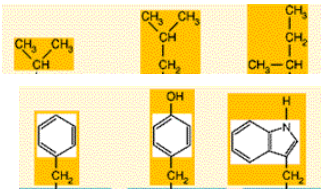
+	Ala	A	0.0780
	Arg	R	0.0512
	Asn	N	0.0448
	Asp	D	0.0536
-	Cys	C	0.0192
	Gln	Q	0.0426
+	Glu	E	0.0629
+	Gly	G	0.0737
-	His	H	0.0219
	Ile	I	0.0514
+	**Leu**	**L**	**0.0901**
	Lys	K	0.0574
-	Met	M	0.0224
-	Phe	F	0.0385
	Pro	P	0.0520
+	Ser	S	0.0711
	Thr	T	0.0584
-	***Trp***	***W***	***0.0132***
-	Tyr	Y	0.0321
+	Val	V	0.0644

Robinson and Robinson,
PNAS (1991) 88:8880

Amino acid Hydropathicity/Hydrophobicity

Hopp T.P., Woods K.R. (1981) PNAS. 78:3824-3828.
Kyte J., Doolittle R.F. (1982). J. Mol. Biol. 157:105-132
D. M. Engelman, T. A. Steitz, A. Goldman, (1986) Annu. Rev. Biophys. Biophys. Chem. 15, 321





Hopp/ Woods		Kyte/ Doolittle		GES	
Arg:	3.0	Arg:	-4.5	Arg:	12.3
Lys:	3.0	Lys:	-3.9	Asp:	9.2
Asp:	3.0	Asp:	-3.5	Lys:	8.8
Glu:	3.0	Glu:	-3.5	Glu:	8.2
Ser:	0.3	Gln:	-3.5	Asn:	4.8
Gln:	0.2	Asn:	-3.5	Gln:	4.1
Asn:	0.2	His:	-3.2	His:	3.0
Pro:	0.0	Pro:	-1.6	Tyr:	0.7
Gly:	0.0	Tyr:	-1.3	Pro:	0.2
Thr:	-0.4	Trp:	-0.9	Ser:	-0.6
His:	-0.5	Ser:	-0.8	Gly:	-1.0
Ala:	-0.5	Thr:	-0.7	Thr:	-1.2
Cys:	-1.0	Gly:	-0.4	Ala:	-1.6
Met:	-1.3	Ala:	1.8	Trp:	-1.9
Val:	-1.5	Met:	1.9	Cys:	-2.0
Leu:	-1.8	Cys:	2.5	Val:	-2.6
Ile:	-1.8	Phe:	2.8	Leu:	-2.8
Tyr:	-2.3	Leu:	3.8	Ile:	-3.1
Phe:	-2.5	Val:	4.2	Met:	-3.4
Trp:	-3.4	Ile:	4.5	Phe:	-3.7

Amino-acid classes from evolution/mutation

Given a set of (closely) related protein sequences...

```
GSTM1_HUMAN     MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM2_HUMAN     MPMTLGYWNIRGLAHSIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM4_HUMAN     MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGGAPDYDRSQWLNEKFKLGLD
GSTM5_HUMAN     MPMTLGYWDIRGLAHAIRLLLEYTDSSYVEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM1_MOUSE     MPMILGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRSQWLNEKFKLGLD
GSTM2_MOUSE     MPMTLGYWDIRGLAHAIRLLLEYTDTSYEDKKYTMGDAPDYDRSQWLSEKFKLGLD
GSTM3_MOUSE     MPMTLGYWNTRGLTHSIRLLLEYTDSSYEEKRYVMGDAPNFDRSQWLSEKFNLGLD
GSTM4_MOUSE     MSMVLGYWDIRGLAHAIRMLLEFTDTSYEEKRYIQGEAPDYDRSQWLDVKFKLDLD
GSTM3_RABIT     MPMTLGYWDVRGLALPIRMLLEYTDISYEEKKYTMGDAPNYDQSKWLSEKFTLGLD
```

… how often is one amino-acid replaced by another?

REVIEW

Central dogma, databases, and amino-acids

- DNA, RNA, and proteins are template driven bio-polymers (what is the template for each?)
- Today, secondary, curated databases provide much more biological information than primary databases
- The 20 amino acids can be divided into different functional/chemical classes (they are not equally frequent)

Basic biology you should know

- Central dogma DNA transcribed into RNA translated into protein
- Prokaryotes – bacteria, archaea
  - no nuclei or mitochondria
  - small genomes (1,000 – 5,000 genes), >90% of genome is protein coding
  - RNA transcript = mRNA (unspliced)
- Eukaryotes – higher organisms (yeast, plants, people)
  - nuclei, mitochondria, chloroplasts (plants)
  - small (yeast) to large (plants, metazoa) genomes
  - large genomes similar numbers of genes (10,000 – 20,000), but < 5% of genome codes for protein
  - RNA transcripts can be spliced into mRNA
- proteins – (20 amino acids)
  - average size ~400 amino acids, range from 10 – 40,000 amino acids
  - are directional (start at N-terminus, initiation codon, AUG, end at C-terminus, stop codon, UAA, UAG, UGA)
  - fold into distinct 3-D structures, characterized by alpha-helices, beta-sheets
- mRNA – (4 nucleotides, 61 codons for amino acids + 3 termination)
  - average size ~2000 nucleotides (1200 nt code for protein, remainder short 5'-untranslated, long 3-untranslated), end with poly-A (added after transcription)
  - in prokaryotes, same as transcript
  - in eukaryotes, built from exons (separated by introns) from a much longer transcript
  - RNAs differ in abundance (>1000-fold) in different tissues

What should you do to reinforce the lecture material?

- Luscombe et al. (2001) "What is Bioinformatics? An introduction and overview" Methods Inf Med. 40:346-58. PMID: 11552348
- Pevsner, Ch. 1, 2
- Recombinant DNA, Ch. 1,2
- Basic Biology:
  - what is the DNA alphabet? the protein alphabet?
  - what is an "exon"? an "intron"? which sequences make mRNA?
  - what is an initiation codon (how many are there)? a termination codon (how many)?
- Visit the NCBI website (www.ncbi.nlm.nih.gov), and look up the plant protein alpha amylase in rice. (`alpha amylase AND rice[organism]`)
  - How many proteins are there? How many in RefSeq? What is the longest? The shortest? How many genes?
  - Pick a single rice alpha amylase (one longer than 400 aa) at the NCBI and check its domains (how many?), and gene structure (how many exons?, how many code for protein?).
- Look for rice alpha-amylase proteins at Uniprot (www.uniprot.org).
  - How many alpha-amylases are in SwissProt? In Trembl?
  - Can you find a long (>400 aa) rice alpha amylase in RefSeq that is not found in SwissProt? Can you find it in Trembl?
  - What information is available at the NCBI that is not available at Uniprot?
  - At Uniprot but not NCBI?

Before Unix Lab (tomorrow, Friday)

1. Make certain your laptop can use the "Cavalier" wireless
2. Windows: download and install SecureCRT
3. Know/reset your "`its`" eservices password
   `its.virginia.edu/accounts/createacct.html`
4. (For work outside UVA) Install UVA Anywhere VPN
5. Try to connect (ssh) to `interactive.hpc.virginia.edu`