

## Bioinformatics and Functional Genomics

### *Course Overview, Introduction of Bioinformatics, Biology Background*

Biol4230 Thurs, Jan 17, 2017

Bill Pearson [wrp@virginia.edu](mailto:wrp@virginia.edu) 4-2818 Jordan 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

1

## 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](http://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](http://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?

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

2

## 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

3

## 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"?

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

4

## Bioinformatics and Functional Genomics – What will you learn?

- Similarity searching, from the command line, and using scripts
- Multiple sequence alignment and phylogeny 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?

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

5

## 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

6

# What is Bioinformatics?

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

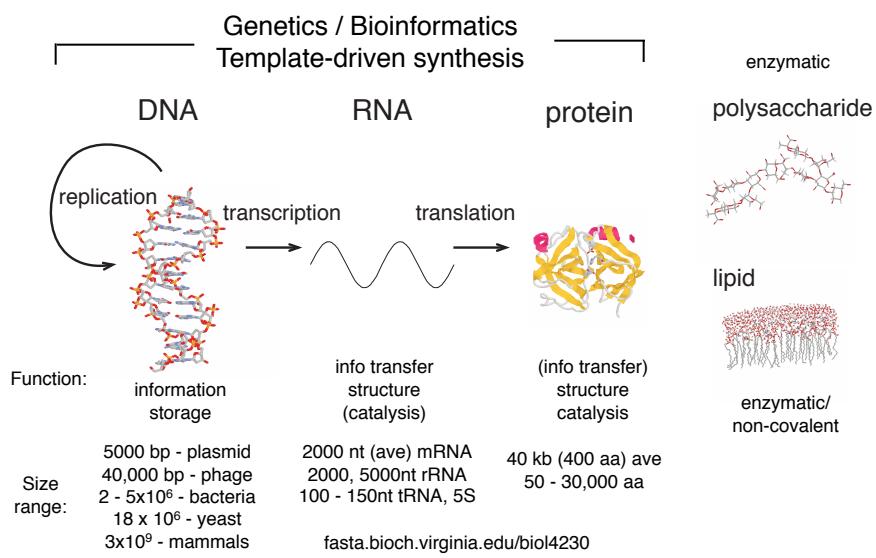
- genetic information – molecules made from a genetic template
  - changes in DNA: variation
    - mutation (single site) or copy number variation (gene or multigene regions)
    - no changes in abundance, all cells have (almost) the same DNA content
  - changes in RNA: structure and abundance
    - 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

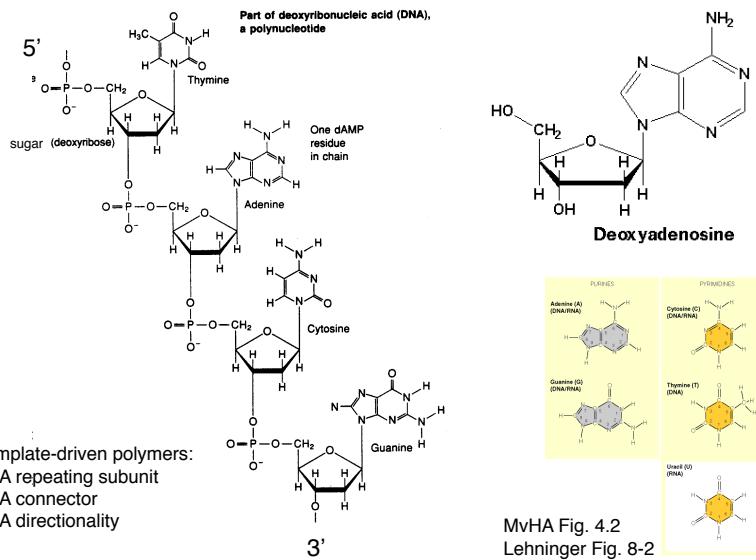
7

# The Central Dogma of Molecular Biology

## Molecules for Information transfer, storage, and function



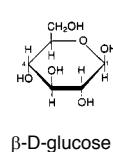
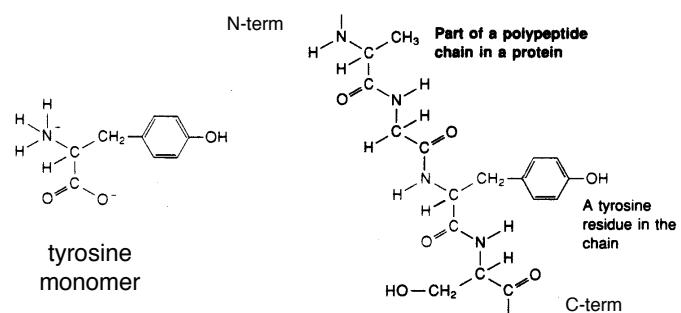
## Polymers and Monomers - DNA



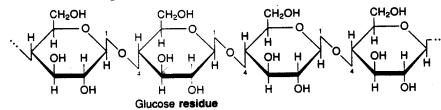
[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

9

## Monomers and Polymers - proteins



### carbohydrates



[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

10

## Basic biology you should know

- Central dogma DNA transcribed into RNA translated into protein
- Prokaryotes – bacteria, archaea
  - no nuclei
  - 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

11

## 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](http://nar.oxfordjournals.org/content/45/D1/D1.abstract)

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

12

www.ncbi.nlm.nih.gov

The screenshot shows the NCBI homepage with a sidebar containing links like NCBI Home, Resource List (A-Z), and a detailed list of databases such as Assembly, BioProject, BioCatalog, BioSystems, Books, ClinVar, Disease, Conserved Domains, dbGaP, Bioneer, EST, Gene, Genomes, GEO Datasets, GEO Profiles, GSS, GTR, HomoloGene, MedGen, Mesh, NCBI Web Site, NLM Catalog, Nucleotide, OMIM, PMS, PopSet, Probe, Protein Clusters, PubChem Compound, PubChem Substance, PubMed Health, SRS, Sparcle, SRA, Structure, Taxonomy, Toolkit, ToolkitAll, ToolkitBook, UniGene. The main content area features sections for Welcome to NCBI, Submit, Download, Learn, Develop, Analyze, and Research, along with Popular Resources and NCBI Announcements.

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

- 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

fasta.bioch.virginia.edu/biol4230

15

## Finding protein sequences with Entrez/Proteins

The screenshot shows the NCBI Entrez Proteins search interface. The search query in the top bar is circled in red: "glutathione s-transferase AND human[orgn] AND srchb\_refseq[prop]". The search results page displays 1 to 20 of 133 entries, with the first five listed:

1. glutathione S-transferase Mu 2 [Homo sapiens]
  - Accession: NP\_000840.2 GI: 23065552
  - GenPept FASTA Graphics Related Sequences Identical Proteins
2. glutathione S-transferase Mu 2 isoform 2 [Homo sapiens]
  - Accession: NP\_001135840.1 GI: 215277000
  - GenPept FASTA Graphics Related Sequences
3. glutathione S-transferase Mu 2 isoform 1 [Homo sapiens]
  - Accession: NP\_000839.1 GI: 4504175
  - GenPept FASTA Graphics Related Sequences Identical Proteins
4. glutathione S-transferase A2 [Homo sapiens]
  - Accession: NP\_000837.3 GI: 215276987
  - GenPept FASTA Graphics Related Sequences Identical Proteins
5. glutathione S-transferase omega-2 isoform 4 [Homo sapiens]
  - Accession: NP\_001177944.1 GI: 300380571
  - GenPept FASTA Graphics Related Sequences Identical Proteins

The right side of the interface includes a "Filters: Manage Filters" panel with a dropdown menu for "Database: Select" (circled in red), a search bar, and a "Recent act" section.

fasta.bioch.virginia.edu/biol4230

16

ncbi.nlm.nih.gov

Protein Protein Advanced Search Help

Display Settings: GenPept Send to: Change region shown

Customize view Analyze this sequence

Run BLAST Identify Conserved Domains Highlight Sequence Features Find in this Sequence

**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 NP\_000552 [Domain Mapping of Disease Mut...]

Aflatoxin B1 metabolism Ensembl Related Sequences

Estrogen metabolism Ensembl Identical Proteins

BioAssay by Target (Identical Proteins, List)

See all... A selection of literature about the proteins BioAssay by Target (Identical Proteins, Summary)

Protein expression data [Model Organism Protein Express...]

LinkOut to external resources BioProject

Related information BLINK

Reference sequence information Transcript/Protein Information BioSystems

RefSeq genomic sequence [PANTHER Classification System] CDDS

See the genomic reference sequence for the GSTM1 gene (NG\_009246.1). PSI Structural Biology Knowledgebase

RefSeq mRNA [PSI Structural Biology Knowledgebase]

See reference mRNA sequence for the GSTM1 gene (NM\_000551.3).

RefSeq protein isoforms antibody review [ExactAntigen/Labome]

See 4 reference sequence protein isoforms for the GSTM1 gene.

biochemicals [ExactAntigen/Labome]

protein and peptide [ExactAntigen/Labome]

Domain Relatives Encoding mRNA

Full text in PMC

More about the GSTM1 gene antibody Gene (circled)

Cytosolic and membrane-bound forms of glutathione S-transferase are encoded by two distinct supergene families. At present, eight distinct c... Gene Genotype

Also Known As: GST1, GSTM1-1, GSTM1-1... cDNA clone GeneView in dbSNP

Others siRNA and shRNA Genome

Homologs of the GSTM1 gene Evolutionary Trace of Functional Site [Evolutionary Trace of Functio...]

The GSTM1 gene is conserved in chimpanzee, siRNA and shRNA [ExactAntigen/Labome] OMIM

Rhesus monkey, cow, mouse, rat, and frog. others HomoloGene

Map Viewer Nucleotide

Protein (UniProtKB) See all 5 structures...

PubMed (RefSeq)

fastabioch.virginia.edu/biol4230 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; Haplorhini; Catarrhini; Hominidae; Homo

Also known as MU; H-8; GST1; GTH4; GTM1; MU-1; GSTM1-1; GSTM1-α-1; 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

fastabioch.virginia.edu/biol4230 18

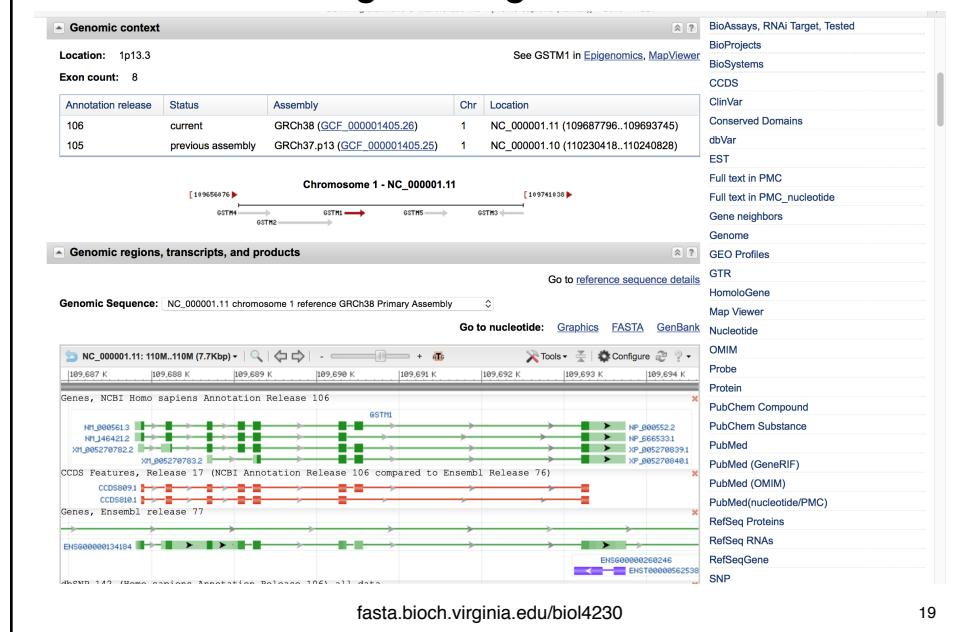
Table of contents

- 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

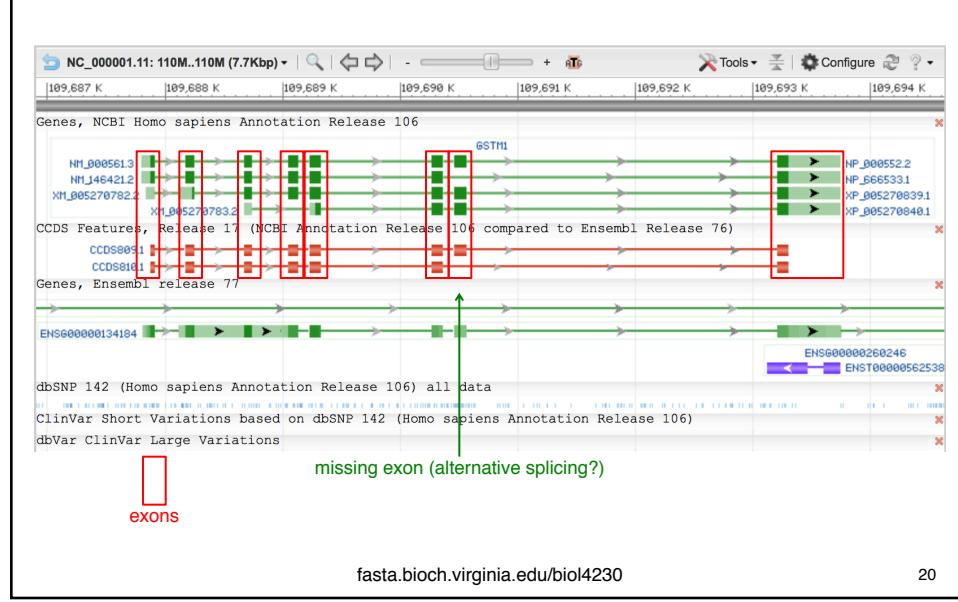
Related information

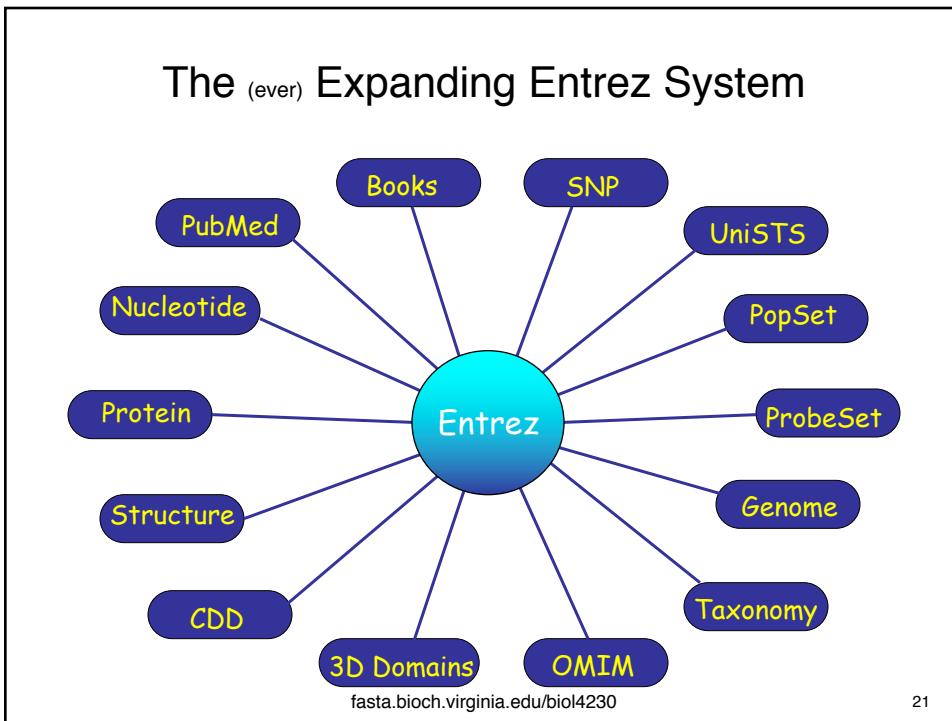
- Order cDNA clone
- 3D structures
- BioAssay
- BioAssay by Target (List)
- BioAssay by Target (Summary)
- BioAssay by Gene target
- BioAssays, RNAi Target, Tested
- BioProjects
- BioSystems
- CCDS
- ClinVar

## Entrez Gene: genetic/genomic information



## Entrez Gene: Genomic/transcript structure





## Uniprot/SwissProt ([uniprot.org](http://uniprot.org)) Comprehensive (inclusive) Database links

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.

**UniProtKB**

- Swiss-Prot (547,357) Manually annotated and reviewed.
- TrEMBL (89,451,166) Automatically annotated and not reviewed.

**UniRef** Sequence clusters

**UniParc** Sequence archive

**Proteomes**

**Supporting data**

- Literature citations
- Taxonomy
- Subcellular locations
- Cross-ref. databases
- Diseases
- Keywords

**News**

Thalidomide, the pharmacological version of yin and yang | Cross-references to DEPOD, MoonProt and Proteomes  
Uniprot release 2015\_01

Higher and higher | New mouse and zebrafish variation files | Structuring of 'cofactor' annotations  
Uniprot release 2014\_11

**Getting started**

- Text search** Our basic text search allows you to search all the resources available
- BLAST** Find regions of similarity between your sequences

**UniProt data**

- Download latest release** Get the UniProt data
- Statistics** View Swiss-Prot and TrEMBL statistics
- [fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)
- Forthcoming changes**

**Protein spotlight**

**The Hidden Things** December 2014

Nature has its secret ways. During the course of the 19th century, the Augustinian friar Gregor Mendel worked out the basics of genetic inheritance as he crossbred pea plants. About a century

22

UniProtKB glutathione "S transferase" Advanced Search Help Contact Show help for UniProtKB Basket

## Results

**High quality (reviewed) annotations**

**Filter by:** Reviewed (349), Unreviewed (64,350) (TREMBL)

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

fasta.bioc.virginia.edu/biol4230

23

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**: Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles.

**Catalytic activity**: RX + glutathione = HX + R-S-glutathione.

**Sites**

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

**GO - Molecular function**: enzyme binding (Source: BHF-UCL), glutathione binding (Source: BHF-UCL), glutathione transferase activity (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), xenobiotic catabolic process (Source: BHF-UCL)

fasta.bioc.virginia.edu/biol4230

24

**Structure<sup>i</sup>**

**Secondary structure**  
 1  218  
 Legend: █ Helix █ Turn █ Beta strand  
[Show more details](#)

<b>3D structure databases</b>		<b>Family and domain databases</b>	
Select the link destinations: <input checked="" type="radio"/> PDB <i><sup>i</sup></i> <input type="radio"/> RCSB PDB <i><sup>i</sup></i> <input type="radio"/> PDBj <i><sup>i</sup></i>		<b>Entry</b> <b>Metho</b> <b>1GTU</b> X-ray <b>1XW6</b> X-ray <b>1XWK</b> X-ray <b>1YJ6</b> X-ray <b>2F3M</b> X-ray	<b>Gene3D<sup>i</sup></b> <a href="#">1.20.1050.10.</a> 1 hit. <a href="#">3.40.30.10.</a> 1 hit.  <b>InterPro<sup>i</sup></b> <a href="#">IPR010987.</a> Glutathione-S-Trfase_C-like. <a href="#">IPR004045.</a> Glutathione_S-Trfase_N. <a href="#">IPR004046.</a> GST_C. <a href="#">IPR003081.</a> GST_mu. <a href="#">IPR012336.</a> Thioredoxin-like_fold. <a href="#">[Graphical view]</a>  <b>Pfam<sup>i</sup></b> <a href="#">PF00043.</a> GST_C. 1 hit. <a href="#">PF02798.</a> GST_N. 1 hit. <a href="#">[Graphical view]</a>  <b>PRINTS<sup>i</sup></b> <a href="#">PR01267.</a> GSTRNSFRASEM.
<b>ProteinModelPortal<sup>i</sup></b> <a href="#">P09488.</a>  <b>SMR<sup>i</sup></b> <a href="#">P09488.</a> Position: <input type="text" value="Search..."/>		<b>SUPFAM<sup>i</sup></b> <a href="#">SSF47616.</a> SSF47616. 1 hit. <a href="#">SSF52833.</a> SSF52833. 1 hit.	  <b>PROSITE<sup>i</sup></b> <a href="#">PS50405.</a> GST_CTER. 1 hit. <a href="#">PS50404.</a> GST_NTER. 1 hit. <a href="#">[Graphical view]</a>
<b>Miscellaneous databases</b>  <b>EvolutionaryTrace<sup>i</sup></b> <a href="#">P09488.</a>		<a href="http://fasta.bioch.virginia.edu/biol4230">fasta.bioch.virginia.edu/biol4230</a>	

25

## Glutathione S-transferase GSTM1

```
>sp|P09488|GSTM1_HUMAN Glutathione S-transferase Mu 1 GN=GSTM1
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKGLDFPNL
PYLIDGAHKITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEF
EKLKPYLEELPEKLKLYSEFLGKRPFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPN
LKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVGNGK
```

Sequence in "FASTA" format

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

26

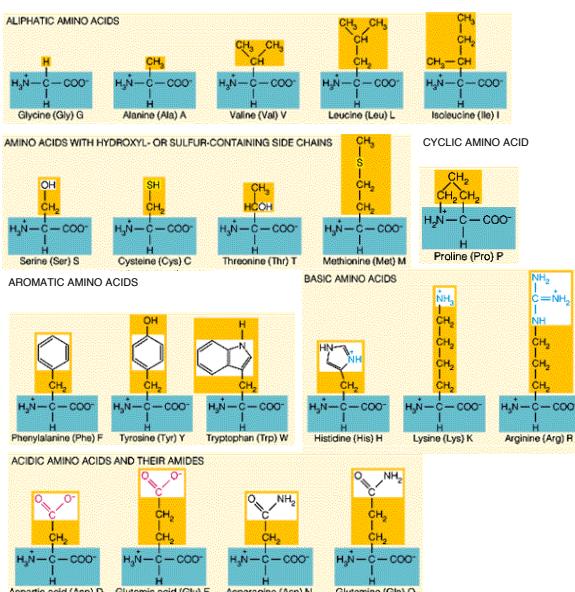
## Structure and properties of Amino-acids

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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

27

Figure 5.3: The amino acids found in proteins.



[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

28

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
+	<b>Leu</b>	<b>L</b>	<b>0.0901</b>
	Lys	K	0.0574
-	Met	M	0.0224
-	Phe	F	0.0385
	Pro	P	0.0520
+	Ser	S	0.0711
	Thr	T	0.0584
-	<b>Trp</b>	<b>W</b>	<b>0.0132</b>
-	Tyr	Y	0.0321
+	Val	V	0.0644

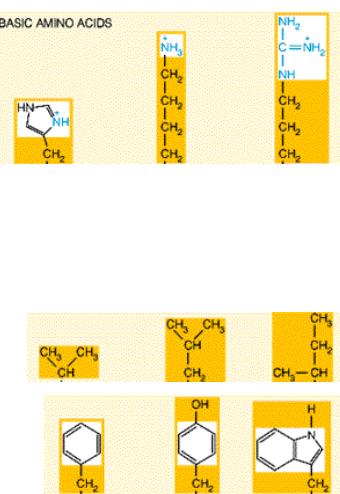
Robinson and Robinson,  
PNAS (1991) 88:8880

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

29

## 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

30

## Amino-acid classes from evolution/mutation

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

```

GSTM1_HUMAN      MPMIILGYWDIIRGLAHPIRILLLEYTDSSYEEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM2_HUMAN      MPMTLGYWNIRGLAHPIRILLLEYTDSSYEEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM4_HUMAN      MPMIILGYWDIIRGLAHPIRILLLEYTDSSYEEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM5_HUMAN      MPMTLGYWNVRGLTHPIRMLLEYTDSSYVEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM1_MOUSE      MPMIILGYWDIIRGLAHPIRILLLEYTDSSYEEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM2_MOUSE      MPMTLGYWDIIRGLAHPIRILLLEYTDSSYEEKKYIMGDAAPDYDRSOWLNEKFKLGLD
GSTM3_MOUSE      MPMTLGYWNVRGLTHPIRMLLEYTDSSYEEKKYIMGDAAPDYDRSOWLSEKFNLGLD
GSTM4_MOUSE      MSMVILGYWDIIRGLAHPIRMLLEFTDSSYEEKKYIMGDAAPDYDRSOWLDFKFLGLD
GSTM3_RABBIT     MPMTLGYWDVVRGLALPIRMLLEYTDSSYEEKKYIMGDAAPNYDQSOWLSEKFNLGLD

```

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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

31

## 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)

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

32

## Basic biology you should know

- Central dogma DNA transcribed into RNA translated into protein
- Prokaryotes – bacteria, archaea
  - no nuclei
  - 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

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

33

## 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](http://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](http://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?

[fasta.bioch.virginia.edu/biol4230](http://fasta.bioch.virginia.edu/biol4230)

34

## 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](https://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