Bioinformatics and Functional Genomics
Course Overview, Introduction of Bioinformatics, Biology Background
Biol4230 Thurs, Jan 18, 2018
Bill Pearson 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

What should you do to reinforce the lecture material?
• 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**

- **Genetics / Bioinformatics**: Template-driven synthesis

  - DNA
    - replication
    - function: information storage
    - size range: 5000 bp - plasmid
  - RNA
    - transcription
    - function: information transfer (catalysis)
    - size range: 2000 nt (ave) mRNA, 2000, 5000 nt rRNA, 100 - 150 nt tRNA, 5S
  - protein
    - translation
    - (information transfer) structure (catalysis)
    - size range: 40 kb (400 aa) ave, 50 - 30,000 aa

Enzymatic: polysaccharide, lipids, enzymatic/non-covalent interactions

fasta.bioch.virginia.edu/biol4230
Polymers and Monomers - DNA

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

MvHA Fig. 4.2
Lehninger Fig. B-2

Monomers and Polymers - proteins

N-term
C-term
carbohydrates

β-D-glucose

fasta.bioch.virginia.edu/biol4230
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
NCBI Databases and Services

- **GenBank** primary sequence database
- Free public access to biomedical literature
  - 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
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

![Finding protein sequences with Entrez/Proteins](image-url)
Entrez Gene: genetic/genomic information

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

Gene ID: 2944, updated on 4-Jan-2019

**Summary**

- **Official Symbol**: GSTM1
- **Official Full Name**: glutathione S-transferase mu 1
- **Primary Source**: HGNC

**Related Gene**

- protein coding

**Organism**

- Homo sapiens

**Lineage**

- Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Primates, Haplorrhini, Catarrhini, Hominidae, Homo

**Also known as**

- CYP1C1, GST1, GSTM1, GSTM4, GSTM1/2, GSTM4/5

**Cytosolic and membrane-bound forms of glutathione S-transferase are encoded by two distinct supergene families. At present, eight distinct classes of the single cytosolic mammalian glutathione S-transferases have been identified: alpha, kappa, mu, sigma, alpha, theta, and zeta. The 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 gene 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. Full names of the genes may 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.**

**Genetic context**

- **Location**: 1p13.3
- **Exon count**: 17

**Related information**

- 3D structures
- BioBrowse
- BioSearch by Gene (Summary)
- BioSearch by Gene (List)
- BioSearch by Gene (Detailed)

**Reference**

fasta.bioch.virginia.edu/bio/4230
Entrez Gene: genetic/genomic information

entrez gene: genomic/transcript structure

missing exon (alternative splicing?)
The (ever) Expanding Entrez System

Uniprot/SwissProt (uniprot.org)
Comprehensive (inclusive) Database links
Glutathione S-transferase GSTM1

>sp|P09488|GSTM1_HUMAN Glutathione S-transferase Mu 1 GN=GSTM1
MPMILGYWDIGLAHAILRLLLEYTDSSYEEKTTYMGDAPYDQWLNEKFKLGDFFNL
PYLIDGAKSNAILCYAKNHNLGCGETEERIKVDILENGTMDHMQVLMQIYNPEF
EKLKPKYELPEKLYSEFLGKRPFAGKTFVDFLVYDVLDDLHRIFPKCLDAFPN
LKDFISRFEGLEKISAYMKSSRFLPRPVFSRMHVGNK

Sequence in "FASTA" format
Structure and properties of Amino-acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbrev</th>
<th>Initial</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>Leucine</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>Lysine</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>Methionine</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>Proline</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Glu</td>
<td>Q</td>
<td>Serine</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td>Threonine</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>Valine</td>
</tr>
<tr>
<td>Asp/Asn</td>
<td>Asx</td>
<td>B</td>
<td>Glu/Gln</td>
</tr>
<tr>
<td>Asp/Asn</td>
<td>Asx</td>
<td>B</td>
<td>Glu/Gln</td>
</tr>
</tbody>
</table>

Figure 5.3: The amino acids found in proteins.
Some amino acids are more common than others:

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Hydropathicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Ala</td>
<td>0.0780</td>
</tr>
<tr>
<td>Arg</td>
<td>0.0512</td>
</tr>
<tr>
<td>Asn</td>
<td>0.0448</td>
</tr>
<tr>
<td>Asp</td>
<td>0.0536</td>
</tr>
<tr>
<td>Cys</td>
<td>0.0192</td>
</tr>
<tr>
<td>Gln</td>
<td>0.0426</td>
</tr>
<tr>
<td>+ Glu</td>
<td>0.0629</td>
</tr>
<tr>
<td>+ Gly</td>
<td>0.0737</td>
</tr>
<tr>
<td>His</td>
<td>0.0219</td>
</tr>
<tr>
<td>Ile</td>
<td>0.0514</td>
</tr>
<tr>
<td>+ Leu</td>
<td>0.0901</td>
</tr>
<tr>
<td>Lys</td>
<td>0.0574</td>
</tr>
<tr>
<td>Met</td>
<td>0.0224</td>
</tr>
<tr>
<td>Phe</td>
<td>0.0385</td>
</tr>
<tr>
<td>Pro</td>
<td>0.0520</td>
</tr>
<tr>
<td>+ Ser</td>
<td>0.0711</td>
</tr>
<tr>
<td>Thr</td>
<td>0.0584</td>
</tr>
<tr>
<td>Trp</td>
<td>0.0132</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.0321</td>
</tr>
<tr>
<td>+ Val</td>
<td>0.0644</td>
</tr>
</tbody>
</table>


Amino acid Hydropathicity/Hydrophobicity

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

Robinson and Robinson, PNAS (1991) 88:8880
Amino-acid classes from evolution/mutation

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

GSTM1_HUMAN     MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM2_HUMAN     MPMTLGYWNIRGLAHSIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM4_HUMAN     MPMILGYWDIRGLAHAIRLLLEYTDSSYVEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM5_HUMAN     MPMTLGYWDIRGLAHAIRLLLEYTDSSYVEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM1_MOUSE     MPMILGYWNIRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRSQWLNEKFKLGLD
GSTM2_MOUSE     MPMTLGYWDIRGLAHAIRLLLEYTDTSYEDKKYTMGDAPDYDRSQWLSEKFKLGLD
GSTM3_MOUSE     MPMTLGYWNIRGLTHSIRLLLEYTDSSYEEKRYVMGDAPNFDRSQWLDVKFKLD
GSTM4_MOUSE     MSMLGYWDGRLATILLEYDSYVEKKYTMGDAPFDRSQWLDVKFKLGLD
GSTM3_RABIT     MPMTLGYWDGRVIRGLALPIRMLLEYTDTSYEEKKYTMGDAPNYDSQWLSEKFTKLGLD

... 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?

- 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)?
  - 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?
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