This course will introduce the concepts and programs needed to apply bioinformatics in molecular biology and biotechnology research.

COURSE DESCRIPTION:

• **Requirements**: ECS 10 or 15 or AMR 21; Biological Sciences 101 and 104; AMR 120 or Statistics 13 or 100.
• **Lecture**: T & R 10:00 am to 10:50 am **Wellman 103**.
• **Laboratory**: R 4:00 pm to 6:00 pm. 1137 PES (door code **93175**)  
• **TA office hours**: Monday and Wednesday 10:30 to noon 1137 PES
• **My office hours**: Friday 2:00 pm - 4:00 pm. 281 Hunt Hall. Phone: 752-5159


**GRADING %**: Laboratory problem sets and homework (20% Dubcovsky’s section, 20% Neale’s section), a mid-term examination Dubcovsky’s section (15% in class, 15% take home excercises) and a final examination of Neal’s section (30%).
Computational Biology and Bioinformatics

- **Computational biology**
  - Development of algorithms to solve problems in biology

- **Bioinformatics**
  - Application of computational biology to the analysis and management of biological data

- **Applied bioinformatics**
  - Intelligent use of tools to navigate the sequence space

Better experiments can be designed by a careful Bioinformatics analysis before the bench work
DNA sequencing

- DNA sequences are the most abundant type of sequences (213 billion bp)
- Generated by the chain termination method (Sanger sequencing)
- Based on the action of a DNA polymerase that adds nucleotides to complementary strand
- Fluorescently labeled ddNTP (Dideoxynucleotides) stop synthesis acting as chain terminators. They are included in amounts so as to terminate every time the base appears in the template
- Requires template DNA, primers and one ddNTP for each base: A, C, G, and T
- Products are separated by electrophoresis
In the past, four different reactions, one for each ddNTP, were separated on a gel that could resolve one-base differences. The sequence was then read from the bottom of the gel to the top.
Capillary electrophoresis

- Automated sequencers use very thin capillary tubes
- Run all four fluorescently tagged reactions in same capillary
- Can have 96 capillaries running at the same time
Sequence reading of fluorescently labeled reactions

- Fluorescently labeled reactions scanned by laser as particular point is passed
- Color picked up by detector
- Output sent directly to computer

In its simplest form a sequence can be represented as a string of nucleotides with a basic tag or identifier after a greater than character “>”: FASTA format

```
>U54469.1
CGTTGCTTGGTTTATACACAGTCTGACAGGCTTTCCAGAGTTGGCTGTCAACATCGATGCATGCCTTTGGCCACCAAAATCCCAAACTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCAGTAACCTGCTCAAACTTGGCTAACCAGATCTACATTCGATTCAGTAACTGATGCTGGATGCTGGGTACCATGGGTTCGATTTGCGCTGAGCCGTGGCAGGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGCTCTCTTGTCAAGACATCGCGCGCGTGTGTGTGGGTGTGTCTCTAGCACATATACATAAATAGGAGAGCGG
```

More information can be added to the FASTA definition line

```
>gb|U54469.1|DMU54469 Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene.
```

GenBank
Accession.version
Locus
Description
Primer walking to sequence stretches of few kb

Animation sequencing: [http://www.jgi.doe.gov/education/how/how30minflash.html](http://www.jgi.doe.gov/education/how/how30minflash.html)
Editing Errors

Query: 544 atgcctaggtgtacgaagtaaagccsaaggaagccgacatgacatgagaagaatca 703

Sbjct: 11 atgcctaggtgtacgaagtaaagccsaaggaagccgacatgacatgagaagaatca 69

Query: 704 tgggcatagactctgtcatctctctgtcatatgtctctctgatcctacttttgtaaga 763

Sbjct: 70 tgggcatagactctgtcatctctctgtcatatgtctctctgatcctacttttgtaaga 129

Query: 764 caactctgtgcattcttatgtgacatgcatactacatccatatcttgcgacacac-at 822

Sbjct: 130 caactctgtgcattcttatgtgacatgcatactacatccatatcttgcgacacac-at 189

Query: 823 cgccacccaaatccaaatctataattctagattgttggag-ccacagattctt 881

Sbjct: 190 cgccacccaaatccaaatctataattctagattgttggag-ccacagattctt 249

Query: 882 a-ttcattgtgtatgccgtacacctgtgaaacgccactttgtc 925

Sbjct: 250 a-ttcattgtgtatgccgtacacctgtgaaacgccactttgtc 294
New sequencing technologies

• **454 Life Sciences**: FLX titanium: 1,000,000 reads 400 bp = 400,000,000 high quality bp in 10h run

  - **Sample Input and Fragmentation**: DNA and BACs are fractionated into small, 400- to 800-basepair fragments. For smaller samples, such as small non-coding RNA fragmentation is not required.

  - **Library Preparation**
    Short adaptors (A and B) - specific for both the 3’ and 5’ ends - are added to each fragment. The adaptors are used for purification, amplification, and sequencing steps.

  - **One Fragment = One Bead**
    The single-stranded DNA library is immobilized onto specifically designed DNA Capture Beads. Each bead carries a unique single-stranded DNA library fragment. The bead-bound library is **emulsified** with amplification reagents in a water-in-oil mixture resulting in microreactors containing just one bead with one unique seq.

  - **emPCR (Emulsion PCR) Amp.**
    (also used in SOLiD). Each unique fragment is amplified within its own microreactor. PCR amplification is done in parallel. 1,000,000 seq. at a time! The emPCR is broken while the amplified fragments remain bound.

  - **One Bead = One Read**
    The beds are loaded onto individual wells in a PicoTiterPlate device. Sequencing enzymes are added. The instrument flows individual nucleotides in a fixed order and the addition of a nucleotide complementary to the template strand results in a chemiluminescent signal (pyrosequencing)

  - **Data Analysis**
    The signal intensity and position is used to determine the sequence of 1,000,000 individual reads per 10-hour instrument run simultaneously. Different bioinformatics tools are available for *de novo* assembly up to 400 Mb; resequencing; & SNPs
ILLUMINA / SOLEXA

• 2 billion bases (2 Gb) per lane (Full run = 7 lanes = 14G).

• $1000 per lane. 1 cent = 20,000 bp

• Short runs 80-100 bp (shorter seq. can be used as transcript tags.

• Data has to be moved off the instrument for further analysis; currently >1 Terabyte per run.

• Allows: paired-reads and bar-coding (48 barcoded samples per run)
Present and Future of sequencing

- Sequencing costs
  - Dropping each year

GenBank 2010 NAR article (see link)

255 billion bp from 300,000 organisms

We need databases!

EST (Expressed Sequence TAGs)
34 billion bp

GSS (Genome Survey Sequences) BAC ends
16.7 billion bp

WGS (Whole Genome Shotgun) 4letter 6 digit
148 billion bp

HTG (High throughput Genomics)
23.9 billion bp (unfinished in transition)

- 454: $ 0.00004 per base
  - 200 MB =$8000.

- Illumina: $0.0000005 per base
  - 2 Gb =$1000. 1 cent= 20,000 bp
  - 1000 human genome project...
Sequence databases

• What is a database?
  – An indexed set of records
  – Records retrieved using a query language

• Examples of sequence databases
  – Primary databases (archival)
    • GenBank
    • EMBL (European Molecular Biology Laboratory)
    • DDBJ (DNA Data Bank of Japan)
  – Secondary databases (curated)
    • RefSeq
    • EMBL Genome Reviews
    • Protein databases
    • TPA (Third party annotations)
Data flow of submissions between primary databases (Chp. 1)

- **Entrez**: Integrated information retrieval system
  - Submissions
  - Updates (Sequin)
  - [Website](http://www.ncbi.nlm.nih.gov/sites/entrez)

- **NCBI**: National Center for Biotechnology Information
  - Submissions
  - Updates (Sequin)

- **GenBank**: International Nucleotide Sequence Database Collaboration Updated every 24 hs

- **EMBL**: European Molecular Biology Laboratory
  - Submissions
  - Updates
  - [Website](http://www.ensembl.org/index.html)

- **DDBJ**: DNA Data Bank of Japan
  - Submissions
  - Updates (Sakura)

- **EBI**: European Bioinformatics Institute
  - Submissions
  - Updates

- **CIB**: Center for Information Biology
  - Submissions
  - Updates (Sakura)
  - [Website](http://getentry.ddbj.nig.ac.jp/getstart-e.html)

- **NIG**: National Institute of Genetics (JAPAN)
  - Submissions
  - Updates (Sakura)
  - [Website](http://getentry.ddbj.nig.ac.jp/getstart-e.html)
Nucleotide sequence flatfiles

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>DROS0460</th>
<th>1081 bp</th>
<th>DNA</th>
<th>linear</th>
<th>INV 22-FEB-1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene, alternative splice products, complete cds.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>US4469</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>US4469.1</td>
<td>GI:1322209</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Drosophila melanogaster (fruit fly)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Drosophila melanogaster</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCES</td>
<td>1 (bases 1 to 3881)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Lavoie, C.A., Lachance, P.K., Sennemberg, N., and Lasio, P.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Alternatively spliced transcripts from the Drosophila eIF4E gene produce two different cap-binding proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>J. Biol. Chem. 271 (27), 16999-16998 [1996]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEATURES</td>
<td>Location/Qualifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>source 1.2861</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/organism=&quot;Drosophila melanogaster&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/mol_type=&quot;genomic DNA&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/da_xref=&quot;taxon:7627&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/da_xref=&quot;chromosome:3&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/map=&quot;678A-B2&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50..12861</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;eIF4E&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td>join(50..224,1556..1565,1960..1966,2086,2317..2404,2466..26081)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;eIF4E&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/product=&quot;eukaryotic initiation factor 4E-1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td>join(50..224,1556..1565,1960..1966,2086,2317..2404,2466..26081)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;eIF4E&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/product=&quot;eukaryotic initiation factor 4E-1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA</td>
<td>join(50..224,1556..1960,1966..2086,2317..2404,2466..26081)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;eIF4E&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/product=&quot;eukaryotic initiation factor 4E-1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDS</td>
<td>join(201..324,1455..1520,1960..2086,2317..2404,2466..26081)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;eIF4E&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/note=&quot;Method: conceptual translation with partial peptide sequencing&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/codon_start:1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/product=&quot;eukaryotic initiation factor 4E-1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/protein_id=&quot;AC0324.1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/db_xref=&quot;GI:1322209&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/translation=&quot;KTVLLETVKSAAPTRGAPSPFPSAASKAAMNNSKPEPQDGCGAPNQNTAPADASTYSLTQPSLQKILPSQGKVSSPSVNLKSSSINLDVRLSLHQCQPSFDPEQDQCGAVGVRGKSNKRFISIUTADONHEAASLIEHFLPRTIQLGNSLQYDLRCRQVFXQGSSVSY&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEQUENCE</td>
<td>actual sequence ending in //</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Secondary databases


- Comprehensive, integrated, and non-redundant set of sequences
- Genomic DNA – RNA – Protein explicitly linked
- 2_6 Format (the underscore is never present in GenBank accessions):
  - **Experimental**
  - NT_123456 (genomic contig)
  - NM_123456 (mRNAs) [XM_123456 (model mRNAs)]
  - NP_123456 (Proteins) [XP_123456 (model protein)]
- Undergo continuous curation: most up to date sequence
- Each RefSeq is a synthesis of information, not a piece of a primary research: equivalent to a “review article”
- Message for
  - Removed 
  - Secondary

---

**Example Entry**

```
(a) 1: NM_032931
ref NM_032931.1[14249727]
This record was removed at the submitters request.

(b)
LOCUS  NM_002729  1021 bp  mRNA  linear  PRI 19-FEB-2002
DEFINITION Homo sapiens hematopoietically expressed homeobox (HHEX), mRNA.
ACCESSION NM_002729 NM_081529
VERSION  NM_002729.2  GI:17978673
KEYWORDS
SOURCE Homo sapiens
ORGANISM Homo sapiens
```
Secondary databases

Third Party Annotation (TPA)

- Includes
  - Reannotations,
  - Combinations of novel and existing primary entries
  - Annotations of trace archives
  - Whole genome Shotgun data

- Provides
  - GenBank accession. Version numbers and nucleotide locations for all primary entries to which the TPA sequence relates

EMBL Genome reviews

- Includes
  - Add information from UniProt knowledgebase, Gene Ontology Annotation, InterPro, and others
  - Curated versions of entries representing complete genomes
  - Standardize annotations
Protein databases

- **GenPept**: NCBI translations of all CDS. Not curated
- **Uniprot (Swiss-Prot/TrEMBL/PIR-PSD)**
  - **UniParc**: most comprehensive, public nonredundant protein database
    - Swiss-Prot (manual)/TrEMBL(computer)/PIR-PSD
    - GenBank, Patents, Int. Pr. Index (IPI)
    - Protein Data Bank
  - **UniProt Knowledgebase**: curated subset of UniParc
    - Function
    - Postranslational modifications
    - Domains
    - Catalytic sites
    - Structures
    - Associated diseases
    - Pathways
    - Etc.
  - **UniRef**: UniProt nonredundant reference database: 95%, 90% and 50% sets.
- **Functional groups**
  - Pfam
  - Prosite
  - InterPro

![Diagram of protein databases and their connections](http://www.ebi.ac.uk/Tools/sss/)

---

**Protein ROA1_HUMAN**

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Heterogeneous nuclear ribonucleoprotein A1</td>
</tr>
<tr>
<td>Synonyms</td>
<td>Helix-stabilizing protein | Single-strand binding protein | hHRNAP core protein A1</td>
</tr>
<tr>
<td>Gene name</td>
<td>HNRNPA1</td>
</tr>
<tr>
<td>Species</td>
<td>Homo sapiens (Homo sapiens)</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Eukaryota; Metazoa; Chordata; Gymnophiona; Vertebrata; Subclass: Aves; Class: Aves; Order: Galliformes; Family: Phasianidae; Genus: Phasianus; Species: Phasianus colchicus; Common name: Red junglefowl</td>
</tr>
</tbody>
</table>

**References**