CENG 734
Advanced Topics in Bioinformatics

Fall 2008-2009

Instructor Info

- Tolga Can
  - e-mail: tcan@ceng.metu.edu.tr
  - Office: B-109
  - Office hours: E-mail me to schedule an appointment
  - Preferred way of resolving any course related problem/question: by e-mail

Class Web Page & E-mail list

http://www.ceng.metu.edu.tr/~tcan/ceng734/

- Lecture slides
- Syllabus
- Reading material (next week's reading material is going to be posted tonight, quiz after the bayram)

I will also maintain an e-mail list for announcements

- Sign-up for the e-mail list

Prerequisites

- No formal prerequisites. However, some familiarity with Bioinformatics will help the students get the most benefit out of the course.
- Programming: may be required for the project
- Algorithms and Complexity Analysis
- Basic probability and statistics
- Some molecular and cellular biology terminology is required
- If you are new to Bioinformatics, I encourage you to take CENG 465 offered in Spring

Course Objectives

- The primary objectives of this course are to expose students to recent developments in the field of bioinformatics and to enable students initiate research in this area. Upon completion of this course the students will:
  - be aware of the current challenges in Bioinformatics,
  - have learnt the state-of-the-art methods to tackle important biological problems,
  - and be able to initiate and conduct research in the area of Bioinformatics.

Reading Material

- Reading material will be provided online on the web or by e-mail
- Papers from recent conferences or journals

Grading

- Reading : 40%
  - 8-10 quizzes about reading material
- Term project: 40%
- Final exam: 20%

Project

- May be related to your current research or what you may want to do for research
- Groups of 1-4 students
- You are free to choose project topics but will discuss details/goals/work plan with the instructor before starting to work on the project
- Project topic examples:
  - Small improvements on techniques/algorithms discussed in class
  - Application of a technique on a different data set.

Outline of the course

- This week: Introduction and characteristics of biological data. Who is working on what?
- Challenges in genome analysis
- Evolution and phylogeny
- Protein structure, functional classification, genome annotation
- Gene regulation and transcriptomics
- Text mining in bioinformatics
- Protein interactions and molecular networks

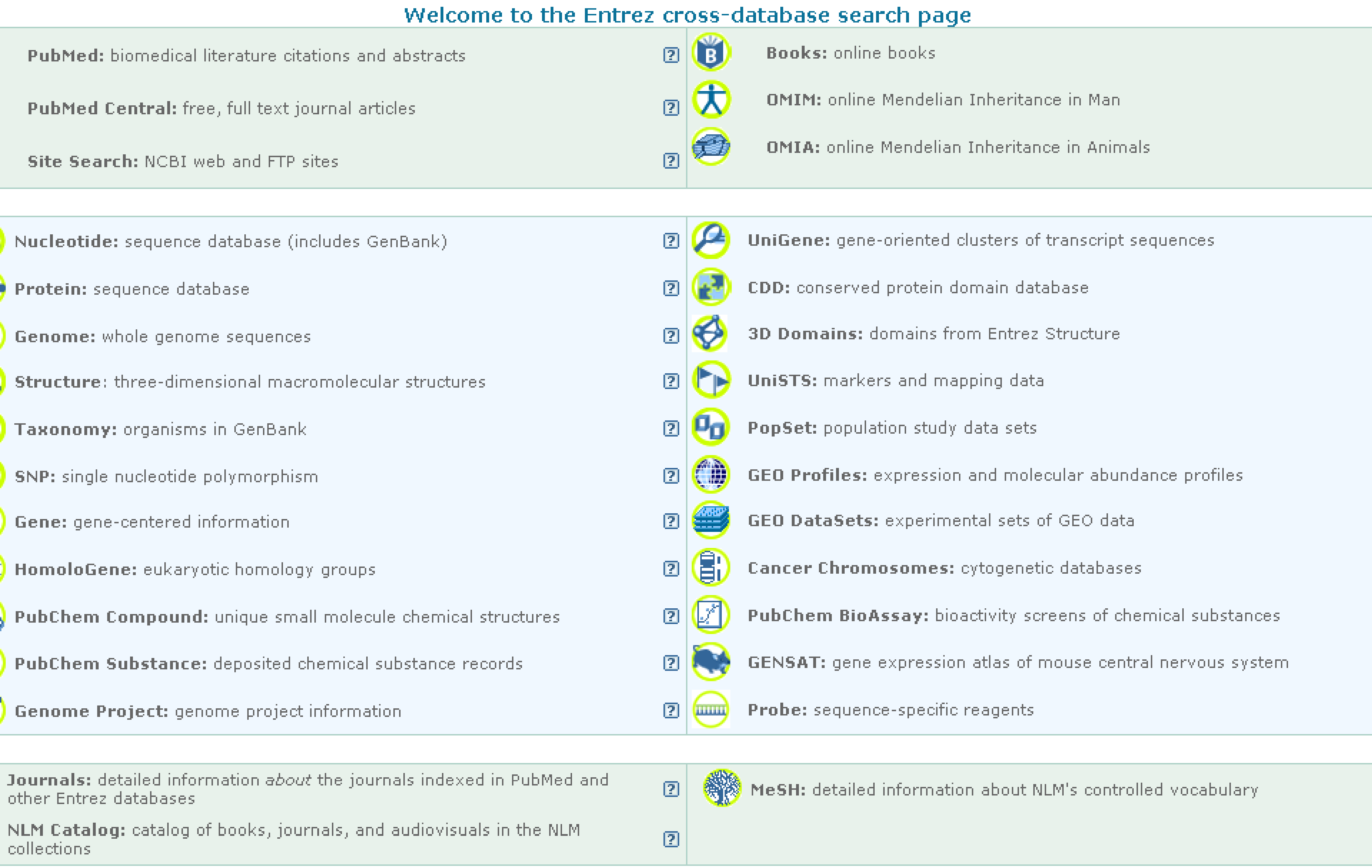
....

- Challenges in heterogeneous data integration
- Bioimage informatics
- Multiscale Modeling and Simulation: from Molecules to Cells to Organisms (PSB 2008 Session)

Biological Data

- Comes in many different forms
  - Sequence Databases:
    - Nucleotide (GenBank), SWISS-PROT, Whole genome databases
  - Structure databases
    - Protein Data Bank
  - Expression data
    - NCBI GEO
  - Interaction data, Pathways
  - Bioimages
  - Published biological papers (PubMed)
  - Domain, annotation information

NCBI Entrez



Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Site Search: NCBI web and FTP sites

Books: online books

OMIM: online Mendelian Inheritance in Man

OMIA: online Mendelian Inheritance in Animals

Nucleotide: sequence database (includes GenBank)

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

Taxonomy: organisms in GenBank

SNP: single nucleotide polymorphism

Gene: gene-centered information

HomoloGene: eukaryotic homology groups

PubChem Compound: unique small molecule chemical structures

PubChem Substance: deposited chemical substance records

Genome Project: genome project information

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

3D Domains: domains from Entrez Structure

UniSTS: markers and mapping data

PopSet: population study data sets

GEO Profiles: expression and molecular abundance profiles

GEO DataSets: experimental sets of GEO data

Cancer Chromosomes: cytogenetic databases

PubChem BioAssay: bioactivity screens of chemical substances

GENSAT: gene expression atlas of mouse central nervous system

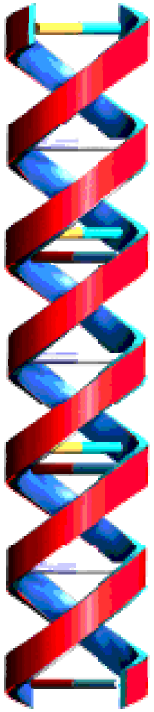
Probe: sequence-specific reagents

Journals: detailed information *about* the journals indexed in PubMed and other Entrez databases

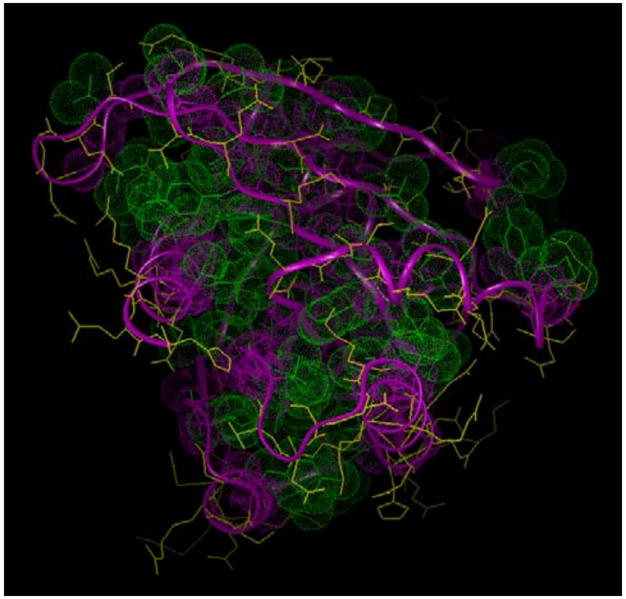
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

MeSH: detailed information about NLM's controlled vocabulary

Introductory Biology



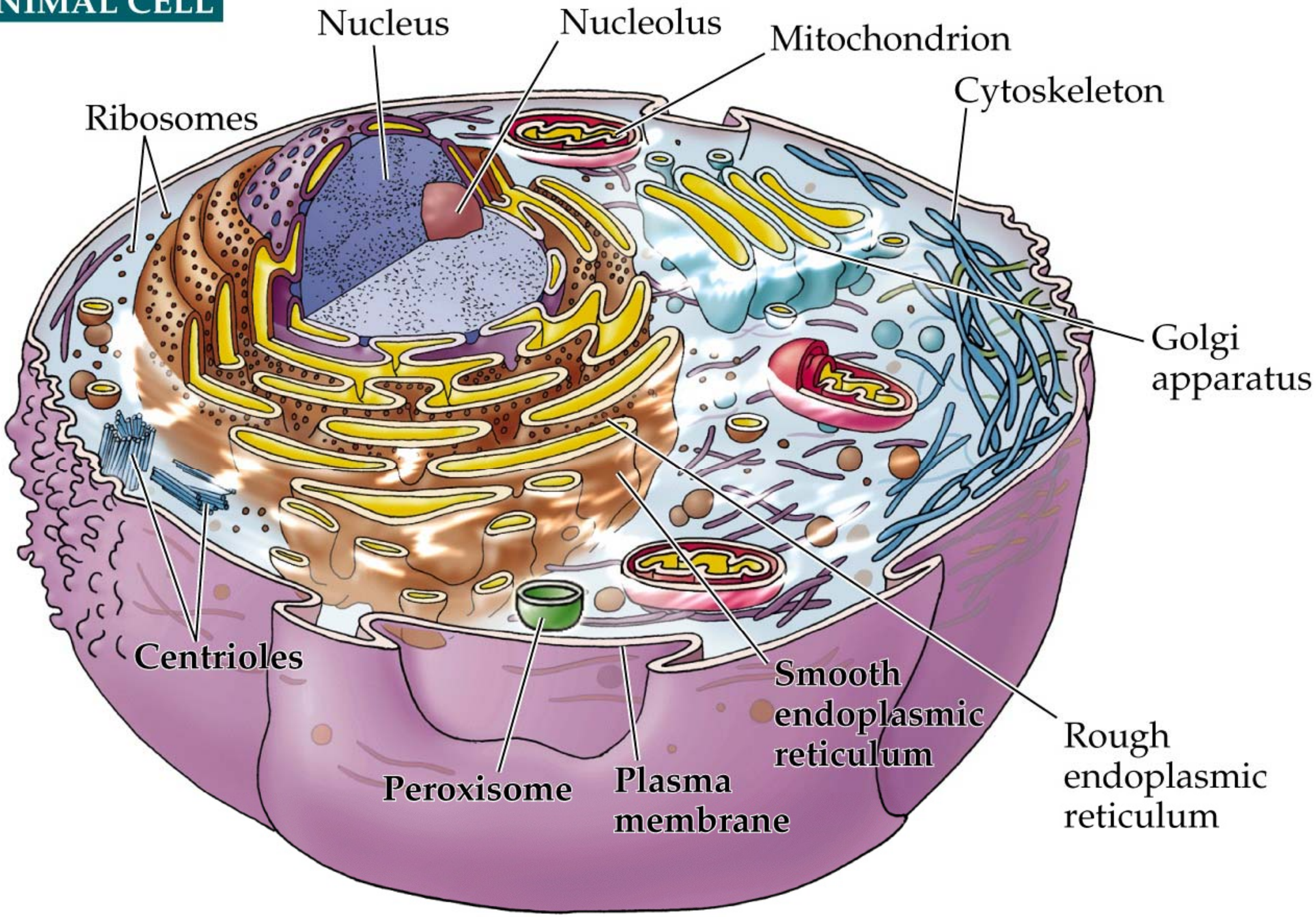
DNA
(Genotype)



Protein

Phenotype

AN ANIMAL CELL



DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 Kb in a gene, ~2 Mb in genome
  - ~3 Gb Human

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .


. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccattttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

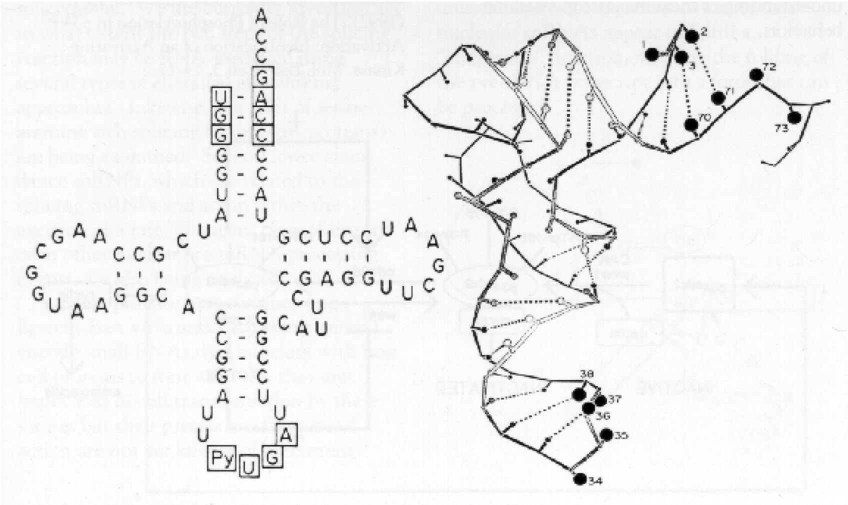
Protein Sequence

- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria),

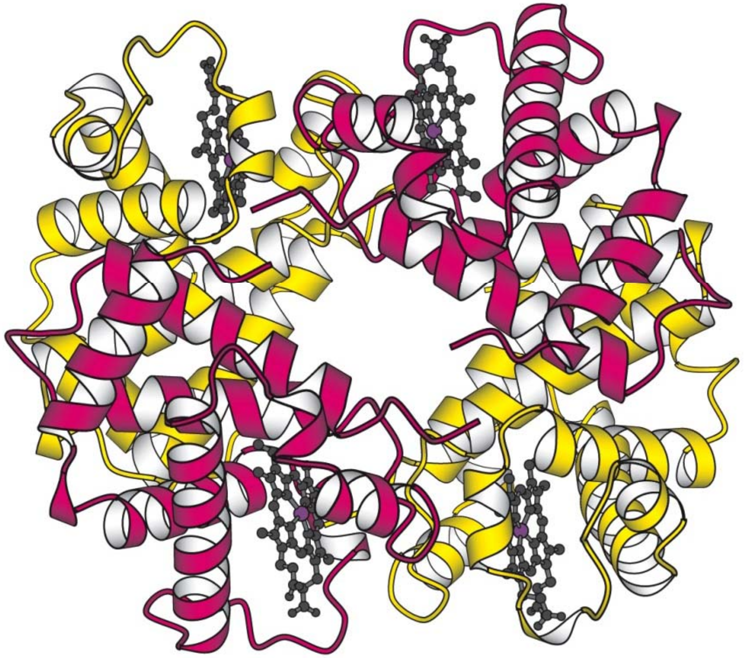
  ~200 aa in a domain
- >6M known protein sequences
- Uniprot
  - **UniProtKB/Swiss-Prot**: proteins with high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.) 397539 entries as of September 02.
  - **UniProtKB/TrEMBL**: a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. 6212793 entries as of September 02.

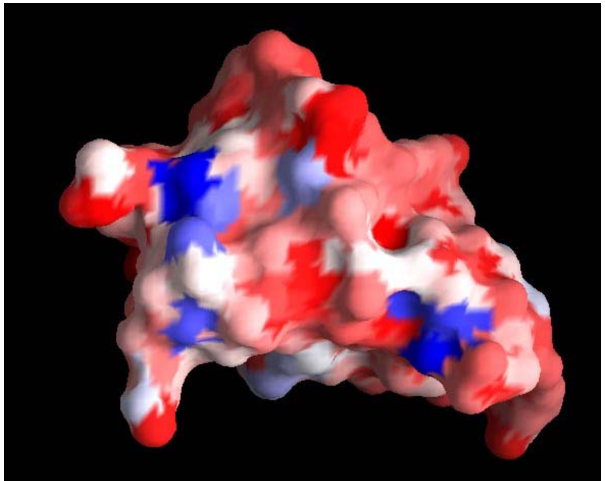
Structures

- DNA/RNA/Protein
  - Mostly protein structures at PDB



'Identity elements' in *Escherichia coli* glutamine tRNA.





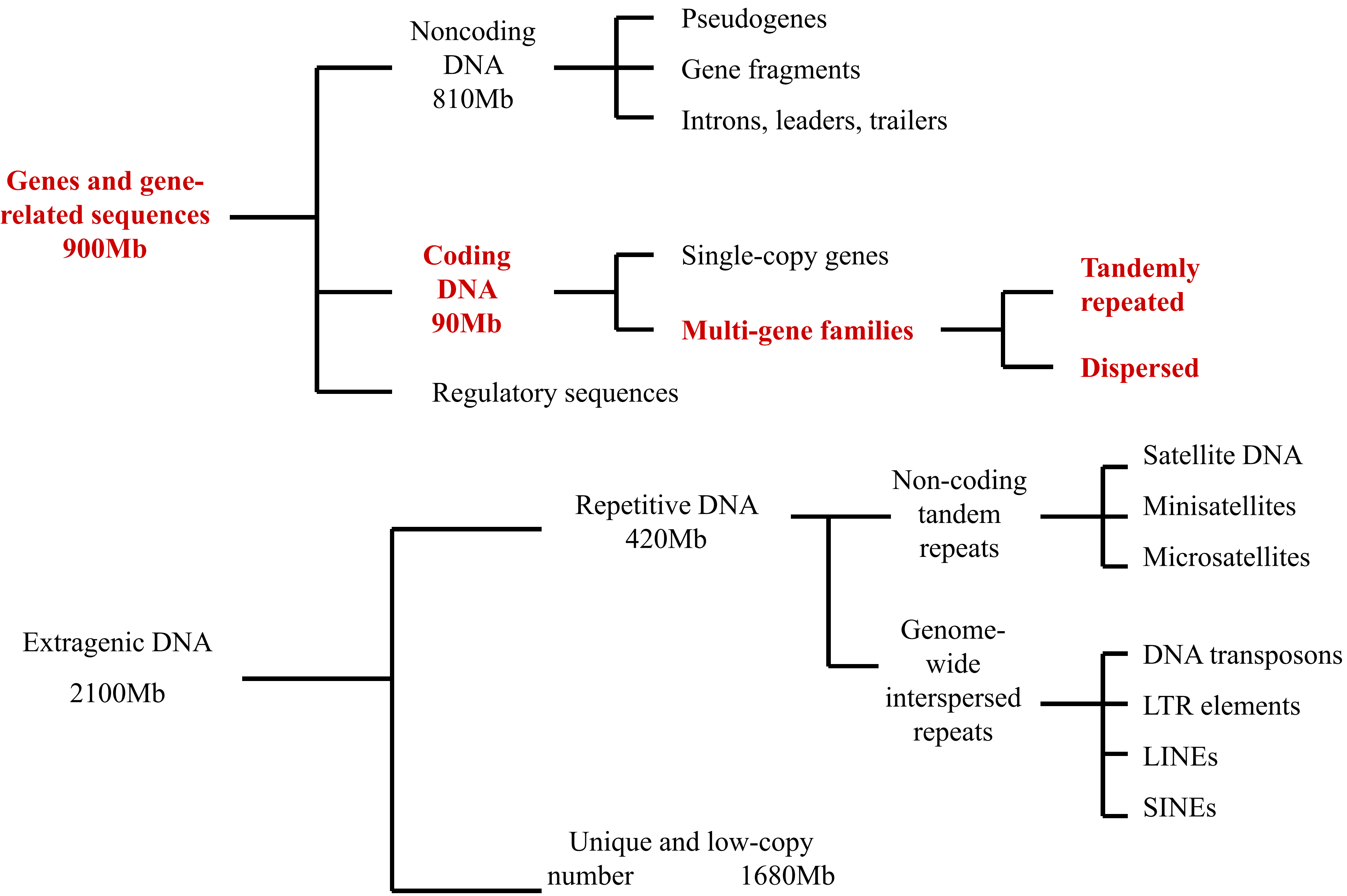
Genes and Proteins

- One gene encodes one* protein.
- Like a program, it starts with start codon (e.g. ATG), then each three code one amino acid. Then a stop codon (e.g. TGA) signifies end of the gene.
- Sometimes, in the middle of a (eukaryotic) gene, there are introns that are spliced out (as junk) during transcription. Used parts are called exons. This is the task of gene finding.

Genomes

- NCBI Entrez Genome Database
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
- 4707 organisms sequenced
  - Achaea: 67
  - Bacteria: 926
  - Eukaryote: 1630
  - Viral genomes: 2007
  - Viroids: 39
  - Plasmids: 38

Human genome



- **Genes and gene-related sequences 900Mb**
  - Noncoding DNA 810Mb
    - Pseudogenes
    - Gene fragments
    - Introns, leaders, trailers
  - **Coding DNA 90Mb**
    - Single-copy genes
    - **Multi-gene families**
      - **Tandemly repeated**
      - **Dispersed**
  - Regulatory sequences
- Extragenic DNA 2100Mb
  - Repetitive DNA 420Mb
    - Non-coding tandem repeats
      - Satellite DNA
      - Minisatellites
      - Microsatellites
    - Genome-wide interspersed repeats
      - DNA transposons
      - LTR elements
      - LINEs
      - SINEs
  - Unique and low-copy number 1680Mb

Gene expression data

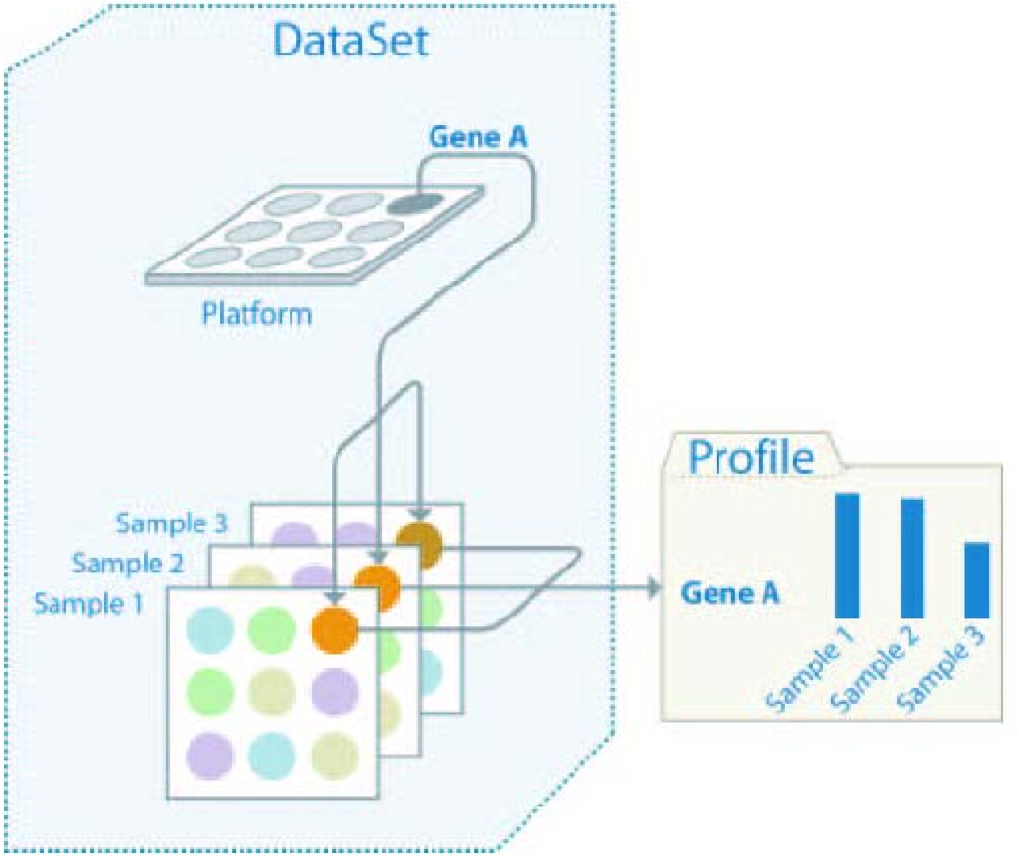
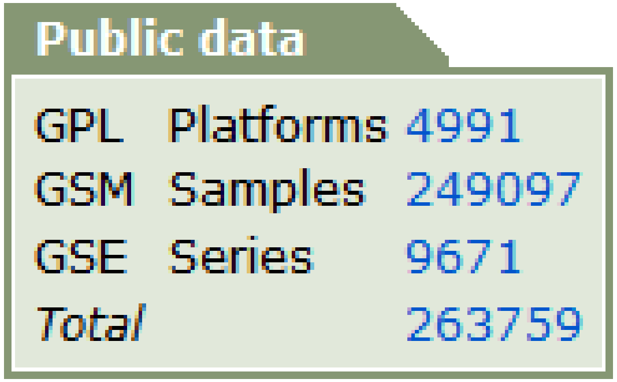


Figure 1. Schematic diagram of the relationships between GEO Platform, Sample, DataSet and Profiles. For each gene on a Platform (e.g. Gene A), multiple Sample measurement values are generated (Sample1–Sample3). Related Samples make up a DataSet, from which multiple, individual gene profile entities are generated.

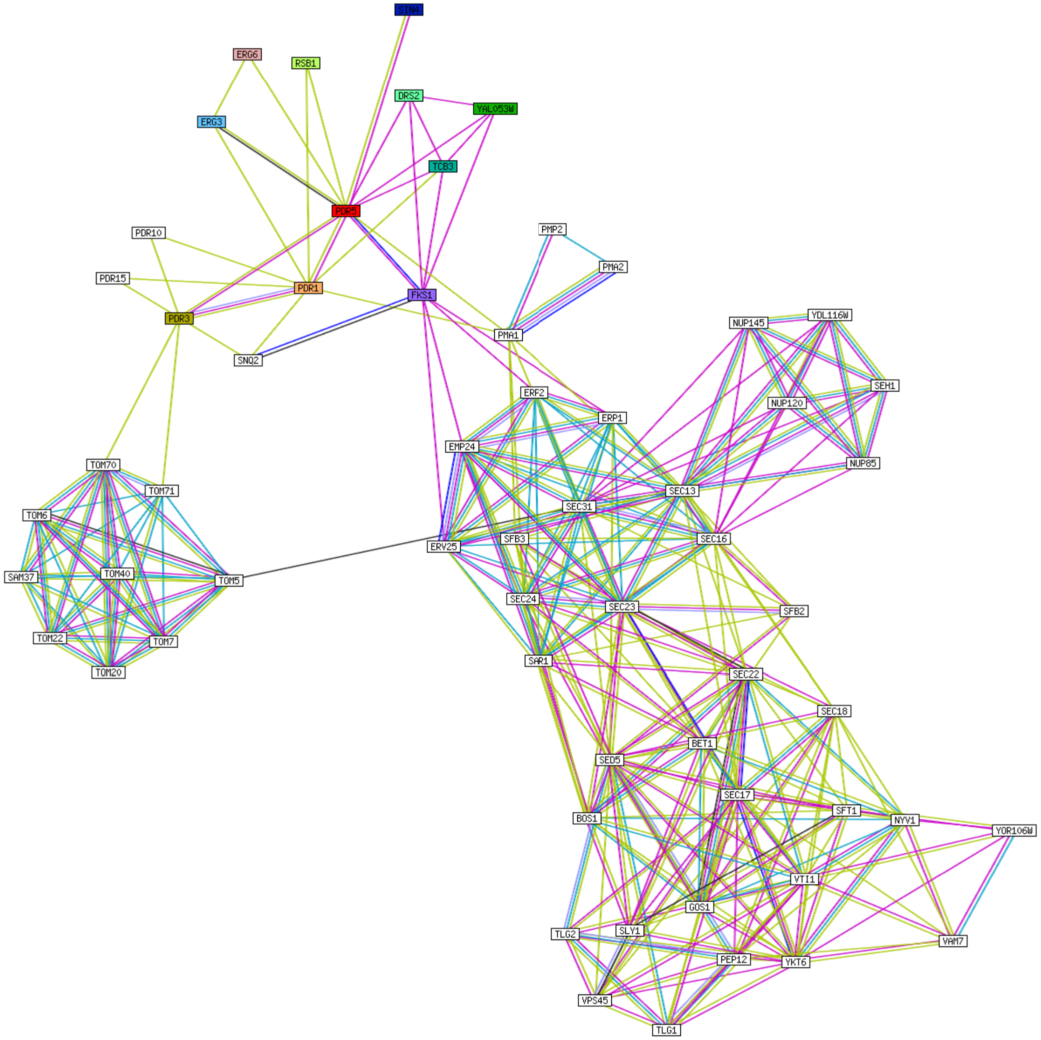


Public data

GPL	Platforms	4991
GSM	Samples	249097
GSE	Series	9671
Total		263759

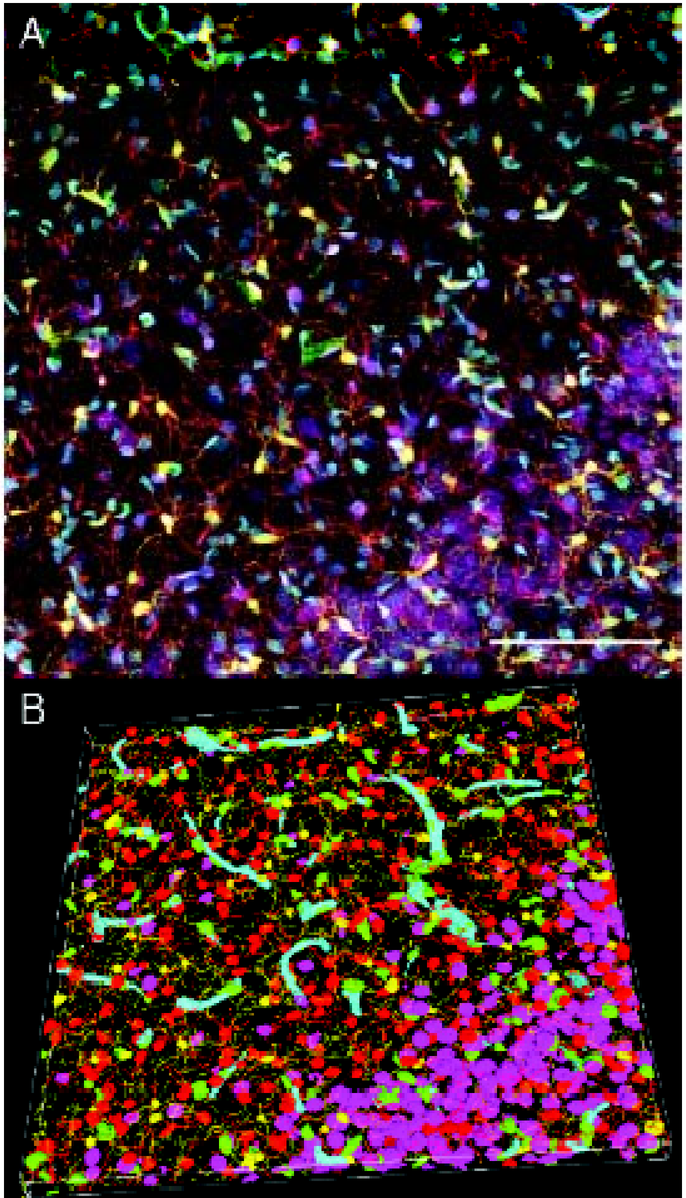
from NCBI GEO NAR 2005 paper

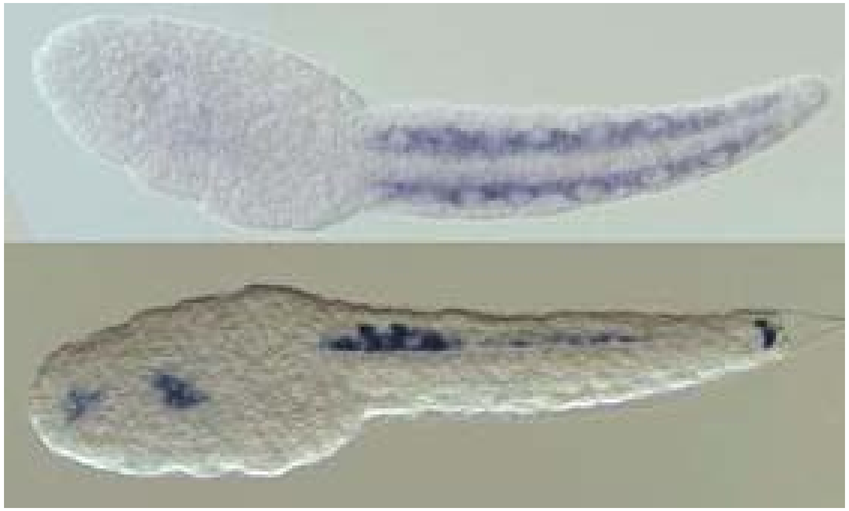
Protein Network Data

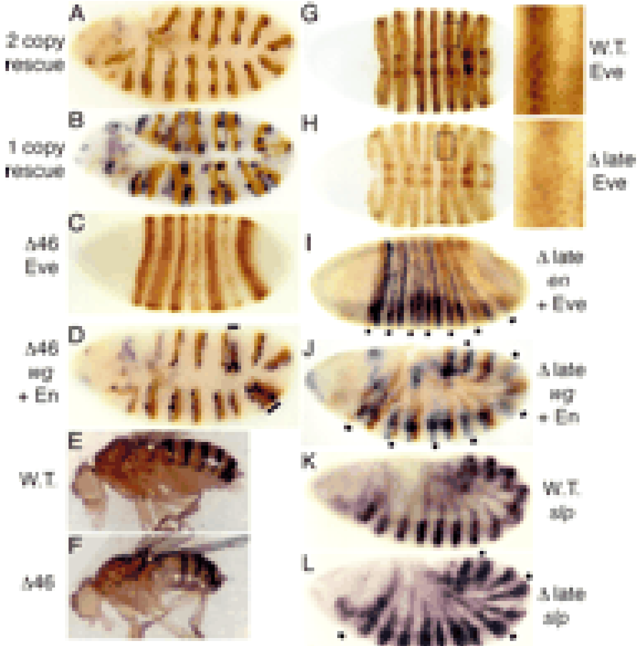


from STRING database

Bioimages

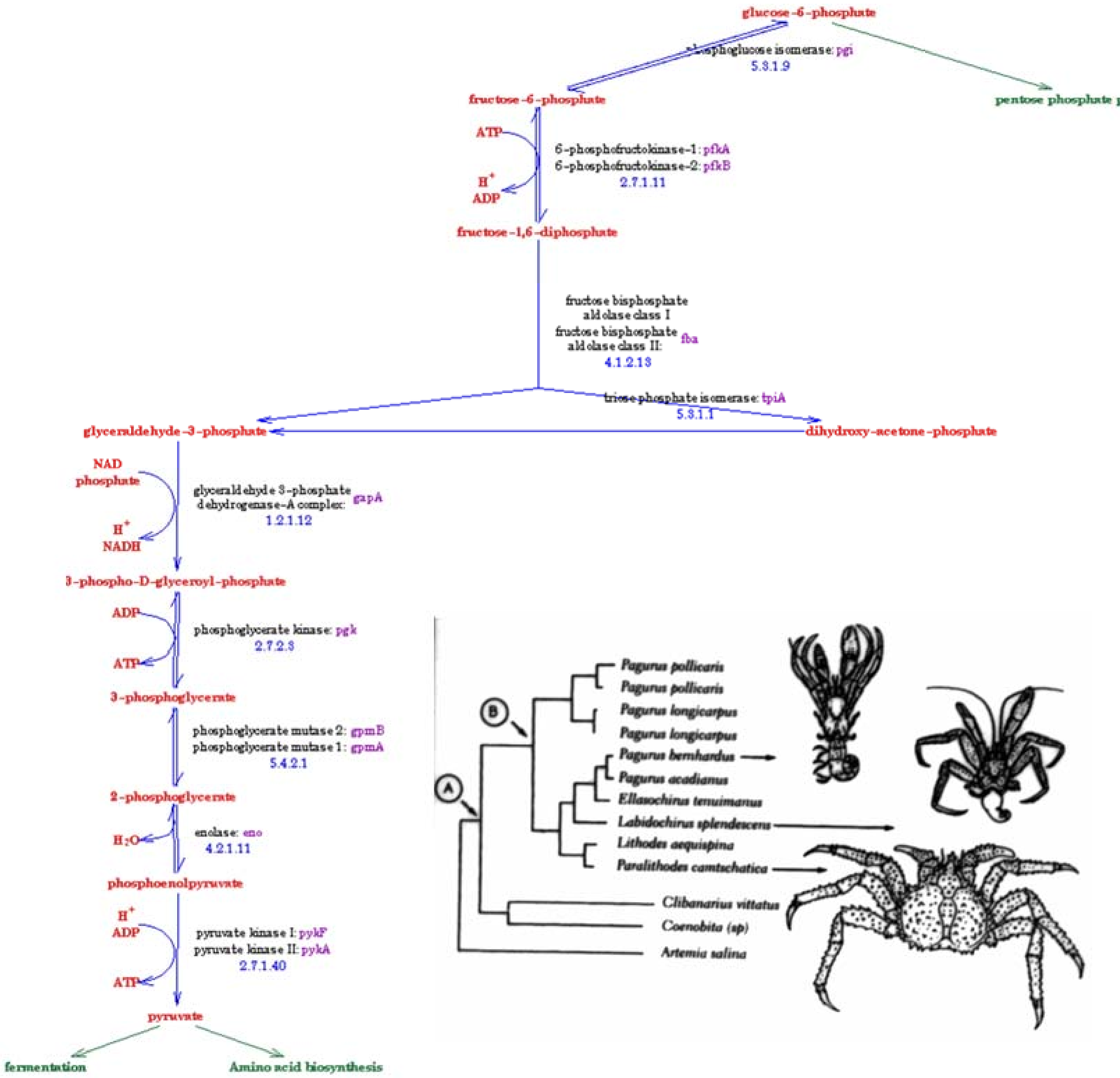

A
B




A
2 copy
rescue
B
1 copy
rescue
C
Δ46
Eve
D
Δ46
wg
+ En
E
W.T.
F
Δ46
G
W.T.
Eve
H
Δ late
Eve
I
Δ late
en
+ Eve
J
Δ late
wg
+ En
K
W.T.
slp
L
Δ late
slp

Other Types of Data

- Information to understand genomes
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature (PubMed)



Data sources

- NAR (Nucleic Acids Research) journal maintains a list of data collections (1078 databases with the 2008 Database Issue update)
- http://www.oxfordjournals.org/nar/database/c/
  - Sequence
    - Genomes, ESTs, Promoters, transcription factor binding sites, repeats, ..
  - Structure
    - Domains, motifs, classifications, ..
  - Others
    - Microarrays, subcellular localization, ontologies, pathways, SNPs, ..

Challenges of working in bioinformatics

- Need to feel comfortable in an interdisciplinary area
- Depend on others for primary data
- Need to address important biological *and* computer science problems

Skill set

- Artificial intelligence
- Machine learning
- Statistics & probability
- Algorithms
- Databases
- Programming
- Molecular and Cellular Biology
- More?

Challenging sequence related problems

- More sensitive pairwise alignment
  - Dynamic programming is O(mn)
    - m is the length of the query
    - n is the length of the database
- Scalable multiple alignment
  - Dynamic programming is exponential in number of sequences
  - Currently feasible for around 10 protein sequences of length around 1000
- Shotgun alignment
  - Current techniques will take over 200 days on a single machine to align the mouse genome

Challenging structure related problems

- Alignment against a database
  - Single comparison usually takes seconds.
  - Comparison against a database takes hours.
  - All-against-all comparison takes weeks.
- Multiple structure alignment and motifs
- Combined sequence and structure comparison
- Secondary and tertiary structure prediction

- And many more other challenging problems in other areas of bioinformatics....

Top journals

- Science
- Nature (Nature Genetics, Nature Biotechnology)
- PNAS (Proceedings of the National Academy of Sciences)
- NAR (Nucleic Acids Research)
- Bioinformatics
- JCB (Journal of Computational Biology)
- BMC Bioinformatics
- Genome Research
- Proteins: Structure and Function, and Bioinformatics
- PLoS Computational Biology

Top conferences

- RECOMB: Research in COmputational Molecular Biology
- ISMB: Intelligent Systems for Molecular Biology
- ECCB: European Conference on Computational Biology
- PSB: Pacific Symposium on Biocomputing
- CSB: Computational Systems Bioinformatics
- CIBCB: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology
- BIBE: IEEE International Conference on Bioinformatics and Bioengineering

Current research?

- Bioinformatics recent issue
- BMC Bioinformatics web site
- Journal of Computational Biology recent issue
- Proteins: Structure, Function and Bioinformatics recent issue
- PLoS Computational Biology recent issue
- RECOMB 2008 accepted papers
- ISMB 2008 accepted papers
- ECCB 2008 accepted papers
- CSB 2008 accepted papers

Two weeks from now

- Genome Analysis/Comparative Genomics
- Reading:
  - One paper from ECCB 2008

    HapCUT: an efficient and accurate algorithm for the haplotype assembly problem by Bansal and Bafna
  - One paper from ISMB 2008

    Classification and feature selection algorithms for multi-class CGH data by Liu et al.