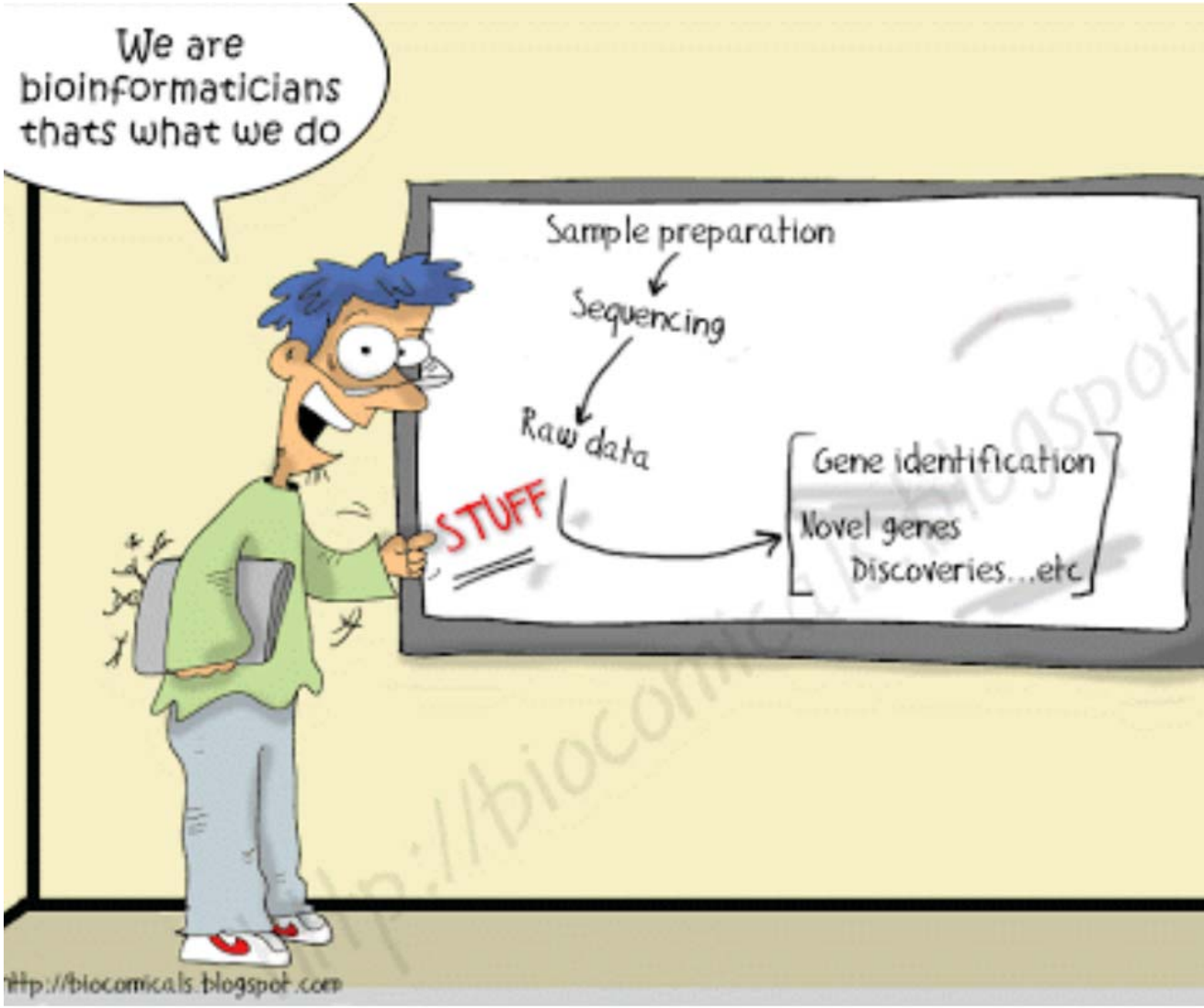
CS123A Bioinformatics Module 1 – Week 2 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ



ttp://www.genomicglossaries.com/images/shenemangenome.gif

Agenda

- Introduce you to the field of bioinformatics
- Describe two main examples that have been researched extensively,
  - hemoglobin and RBP4
- Introduce the key bioinformatics websites in use today

Summary

- Bioinformatics is a relatively new field which merges molecular biology with computers. Computer databases and algorithms are being used to characterize genes, proteins, mRNAs, and entire genomes.

- The biggest challenge is to deal with the enormous amount of data and to reveal fundamental structural and functional characteristics of the sequences identified and organisms overall. Throughout the semester we will look at both theory of bioinformatics and the practical use of databases and algorithms available.

The field of bioinformatics can be viewed from three perspectives:

1) The cell

- The cell contains the genome, the transcriptome, and the proteome specific for that cell. These sequences are derived from the central dogma of molecular biology and allow us to investigate cellular functions. From the cell we can widen our perspective to the organism.

2) The organism

- The organism changes over developmental time and different tissues will exhibit different characteristics. Organisms can also change depending on the environment or a particular disease state.

3) The tree of life.

- At the next perspective, the tree of life, we can group all known species into three major branches: bacteria, archaea, and eukaryotes. At this level, we can appreciate the power of comparative genomics and how organisms and chromosomes evolve by way of chromosomal duplications, deletions, and rearrangements.

Myoglobin and Hemoglobin Example

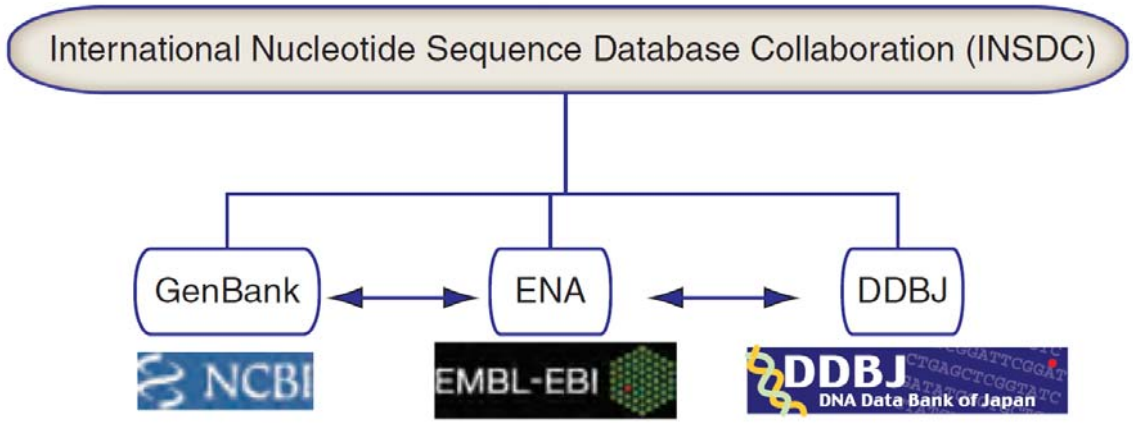
- Hemoglobin is one of the best characterized sequences in biology and lends itself to bioinformatics studies for the following reasons:
  - Hemoglobin was one of the first proteins to be studied, therefore, a lot of information available
  - Myoglobin was the first protein to have a 3D structure solved
  - Hemoglobin is a tetramer of 4 globin subunits
  - The globin loci in humans was the first to be sequenced
  - The globin family extends to plants, invertebrates, bacteria, archaea, and fungi

Retinol-Binding Protein Example

- Retinol-binding protein (RBP4) is another consistent example for several reasons:
  - Many proteins are homologous to RBP4 in a variety of species
  - Human proteins are closely related to RBP4, called lipocalins
  - Bacterial lipocalins may have entered eukaryotes by way of lateral gene transfer
  - The biochemical properties of lipocalins have been characterized in detail
  - Some lipocalins are associated with disease

Key Bioinformatics Websites

- The field of bioinformatics relies heavily on computer based programs, either locally or web based.
- At these web based sites, sequences can be accessed, software can be downloaded for analysis of sequences, sequences can be uploaded, and information can be integrated between different types of resources.
- Initially we will focus on the three main repositories of DNA and protein data; NCBI (National Center for Biotechnology Information), EMBL (European Molecular Biology Laboratory), and DDBJ (DNA Database of Japan). These research teams share data on a daily basis. We will also discuss several genome browsers including UCSC and Ensembl.



Biological Databases

- The genomes of thousands of organism, including viruses, have been sequenced over the past several years. Publicly available databases house these sequences in addition to many other sequences collected from over 260,000 organisms. The GenBank at NCBI, EMBL at EBI (European Bioinformatics Institute, and DDJB at the National Institute of Genetics in Mishima) are coordinated by the International Nucleotide Sequence Database Collaboration (INSDC).
- Although these are the main databases, there are several others that concentrate on genome sequences, specific nucleic acid type (i.e. tRNA), sequences related to one organism or protein, protein 3D structure data, or literature.

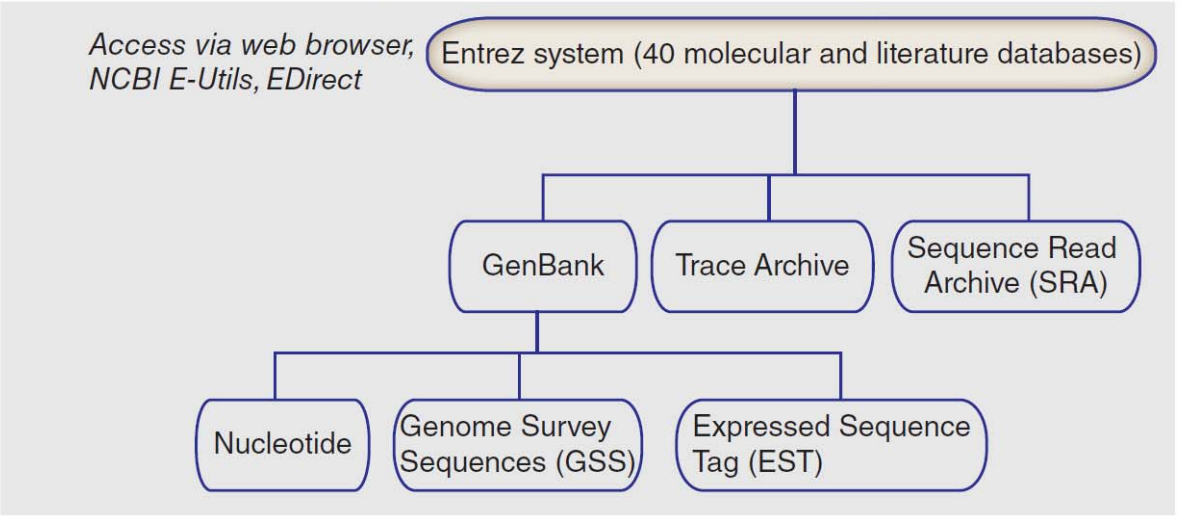
Pitfalls of Biological Databases

- There are inherently many errors in sequence databases. Many of the errors are caused by sequencing errors, but there are more common prior to the 1990's.
- Redundancy is another problem and increases the challenge of searching through a database for specific sequences or information.
- RefSeq attempts to combat this problem since it is a 'nonredundant' database where all duplicates are merged into one record.
- The other major problem is annotation errors when the same gene has a different name or two different genes acquire the same name.

Entrez: Entry Into Multiple DBs

- Entrez is one of the best information retrieval systems available today. Within Entrez you can search with keywords of sequence identifiers to locate information. When using keywords, you can use Boolean operators to narrow your search.
  - AND - search result must contain both/all search terms
  - NOT - excludes results containing a specfic search term
  - OR - results contain one or more of the search terms

(a) National Center for Biotechnology Information

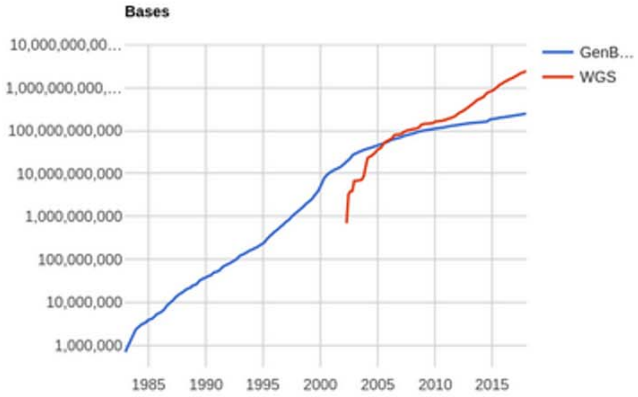


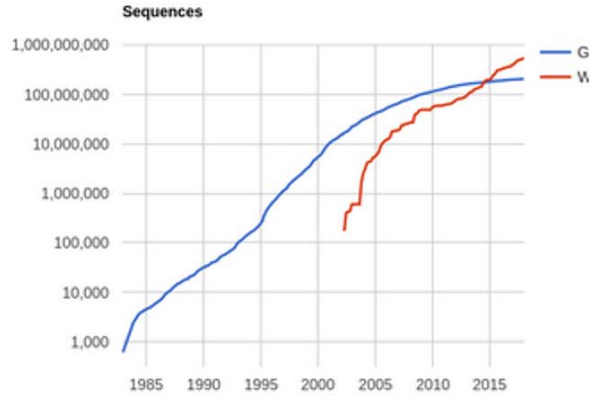
GenBank

- GenBank contains over 249 billion bases and represents over 260,000 organisms.
- The increase in sequence number is due to enhancement of sequencing technology over the years.
- Sequence data housed in GenBank comes in many forms including genomic, cDNA (complementary DNA), protein, ESTs (Expressed Sequence Tags), STSs (Sequence-Tagged Sites), GSSs (Genome Survey Sequences), and HTGSs (High-Throughput Genomic Sequence).

GenBank *(cont.)*

- The whole genome shotgun sequences includes sequences generated by high-throughput sequencing and includes 2,466,098,053,327 bases in 551,063,065 sequence records, as of December 2017.
- Since 1982 the number of bases in GenBank has almost doubled every 18 months. The figure below illustrates the growth in the number of bases and sequences in GenBank and the WGS, over the past 35 years.





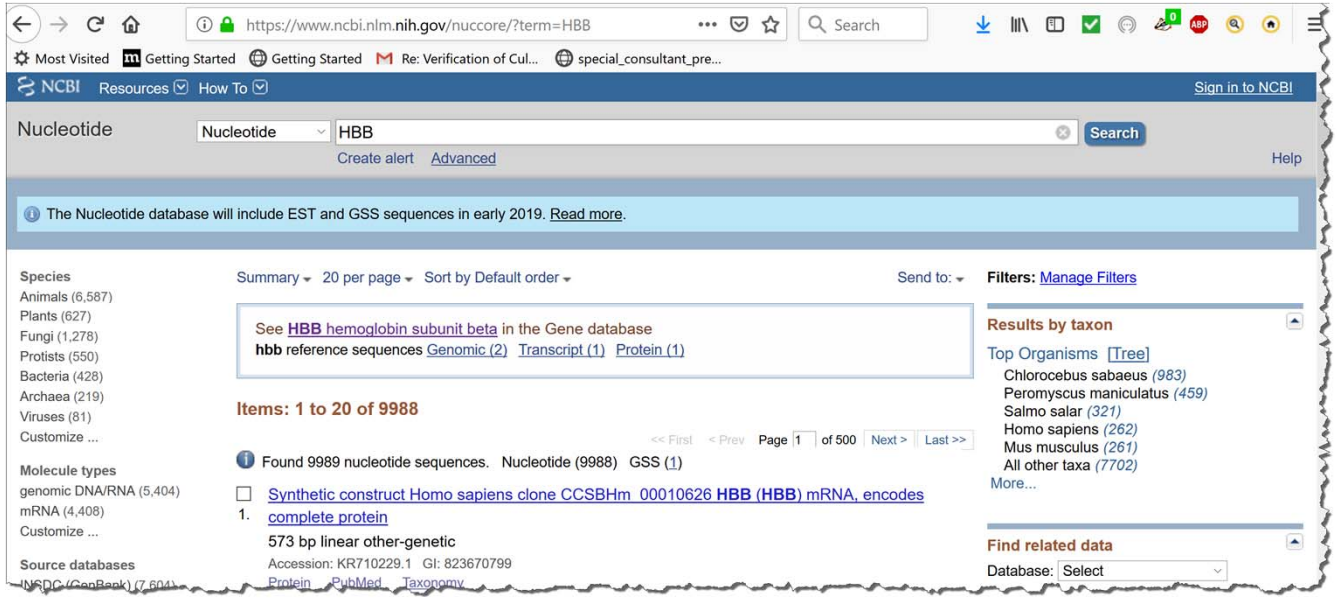
This increase in sequence number has largely been due to advances in sequencing technology. Next generation sequencing can produce 1 billion bases of sequence in one experiment that is completed in a matter of days. This increase in data is astounding.

Types Of GenBank Data

- GenBank will contain genomic DNA, cDNA corresponding to expressed genes or mRNA, expresses sequence tags (ESTs), sequence tagged sites (STSs), genome survey sequences (GSSs), high throughput genomic sequences (HTGS), and protein seqeuences.
- The genomic DNA of a specific gene, for example beta globin, may be found as part of a chromosome; a large fragment of DNA in a cosmid, BAC, or YAC; a gene; or as a STS or small fragment of DNA.
- Beta globin will also be found in GenBank as a cDNA and an EST. The cDNA may be the full coding sequence, or it may be partial. By partitioning ESTs into nonredundant sets, gene-oriented clusters are created for the UniGene (unique gene) project.
- If a gene is expressed at a low rate, there may be only one EST in a cluster. However, if the gene is expressed at a high rate, there may be thousands of ESTs in the cluster.
- STSs are short genomic landmark sequences that can be used as genomic markers if they are polymorphic.
- GSSs are similar to ESTs except that they originate from genomic sequences, rather than mRNA. The HTGS division contains unfinished genomic sequence data that was generated by high throughput sequencing centers such as Wellcome Trust Sanger Institute. Protein data can be found in the nonredundant (nr) database of GenBank along with SwissProt and UniProt.

Lets Get Info About HBB

- NCBI: http://www.ncbi.nlm.nih.gov/
- GenBank: https://www.ncbi.nlm.nih.gov/nucleotide/
- Enter HBB into the search box …



- …. in class explanation of various fields & tables.

Find The following Info For RBP4 In GenBank

- Gene ID:
- Nucleotide #:
- Last Updated:
- Location:
- Number Exons:
- Any Synonyms, If so, what are they:
- From FASTA file, what is the start position of the first coding region, i.e., how many nucleotides from the first nucleotide does AUG start: