CS123A Bioinformatics Module 2 – Week 4 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ

Agenda

- Introduction To Sequence Alignment
  - Local, Global, BLAST
  - BLAST example and exercise

What Is Sequence Alignment?

- One of the most basic questions about a gene or protein is whether it is related to any other gene or protein.
- Relatedness of two DNA or proteins at the sequence level suggests that they are homologous. Relatedness also suggests that they may have common functions. By analyzing many DNA and protein sequences, it is possible to identify domains or motifs that are shared among a group of molecules.
- Two flavors: Pairwise and Multiple sequence alignment
  - Pairwise: Compare two DNA/protein sequences for relatedness.
  - Multiple: Compare three or more DNA/protein sequences for relatedness.

Pairwise Alignment

- Pairwise sequence similarity searches are the backbone of many bioinformatics tasks. Sequences can be analyzed at the structural, functional, and evolutionary levels.

- The alignment provides information about how two sequences are related (or not related), and if they may be homologous. Through local and global alignment methods and three main algorithms, sequences can be evaluated.

- The three algorithms include the dot matrix method, dynamic programming, and the word method.

- Scoring matrices are used to describe the statistical probabilities of one residue or nucleotide being substituted for another.

Important Definitions

- Similarity - a quantitative measure based on sequence identity and pairwise alignment.

- Homology – *common ancestral gene/protein* - extrapolated from similarity and usually implies an evolutionary link. (*e.g., genetic codes underlying a bat wing and a bear arm. Both retain similar features and are utilized in similar manners.*)

- Orthologs – *same gene/protein in different species* - genes separated by speciation and typically have the same function, 3D structure, and domain structure. (e.g., electron transport proteins NADH, FADH2, cytochrome c, …)

- Paralogs - *same genes/proteins within the genome of a species* - genes that are separated by genetic duplication and typically do not have a similar function. That is, a copy of the gene is made in the genome and evolved to have another function. (e.g., Hemoglobin & Myoglobin)

Important Definitions *(cont.)*

- **Global Alignment:** Global alignment looks at full length sequences and attempts to make the best alignment over the full length of both (or several) sequences. This method is most useful when the sequences being aligned are the same length. The most general global alignment method was devised by Needleman & Wunsch, called the Needleman-Wunsch algorithm, and is based on dynamic programming. Global alignments may overlook important, smaller similarities such as functional domains.

- **Local Alignment:** Local alignment techniques attempt to align subsections of sequences and typically return many alignments for one sequence. A general algorithm used for local alignments is the Smith-Waterman algorithm. This algorithm is also based on dynamic programming. Local alignments are useful in finding small stretches of similarity in sequences of varying length. If sequences are very similar, you will see little difference between a global and local alignment. However, in the example on the following slide, you can see how global and local alignments differ when the sequences are not very similar. In this case, the global alignment inserts many gaps and reduces the quality of the alignment .

Local vs Global Alignment

- Suppose we wish to align the sequence FTALLLAAV …

- Global: FTFTALILLAVAV

  F---TAL-LLA-AV

- Local: FTFTALILLAVAV

  --FTAL-LLAAV

Aligning Human Myoglobin and Hemoglobin

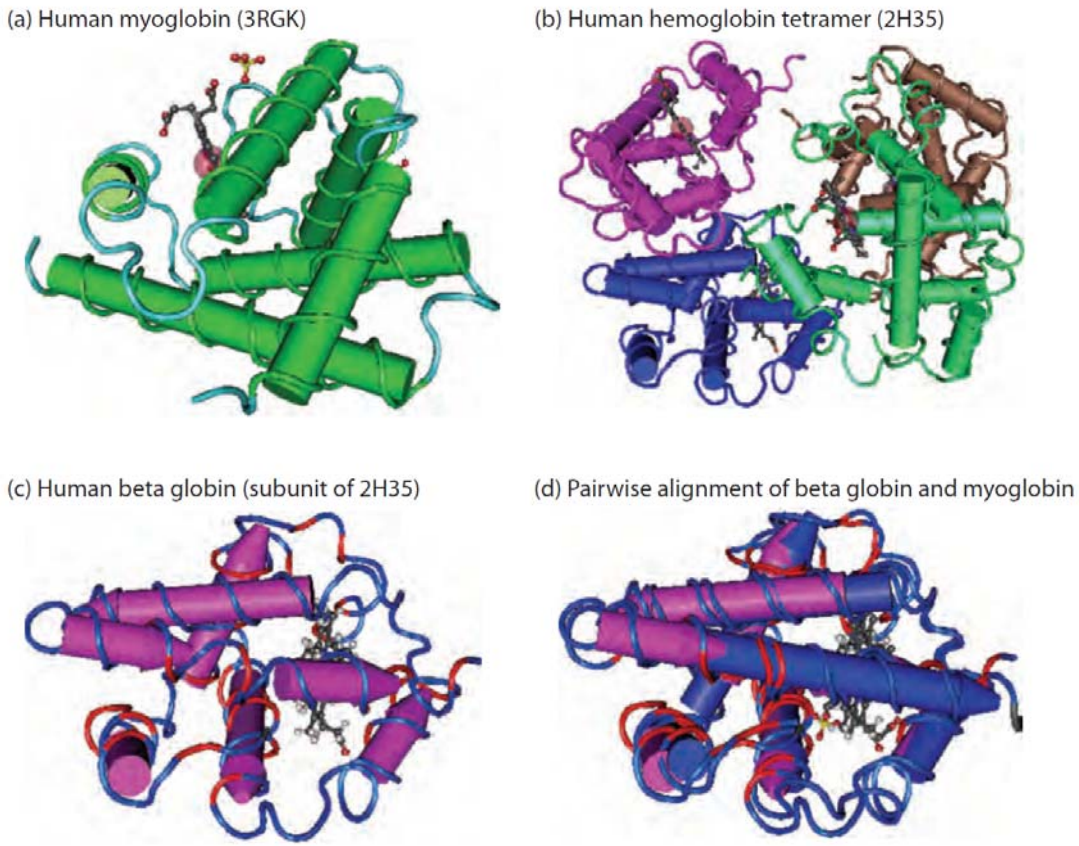


Fig 3.1 in textbook

Phylogenetic Tree Based On Hemoglobin & Myoglobin Alignment

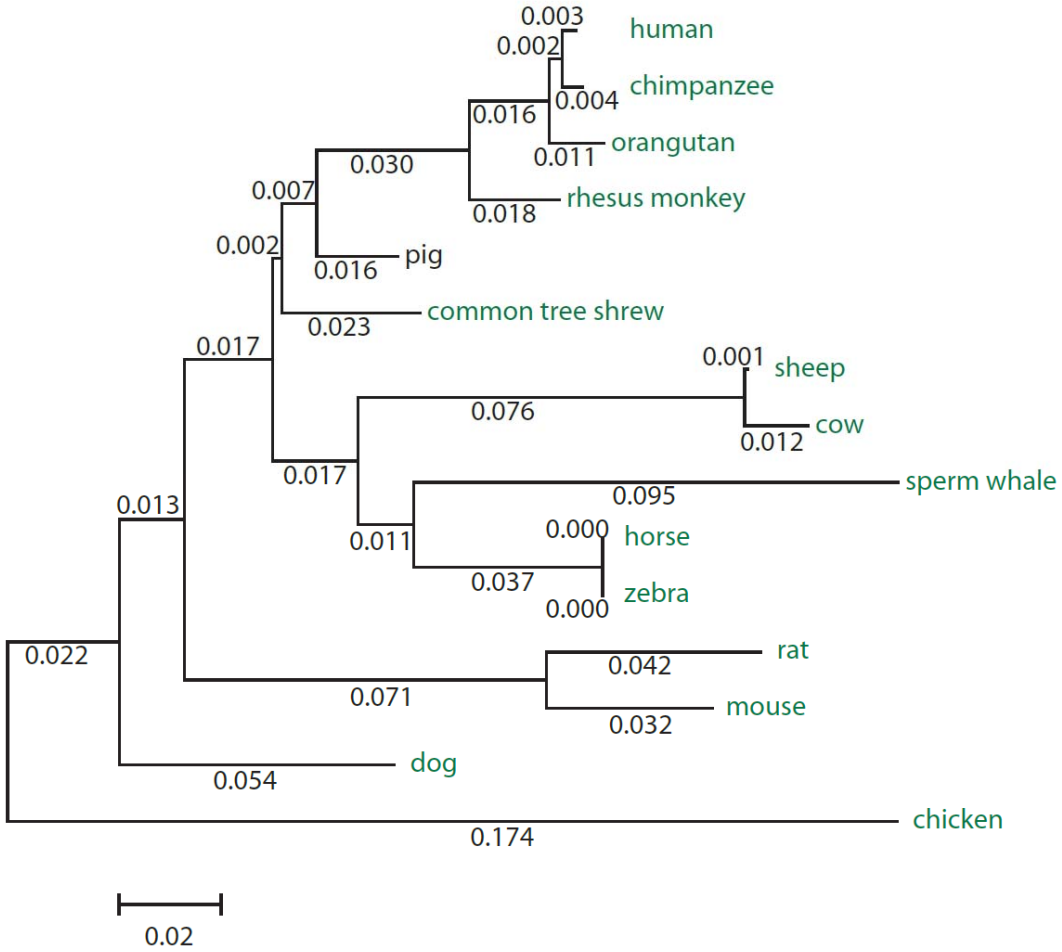


Fig 3.2 in textbook.

Alignment Algorithms

- Algorithms used for global and local alignments are fundamentally similar, and differ only in the optimization steps. The three types of methods used to produce a pairwise alignment are the dot-matrix or dot plot method, dynamic programming, and word methods.

Cytochrome C Alignment Example Using BLAST

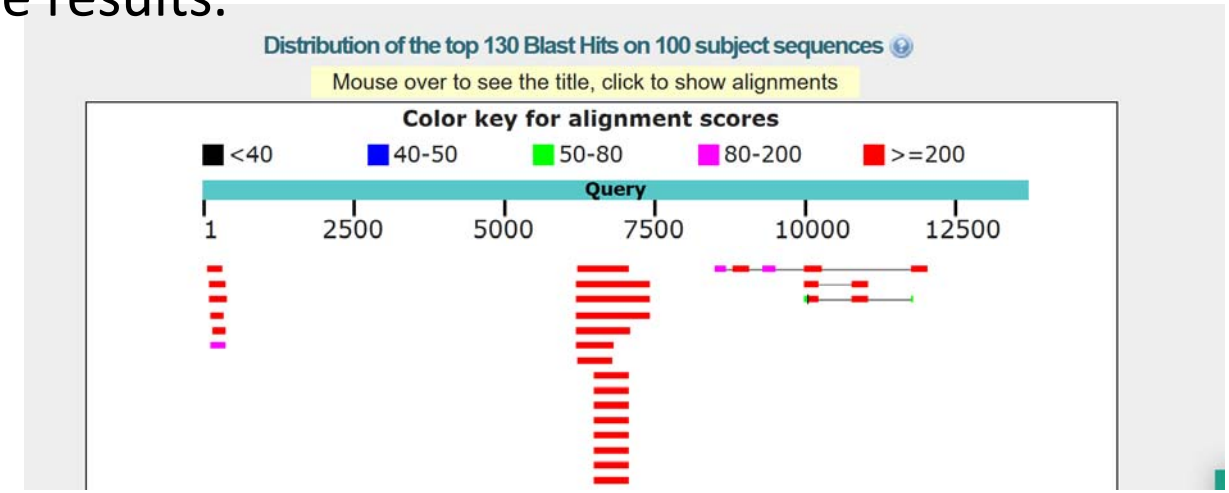
- What is cytochrome C?
- Go to NCBI www.ncbi.nlm.nih.gov and select nucleotide DB
- Enter "human cytochrome c" into the search box and click SEARCH
- Click on the "RefSeq Gene (1)" link in the CYCS – cytochrome c, somatic" box area.
- Click on the "FASTA" link in the upper left of the page.
- Click the drop down button to the right of the FASTA text at the top left.
- Copy just the ACC number in the "> …" line or the entire page.

Cytochrome C Alignment Example Using BLAST *(cont.)*

- Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Click on the Nucleotide BLAST box on the left.
- In the "Enter accession number(s) …" area paste the copied ACC number or the entire FASTA text you previously copied.
- Just click in the Job Title area to give your search a title.
- Click the "Others" button on the Database line
- In the Organism window, enter "mouse" and select a mouse entry of interest to you. Click the "+" to add one more organism. Enter Rat and select Rattus.

Cytochrome C Alignment Example Using BLAST *(cont.)*

- Then scroll down until you see the “BLAST” button. Select the box next to the “Show results in a new window” text.
- Click on BLAST and wait for results.
- At top right of results page is a link to a YouTube video that describes how to read/interpret the results.



Cytochrome C Alignment Example Using BLAST *(cont.)*

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

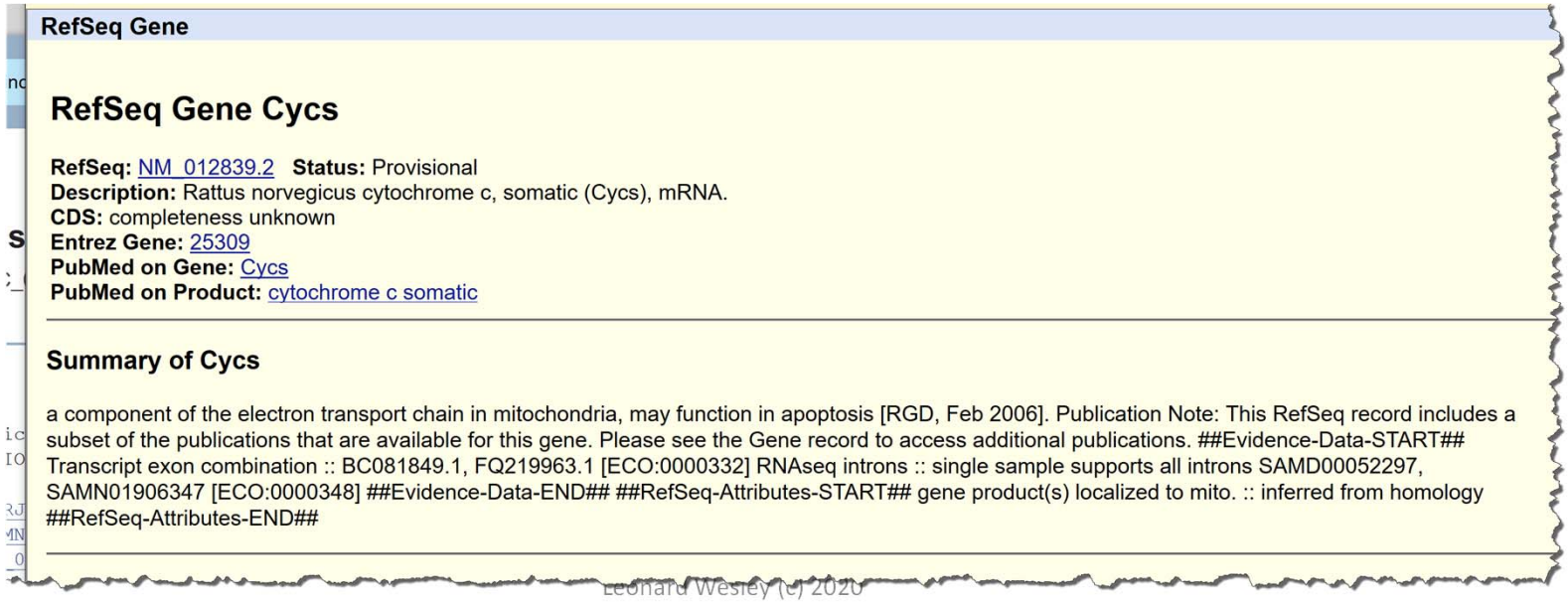
Description	Max score	Total score	Query cover	E value	Ident	Accession
Rattus norvegicus cytochrome c gene, complete cds; nuclear gene for mitochondrial product	586	586	6%	2e-163	79.81%	K00750.1
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Cycs:tm1a(EUCOMM)Hmgu; transgenic	566	566	9%	3e-157	76.09%	JN963523.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele Cycs:tm1e(EUCOMM)Hmgu; transgenic	566	566	9%	3e-157	76.09%	JN951646.1
Mus musculus 6 BAC RP23-198D7 (Roswell Park Cancer Institute (C57BL/6J Female) Mouse BAC Library) complete sequence	566	566	9%	3e-157	76.09%	AC153386.3
Mouse cytochrome c gene (MC1)	564	564	6%	1e-156	78.95%	X01756.1
Mus musculus cytochrome c (CYCS) gene, complete cds	499	499	4%	3e-137	82.02%	JF919281.1
Rattus sp. cytochrome c (CYCS) gene, complete cds	484	484	4%	9e-133	82.34%	JF919282.1
Rattus norvegicus TUAGA43YA12 mRNA sequence	405	405	4%	7e-109	80.60%	FQ216426.1

Cytochrome C Alignment Example Using BLAST *(cont.)*

- Click on the first link “Rattus norvegicus cytochrome c …”
- Under “Related Information” on the right, click on the “Gene” link.
- Copy the REGION: complement location range.
- Go to UCSC Genome browser, select Rat species.
- In the Position/Search Term window enter the chromosome number followed by a colon (:) and the copied complement location range. NOTE: Replace the “..” with a “-” in the range.
- On the left, go to the line with “Cycs” and hover over the left most purple region. What pops up?

Cytochrome C Alignment Example Using BLAST *(cont.)*

- At the bottom are the results of alignments with other organisms.
- Hover over the left most exon, then hold down the Shift key and click. You should see a window like



BLAST The Following DNA Sequence

- Go to the “Slides” folder of Module 2 Week 4 on Canvas and download the dna_seq.txt file.

- BLAST the sequence against humans, mice, and rats. Describe the results of the BLAST. What is the sequence, location, length, chromosome #, # exons for each organism, function (from the description)?

Next Class How BLAST Works