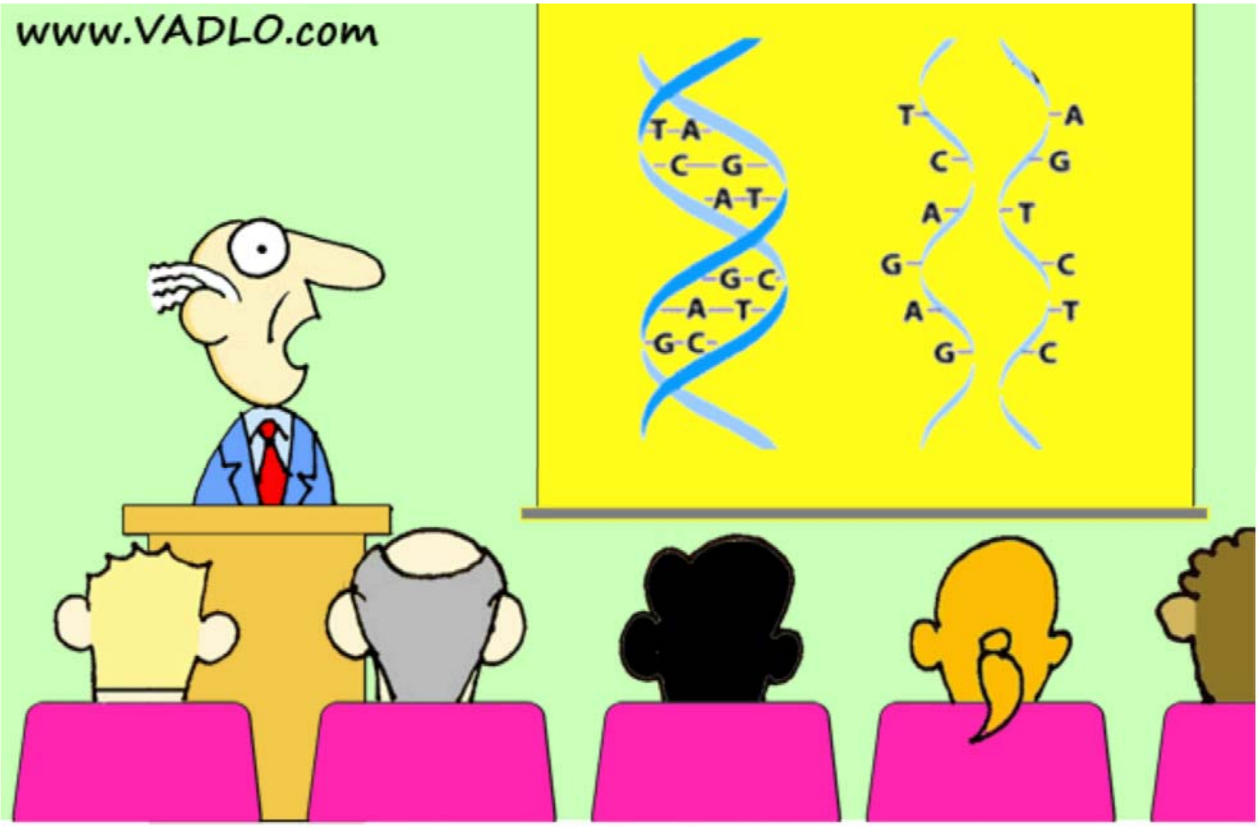
CS123A Bioinformatics Module 2 – Week 4 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ



"Then Crick asked me,
'Watson, should we name them A,C,T,G or A,B,C,D?"

Agenda

- Quiz 1 (next Thursday)
  - Study guide
  - Format (Canvas)

- Bioinformatics DB Tutorials

- Bioinformatics DB Example Question

- Sequence Alignment

Quiz 1 Study Guide

- Biology Basics
  - Central Dogma, replication, transcription, translation, where the steps of the central dogma take place in a eukaryote cell.
  - DNA, RNA structure, nucleotides, nucleosides, major components of and distinction between eukaryotes & prokaryotes, purines, pyrimidines, type of bonding between nucleotides, types of mutations and their impact/consequences.
- Bioinformatics DBs
  - The three major bioinformatics DBs in the world, NCBI, EMBL/EBI, (Don’t worry about the Japan DB for now), how to find the type of information searched for and found in the in-class exercise, lectures, and HWs and short vides.

Quiz 1 Format

- On Canvas ~35 to 45mins
- Open book, open notes, open web, and open mind. However, you MUST NOT communicate with or channel to any other living or dead organism or object in or outside of the known universe during the exam/quiz.
- Will start at 10:35AM sharp. Door closes at 10:40AM. If you are not in the classroom by 10:40AM you will need to wait until the quiz is over to enter. Go to the restroom BEFORE the exam starts. Once you leave during the exam/quiz, you cannot get back in until it is over.

Bioinformatics DB Tutorials

- Sample GenBank Record With Field Definitions (not a video):
  - https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

- GenBank Tutorial:
  - https://www.youtube.com/watch?v=g5a__okj5Zs 4:59

- NCBI Gene DB: www.youtube.com/watch?v=IqhkhnplR38 5:47
  - Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

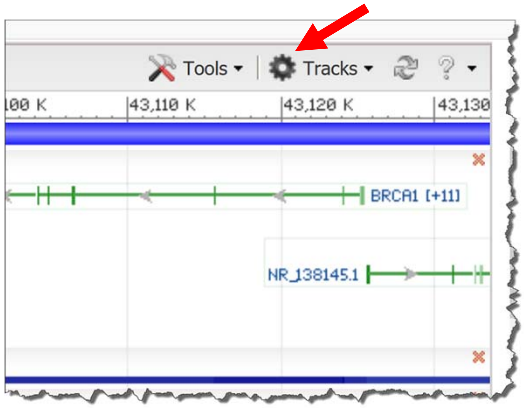
- NCBI Nucleotide: https://www.youtube.com/watch?v=2gsGEXsOAlI 6:39
  - The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

- NCBI ClinVar: https://www.youtube.com/watch?v=A8G3ej83ZgU 6:25
  - A DB of genomic variance and their relevance to clinical diseases.

- NCBI Variant Viewer: https://www.youtube.com/watch?v=rnWZ9MFBwUM ~6:00
  - Variation Viewer allows you to view, search, and navigate variations in genomic context. You can review data from dbSNP, dbVar and ClinVar, or upload your own data. You can search based on chromosomal location, gene, variant IDs from dbSNP and dbVar, or phenotype; and review results both as sequence annotation tracks and in a filterable table.

NCBI Variant Viewer

- Go to https://www.ncbi.nlm.nih.gov/variation/
  - Notice there are several variation DBs of potential interest, e.g., dbSNP, dbVar, CLinVar …etc.
- Click on the "Variation Viewer" link. <- NOTE: This link was there ~Spring2020.
- Under "Pick Assembly" select GRCh38.p12 Annotation Release 109
- Click on "Search examples" to see the format & type of entries you can enter to search.
- Enter BRCA1 in the Search window.
- On the "Genes" tab on the left, select the desired gene, e.g., BRCA1, and then click on the blue arrow/tab on the right.
- Several tracks are displayed below the green track that is centered around the BRCA1 gene
  - ClinVar short/large variations ..etc.
  - Click on the Configure Icon (upper right) to select which tracks to display.
  - There is also a Help menu ("?") if needed.



NCBI Variant Viewer *(cont.)*

- Use the “Region” link (in blue band area near top) to view the gene with or without padding.
- To the right of the “Region” link is a “Gene” menu. You can select to view the gene, or (if listed) RPLs (Recurrent Pregnancy Loss) versions of the gene.
- To the right of “Gene” is a “Transcript“ selection box.
- Further right is a series of dots where each dot corresponds to an exon in the gene. Clicking a dot will display a red line on/around the area where the exon is located within the gene.
- Hovering over the BRCA1 symbol to the right of the green gene track will display some additional info, e.g., length, gene id, GenBank view, …etc.

NCBI Variant Viewer *(cont.)*

- Click on one of the exon regions, lets try the second dot. We now zoom into the exon region and below are the ClinVar and/or dbSNP links.
- Hover over the top dbSNP (nsv2769779, nsv = nucleotide snp variation). We are taken to a page with links to the study that produced the presented results, whether it has been validated, publication links, visual display/icon of the relevant chromosome, and on the right are one or more links to ClinVar info if such is available.
- Click on the top ClinVar link. We are taken to a page with the same type of ClinVar information that you saw in the in-class exercise. Notice under Browser views" to the right is a link to related info in the UCSC DB.

NCBI Variant Viewer *(cont.)*

- Now select a RPL number under the Gene pull-down menu in the blue band near the top. Zoom in to see the various SNPs.
- Hover over a SNP entry to display additional info and links.

Example Question

- What is the CFTR gene?
- What is the location of the CFTR gene?
- What is the length of the CFTR gene?
- Name a publication, author(s) and date related to this Gene.
- What is the ACC number for this gene?
- When was info about this gene last updated to GenBank?
- Is there any variation info about this gene? If so, what is the “rs” number for the first entry displayed in the Clinical, dbSNP aree of the display?
- Is there clinical variation info available? If so, what is the ACC number, location of the variation, and type of variation (e.g., Insertion, deletion, nonsense,…)? What is the clinical significance of the variation.

Sequence Alignment

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 - extrapolated from similarity and usually implies an evolutionary link.

- Orthologs - genes separated by speciation and typically have the same function, 3D structure, and domain structure.

- Paralogs - genes that are separated by genetic duplication and typically do not have a similar function.

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 against other sequences in a DB.

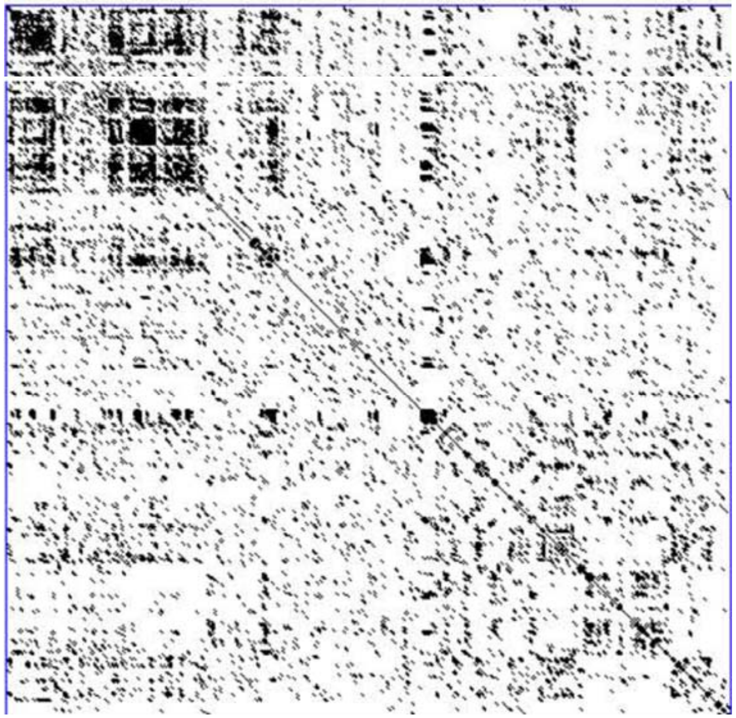
- Global:
```
FTFTALILLAVAV
F---TAL-LLA-AV
```

- Local:
```
FTFTALILLAVAV
--FTAL-LLAAV
```

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.

Dotplots



The dot plot is a visual method used to identify regions of local alignment, direct or inverted repeats, insertions, deletions, or low-complexity regions. To construct a dot plot, two sequences are written along the top row and the most left-hand column, of a two-dimensional matrix or table. A dot is placed at the point where the characters match at the intersection of the x and y axis. If the sequences are very similar, a straight line will be visualized.Dotplots are very useful in identifying regions of biological importance, but they do not provide a statistical view.The Dotie! program is the most user friendly, web-based software available.Dotie! can be used to compare a sequence to itself in order to identify tandem repeats and low-complexity regions.The figure below shows an example of a DNA dot plot of a human zinc finger transcription factor (NM_002383). The main diagonal represents the sequence homology with itself. Lines off the main diagonal represent repetitive elements within the sequence.

Dynamic Programming & Word Method

- **Dynamic Programming**

  Dynamic programming can be used with either global or local alignment algorithms.
  Dynamic programming also creates a two dimensional alignment matrix, but the alignment is more quantitative. Typically, protein alignments use a substitution matrix to assign scores or values to amino acid matches or mismatches. A gap penalty is also assigned either for opening **a** gap or extending a gap. Given the scoring method, dynamic programming is guaranteed to find an optimal alignment based on **a** specific scoring matrix. However, the more difficult parl is to identify the correct scoring matrix.

- **Word Methods**

  Also known as k-tuple methods, are heuristic methods that are not guaranteed to find an optimal alignment, but are significanly faster than dynamic programming. These methods are especially useful in large-scale database searches. The two best known word method search tools are BLAST and FASTA. Word methods identify a short, non.overlapping sequence or word in the query sequence. These words are then matched to sequences in the database. The word method reduces the number of unnecesssary comparisons with sequences with little similarity, thus increasing speed.

Continued Next Class