EECS 730
Introduction to Bioinformatics

Luke Huan
Electrical Engineering and Computer Science
http://people.eecs.ku.edu/~jhuang/
Outline

- Why sequence alignment and definitions
- What is sequence alignment
- How to score an alignment
  - Substitution matrix
  - Gap penalty
Access Sequence Database

- Query by sequence
  
  TCCAGCAGTGATCTACTGGAGAAATATATATGCAGCAAGGGGA
  AAAGACAGAGAGAGCAGAGAGAGAGGAGACATCAAGGGGAGGT
  GACAGAGGACTTACTGCACCTCGAGCAGAGGGAGACACCAT
  ACAGGGGAGCCACCAACAGAGAGACTTGCTGCACCTCAATTCT
  CTCTTTGGAAAAGACCAGTAGTCACACGATACATATTGAGGGT
  CAGCCAGTAGAAGTTTTTGTAGACACCGGGAGCTGACGACTC
  AATAGTAGCAGGAATAGAGTTAGGAAACAATTATAGCCAAAA
  AATAGTAGGGGGGAATAAGGGGGGATTTCATAAATACCAAGGAAT
  ATAAAAATGTAGAGATAGAAAGTTCTAAATAAAAGGTACGGG
  CCACCATAATAGACAGGCGACACCCCCCAATCAACATTTTTGGCA
  GAAATATTCTGACAGGCCTTGGCATGTCATTAAATCTA
Aligning biological sequences

- **Nucleic acid (4 letter alphabet + gap)**
  
  TT–GCAC
  TTTACAC

- **Proteins (20 letter alphabet + gap)**
  
  RKVA--GMAKPNM
  RKIAVAAAASKPAV
What is sequence alignment

Given two strings, and a scoring scheme for evaluating matching letters, find the optimal pairing of letters from one sequence to letters of the other sequence

Align:
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
THIS IS A SHORT SENTENCE

or

THIS IS A RATHER LONGER - SENTENCE THAN THE NEXT
| | | | --* | -- -- |---|---| | | --- --- --- ---
THIS IS A --SH-- -O---R T SENTENCE ---- --- ----

or

THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
| | | | ------- ------- | | | | | | | | | | --- --- --- ---
THIS IS A ------- -SHORT SENTENCE ---- --- ----
Why sequence alignment

- Lots of sequences with unknown structure and function vs. a few (but growing number) sequences with known structure and function.
- If they align, they are “similar”
- If they are similar, then they might have similar structure and/or function. Identify conserved patterns (motifs).
- If one of them has known structure/function, then alignment of other might yield insight about how the structure/functions works. Similar motif content might hint to similar function.
- Define evolutionary relationships.
Problems!

- How much is “similar”
  - 95% similarity in proteins is ~ identical
  - 80% similarity is a lot in proteins
  - Less similarity than that needed for DNA

- Database techniques inadequate – they are too precise!

- Datasets very large to search
Pair-wise sequence alignment is the most fundamental operation of bioinformatics

- It is used to decide if two proteins (or genes) are related structurally or functionally
- It is used to identify domains or motifs that are shared between proteins
- It is used in the analysis of genomes
Similar (or not)

- **Similarity**
  The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

- **Identity**
  The extent to which two sequences are invariant.

- **Conservation**
  Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP:</td>
<td>26</td>
<td>RVKENFDKARFSGTWYAMAKKDPEGFLQDNIVA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ K++ + ++ GTW++MA + L + A</td>
</tr>
<tr>
<td>Glycodelin:</td>
<td>23</td>
<td>QTQDLELPKLAGTWHSMA-TNNISLMA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LK'A 55</td>
</tr>
</tbody>
</table>
retinol-binding protein 4  
(NP_006735)  

β-lactoglobulin  
(P02754)
Pairwise alignment of retinol-binding protein 4 and β-lactoglobulin

1 MKWVWALLLAAWAAAERDCRVSFRVKENFDFSGTNYAMAKKDPEG 50 RBP
   . ||| | . . . | :: ||||.:| :
1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
 : | | | | :: | . | . | | | :: | |
45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECALKKKIAEKT 93 lactoglobulin

98 DPAKFKMKYWGASFLQKGNDDHWIVDGDYDTYAV...........QYSC 136 RBP
 || ||. ||| :: ||||| | .
94 IPAVFKIDALNENKVL........VLDTDYKYYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADYSFVFSRDPNGLPPEAKIVRQRQ.EELCLARQYRLIV 185 RBP
 . | | | | :: | . | | |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
Pairwise alignment of retinol-binding protein and \( \beta \)-lactoglobulin

| 1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP |
| . ||| | . | . . | : ||| . . . .:| : |
| 1 ...MKCLLLLALALTCAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin |
| 51 LFLQDNIVAEFSVDETQMSATAKGRVR.LLNWD.AADMVGTFTDTE 97 RBP |
| 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENCQKKKIAEKT 93 lactoglobulin |
| 98 DPAKFKMKYWGVASFLQKGNDDHWIVTDYDTYAV..........QYSC 136 RBP |
| ||| | | : ||| ||| | . | . . |
| 94 IPAVFKIDALNENKVL........VLTDYKKYLLEGNSAEPEQSLAC 135 lactoglobulin |
| 137 RLLNLGDTCADSYSFVSRDNPGLPPAEAQKIVVERQYRLIV 185 RBP |
| . | | | | : | | |
| 136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSF........... 178 lactoglobulin |
Pairwise alignment of retinol-binding protein and β-lactoglobulin

Somewhat similar (one dot)

Very similar (two dots)
Pairwise alignment of retinol-binding protein and β-lactoglobulin

1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDEPG 50 RBP
   . ||| | . | . . | : .|||.:| : |
1 ...MKCLLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

51 LFLQDNIVAEFSVDETGQMSATAKGVR.LLNNWD..VCADMVGTFTDTE 97 RBP
 : | | | :: :.| . | | | : || |
45 ISLLDAQSAPLRV.YVEELKPTPDEEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV........QYSC 136 RBP
 || || | :: :||| | . |
94 IPAVFKIDALNENKVL........VLDTYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCSYFVFSDEGLPPEAQKIVRQRO.EELCLARQYRLIV 185 RBP
 . | | | : || . | || |
136 QCLVRTPEVDREALEDKFKV.KALPMHIRLSFNPTQLEEQCHI........ 178 lactoglobulin

Internal gap

Terminal gap
Biological Interpretation of sequence alignment

- **Homology**
  Similarity attributed to descent from a common ancestor.

- **Two types of homology**
  - **Orthologs**
    Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.
  - **Paralogs**
    Homologous sequences within a single species that arose by gene duplication.
Orthologs:
members of a gene (protein) family in various organisms.

This tree shows Retinol binding protein (RBP) orthologs.
Paralogs: members of a gene (protein) family within a species

- apolipoprotein D
- retinol-binding protein 4
- Alpha-1 Microglobulin /bikunin
- progestagen-associated endometrial protein
- Odorant-binding protein 2A
- Lipocalin 1
- Complement component 8
- prostaglandin D2 synthase
- neutrophil gelatinase-associated lipocalin

10 changes
Similarity versus Homology

- Similarity refers to the likeness or % identity between 2 sequences
- Similarity means sharing a statistically significant number of bases or amino acids
- Similarity does not imply homology

- Homology refers to shared ancestry
- Two sequences are homologous if they are derived from a common ancestral sequence
- Homology usually implies similarity
Similarity versus Homology

- Similarity can be quantified
- It is correct to say that two sequences are X% identical
- It is correct to say that two sequences have a similarity score of Z
- It is generally incorrect to say that two sequences are X% similar
Aligning biological sequences

- Any two sequences can always be aligned
- There are many possible alignments
- Nucleic acid (4 letter alphabet + gap)
  - TT-GCAC  TTG-CAC
  - TTTACAC  TTTACAC
- Proteins (20 letter alphabet + gap)
  - RKVA----GMAKPNM  RK----VAGMAKPNM
  - RKIAVA--AAASKPAV  RKIAVAAAASKPAV
- Sequence alignment needs to be scored to find the "optimal" alignment
Statement of problem

- Given:
  - 2 sequences
  - Scoring system for evaluating match (or mismatch) of two characters (simple for nucleic acids / difficult for proteins)
  - Penalty function for gaps in sequences

- Produce:
  - Optimal pairing of sequences that retains the order of characters in each sequence, perhaps introducing gaps, such that the total score is optimal.
Pairwise alignment: protein sequences or DNA sequence?

- Protein is more informative (20 vs 4 characters); many amino acids share related biophysical properties.
- Codons are degenerate: changes in the third position often do not alter the amino acid that is specified.
Pairwise alignment: protein sequences can be more informative than DNA

- DNA sequences can be translated into protein, and then used in pairwise alignments
- DNA can be translated into six potential proteins

```
5' CATCAA
5' ATCAAC
5' ATCAAC
5' TCAACT
5' CATCAACTACAACCTCCCAAGACACCCCTACACATCAACAAACCTACCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 5'
5' GTGGGT
5' TGGGTA
5' GGGTAG
```
Scoring Function

- Positive score for identities
- Some partial positive score for conservative substitutions
- Gap penalties

\[ S = \sum_{\text{identities, mismatches}} - \sum \text{ (gap penalties)} \]

Score = \text{Max}(S)
Parameters of Sequence Alignment

**Scoring Systems:**

• Each symbol pairing is assigned a numerical value, based on a symbol comparison table.

**Gap Penalties:**

• Opening: The cost to introduce a gap
• Extension: The cost to elongate a gap
Recap

- Sequence alignment reveals the similarity between two sequences.
- Similar sequences might be homolog sequences and (due to the evolutionary connection) have similar function.
- The sequence alignment problem is an optimization problem: produce the best alignment according to a scoring function.
- A scoring function provides numeric values for each possible symbol pairing and for gaps in an alignment.
Next time

- Algorithms to solve the sequence alignment problem.
Acknowledge