Identifying and Aligning Homologs

Fran Lewitter
Bioinformatics and Research Computing @ Whitehead Institute
6/6/12
Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Table 1. Sequence similarity between p28sis and PDGF. The p28sis sequence is from (10); the PDGF sequences are from (18, 20). Residue identity between the p28sis and PDGF sequences is indicated by the solid lines between the sequences. A question mark indicates that no amino acid sequence assignment has yet been made for that position; the brackets indicate no sequence is yet available for the included segments. The box around p28sis positions 65 and 66 indicates a possible proteolytic processing position for generation of a fragment of p28sis corresponding to PDGF-2. Single letter abbreviations for the amino acid residues are as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.
What I hope you’ll learn

• What we can learn from sequence alignments
• Fundamentals of alignments
• Tools for building alignments
Topics to Cover

• Introduction
  – Why do alignments?
  – Definitions
  – Scoring alignments

• Pairwise Alignment methods

• Multiple sequence alignments

• Pre-computed alignment resources

Identifying and Aligning Homologs (Whitehead Institute)
Why do alignments

- Use sequence similarity to infer homology and/or structural similarity between 2 or more genes/proteins
- Identify more conserved regions of a protein, potentially identifying regions of most functional importance
- Compare and contrast homologs (perhaps into groups) based on shared positions or regions
- Infer evolutionary distance from sequence dissimilarity
Evolutionary Basis of Sequence Alignment

- **Similarity** - observable quantity, such as percent identity
- **Homology** - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this
  - *Paralog* – genes related by duplication
  - *Ortholog* – genes related by speciation
More Definitions

• An alignment is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.

• An optimal alignment is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.
Alignment Concepts

• **Global alignment** - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.

• **Local alignment** - Smith-Waterman (1981) produces the highest scoring regional match between two sequences.

• **Insertion and Deletions** (indels)

• **Affine gap costs** - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap’s length
Global vs Local Alignment

GLOBAL

LOCAL

From Mount, Bioinformatics, 2004, pg 71

Identifying and Aligning Homologs (Whitehead Institute)
Possible Alignments

A: T C A G A C G A G T G
B: T C G G A G C T G

I. T C A G A C G A G T G
   T C G G A – – G C T G

II. T C A G A C G A G T G
    T C G G A – G C – T G

III. T C A G A C G A G T G
     T C G G A – G – C T G

Identifying and Aligning Homologs (Whitehead Institute)
Nucleotide vs Protein

• If comparing protein coding genes, use protein sequences because of less noise
• If protein sequences are very similar, it might be more instructive to use DNA sequences
• If interested in DNA alignment of coding sequences, first do a protein alignment and use it as a template for aligning DNA sequences

Identifying and Aligning Homologs (Whitehead Institute)
**AA Scoring Matrices**

**Part of PAM 250 Matrix**

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-4</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Part of BLOSUM 62 Matrix**

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>-1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

- **PAM** - point accepted mutation based on *global* alignment [evolutionary model]

- **BLOSUM** - block substitutions based on *local* alignments [similarity among conserved sequences]

Log-odds = \( \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}} \)

Identifying and Aligning Homologs (Whitehead Institute)
Substitution Matrices

- BLOSUM 30
- BLOSUM 62
- BLOSUM 80

- PAM 250 (80)
- PAM 120 (66)
- PAM 90 (50)

Increasing similarity

% identity

% change

Identifying and Aligning Homologs (Whitehead Institute)
Scoring for BLAST Alignments

Score = 94.0 bits (230), Expect = 6e-19
Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTLSGAPCQPWASEATYRNVTAEQ---AR 256
Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNLSILIGKYYTAQNPAS 257

Query: 257 NWGLGGHAFCRNDIRPWCFVLNRDSLWEYCDLAQCQT 297
Sbjct: 258 ALGLGKHNYCRNDGAKPWCHVLKNRRFLWEYCDVPSCST 298

Position  1: Y - Y =   7
Position  2: T - S =   1
Position  3: G - S =   0
Position  4: P - E =  -1
   ...  
Position  9: - - P = -11
Position 10: - - A = -1
   ...  
Sum     230

Based on BLOSUM62
What's significant?

• High confidence - >40% identity for long alignments (Rost, 1999 found that sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity >40%)

• “Twilight zone” – blurry - 20-35% identity

• “Midnight zone” - <20% identity
Topics to Cover

• **Introduction**

• **Pairwise Alignment methods**
  – Dot plot analysis
  – Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  – Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST, BLAT)

• **Multiple sequence alignments**

• **Pre-computed alignment resources**
Comparing two sequences

- **DOTLET – Dot Plot** ([http://myhits.isb-sib.ch/cgi-bin/dotlet](http://myhits.isb-sib.ch/cgi-bin/dotlet))

- **NCBI**

- **EBI** — ([http://www.ebi.ac.uk/Tools/psa/](http://www.ebi.ac.uk/Tools/psa/))
  - **GLOBAL**
    - needle (EMBOSS) - Needleman-Wunsch
    - stretcher (EMBOSS) – modification of N-W
  - **LOCAL**
    - water (EMBOSS) - Smith-Waterman
    - matcher (EMBOSS) - uses algorithm based on LALIGN

Identifying and Aligning Homologs (Whitehead Institute)
Dot Plot

- Graphical way of looking at alignment of 2 sequences
- Look at structure of a sequence by doing a self comparison
- Method first described by Gibbs and McIntyre (1970)
- Can find direct or inverted repeats in sequences
Dot Matrix Comparison

http://myhits.isb-sib.ch/cgi-bin/dotlet
Identifying and Aligning Homologs (Whitehead Institute)
NCBI - Blast2seq

Identifying and Aligning Homologs (Whitehead Institute)
NCBI - Blast2seq

Identifying and Aligning Homologs (Whitehead Institute)
Identifying and Aligning Homologs (Whitehead Institute)
### Pairwise Sequence Alignment

**Pairwise Sequence Alignment** is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, **Multiple Sequence Alignment (MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

#### Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

- **Needle** (EMBOSS): EMBoss Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.
  - [Protein](#), [Nucleotide](#)

- **Stretcher** (EMBOSS): EMBoss Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.
  - [Protein](#), [Nucleotide](#)

#### Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

- **Water** (EMBOSS): EMBoss Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.
  - [Protein](#), [Nucleotide](#)

- **Matcher** (EMBOSS): EMBoss Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.
  - [Protein](#), [Nucleotide](#)

- **LALIGN** (EMBOSS): LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.
  - [Protein](#), [Nucleotide](#)

---

http://www.ebi.ac.uk/Tools/psa/
Needle (global)

Identifying and Aligning Homologs (Whitehead Institute)
Identifying and Aligning Homologs (Whitehead Institute)

Needle (global)
Identifying and Aligning Homologs (Whitehead Institute)
Identifying and Aligning Homologs (Whitehead Institute)
Topics to Cover

• Introduction
• Pairwise Alignment methods
• Multiple sequence alignments
• Pre-computed alignment resources
Multiple Sequence Alignment

- Additional sequences can help resolve ambiguities found in a pairwise comparison
- Remove uninformative sequences
- Dynamic programming techniques require prohibitively large computer resources
- Tree or hierarchical methods (successive pairwise alignments), Consistency-based methods, Template-based methods
Multiple Sequence Alignment

http://www.ebi.ac.uk/Tools/msa/

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, Pairwise Sequence Alignment tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

- **Clustal Omega**: New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments (protein only). Suitable for medium-large alignments.
  - Launch Clustal Omega

- **ClustalW2**: Popular MSA tool that uses tree-based progressive alignments. Suitable for medium alignments.
  - Launch ClustalW2

- **DbClustal**: Create a Multiple Sequence Alignment from a protein BLAST result using the DbClustal program.
  - Launch DbClustal

- **Kalign**: Very fast MSA tool that concentrates on local regions. Suitable for large alignments.
  - Launch Kalign

- **MAFFT**: MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.
  - Launch MAFFT

- **MUSCLE**: Accurate MSA tool, especially good with proteins. Suitable for medium alignments.
  - Launch MUSCLE

- **MView**: Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.
  - Launch MView

- **T-Coffee**: Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.
  - Launch T-Coffee

Identifying and Aligning Homologs (Whitehead Institute)
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile techniques to generate alignments.

**Use this tool**

**STEP 1 - Enter your input sequences**

Enter or paste a set of PROTEIN sequence in any supported format.

**STEP 2 - Set your parameters**

**DEALIGN INPUT SEQUENCES**

**OUTPUT ALIGNMENT FORMAT**

**MBED-LIKE CLUSTERING TREE**

**NUMBER OF COMBINED ITERATIONS**

**MAX GUIDE TREE ITERATIONS**

**MAX HMM ITERATIONS**

**STEP 3 - Submit your job**

Be notified by email (Tick box if you want to be notified by email)

Submit
Muscle

Identifying and Aligning Homologs (Whitehead Institute)
T-Coffee

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

Use this tool

STEP 1 - Enter your input sequences
Enter or paste a set of sequences in any supported format:

Or upload a file: Choose File No file chosen

STEP 2 - Set your Parameters

<table>
<thead>
<tr>
<th>MATRIX</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>aligned</td>
</tr>
</tbody>
</table>

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

Identifying and Aligning Homologs (Whitehead Institute)
M-Coffee

Aligns DNA, RNA or Proteins by combining the output of popular aligners

Sequences input
Paste or upload your set of sequences in FASTA format

Sequences to align
>gi|187936925|ref|NP_001120691.1| 1-aminoacyclopropyl
synthase-like protein 1 (Homo sapiens)
MFLPKQDPRAITQCIPGQGLSBRQHELGRKSKKLKDQKLPELQVD
KSPFKEEAEGRTVRQGTVYSGEHGKQGGLINILTSEQKLCFQLLSLLSQRN
LREERKFAKFLSFYCKHPVPLPVPVLQVNLGAGFLSALATCFMLCEAGAFILI
PANYLDSSTLCLPNPVQFLKLYELEKIGVSEKPRYGKQGVLGL
LSIVVISWLNSVKEQDGVSFRLSKLPDPQKRVWSAEKRDPCIGELAS
RYRELILQTVQNAQLLRDKDKQYNLVYPNHKLXAARFTYVSEKHALGII
- OR - Click here to upload a file

Output options
Use this section to control the output format

Alignment format
- score_html
- clustalw_align
- pir_align
- pir_seq
- ggd
- fasta_align
- score_ascii
- msf_align
- phylip

Case
- upper

Residue number
- on

Order
- input

Alignment length
- 80

Your email address

Submit
Reset

Identifying and Aligning Homologs (Whitehead Institute)
M-Coffee

Identifying and Aligning Homologs (Whitehead Institute)

http://tcoffee.crg.cat/
Other Considerations

Web vs Command Line
  – More options to change parameters
  – Process lots of alignments in one command

Which web page to use?
  – Alignment home page vs uniform interface

Alternatives
  – Use subsequences or subset of sequences

Realign by hand

Identifying and Aligning Homologs (Whitehead Institute)
Doing Lots of Alignments

Our favorite method is to use the T-COFFEE suite (more specifically, M-Coffee) to run multiple alignment methods and then create a consensus alignment, a sort of a meta-alignment. This can be done with a single command like

```
t_coffee my_proteins.fa
   -method=t_coffee_msa,mafft_msa,probcons_msa,muscle_msa
   -output=fasta_aln
```

The final consensus alignment will appear in the file `my_proteins.fasta_aln`, which can then be viewed in JalView.

Identifying and Aligning Homologs (Whitehead Institute)
Jalview (http://www.jalview.org)

Identifying and Aligning Homologs (Whitehead Institute)
Topics to Cover

• Introduction
• Pairwise Alignment methods
• Multiple sequence alignments
• Pre-computed alignment resources
  – Homologene (NCBI)
  – Ensembl (EBI)
Ensembl (www.ensembl.org/)
Ensembl

Gene: MSH2 ENSG00000095002

Description: mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) [Source:HGNC Symbol:Acc:7325]


Transcripts: This gene has 13 transcripts

Identifying and Aligning Homologs (Whitehead Institute)
What I hope you learned

• What we can learn from sequence alignments
• Fundamentals of alignments
• Tools for building alignments
References

1. Web links within handout plus
   – Local NCBI - http://tak.wi.mit.edu/blast/
2. SOPs - https://gir.wi.mit.edu/trac/wiki/barc/SOPs
3. What’s on tak - https://tak.wi.mit.edu/trac/wiki/Packages
4. Help pages for each application
   – Click, click, click

Ensembl- NAR 2012, PMID: 22086963

Homologene – NAR 2012, PMID: 22140104

Clustal Omega – Mol Syst Biol, PMID: 21988835

T-Coffee – NAR 2011, PMID: 21558174

Challenges for MSA in hi-thruput era – Bioinformatics 2009, PMID: 19648142

MSA - Curr Opin Struct Biol. 2006, PMID: 16679011

Identifying and Aligning Homologs (Whitehead Institute)