

Sequence Analysis Tools

Erik Arner edit Master subtitle style
Omics Science Center, RIKEN
Yokohama, Japan
arner@gsc.riken.jp

5/26/09

Outline

- Sequence analysis
 - Homology/similarity
- Basics of sequence alignment
 - Global vs. local
 - Computing/scoring alignments
 - Substitution matrices
- Similarity search
 - BLAST
- Multiple alignment
 - ClustalW

5/26/09

Aim of lecture

- Why align sequences?
- How are sequences aligned to each other?
 - Variants
 - Limitations
- Basic understanding of common tools for
 - Similarity search
 - Multiple alignment

5/26/09

Sequence analysis

- Sequence analysis
 - Inferring biological properties through
 - Similarity with other sequences
 - Properties intrinsic to the sequence itself
 - Combination
- Sequence analysis often (always?) includes sequence alignment
- Sequence alignment methods fundamental part of bioinformatics

5/26/09

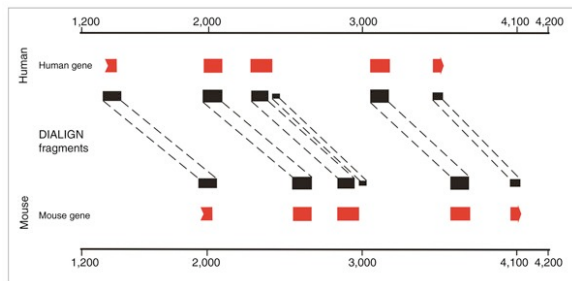
Sequence analysis

Why aligning sequences?

- Similarity in sequence → similarity in function
- Similarity in sequence → common ancestry
 - Homology = similarity due to shared ancestry
- Similar → important
 - Selective pressure

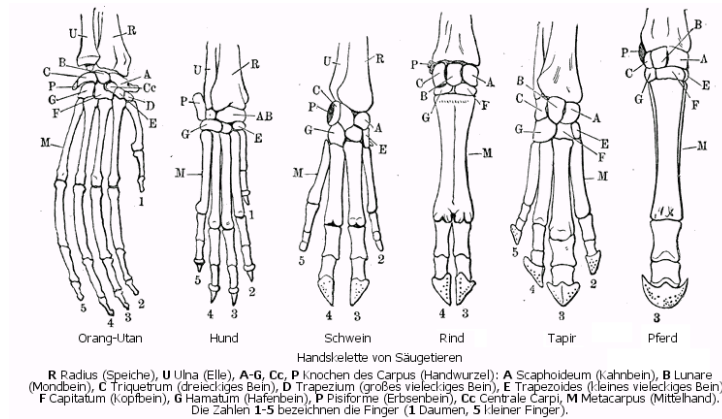
5/26/09

Sequence analysis



5/26/09

Sequence analysis



5/26/09

Sequence analysis

Similarity ≠ homology

- Similarity = factual (% identity)
- Homology = hypothesis supported by evidence

5/26/09

Sequence analysis

- Similarity \neq homology
 - Similarity = factual (% identity)
 - Homology = hypothesis supported by evidence
- ... but in many cases, similarity is the only tool we have accessible
 - Need a measure of the significance of the similarity

5/26/09

Basics of sequence alignment

- Global vs. local alignment
 - Global
 - Assumes sequences are similar across entire length
 - Local
 - Allows locally similar sub-regions to be pinpointed

```
Introns/exons
Global FTFTALILLAVAV
      F--TAL-LLA-AV
Local  FTFTALILL-AVAV
      --FTAL-LLAAV--
```

5/26/09

Basics of sequence alignment

- Which one is correct?

```
Global FTFTALILLAVAV
      F--TAL-LLA-AV
Local  FTFTALILL-AVAV
      --FTAL-LLAAV--
```

5/26/09

Basics of sequence alignment

- Which one is correct?
 - Both?
 - None?
 - In sequence alignment, you get what you ask for

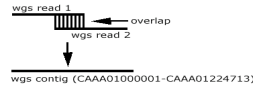
```
Global FTFTALILLAVAV
      F--TAL-LLA-AV
Local  FTFTALILL-AVAV
      --FTAL-LLAAV--
```

5/26/09

Basics of sequence alignment

Other types of alignment

- Glocal
 - Overlaps in shotgun sequencing
- Structural



5/26/09

Basics of sequence alignment

Computing alignments

- Dynamic programming
- Needleman – Wunsch (global alignment)
- Smith – Waterman (local alignment)
- For a given pair of sequences and a scoring scheme, find the optimal alignment
 - Several may exist

5/26/09

Basics of sequence alignment

Scoring alignments

- Simple example
 - Match = +1
 - Mismatch = -1
 - Gap = -1

ATGC **AGTC**

ATGC
AGTC = 0

ATG-C
A-GTC = 1

	A	T	G	C
A	+1	-1	-1	-1
T	-1	+1	-1	-1
G	-1	-1	+1	-1
C	-1	-1	-1	+1

5/26/09

Basics of sequence alignment

Scoring alignments

- Simple example
 - Match = +1
 - Mismatch = -1
 - Gap = -2

ATGC **AGTC**

ATGC
AGTC = 0

ATG-C
A-GTC = -1

	A	T	G	C
A	+1	-1	-1	-1
T	-1	+1	-1	-1
G	-1	-1	+1	-1
C	-1	-1	-1	+1

5/26/09

Basics of sequence alignment

- In sequence alignment, you get EXACTLY what you ask for
 - Heavily penalized gaps → less gaps in alignment
 - Heavily penalized mismatches → more gaps in alignment

5/26/09


Basics of sequence alignment

- Substitution matrices
 - DNA scoring mostly straightforward
 - More clever scoring for protein sequences
 - Biochemical properties
 - Lower penalties for substitutions into amino acids with similar properties
 - Low penalty for isoleucine(I) → valine(V) substitution – both hydrophobic
 - Observed substitution frequencies
 - Multiple alignments of proteins known to share ancestry and/or function

5/26/09

Basics of sequence alignment

- Common substitution matrices
 - PAM
 - BLOSUM
- BLOSUM62 most widely used
 - Default in BLAST
- What matrix to use?

BLOSUM 45	BLOSUM 62	BLOSUM 90
PAM 250	PAM 160	PAM 100
<i>More Divergent</i>		<i>Less Divergent</i>

5/26/09

Basics of sequence alignment

- Gap penalties
 - Gaps generally considered to cause greater disruption of function than mismatches
 - Gap open penalty
 - Gap extension penalty

5/26/09

Similarity search

· Premise:

- The sequence itself is not informative; it must be analyzed by comparative methods against existing databases to develop hypothesis concerning relatives and function.
- Abundance of biological sequence data forbids extensive searches
 - All nucleotides/amino acids in query sequence cannot be compared to all aa:s/nt:s in database
 - Fast searches are achieved using methods that trade off sensitivity for speed and specificity

5/26/09

Similarity search

· General approach:

- A set of algorithms (e.g. BLAST) are used to compare a query sequence to all the sequences in a specified database
- Comparisons are made in a pairwise fashion
- Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared
 - The higher the score, the greater the degree of similarity
- Alignments can be global or local (BLAST: local)
 - Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance
 - Similarity, by itself, cannot be considered a sufficient indicator of function

5/26/09

Similarity search – BLAST

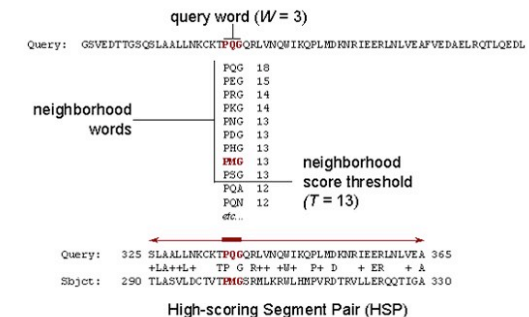
· BLAST

- A set of sequence comparison algorithms introduced in 1990
- Breaks the query and database sequences into fragments ("words"), initially seeks matches between fragments
- Initial search is done for a word of length "W" that scores at least "T" when compared to the query
 - using a given substitution matrix
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S"
- "W" parameter dictates the speed and sensitivity of the search

5/26/09

Similarity search – BLAST

The BLAST Search Algorithm



5/26/09

Similarity search – BLAST

- Word size
 - Short word size → many hits in database → many candidates to evaluate
 - Slow
 - Sensitive
 - False positives
 - Long word size → few hits in database → few candidates
 - Fast
 - Specific
 - False negatives

5/26/09

Similarity search – BLAST

- Scoring
 - Unitary matrix used for DNA
 - Only identical nucleotides give positive score
 - Substitution matrices are used for amino acid alignments
 - BLOSUM62 is default
 - Non-identical amino acids may give positive score
- Gaps
 - Gap scores are negative
 - The presence of a gap is ascribed more significance than the length of the gap
 - A single mutational event may cause the insertion or deletion of more than one residue
 - Initial gap is penalized heavily, whereas a lesser penalty is assigned to each subsequent residue in the gap
 - No widely accepted theory for selecting gap costs

It is rarely necessary to change gap values from the default

5/26/09

Similarity search – BLAST

- Significance of hits
 - P value
 - Given the database size, the probability of an alignment occurring with the same score or better
 - Highly significant P values close to 0
 - Expectation value – E value
 - The number of different alignments with equivalent or better scores that are expected to occur in a database search by chance
 - The lower the E value, the more significant the score
- Human judgment

5/26/09

Similarity search – BLAST

- BLAST at NCBI
 - <http://blast.ncbi.nlm.nih.gov>

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- | | | |
|---|--|--|
| <input type="checkbox"/> Human | <input type="checkbox"/> Oryza sativa | <input type="checkbox"/> Gallus gallus |
| <input type="checkbox"/> Mouse | <input type="checkbox"/> Bos taurus | <input type="checkbox"/> Pan troglodytes |
| <input type="checkbox"/> Rat | <input type="checkbox"/> Danio rerio | <input type="checkbox"/> Microbes |
| <input type="checkbox"/> Arabidopsis thaliana | <input type="checkbox"/> Drosophila melanogaster | <input type="checkbox"/> Apis mellifera |

5/26/09

Similarity search – BLAST

- BLAST at NCBI
 - <http://blast.ncbi.nlm.nih.gov>

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

5/26/09

Similarity search – BLAST

- BLAST at NCBI
 - <http://blast.ncbi.nlm.nih.gov>

Specialized BLAST

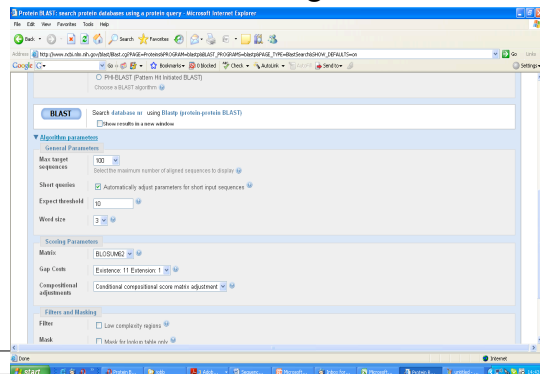
Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay

5/26/09

Similarity search – BLAST

- BLAST at NCBI
 - <http://blast.ncbi.nlm.nih.gov>



5/26/09

BLAT

- BLAST-Like Alignment Tool
 - Faster than BLAST
 - Trades off sensitivity for specificity and speed
 - Optimized for aligning spliced mRNAs to large genome sequences

5/26/09

Multiple alignment

- Why align multiple sequences?
 - Determine evolutionary relationship between sequences → species
 - Phylogenetics
 - Identify domains
 - Pinpoint functional elements
 - Highly conserved amino acids among more divergent ones → catalytic activity?

5/26/09

Multiple alignment

- Multiple alignment algorithms
 - Finding optimal alignment is very time consuming
 - Exponential complexity
 - Approximations and heuristics used for speeding up
 - Heuristics: "rules of thumb", educated guesses, intuitive judgments or simply *common_sense* (from Wikipedia)
 - Progressive alignment

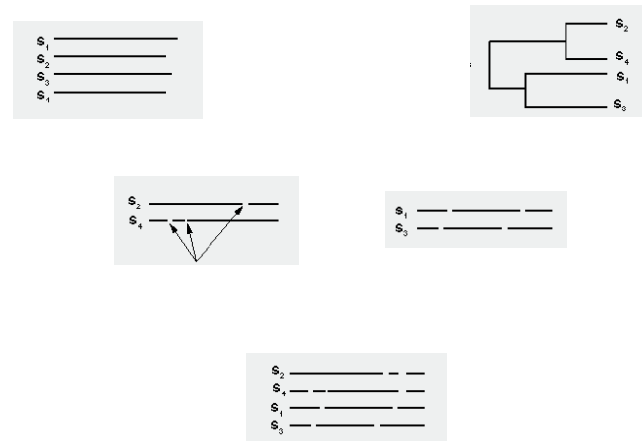
5/26/09

Multiple alignment – ClustalW

- Basics of progressive algorithm
 - All sequences are compared to each other pairwise
 - A guide tree is constructed, where sequences are grouped according to pairwise similarity
 - The multiple alignment is iteratively computed, using the guide tree

5/26/09

Multiple alignment – ClustalW



5/26/09

Multiple alignment – ClustalW

- Heuristics
 - Individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones
 - Amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned
 - Residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure
 - Positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions

5/26/09

Summary

- Know your parameters
 - Defaults are good choices in most cases
 - However, be aware of what they mean
 - You get what you ask for

5/26/09