Master Course
Algorithms in Sequence Alignment
Lecture 9
Database searching (2)

What is the statistical significance of an alignment (cnt.)
- extracting local alignments from random sequences:
  - Can scramble all sequences (preserving residue composition) in database to get a statistical background
  - Can do this for each sequence (e.g. make 100-1000 randomisations), so calculating alignment score significance against ‘self’.

What is the statistical significance of an alignment

- Score should optimise the chance to select proper hits (True Positives)
- Scoring alignments is dependent on
  - The scoring system used (residue exchange matrix and gap penalty regime)
  - Characteristics of the sequence database (size, residue composition)
- The BLAST way of scoring has been adopted by other methods as well; e.g., some implementations of FASTA, etc.
  - Bit-score
  - E-value

Scoring BLAST alignments

Divergent evolution allows function prediction

Structure (and function) more conserved than sequence

Z = (x – Mean) / SD

What is the statistical significance of an alignment

- To get a null model: extract local alignments from random sequences
- Use alignment method to make alignments of random sequences and compare results
- Using multiple randomisations:
  - Z = (x – Mean) / SD
  - In practice: Z should be >4 to >6 SD (significance threshold)
### Alignment Bit Score
\[ B = (\lambda S - \ln K) / \ln 2 \]
- \( S \) is the raw alignment score
- The bit score (‘bits’) \( B \) has a standard set of units
- The bit score \( B \) is calculated from the number of gaps and substitutions associated with each aligned sequence. The higher the score, the more significant the alignment
- \( \lambda \) and \( K \) are statistical parameters associated with a given scoring system (e.g. BLOSUM62 in BLAST)
  - See Altschul and Gish (1996) for a collection of values for \( \lambda \) and \( K \) over a set of widely used scoring matrices.
- Because bit scores are normalized with respect to the scoring system, they can be used to compare alignment scores from different searches based on different scoring schemes (a.a. exchange matrices)

### What is the statistical significance of an alignment
- Using a null model based on local alignments from random sequences
- **P-value**
  - The probability of obtaining the result by pure chance
  - An alignment giving a lower P-value than a threshold value set by the user is considered a hit.

### Normalised sequence similarity
- The **p-value** is defined as the probability of seeing at least one unrelated score \( S \) greater than or equal to a given score \( x \) in a database search over \( n \) sequences.
  - This probability follows the Poisson distribution (Waterman and Vingron, 1994):
    \[ P(x, n) = 1 - e^{-nP(S \geq x)} \]
  - where \( n \) is the number of sequences in the database
  - Depending on \( x \) and \( n \) (fixed)

### E-value
- The concept of P-value applies to single comparisons
- **What about searching in a large database?**

**Task:**
- Having a protein, we want to find similar ones in a large database (>4 million sequences in NCBI’s NR database - used in BLAST).
- We are interested in P-value < 0.01
- Count the number of hits we’ll get by chance alone.

### Normalised sequence similarity

**Statistical significance**
- The **E-value** is defined as the expected number of non-homologous (random) sequences with score greater than or equal to a score \( x \) in a database of \( n \) sequences:
  \[ E(x, n) = nP(S \geq x) \]
- For example, if E-value = 0.01, then the expected number of random hits with score \( S \geq x \) is 0.01, which means that this E-value is expected by chance only once in 100 independent searches over the database.
- If the E-value of a hit is 5, then five fortuitous hits with \( S \geq x \) are expected within a single database search, which renders the hit not significant.

### A model for database searching score probabilities
- Scores resulting from searching with a query sequence against a database follow the Extreme Value Distribution (EVD) (Gumbel, 1955).
- Using the EVD, the raw alignment scores are converted to a statistical score (E value) that keeps track of the database amino acid composition and the scoring scheme (a.a. exchange matrix)
Database searching is commonly Scrambling sequences allows Z
value x, the more significant the score 
approximation 1
In practice, the probability P
For the raw alignment score S becomes

P(S ≥ x) = 1 − exp(−Kmn e−x)

Using the equation for μ (preceding slide), the probability for the raw alignment score S becomes

P(S ≥ x) = 1 − exp(−Kmn e−x)

In practice, the probability P(S ≥ x) is estimated using the approximation 1 − exp(−e−x) = e−x, which is valid for large values of x. This leads to a simplification of the equation for P(S ≥ x):

P(S ≥ x) = e−Kmn e−x

The lower the probability (E value) for a given threshold value x, the more significant the score S.

Normalised sequence similarity Statistical significance
• Database searching is commonly performed using an E-value in between 0.1 and 0.001.
• Lower E-values decrease the number of false positives in a database search, but increase the number of false negatives, thereby lowering the sensitivity of the search (see later slides).

Approximating statistical significance
• Scrambling sequences allows Z-score calculations that are slow but independent of the database size and composition
• E-value calculations based upon the EVD are much faster but do depend upon the size of the database: an E-value score for a given query and DB sequence can change upon a next release of the sequence database.

Extreme Value Distribution

Extreme Value Distribution (EVD)

Not a normal (Gaussian) distribution

You know that an optimal alignment of two sequences is selected out of many suboptimal alignments, and that a database search is also about selecting the best alignment(s) out of many database sequences. This double selection bodes well with the EVD which has a right tail that falls off more slowly than the left tail. Compared to using the normal distribution, when using the EVD an alignment has to score further away from the expected mean value to become a significant hit.

The probability of an unrelated score S to be larger than a given value x can be calculated following the EVD as:

\[ E-value: P(S ≥ x) = 1 − \exp(-e^{-\frac{x}{\lambda}}), \]

where \( \mu = \frac{\ln(Kmn)}{\lambda} \), with m and n query and database sequence length and K a constant that can be estimated from the background amino acid distribution and scoring matrix (see Altschul and Gish, 1996 (Methods in Enzymology), for a collection of values for \( \lambda \) and K over a set of widely used scoring matrices).

Extreme Value Distribution

Sequence Analysis

Statistical significance

Statistical significance

Statistical significance
Pair-wise alignment quality versus sequence identity

- Vogt et al., JMB 249, 816-831, 1995

Alignment quality quickly deteriorates with more distant sequences:

What does this mean for alignment?

- Alignments need to be able to match distantly related sequences, skip secondary structural elements to complete domains (i.e. putting gaps opposite these motifs in the shorter sequence).
- Depending on the residue exchange matrix and gap penalties chosen, the algorithm might have difficulty with aligning distant homologs or inserting long gaps (for example when using high affine gap penalties), resulting in incorrect alignment.

What does this mean for homology searching?

- Database searching algorithms just need to decide if the alignment score is good enough for inferring homology.
- Sometimes, alignments can be incorrect but the score can be close enough for the database searching method to correctly identify the DB sequence as a homolog (or not).
- However, for distant hits alignments become crucial.

Sequence Analysis/Database Searching

Finding relationships between genes and gene products of different species, including those at large evolutionary distances

Some Known Genomes

Eukaryotes
- Human, Rat, Mouse
- Anopheles gambliae, Drosophila melanogaster, Arabidopsis thaliana, Oryza sativa, Solanum tuberosum

Parasites
- Brugia malayi, Entamoeba histolytica, Plasmodium falciparum, Plasmodium yoelli, Toxoplasma gondii, Trypanosoma brucei, Trypanosoma cruzi, Schistosoma mansoni

Bacteria
- Chlamydia pneumoniae, Haemophilus influenzae, Helicobacter pylori, Mycobacterium tuberculosis, Mycoplasma genitalium, Vibrio cholerae

Fungal Projects
- Cryptococcus neoformans, Aspergillus fumigatus

(http://www.tigr.org/tdb)
(http://www.ebi.ac.uk/

Why are genome sequences so important?

The cellular machinery is extremely complex.

The transformation of genomic information from the cell to a text sequence is a reduction in complexity.

This formalisation makes genomic information accessible.

Compare to other formalisations:
- Carl Linnaeus (1707-1778): systematic classification
- Charles Darwin (1809-1882): Evolution
- Alan Turing (1912-1954): Turing machine
Reductionism

"Ceci n’est pas une pipe" (René Magritte).

This is not a protein.

Compared to the preceding plot, RMSD is better able to pin-point relationships between more divergent sequences (RMSD stays relatively small for a longer time as compared to PAM distance) – Structure more conserved than sequence. Note that the spread around RMSD is larger.

Conclusion: Sequence-based tools use a reduced amount of information.

Structural superpositioning

What does Match mean?

Root mean-square deviation (RMSD): how far are equivalenced Cα atoms separated on average after two structures are superposed?

Two superposed protein structures with two well-superposed helices

Red: well superposed
Blue: low match quality

C5 anaphylatoxin – human (PDB code 1kjs) and pig (1c5a) proteins are superposed

Sequence Searching and Sequence Alignment

- Sequence searching and sequence alignment are different techniques
- Sequence searching uses sequence alignment, but sequence alignment works on a pre-defined sequence set
- Sequence searching tries to extract homologous sequences, sequence alignment tries to correctly identify similarities between homologues
How to assess homology search methods
- We need an annotated database, so we know which sequences belong to what homologous (super)families
- Examples of databases of homologous families are PFAM, Homstrad or Astral
- The idea is to take a protein sequence from a given homologous family, then run the search method, and then assess how well the method has carried out the search (i.e. recognised the family members)
- This should be repeated for many query sequences and then the overall performance can be measured

So what have we got

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>TP</td>
</tr>
<tr>
<td>N</td>
<td>FN</td>
</tr>
</tbody>
</table>

Sensitivity and Specificity

- Positive Predictive Values
  - TP/(TP+FP) = 99.9% (9999/9999)
  - FN/(FN+TN) = 99.9% (9999/9999)

- Negative Predictive Values
  - FP/(FP+TN) = 99.9% (9999/9999)
  - TN/(TN+FN) = 99.9% (9999/9999)

Receiver Operator Curve (ROC)
- Plot Sensitivity (TP/(TP+FN)) against 1-Specificity (1 - TN/(FP+TN)), where the latter is called error

Sensitivity is also called Coverage

Receiver Operator Curve (ROC)
Database Search Algorithms: Sensitivity, Selectivity

- Sensitivity — the ability to detect weak similarities between sequences (often due to long evolutionary separation). Increasing sensitivity reduces false negatives, i.e., those database sequences similar to the query, but rejected.
  Sensitivity (or Coverage) = TP / (TP + FN)

- Selectivity — the ability to screen out similarities due to chance. Increasing selectivity reduces false positives, those sequences recognized as similar when they are not.
  Selectivity (or Positive Prediction Value) = TP / (TP + FP)

- Specificity also describes the ability of the method to select proper hits
  Specificity = TN / (TN + FP)

The Framework

Databank searching can be split into three phases:

1. Matching phase
   The query sequence is compared by (partial) alignment with the databank sequence. Most programs pre-select potential hits at this level. The comparison is usually heuristic and fast (see previous lecture).

2. Scoring phase
   If the database sequence passed the matching phase, the query-hit sequence pair is re-aligned and scored, mostly using a pairwise scoring matrix. (see previous lecture)

3. Selection phase
   Based on a statistical criterion, sequences with a score above a user-defined score cutoff are retained as hits. The hitlist is the result of the databank searching and each hit is a potential homologue.

Methods for Sequence Searching

- Heuristic Alignment
  Matching of short identical words and match extension (Blast, Fasta)

- Dynamic Programming
  Parallelised searching on distributed database (Blascelator)

- Suffix Tree
  Identity matching of substrings (Mummer)

- Pattern Matching
  Matching of conserved motifs using HMMs or regular expressions (HMMer, PROSITE)
Iterative databank searching

Homologously related proteins are classified into families. If the homology can only be derived from similarities in structure or function, they are classified into superfamilies. The ideal sequence search program would find all family members and all super-family members.

In reality many homologous sequences are missed, because the sequence similarity has been lowered by evolution beyond recognition.

The idea of iterative sequence searching is to use the hit list of an initial search as query for a consecutive search. This can be repeated (iterated) until no more sequences are added to the hit-list.

How do we search a database with a list of query sequences?

Extending homology searching using COG - Cluster of Orthologous Groups

Tatusov et al., 1997

• Orthologues found using bi-directional best hit searching with PSI-BLAST

• All COG family members are supposed to have the same function

• Searching with an unknown sequence only needs to hit a single member of a COG family, annotation can then be transferred


Finding orthologs using bi-directional best hit strategy

Database A

Sequence a

Database B

Sequence b

‘Blast’ sequence a against database B and then blast the top-hit (say sequence b) against database A: if now sequence a is the top-hit, this is a strong indication that sequences a and b are orthologous. This strategy is often used as an operational definition of orthology

Structure-based function prediction

• SCOP (http://scop.berkeley.edu/) is a protein structure classification database where proteins are grouped into a hierarchy of families, superfamilies, folds and classes, based on their structural and functional similarities

Structure-based function prediction

• SCOP hierarchy – the top level: 11 classes

Structure-based function prediction

• All-alpha protein, membrane protein, Alpha-beta protein

Structure-based function prediction

• All-beta protein, Coiled-coil
Structure-based function prediction

SCOP hierarchy – the second level: 800 folds

SCOP hierarchy – third level: 1294 superfamilies

SCOP hierarchy – third level: 2327 families

Using sequence-structure alignment method, one can predict a protein belongs to a
- SCOP family, superfamily or fold
- Proteins predicted to be in the same SCOP family are orthologous
- Proteins predicted to be in the same SCOP superfamily are homologous
- Proteins predicted to be in the same SCOP fold are structurally analogous

Some Tricky Problems
Repeats
Multi-domain proteins
Low-complexity regions
Reduncancy
Very short queries
Very distant sequences
Un-annotated sequences → Conserved hypotheticals
Profile wander

Protein structural domain organisation
Multi-domain proteins

Pyruvate kinase
Phosphotransferase
- β barrel regulatory domain
- α/β barrel catalytic substrate binding domain
- α/β nucleotide binding domain
1 continuous + 2 discontinuous domains
Multi-domain Proteins

It can be disadvantageous to search with multi-domain proteins. If a multi-domain protein contains domains A and B, and C is a common domain, many other (multi-domain) proteins containing this domain will be matched. The interesting domain could be A, but the majority of reported hit matches B.

In iterative sequence searching, the multi-domain proteins BC that were detected in the first round will produce hits to other proteins containing CD. The search may drift from AB to CD.

Multi-domain Proteins (cont.)

- A common conserved protein domain such as the tyrosine kinase domain can obscure weak but relevant matches to other domain types (e.g., appearing after 5000 kinase hits).
- Sequences containing low-complexity regions, such as coiled coils and transmembrane regions, can cause an explosion of the search rather than convergence because of the absence of any strong sequence signals.
- Conversely, some searches may lead to premature convergence; this occurs when the PSSM is too strict only allowing matches to very similar proteins, i.e., sequences with the same domain organization as the query are detected but no homologues with different domain combinations.

Detecting Low-Complexity

- SEG and PSEG/NSEG algorithms
  - Wootten and Federhen
  - SEG
  - UNIX Executable available on ncbi servers
    - Command:
      ```
      seg FASTAfile Window Trigger Complexity Extension
      Kf(1) Kf(2)
      ```
    - Longer Window lengths define more sustained regions, but overlook short biased subsequences.

Calculating low-complexity regions in the SEG program

- A sequence of L residues of N types can have \( L! / \left( \sum_{i=0}^{N-1} n_i \right) \) different sequences of that same composition, where the composition vector \( \mathbf{n} = (n_0, n_1, ..., n_{N-1}) \) and \( n_i \) is the number of residues of type \( i \).
  - If \( \mathbf{R} = (r_0, r_1, ..., r_{N-1}) \) is a vector of length \( N \), where the vector numbers correspond to the number of residues with a given frequency (e.g., there are 5 amino acid types with 0 abundance, 3 amino acid types with abundances 1, 2, 3 in the sequence), then the total number of distinct sequences corresponding to a particular complexity state-vector is \( (L! / \Pi_{i} n_i) * (N! / \Pi_{i} r_i) \), where \( \Pi_{i} r_i = r_0! * r_1! * ... * r_{N-1}! \).
  - Based on this, the final complexity score calculated by the SEG program is
    \[
    P_{\text{SEG}} = \frac{1}{N!} * \left( \frac{L!}{\Pi_{i} n_i} \right) * \left( \frac{N!}{\Pi_{i} r_i} \right)
    \]

Low-complexity Regions

Some genome sequences contain low-complexity regions. These can give false-positive hits.

Example:

````
// in SEG program
SEGNNN
| - low-complexity region |
```

Most sequence searching programs use filters to recognize and skip such low-complexity regions. If such regions are by chance included in the hit, the output looks like

````
// in SEG program
SEGNNN
| - low-complexity region |
```

Redundancy

Some databases, like the non-redundant sequence database, contain large number of nearly identical sequences. A typical example is the fusion peptide of hemagglutinin, a protein of the flu virus. This relatively short sequence returns more than 5000 hits to nearly identical (point mutants) sequence.

If one is interested in distant homologues, these redundant hits obscure the results.

One solution is to filter out all similar hits.

Can you think of any filter mechanisms?
Conserved hypotheticals

A substantial fraction of genes in sequenced genomes encode ‘conserved hypothetical’ proteins, i.e. those that are found in organisms from several phylogenetic lineages but that have not been functionally characterized.

Profile wander (or matrix migration)

- Permissive iterative searching using high E-values as a threshold can lead to incorrect hits (false positives) that become included into the profile during search iteration. More incorrect hits can then be added in subsequent iterations, and true homologues can be lost. Also, the search can explode, leading to large numbers of spurious hits.
- A further loss of information can be incurred with PSI-BLAST, because PSI-BLAST PSSMs are trimmed to only use the highest scoring region in a search, ignoring less conserved regions.
- On the other hand, searching with overly strict (low) E-values leads to premature convergence of the search (no further addition of hits to the profile).

Result Processing

The following steps to process result lists are useful:
- Match identification numbers and key words
- Grep sequences out of the database
- Compare results of different searches (on different databases)
- Filter sequences on specified criteria
- Align results using a multiple alignment
- Recognize family patterns or generate a family profile
- Generate a phylogenetic tree and classify hit sequences

Combination of Methods

- Sequence search -> multiple alignment -> phylogenetic tree
- Sequence search -> sequence alignment -> homology modelling
- Sequence search -> sequence alignment -> mutation analysis
- Sequence search -> multiple alignment -> functional annotation
Sequence identity scoring zones

• 25-30% or more: putative homology zone (‘safe zone’)
• 15-25%: twilight zone - many cases of homology but hard to detect using sequence analysis methods
• <15%: midnight zone (Rost, 1999) – not detectable using sequence information

Is midnight zone properly definable?

Recap

• Homology principle
• Alignment score statistical significance
  • Z-scores over scrambled sequences
  • BLAST statistical scheme
    • Extreme value distribution
• Various sequence searching methods
• Low complexity filtering
• Various homology searching pitfalls
  • repeats, multi-domain sequences, database redundancy, short query sequences, distant relatives, profile wander, premature convergence

APPENDIX

Background knowledge:

Sequence notational formalisms and structural features

Sequence notational - Format Conventions

A sequence is composed of a name (often including an accession number) and the residue string. A sequence database is a formatted (and often sorted) list of sequences (here FASTA format).

charge

Proteins are written from the N-terminus to the C-terminus:

('+' ) REN−CH (R1) −CO−NH−CH (R2) − CO− (n) −NH−CH (Ren 2) − COO− ('−')

Nucleotide sequences are defined within a reading frame, because an amino acid is defined by a nucleotide triplet. The notation is from 5’ to 3’:

('−') GPD (O'−')(OH1) − GPD (O'−')(OH2) − (n) − GPD (O'−')(OHn2)
### Sequence Notation - Name Conventions

The bond between two nucleotides is called a 'phosphodiester bond', and the molecule is called a 'nucleotide'.

The bond between two amino acids is called a 'peptide bond', which is chemically an amide bond, and the molecule is called a 'peptide'.

- 2- dipeptide, e.g. GA or Gly-ALA or glycyl-alanine
- 3- tripeptide, e.g. VGA or Tyr-Gly-ALA or tyrosyl-glycyl-alanine
- 4- tetrapeptide
- 5- pentapeptide
- 6 - oligopeptide
- >50 - polypeptide, protein

A protein adopts a folded structure with a hydrophobic core.

### Sequence Notation - Positions and Chains

**Mutation**

- Y30G (human pancreatic trypsin inhibitor) mutation from Tyr to Gly at position 35
- K32R (insulin) mutation from Lys to Pro at position 26 in chain B
- Der(11)I-Insulin-806-carboxamide residues 27 to 30 deleted in chain B and C-terminus amidated

**Chain notation**

- Chains are denoted A, B, C, D in successive order.

**Disulfide bridges**

- Oxidation of proximal Cys-Cys pairs leads to formation of a covalent Cys-Cys disulfide bond (cysteine bridge)

### Protein structure hierarchical levels

- **Primary structure**: amino acid sequence
- **Secondary structure**: helices, strands
- **Tertiary structure**: fold
- **Quaternary structure**: association of several chains

### Hemoglobin

- **Globin fold**
- PDB: 1MBN
- Helices are labelled 'A' (blue) to 'H' (red). D helix can be missing in some globins: what happens with the alignment?
β sandwich  
β protein  
immunoglobulin  
PDB: 7FAB

TIM barrel  
α / β protein  
Triose phosphate  
IsoMerase  
PDB: 1TIM