Bio2

Heuristics, Databases ;
Multiple Sequence Alignment ;
Gene Finding

Heuristic Methods

- FASTA
- BLAST
- Gapped BLAST
- PSI-BLAST

BLAST

Altschul, Gish, Miller, Myers and Lipman (1990) Basic local alignment search tool. J Mol Biol 215:403-410

- Developed on the ideas of FASTA
- Integrates the substitution matrix in the first stage of finding the *hot spots*
- Faster *hot spot* finding

BLAST definitions

- Given two strings S_1 and S_2
- A *segment pair* is a pair of equal lengths substrings of S_1 and S_2 aligned without gaps
- A *locally maximal segment* is a segment whose alignment score (without gaps) cannot be improved by extending or shortening it.
- A *maximum segment pair* (*MSP*) in S_1 and S_2 is a segment pair with the maximum score over all segment pairs.

BLAST Process

- Parameters:
  - *w*: word length (substrings)
  - *t*: threshold for selecting interesting alignment scores

---

BLAST Process

- 1. Find all the *w*-length substrings from the database with an alignment score >*t*
  - Each of these (similar to a hot spot in FASTA) is called a *hit*
  - Does not have to be identical
  - Scored using substitution matrix and score compared to the threshold *t* (which determines number found)
  - Words size can therefore be longer without losing sensitivity: AA - 3-7 and DNA ~12

BLAST Process

- 2. Extend hits:
  - extend each hit to a local maximal segment
  - extension of initial *w* size hit may increase or decrease the score
  - terminate extension when a threshold is exceeded
  - find the best ones (HSP)
- This first version of Blast did not allow gaps….

(Improved) BLAST

Altshul, Madden, Schaffer, Zhang, Zhang, Miller & Lipman (1997) Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402

- Improved algorithms allowing gaps
  - these have superceded the older version of BLAST
  - two versions: Gapped and PSI BLAST

(Improved) BLAST Process

- Find words or hot-spots
  - search each diagonal for two *w* length words such that score >=*t*
  - future expansion is restricted to just these initial words
  - we reduce the threshold *t* to allow more initial words to progress to the next stage

(Improved) BLAST Process

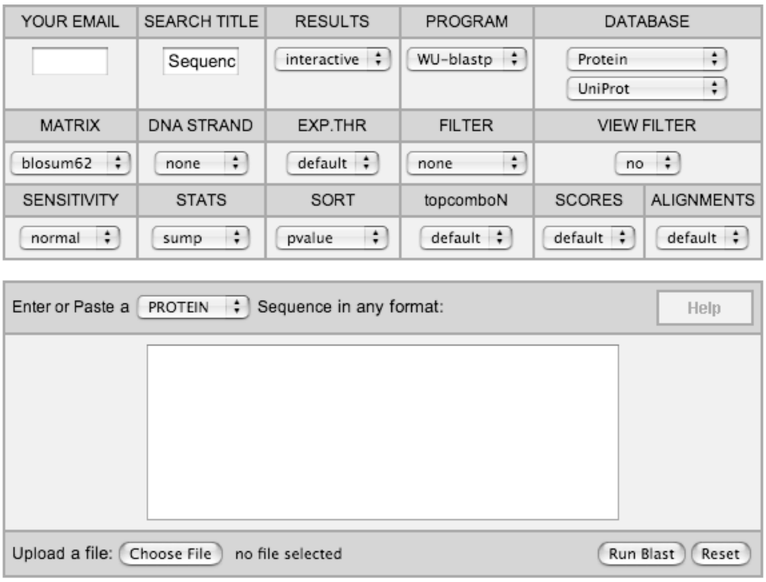
- Allow local alignments with gaps
  - allow the words to merge by introducing gaps
  - each new alignment is comprises two words with a number of gaps
  - unlike FASTA does not restrict the search to a narrow band
  - as only two word hits are expanded this makes the new blast about 3x faster

PSI-BLAST

- Iterative version of BLAST for searching for protein domains
  - Uses a dynamic substitution matrix
  - Start with a normal blast
  - Take the results and use these to ‘tweak’ the matrix
  - Re-run the blast search until no new matches occur
- Good for finding distantly related sequences but high frequency of false-positive hits

BLAST Programs

• blastp	compares an amino acid query sequence against a protein sequence database.
• blastn	compares a nucleotide query sequence against a nucleotide sequence database.
• blastx	compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
• tblastn	compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
• tblastx	compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. (SLOW)



YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASE	
	Sequenc	interactive	WU-blastp	Protein UniProt	
MATRIX	DNA STRAND	EXP.THR	FILTER	VIEW FILTER	
blosum62	none	default	none	no	
SENSITIVITY	STATS	SORT	topcomboN	SCORES	ALIGNMENTS
normal	sump	pvalue	default	default	default

Enter or Paste a PROTEIN Sequence in any format:

Go try them out!

- Links to NCBI and EBI are on the course web site
- Some test sequences will be posted on the course web site

Alignment Heuristics

- Dynamic Programming is better but too slow
- FASTA and BLAST based on several assumptions about good alignments
  - substitutions more likely than gaps
  - good alignments have runs of identical matches
- FASTA good for DNA sequences but slower
- BLAST better for amino acid sequences and pretty good for DNA, fastest.

Biological Databases (sequences)

Biological Databases

- Introduction to Sequence Databases
- Overview of primary query tools and the databases they use (e.g. databases used by BLAST and FASTA)
- Demonstration of common queries
- Interpreting the results
- Overview of annotated ‘meta’ or ‘curated’ databases

DNA Sequence Databases

- Raw DNA (and RNA) sequence
- Submitted by Authors
- Patent, EST, Gemomic sequences
- Large degree of redundancy
- Little annotation
- Annotation and Sequence errors!

Main DNA DBs

• Genbank	US
• EMBL	EU
• DDBJ	Japan
• Celera genomics	Commercial DB

EMBL

- Sources for sequence include:
  - Direct submission - on-line submission tools
  - Genome sequencing projects
  - Scientific Literature - DB curators and editorial imposed submission
  - Patent applications
  - Other Genomic Databases, esp Genbank

International Nucleotide Sequence Database Collaboration

- Partners are EMBL, Genbank & DDBJ
- Each collects sequence from a variety of sources
- New additions to any of the three databases are shared to the others on a daily basis.

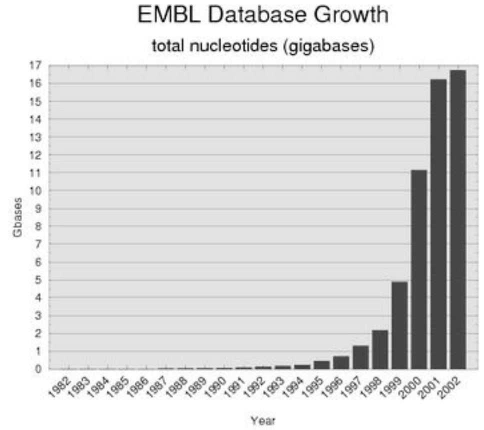
Limited annotation

- Unique accession number
- Submitting author(s)
- Brief annotation if available
- Source (cDNA, EST, genomic etc)
- Species
- Reference or Patent details

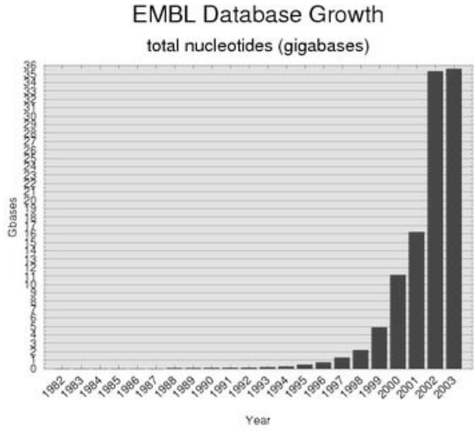
EMBL file tags

```
ID - identification             (begins each entry; 1 per entry)
AC - accession number           (>=1 per entry)
SV - new sequence identifier    (>=1 per entry)
DT - date                       (2 per entry)
DE - description                (>=1 per entry)
KW - keyword                    (>=1 per entry)
OS - organism species           (>=1 per entry)
OC - organism classification    (>=1 per entry)
OG - organelle                  (0 or 1 per entry)
RN - reference number           (>=1 per entry)
RC - reference comment          (>=0 per entry)
RP - reference positions        (>=1 per entry)
RX - reference cross-reference  (>=0 per entry)
RA - reference author(s)        (>=1 per entry)
RT - reference title            (>=1 per entry)
RL - reference location         (>=1 per entry)
DR - database cross-reference   (>=0 per entry)
FH - feature table header       (0 or 2 per entry)
FT - feature table data         (>=0 per entry)
CC - comments or notes          (>=0 per entry)
XX - spacer line                (many per entry)
SQ - sequence header            (1 per entry)
bb - (blanks) sequence data     (>=1 per entry)
// - termination line           (ends each entry; 1 per entry)
```

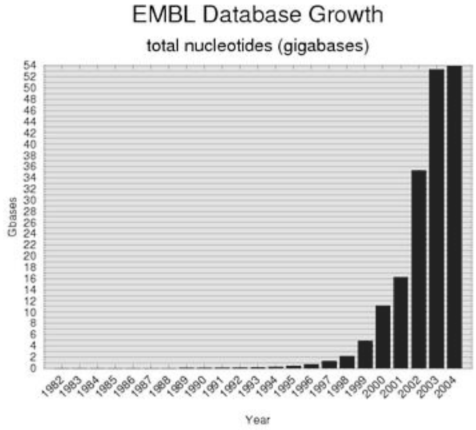
16,759,535,577 bases (27/1/02)



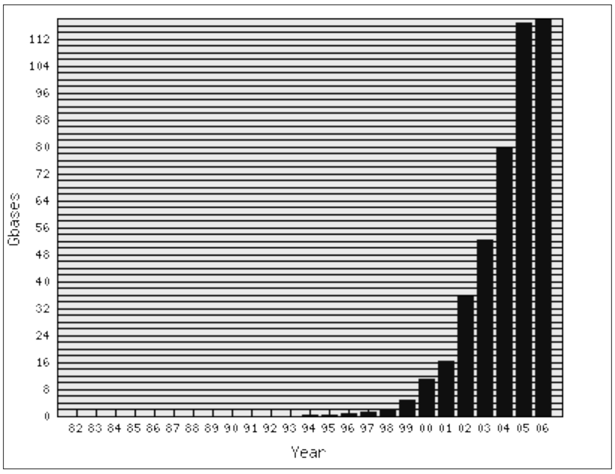
35,602,556,374 bases (17/1/03)



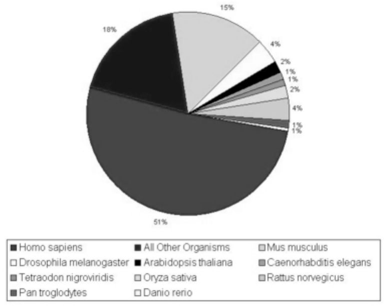
53,958,991,118 bases (24/1/04)



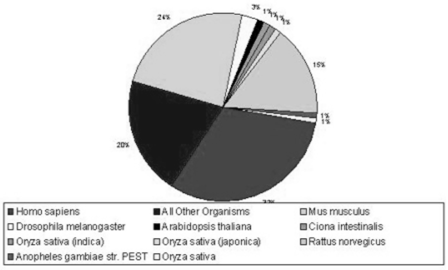
Jan ‘06 117,599,582,673bp



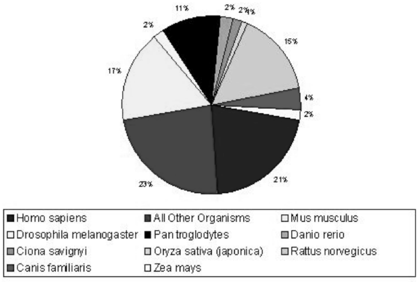
Bases by organism 02



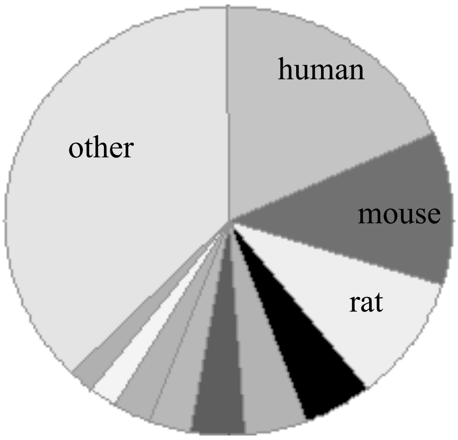
Bases by organism 03



Bases by organism 04



Bases by organism 06



http://www3.ebi.ac.uk/Services/DBStats/

17 Subdivisions

ESTs	EST
Bacteriophage	PHG
Fungi	FUN
Genome survey	GSS
High Throughput cDNA	HTC
High Throughput Genome	HTG
Human	HUM
Invertebrates	INV
Mus musculus	MUS
Organelles	ORG
Other Mammals	MAM
Other Vertebrates	VRT
Plants	PLN
Prokaryotes	PRO
Rodents	ROD
STSs	STS
Synthetic	SYN
Unclassified	UNC
Viruses	VRL

ESTs

- Expressed Sequence Tags
  - short mRNA samples from tissues
  - cloned and sequenced
  - single read
  - approx 1/3 of the database

HTG

- High throughput genomic sequences
  - Partial sequences obtained during genome sequencing.
  - Around 1/3 of the database

Specialist DNA Databases

- Usually focus on a single organism or small related group
- Much higher degree of annotation
- Linked more extensively to accessory data
  - – Species specific:
    - Drosophila: FlyBase,
    - C. elegans: AceDB
  - – Other examples include Mitochondrial DNA, Parasite Genome DB

FlyBase

flybase.bio.indiana.edu

- Includes the entire annotated genome searchable by BLAST or by text queries
- Also includes a detailed ontology or standard nomenclature for *Drosophila*
- Also provides information on all literature, researchers, mutations, genetic stocks and technical resources.
- Full mirror at EBI

Protein DBs

- Primary Sequence DBs
  - Swiss-Prot, TrEMBL, GenPept
- Protein Structure DBs
  - PDB, MSD
- Protein Domain Homology DBs
  - InterPro, CluSTr

UniProtKB/Swiss-Prot

- Consists of protein sequence entries
- Contains high-quality annotation
- Is non-redundant
- Cross-referenced to many other databases
- 104,559 sequences in Jan 02
- 120,960 sequences in Jan 03
- 194,317 sequences in Sep 05 (latest)

Swis-Prot by Species (‘03)

```
------  ---------  -------------------------------------------
Number  Frequency  Species
------  ---------  -------------------------------------------
     1       8950  Homo sapiens (Human)
     2~20%   6028  Mus musculus (Mouse)
     3       4891  Saccharomyces cerevisiae (Baker's yeast)
     4       4835  Escherichia coli
     5       3403  Rattus norvegicus (Rat)
     6       2385  Bacillus subtilis
     7       2286  Caenorhabditis elegans
     8       2106  Schizosaccharomyces pombe (Fission yeast)
     9       1836  Arabidopsis thaliana (Mouse-ear cress)
    10       1773  Haemophilus influenzae
    11       1730  Drosophila melanogaster (Fruit fly)
    12~13%   1528  Methanococcus jannaschii
    13       1471  Escherichia coli O157:H7
    14       1378  Bos taurus (Bovine)
    15       1370  Mycobacterium tuberculosis
```

Swis-Prot by Species (Oct ‘05)

```
------  ---------  -------------------------------------------
Number  Frequency  Species
------  ---------  -------------------------------------------
     1      12860  Homo sapiens (Human)
     2       9933  Mus musculus (Mouse)
     3       5139  Saccharomyces cerevisiae (Baker's yeast)
     4       4846  Escherichia coli
     5       4570  Rattus norvegicus (Rat)
     6       3609  Arabidopsis thaliana (Mouse-ear cress)
     7       2840  Schizosaccharomyces pombe (Fission yeast)
     8       2814  Bacillus subtilis
     9       2667  Caenorhabditis elegans
    10       2273  Drosophila melanogaster (Fruit fly)
    11       1782  Methanococcus jannaschii
    12       1772  Haemophilus influenzae
    13       1758  Escherichia coli O157:H7
    14       1653  Bos taurus (Bovine)
    15       1512  Salmonella typhimurium
```

UniProtKB/TrEMBL

- Computer annotated Protein DB
- Translations of all coding sequences in EMBL DNA Database
- Remove all sequences already in Swiss-Prot
- November 01: 636,825 peptides
- Jan 17th 2003: 728713 peptides
- TrEMBL new is a weekly update
- GenPept is the Genbank equivalent

SNPs

- Biggest growth area right now is in mutation databases
- www.ncbi.nlm.nih.gov/About/primer/snps.html
- Polymorphisms estimates at between 1:100 1:300 base pairs (normal human variation)
- Databases include true SNPs (single bases) and larger variations (microsatellites, small indels)

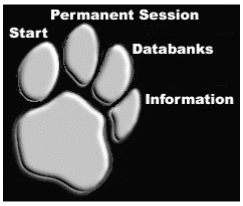
dbSNP

- "The database grows at 90 SNPs per month"
- 125 versions since start in 1998
- Currently 47 million SNPs in latest release
- 15 million added between version 124 and 125

Database Search Methods

- Text based searching of annotations and related data: SRS, Entrez
- Sequence based searching: BLAST, FASTA, MPSearch

SRS



- Sequence Retrieval System
  - Powerful search of EMBL annotation
  - Linked to over 80 other data sources
  - Also includes results from automated searches

SRS data sources

- Primary Sequence: EMBL, SwissProt
- References/Literature: Medline
- Protein Homology: Prosite, Prints
- Sequence Related: Blocks, UTR, Taxonomy
- Transcription Factor: TFACTOR, TFSITE
- Search Results: BLAST, FASTA, CLUSTALW
- Protein Structure: PDB
- Also, Mutations, Pathways, other specialist DBs

Entrez

- Text based searching at NCBI’s Genbank
- Very simple and easy to use
- Not as flexible or extendable as SRS
- No user customisation

Sequence Based Searching

- Queries:

DNA query against DNA db
Translated DNA query against Protein db
Translated DNA query against translated DNA db
Translated Protein query against DNA db
Protein query against Protein db

- BLAST & FASTA

BLAST

Version	Query	DB
Blastn	DNA	DNA
Blastp	Peptide	Peptide
Blastx	DNA	Peptide
tBlastn	Peptide	DNA
tBlastx	DNA	DNA

☐ **`translated`**

FASTA Key Parameters

Database:	Which DNA/Protein db to use.
Program:	fastx3, tfasty3 etc
Matrix:	Substitution score matrix e.g. Blosum50
KTUP	Word length to use in search
Scores:	How many results to summarise
Alignments:	How many full alignments to provide
Open Gap:	Penalty for opening a new gap
Extend Gap:	Penalty for extending a gap by 1

Initial Strategies

- Use a good server with up to date databases
- Run BLAST as a first choice (its quick)
- If appropriate, translated DNA or protein searches are better.
- Refine using FASTA, SW programs or protein prediction packages

Scores

- The raw scores returned by Blast and FASTA are not in themselves all that useful.
- The E-Value (expect) is the number of false positives you would expect to find in that query. A low E-value indicates a higher confidence level

P value

- The Probability of the observed score (probability that it happened by chance) can be calculated:

$$P = 1 - e^{-E}$$

Secondary Databases

- PDB
- Pfam
- PRINTS
- PROSITE
- ProDom
- SMART
- TIGRFAMs

PDB

- Molecular Structure Database (EBI)
- Contains the 3D structure coordinates of ‘solved’ protein sequences
  - X-ray crystallography
  - NMR spectra
- 19749 protein structures

Multiple Sequence Alignment

- What and Why?
- Dynamic Programming Methods
- Heuristic Methods
- A further look at Protein Domains

Multiple Alignment

- Normally applied to proteins
- Can be used for DNA sequences
- Finds the common alignment of >2 sequences.
- Suggests a common evolutionary source between related sequences based on similarity
  - Can be used to identify sequencing errors

Multiple Alignment of DNA

- Take multiple sequencing runs
- Find overlaps
  - variation of ends-free alignment
- Locate cloning or sequencing errors
- Derive a consensus sequence
- Derive a confidence degree per base

Consensus Sequences

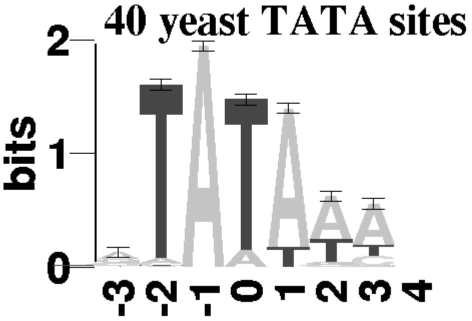
- Look at several aligned sequences and derive the most common base for each position.
  - Several ways of representing consensus sequences
  - Many consensus sequences fail to represent the variability at each base position.
  - Largely replaced by Sequence Logos but the term is often mis-applied

---

Sequence Logos

- Example, from an alignment of the TATA box in yeast genes:

We now have a confidence level for each base at each position



Multiple Alignment of Proteins

- Multiple Alignment of Proteins
- Identify Protein Families
- Find conserved Protein Domains
- Predict evolutionary precursor sequences
- Predict evolutionary trees

Protein Families

- Proteins are complex structures built from functional and structural sub-units
  - When studying protein families it is evident that some regions are more heavily conserved than others.
  - These regions are generally important for the structure or function of the protein
  - Multiple alignment can be used to find these regions
  - These regions can form a signature to be used in identifying the protein family or functional domain.

Protein Domains

- Evolution conserves sequence patterns due to functional and structural constraints.
- Different methods have been applied to the analysis of these regions.
- Domains also known by a range of other names:

motifs patterns prints blocks

Multiple Alignment

- OK we now have an idea WHY we want to try and do this
- What does a multiple alignment look like?
- How could we do multiple alignments
- What are the practical implications

Multiple alignment table

```
dlg_CG1725-PH              ALFDYDPNRDDGLPSRGLPFKH
Sap97_dlgh1                ALFDYDKTKDSGLPSQGLNFRF
chapsyn-110_dlgh2          AMFDYDKSKDSGLPSQGLSFKY
Sap102_dlgh3               ALFDYDRTRDSCLPSQGLSFSY
PSD-95_dlgh4               ALFDYDKTKDCGFLSQALSFHF
                           *:**** .:*  : *:.* * .
```

A consensus character is the one that minimises the distance between it and all the other characters in the column

Conservatived or Identical residues are colour coded

Scoring Multiple Alignments

- We need to score on columns with more than 2 bases or residues:

$$ColumnCost\left(\begin{matrix}S\\C\\A\\P\\P\end{matrix}\right) = 24$$

Multiple alignments are usually scored on cost/difference rather than similarity

Column Costs

- Several strategies exist for calculating the column cost in a multiple alignment
- Simplest is to sum the pairwise **costs** of each base/residue pair in the column using a matrix (e.g. PAM250).
- Gap scoring rules can be applied to these as well.

Scoring Multiple Alignments

- Score = (S,C)+(S,A)+(S,A)+(S,P)+(S,P)+(C,A)+ (C,P)+(C,P)+(A,P)+(A,P)+(P,P)

$$\mathit{ColumnCost}\begin{pmatrix}S\\C\\A\\P\\P\end{pmatrix} = 24$$

Known as the sum-of-pairs scoring method

Sum-of-pairs cost method (SP)

- Score = (S,C)+(S,-)+(S,A)+(S,P)+(S,P)+
  (-,A)+(-,P)+(-,P)+(A,P)+(A,P)+(P,P)

$$\mathit{ColumnCost}\begin{pmatrix} S \\ - \\ A \\ P \\ P \end{pmatrix} = \mathit{24}$$

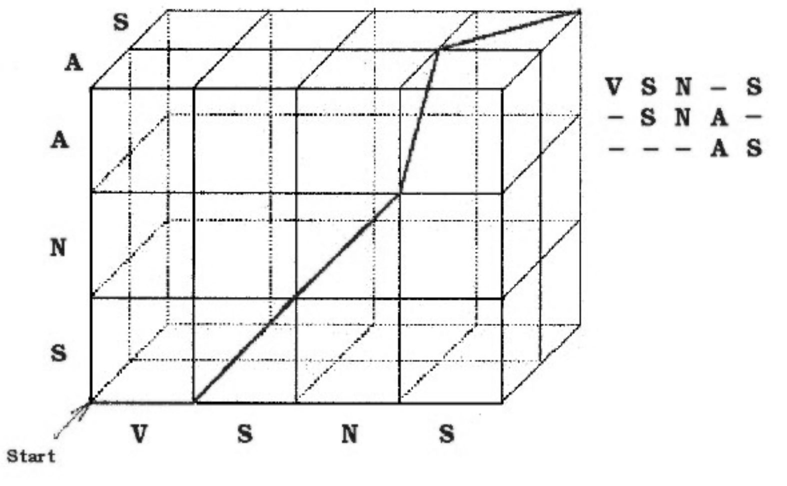
Still works with gaps using whatever gap penalty you want

Multiple Alignment Cost

- Sum of pairs is a simple method to get a score for each column in a multiple alignment
- Based on matrices and gap penalties used for pairwise sequence alignment
- The score of the alignment is the sum of each column

Optimal Multiple Alignment

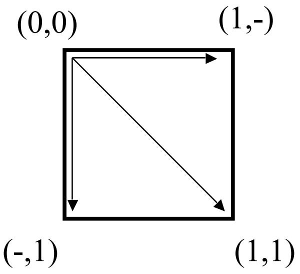
- The best alignment is generally the one with the lowest score (i.e. least difference)
  - depends on the scoring rules used.
- Like pairwise cases, each alignment represents a path through a matrix
- For multiple alignment, the matrix is *n*-dimensional
  - where *n*=number of sequences



(Murata, Richardson and Sussman 1999)

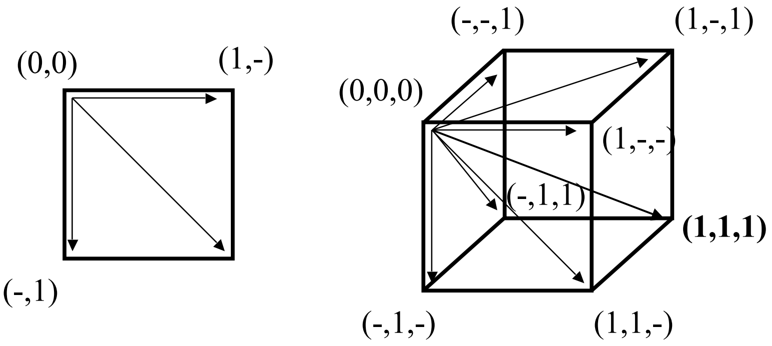
Contrasting pairwise and multiple alignments

Lets compare pairwise with three sequences.



Contrasting pairwise and multiple alignments

Lets compare pairwise with three sequences.



NP-Completeness

- A problem is solvable in polynomial time if an algorithm exists O(n^c)
  - c - some constant
  - *n* - size of the input
- Pairwise alignment is solvable in polynomial time O(n^2)
- More difficult problems are *NP-complete*

Multiple alignment complexity

- For *k* sequences of average length *n*
- *k* dimension matrix has $(n+1)^k$ cells to compute.
- Each entry can be computer in 2^k time
- Running time of the overall algorithm is:

  $O((2n)^k)$
- The real problem hits when considering protein sequences average ~400 residues

MA: Dynamic Programming

- We can use dynamic programming in some small cases.
- For *x* sequences, build an *x* dimensional hypercube.
- Solve as before using gap and substitution penalties but remembering that there are more routes to each cell in the hypercube

MA: Dynamic Programming

- Space complexity is huge:
  - O(sum sequences x ave length)
- Computational complexity is huge
- In practice the DP method is only feasible for small numbers of short strings

Examples

Center Star Method

- Given a set of Strings, define the center string *Sc* as the string that minimises the sum of distances from all other sequences.
  - Found *Sc*
  - Consecutively add on the other sequences so that the alignment of each is optimal.
  - Add spaces where needed to all prealigned sequences
- The center star method is within 2 fold accuracy of true dynamic solution

Iterative pairwise alignment

- In CSA we try to align the chosen center string with all the others in no particular order.
- Often some of the other sequences will be closer to each other and form *clusters*
- Tricky part is deciding how to define close and how to cluster them

How do we cluster sequences?

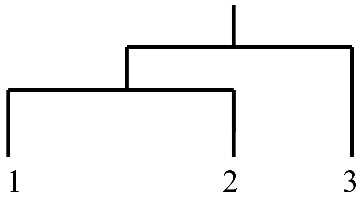
Building trees

- Need to define how the sequences are related to one another.
- Most use the distances between pairs in the set of sequences.
- Key parameter is in defining the distance score.

Clustering Methods

- Unweighted Pair Group Method using Arithmetic averages or UPGMA
- Simple and based on distance pairs
- Each stage joins two clusters creating a new node

An example tree

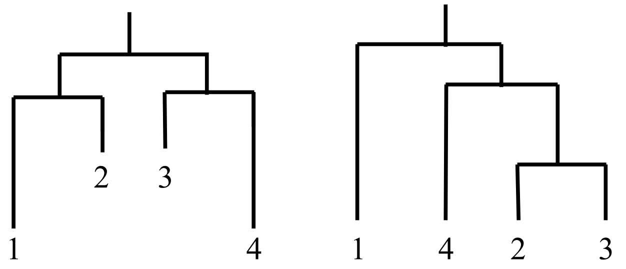


Sequences 1 and 2 are the closest related.
Each sequence lies on its own leaf

UPGMA

- Assign each sequence to its own cluster
- Create a leaf at height zero for each cluster
- Determine the two closest clusters
- Align the two sequences to define a new cluster at the next level up.
- Remove the two pre existing clusters and start over.
- End at two clusters

Drawbacks of UPGMA



Correct tree

Incorrect by UPGMA

Nearest Neighbour

- Similar to the UPGMA algorithm
- UPGMA works on distance between sequence pairs alone
- Nearest Neighbour compensates for the path through the tree to correct situations where distance alone would incorrectly pair two sequences

Back to multiple alignment heuristics

Iterative Pairwise Alignment

- Can be used as a strategy for growing groups of profiles from multiple sequences
- This approach uses pairwise alignment scores to add one additional sequence at a time to a growing multiple alignment.

Iterative Pairwise Alignment

- First align all pairs of strings where one is already in a multiple alignment and one is aligned.
- Find the closest matches.
- Align the unassigned sequence with the family profile of the closest group
- Realign the group and get a new profile.

Feng-Doolittle

- Feng-Doolittle 1987 Journal of Molecular Evolution 25:351-360
- The key principal is that the two most similar sequences in a multiple alignment are the most recently diverged.
- Therefore the pairwise alignment of these two sequences is the most reliable of the entire group
- Gaps present in the alignment should therefore be preserved in the multiple alignment.

Feng-Doolittle

- Calculate the pairwise alignment scores for each sequence
- Construct a tree using these distances
- Traverse the nodes of the tree in order of addition (most similar first)
- Progressively align the sequences starting with the most similar:
- Once a gap is established in the multiple alignment it stays.

ClustalW

- Uses a modification of the Feng-Doolittle algorithm
- Very common software package for multiple alignment

ClustalW

- Starts by calculating pairwise alignments and converting scores to distances
- Uses a neighbour joining algorithm to build a tree from the distances
- Aligns sequences to each other
- Aligns sequences to profiles
- Aligns profiles to profiles
- Can output multiple alignment as well a predicted evolutionary tree

MSA

- Exploits the fact that closely aligned sequence paths will be close to the main diagonal on a DP table.
- Estimates a good solution, removes cells from the hypercube where the score could not feasibly pass through them.

CAP

- Contig Assembly Program
- Designed to optimise alignments between multiple DNA sequences that are suspected to overlap.
- Uses a fast heuristic prescreen then finishes using a dynamic programming approach.

CAP

- Takes all the sequences and split into short fragments
- Eliminate fragment pairs that could not possibly overlap
- The dynamic programming algorithm is used to find the maximal scoring overlaps
- Scores are weighted so that sequencing errors are low cost and mutations higher

Consensus Sequences

- The consensus sequence is the concatenation of the consensus characters
- The alignment error of the multiple alignment is the sum of the distance costs of each consensus character in the consensus sequence.

Scoring Multiple Alignments

- Distance from Consensus
  - In each column, count the number of characters that are different from the consensus sequence.
- Sum of Pairs (covered already)
  - Sum the pairwise distances between all sequence pairs

Scoring Multiple Alignments

- Evolutionary Tree alignment
  - The weight of the lightest tree that can be constructed from the sequences
  - The weight is defined as the the number of changes that correspond to two adjacent nodes in the tree summed over all pairs.

Consensus Sequences

- Given an optimal alignment between >2 sequences, how do we find the consensus sequence?
- Take a multiple alignment in columns of characters

Multiple alignment table

```
dlg_CG1725-PH          ALFDYDPNRDDGLPSRGLPFKH
Sap97_dlgh1            ALFDYDKTKDSGLPSQGLNFRF
chapsyn-110_dlgh2      AMFDYDKSKDSGLPSQGLSFKY
Sap102_dlgh3           ALFDYDRTRDSCLPSQGLSFSY
PSD-95_dlgh4           ALFDYDKTKDCGFLSQALSFHF
                       *:**** .:*  : *:.* * .
```

The consensus character is the one that minimises the distance between it and all the other characters in the column

---

Finally some examples

- We are interested in the protein DLG
  - DLG is a molecular scaffold
  - 1 gene in Drosophila
  - 4 human genes (DLG1-4 with synonyms)
- Tarpey et al 2004 found mutations linking DLG3/Sap102 to Mental Retardation
- Obtained sequences for all 5 proteins
- Run through ClustalW (results on-line)

Another example

- We are also interested in PDE4B
  - PDE4B is a phosphodiesterase
  - 1 gene in Drosophila (dunce) linked to memory
  - multiple human genes closest PDE4B
- Millar et al 2005 found a link between PDE4B and schizophrenia
- A database search funds many possible PDE4B proteins, need to make sense of it all…