Bio2

Heuristics, Databases ;
Multiple Sequence Alignment ;
Gene Finding

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.

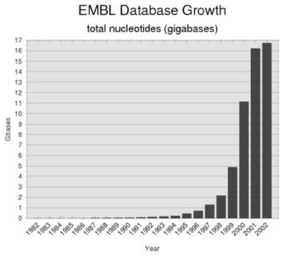
Limited annotation

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

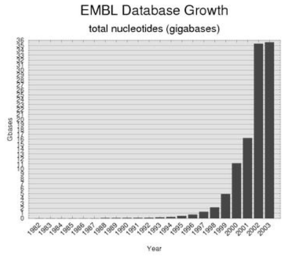
EMBL file tags

```
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)
```

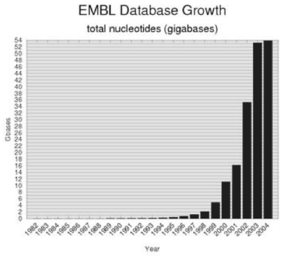
16,759,535,577 bases (27/1/02)

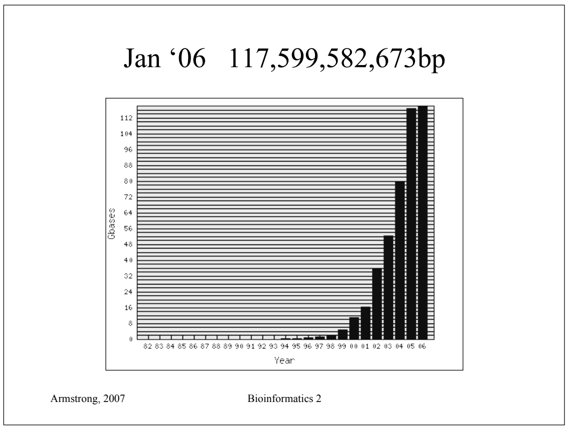


35,602,556,374 bases (17/1/03)



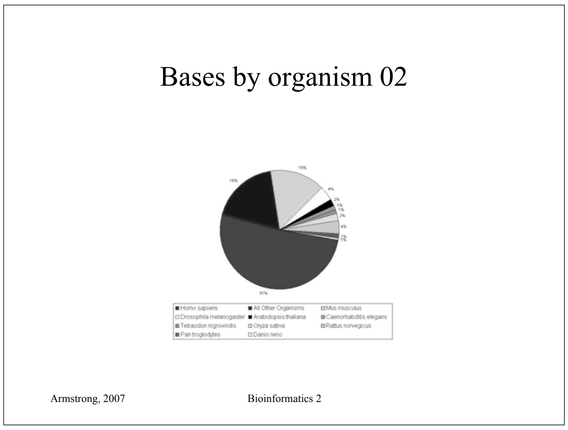
53,958,991,118 bases (24/1/04)





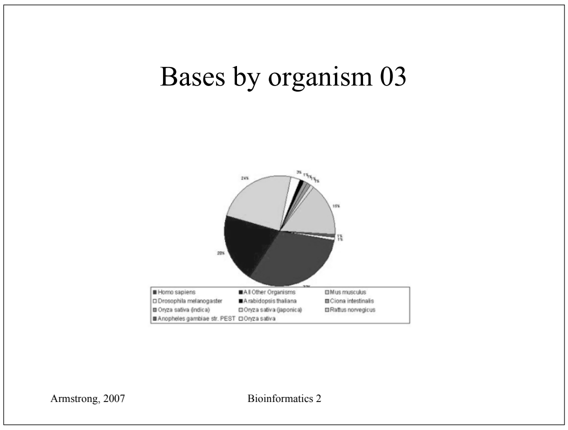
Jan ‘06   117,599,582,673bp

Armstrong, 2007 Bioinformatics 2



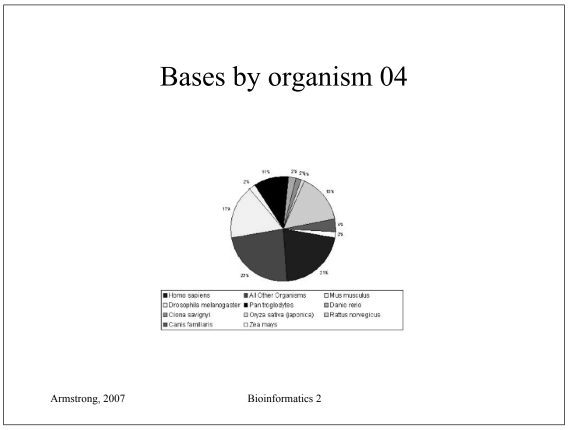
Bases by organism 02

Armstrong, 2007 Bioinformatics 2



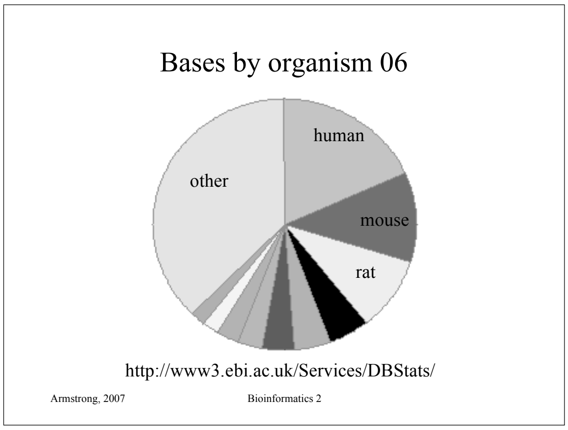
Bases by organism 03

Armstrong, 2007 Bioinformatics 2



Bases by organism 04

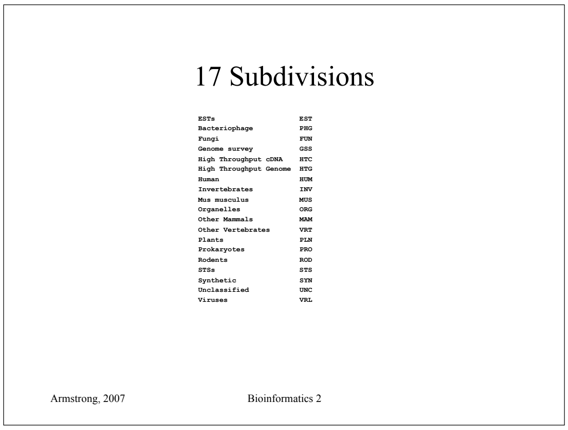
Armstrong, 2007 Bioinformatics 2



Bases by organism 06

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

Armstrong, 2007 Bioinformatics 2



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

Armstrong, 2007 Bioinformatics 2

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)

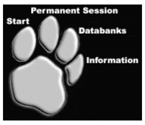
dbSNP

- “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

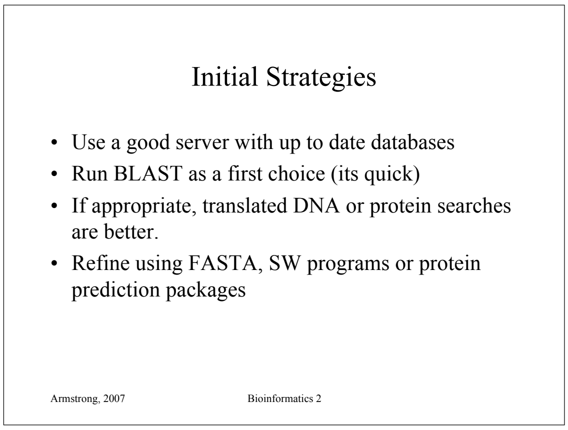
BLAST

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

□ translated

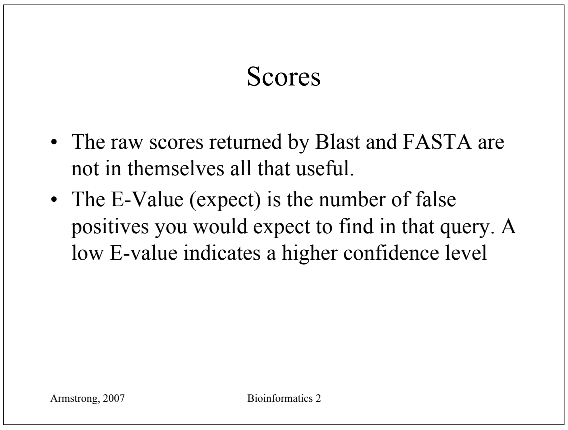
FASTA Key Parameters

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



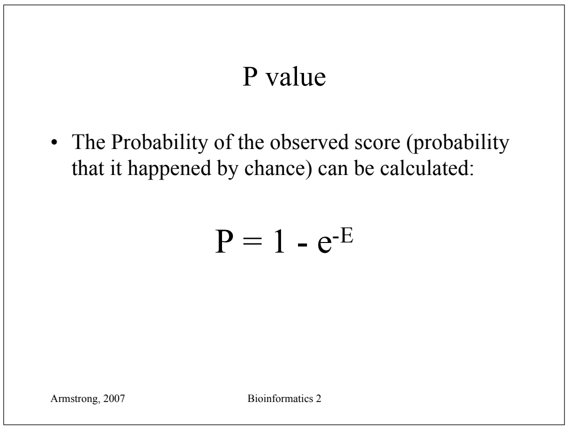
Initial Strategies

- 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

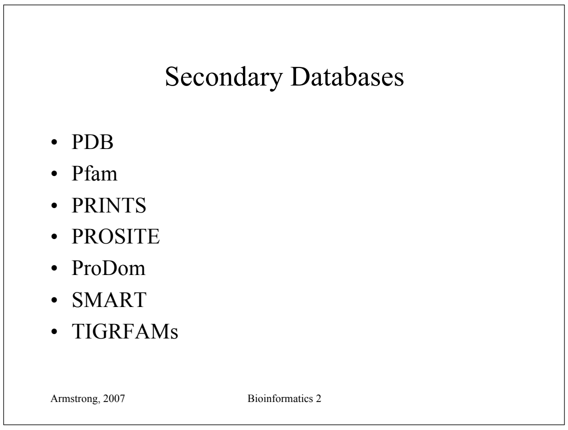
- The raw scores returned by Blast and FASTA are not in themselves all that useful.
- The E-Value (expect) is the number of false positives you would expect to find in that query. A low E-value indicates a higher confidence level



P value

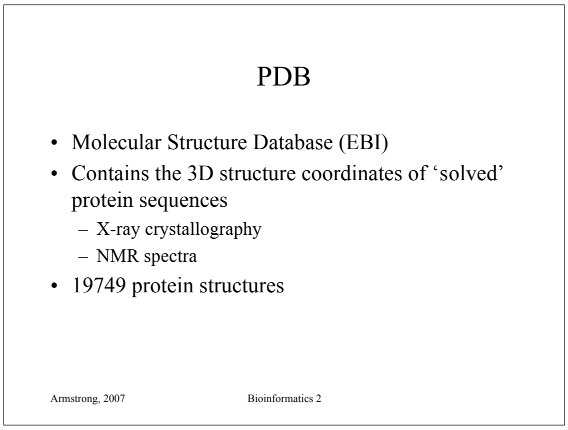
- The Probability of the observed score (probability that it happened by chance) can be calculated:

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



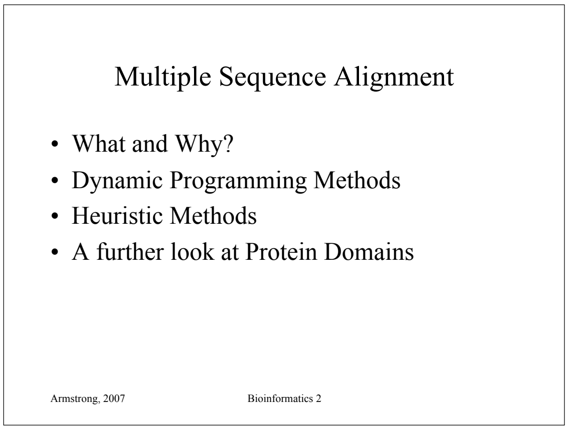
Secondary Databases

- 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

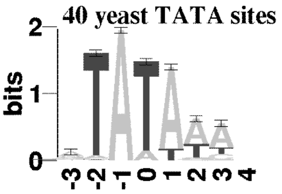
Consensus Sequences

- 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:

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

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)

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

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)

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

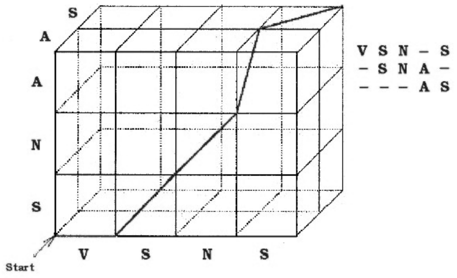
Still works with gaps using whatever gap penalty you want

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

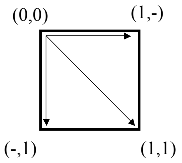
- The best alignment is generally the one with the lowest score (i.e. least difference)
  - depends on the scoring rules used.
- Like pairwise cases, each alignment represents a path through a matrix
- For multiple alignment, the matrix is *n*-dimensional
  - where *n*=number of sequences



(Murata, Richardson and Sussman 1999)

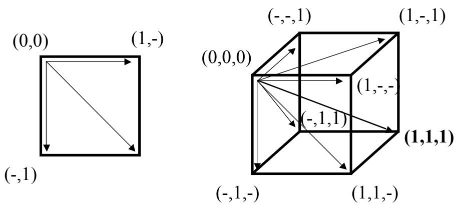
Contrasting pairwise and multiple alignments

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

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

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

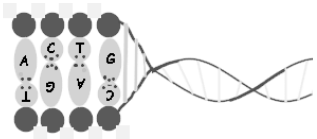
Bio2

Gene and Protein Prediction

Gene prediction

- What is a gene?
  - Simple definition: A stretch of DNA that encodes a protein and includes the regulatory sequences required for temporal and spatial control of gene transcription.
- Characteristics of genes.
  - What genetic features can we use to recognise a gene?

DNA structure



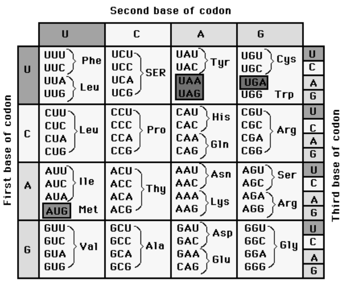
Bases: A,C,G and T

Chemically, A can only pair with T and G with C

Two strands, 5’ and 3’ Genes are encoded along one side of the DNA molecule. The 5’ end being at the left hand side of the gene.

Codons and ORFs

- Three bases that encode an amino acid or stop site.
- A run of valid codons is an Open Reading Frame.
- An ORF usually starts with a Met
- Ends with a nonsense or stop codon.



Second base of codon

First base of codon	U	C	A	G	Third base of codon
U	UUU UUC Phe UUA UUG Leu	UCU UCC UCA UCG SER	UAU UAC Tyr UAA UAG	UGU UGC Cys UGA UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thy	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU GAC Asp GAA CAG Glu	GGU GGC GGA GGG Gly	U C A G

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Predicting ORFs

- 64 total codons
- 3 stop codons, 61 codons for amino acids
- Random sequence 1:21 ratio for stop:coding.
- = 1 stop codon every 63 base pairs
- Gene lengths average around 1000 base pairs.

Finding ORFs

- One algorithm slides along the sequence looking stop codons.
- Scans back until it finds a start codon.
- Fails to find very short genes since it it looking for long ones
- Also fails to find overlaping ORFs
- There are many more ORFs than genes

Amino Acid Bias

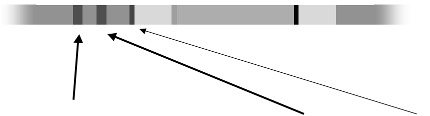
- The amino acids in proteins are not random
  - leucine has 6 codons
  - alanine has 4 codons
  - tryptophan has 1 codon
- The random the ratio would be 6:4:1
- In proteins it is 6.9:6.5:1
  - i.e. it is not random

Gene Prediction

- Take all factors into consideration
- Prokaryotes
  - No Nucleus
  - 70% of the genome encodes protein
  - No introns

Prokaryote gene structure

1. Promoter region



nnnTTGACAnnnnnnnnnnnnnnnnnnTATAATnnnnnnS

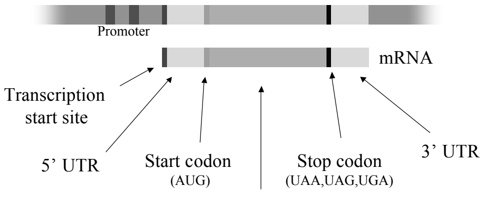
(consensus sequence for *E.coli*.)

Probability matrix for TATA box

Pos.	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

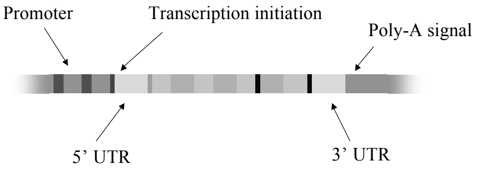
Prokaryote gene structure

2. Transcribed region (mRNA)

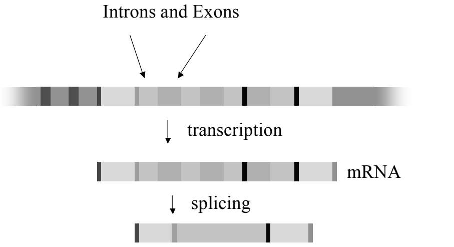


Coding Sequence

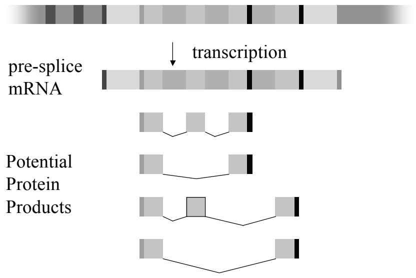
Eukaryote gene structure



Eukaryote gene structure

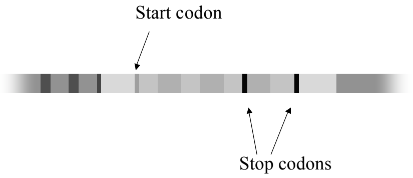


Functional significance of Introns and Exons

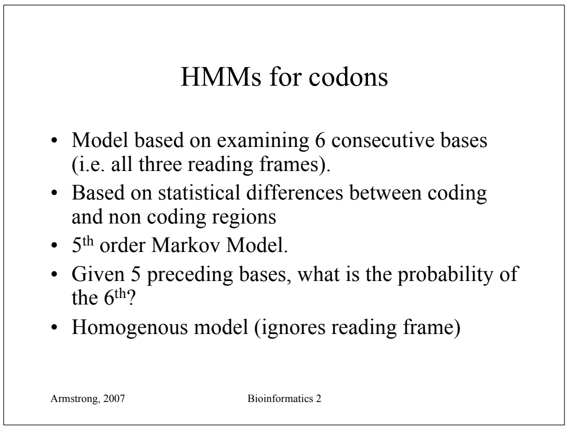


1 gene - 4 protein products

Eukaryote gene structure

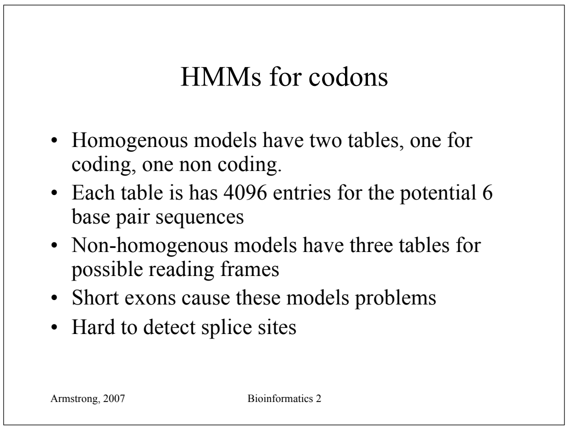


Intron/Exon structure allows multiple start and stop codons



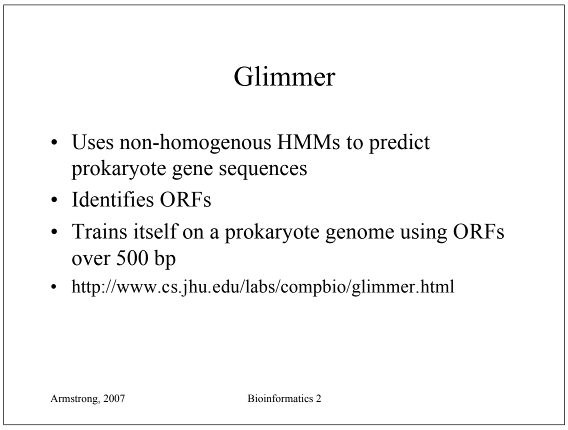
HMMs for codons

- Model based on examining 6 consecutive bases (i.e. all three reading frames).
- Based on statistical differences between coding and non coding regions
- 5th order Markov Model.
- Given 5 preceding bases, what is the probability of the 6th?
- Homogenous model (ignores reading frame)



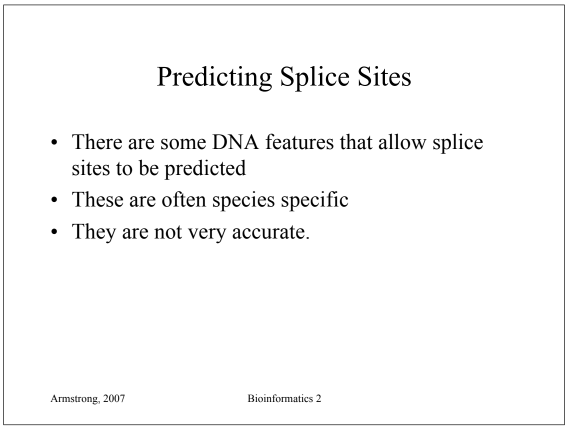
HMMs for codons

- Homogenous models have two tables, one for coding, one non coding.
- Each table is has 4096 entries for the potential 6 base pair sequences
- Non-homogenous models have three tables for possible reading frames
- Short exons cause these models problems
- Hard to detect splice sites



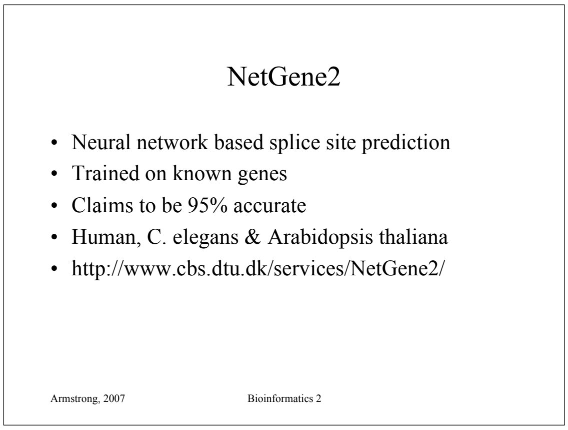
Glimmer

- Uses non-homogenous HMMs to predict prokaryote gene sequences
- Identifies ORFs
- Trains itself on a prokaryote genome using ORFs over 500 bp
- http://www.cs.jhu.edu/labs/compbio/glimmer.html



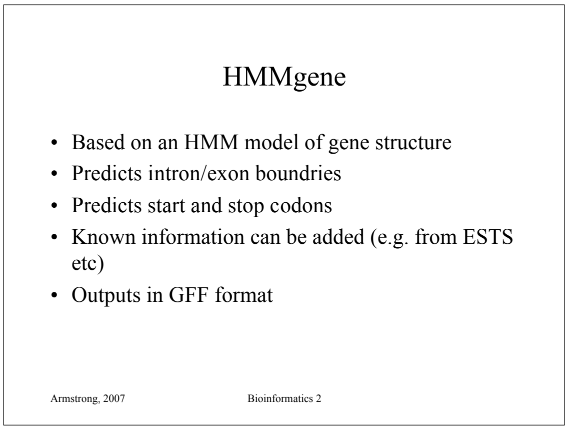
Predicting Splice Sites

- There are some DNA features that allow splice sites to be predicted
- These are often species specific
- They are not very accurate.



NetGene2

- Neural network based splice site prediction
- Trained on known genes
- Claims to be 95% accurate
- Human, C. elegans & Arabidopsis thaliana
- http://www.cbs.dtu.dk/services/NetGene2/



HMMgene

- Based on an HMM model of gene structure
- Predicts intron/exon boundries
- Predicts start and stop codons
- Known information can be added (e.g. from ESTS etc)
- Outputs in GFF format

GFF Format

- Exchange format for gene finding packages
- Fields are:
  - <seqname> name, genbank accession number
  - <source> program used
  - <feature> various inc splice sites
  - <start> start of feature

GFF Format

- <end> end of feature
- <score> floating point value
- <strand> +, - (or .. for n/a)
- <frame> 0,1 or 2

GenScan

- Probabilistic model for gene structure based on a general HMM
- Can model intron/exon boundries, UTRs, Promoters, polyA tails etc
- http://genes.mit.edu/GENSCAN.html

Given a new protein sequence…

- What is the function?
- Where is the protein localised?
- What is the structure?
- What might it interact with?

Given a new protein sequence…

- What is the function?
- Have we seen this protein or a very similar one before?
  - If yes then we can infer function, structure, localisation and interactions from homologous sequence.
- Are there features of this protein similar to others?

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

Profiles

- Given a sequence, we often want to assign the sequence to a family of known sequences
- We often also want to assign a subsequence to a family of subsequences.

Profiles

- Examples include assigning a gene/protein to a known gene/protein family, e.g.
  - G coupled receptors
  - actins
  - globins

Profiles

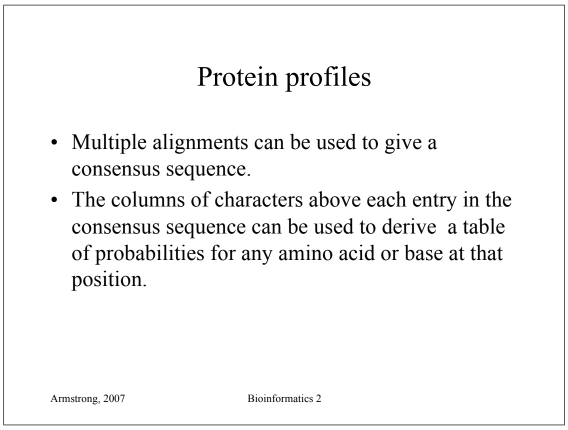
- Also we may wish to find known protein domains or motifs that give us clues about structure and function
  - Phosphorylation sites (regulated site)
  - Leucine zipper (dna binding)
  - EGF hand (calcium binding)

Creating Profiles

- Aligning a sequence to a single member of the family is not optimal
- Create profiles of the family members and test how similar the sequence is to the profile.
- A profile of a multiply aligned protein family gives us letter frequencies per column.

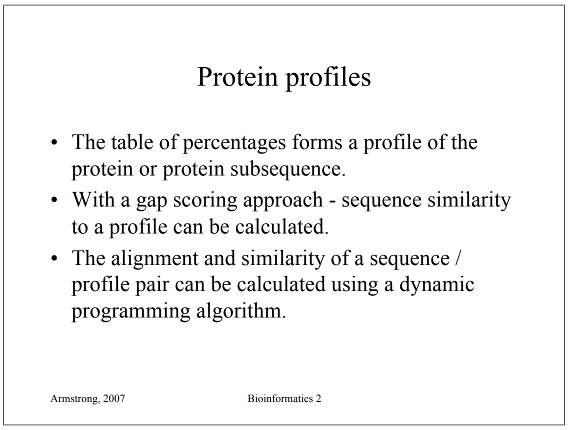
Matching sequences to profiles

- We can define a distance/similarity cost for a base in each sequence being present at any location based on the probabilities in the profile.
- We define define costs for opening and extending gaps in the sequence or profile.
- Therefore we can essentially treat the alignment of a sequence to a profile as a pairwise alignment and use dynamic programming algorithms to find and score the optimal alignments.



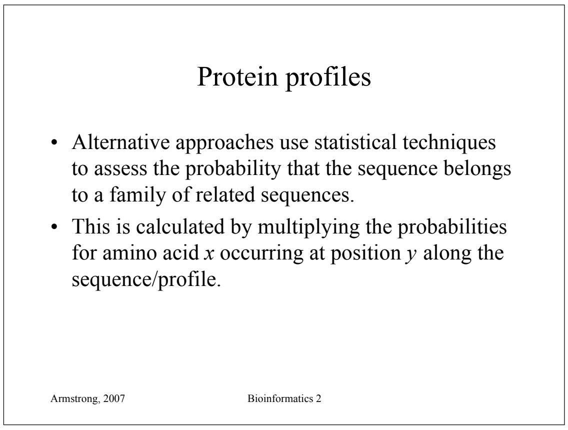
Protein profiles

- Multiple alignments can be used to give a consensus sequence.
- The columns of characters above each entry in the consensus sequence can be used to derive a table of probabilities for any amino acid or base at that position.



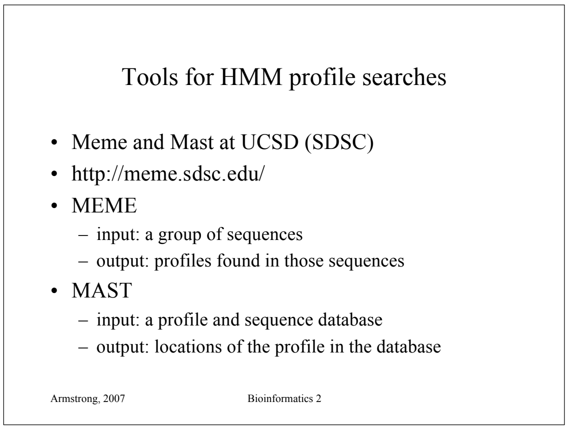
Protein profiles

- The table of percentages forms a profile of the protein or protein subsequence.
- With a gap scoring approach - sequence similarity to a profile can be calculated.
- The alignment and similarity of a sequence / profile pair can be calculated using a dynamic programming algorithm.



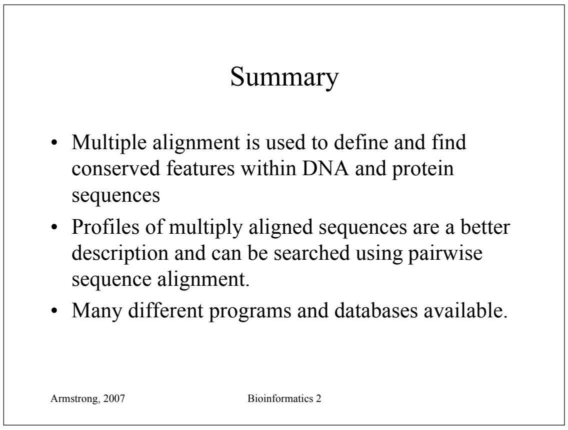
Protein profiles

- Alternative approaches use statistical techniques to assess the probability that the sequence belongs to a family of related sequences.
- This is calculated by multiplying the probabilities for amino acid *x* occurring at position *y* along the sequence/profile.



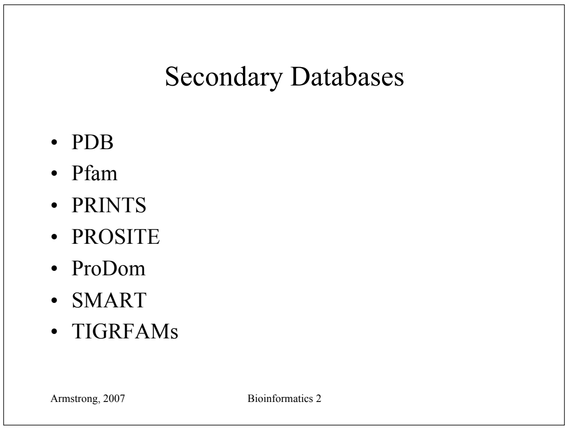
Tools for HMM profile searches

- Meme and Mast at UCSD (SDSC)
- http://meme.sdsc.edu/
- MEME
  - input: a group of sequences
  - output: profiles found in those sequences
- MAST
  - input: a profile and sequence database
  - output: locations of the profile in the database



Summary

- Multiple alignment is used to define and find conserved features within DNA and protein sequences
- Profiles of multiply aligned sequences are a better description and can be searched using pairwise sequence alignment.
- Many different programs and databases available.



Secondary Databases

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

PDB

- Molecular Structure Database
- Contains the 3D structure coordinates of ‘solved’ protein sequences
  - X-ray crystallography
  - NMR spectra
- 29429 protein structures

Superfamily 1.65

HMM library and genome assignments server

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known (based on PDB)

Pfam

- Database of protein domains
- Multiple sequence alignments and profile HMMs
- Entries also annotated
- Swiss-Prot DB all pre-searched
- New sequences can be searched as well.
  - 7973 entries in Pfam last update

PRINTS

- Database of ‘protein fingerprints’
- Group of motifs that combined can be used to characterise a protein family
- ~11,000 motifs in PRINTS DB
- Provide more info than motifs alone

‘linear’ motifs

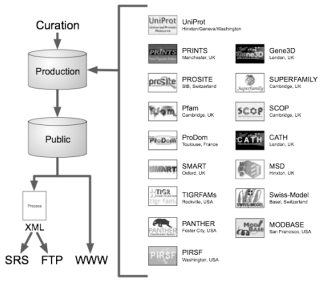
- Not all protein motifs are easy to find
- Linear motifs involved in protein-protein interactions
  - Very degenerate
  - Found in specific regions of proteins
  - Require special treatment
  - Neduva *et al*, PLOS 2005

Linking it all together…

- Database Searches
  - Multiple Alignments
  - Find known motifs and domains
  - Find possible similar folds
- Prediction algorithms
  - Properties of amino acids
  - Predicting folding
  - Finding cysteine bonds

InterPro

- EBI managed DB
- Incorporates most protein structure DBs
- Unified query interface and a single results output.



See http://www.ebi.ac.uk/interpro/

InterPro

DATABASE	VERSION	ENTRIES
SWISS-PROT	48	197228
PRINTS	38	1900
TREMBL	31.1	2342938
PFAM	18	7973
PROSITE	19.10	1882

Currently 15 databases, plans to add 3 new ones this month.

PredictProtein

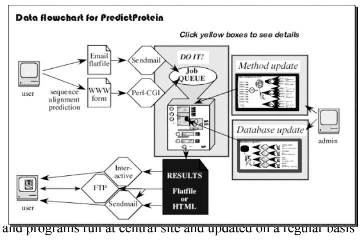
http://www.embl-heidelberg.de/predictprotein/

Database searches:
- generation of multiple sequence alignments (MaxHom)
- detection of functional motifs (PROSITE)
- detection of composition-bias (SEG)
- detection of protein domains (PRODOM)
- fold recognition by prediction-based threading (TOPITS)

PredictProtein

Predictions of:
- secondary structure (PHDsec, and PROFsec)
- residue solvent accessibility (PHDacc, and PROFacc)
- transmembrane helix location and topology (PHDhtm, PHDtopology)
- protein globularity (GLOBE)
- coiled-coil regions (COILS)
- cysteine bonds (CYSPRED)
- structural switching regions (ASP)

Data and methods in PredictProtein



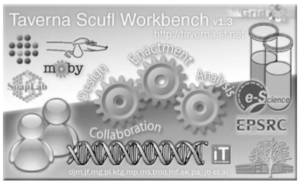
Add data and programs run at central site and updated on a regular basis

Too many programs/databases

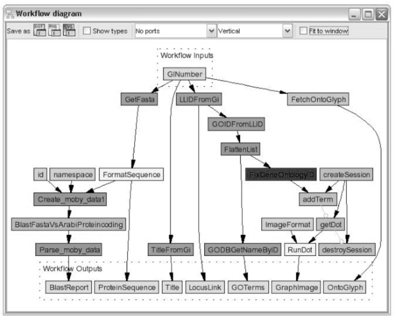
- How do we keep track of our own queries?
  - Repeat an old query
  - Run the same tests on a new sequence
  - Run 100s of sequences..
  - Document the process for a paper or client or for quality assurance

Workflow managers

- Locate and manage connections to software and databases
- Record actions
- Replay a workflow at a later date or against multiple sequences
- Manages redundant external sources (e.g. multiple blast servers)
- Can connect to specialist local sources



- http://taverna.sourceforge.net/
- Open source and free to download
- Runs on PC/linux/mac
- Drag-n-Drop interface to bioinformatics analysis


Workflow diagram
Save as
Show types
No ports
Vertical
Fit to window
Workflow Inputs
GINumber
GetFasta
LLIDFromGi
FetchOntoGlyph
GOIDFromLLID
FlattenList
id
namespace
FormatSequence
FixGeneOntologyID
createSession
Create_moby_data1
addTerm
BlastFastaVsArabiProteincoding
ImageFormat
getDot
Parse_moby_data
TitleFromGi
GODBGetNameByID
RunDot
destroySession
Workflow Outputs
BlastReport
ProteinSequence
Title
LocusLink
GOTerms
GraphImage
OntoGlyph