Contents

1 Cell and Molecular biology

1.1 The Cell

- Minimal unit of life
- Cell is a system of many components enclosed in a series of membranes
- Small organisms such as fungi and bacteria are unicellular
- Plants and animals are generally multicellular

1.1.1 Prokaryotes

- All prokaryotes are single cell
- Smaller than eukaryotic cells
  ($< 1\mu$m in diameter)
- Simple structure
  No inner cellular membranes
- Very adaptable to environment
  Found in almost every habitat
- Approximately $5x10^{30}$ prokaryotic cells in the world
- Essential for healthy life

Prokaryotic cell

- Less complex than eukaryotic cells
- Contain no organelles
- Believed to represent the earliest life on earth and that eukaryotic cells evolved from prokaryotic cells

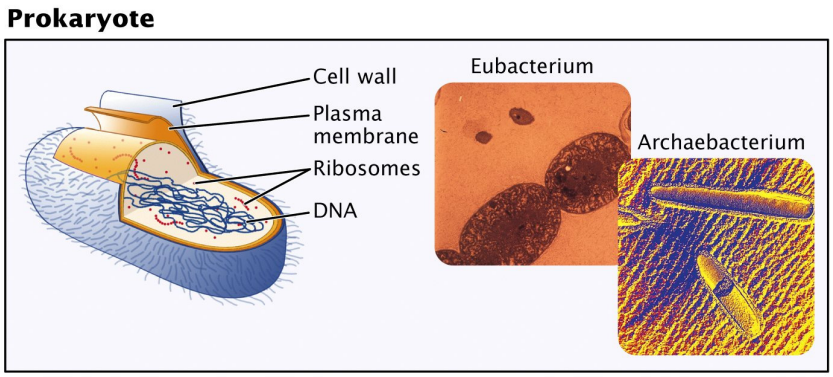


Figure 1: Prokaryotic cell structure

Types of prokaryotes

Types descend from common ancestor.

Eubacteria
: Common bacteria that affects life daily

Archaea
: Tend to exist in extreme habitats (high pressure, temperature, pH)

1.1.2 Eukaryotes

- More structurally and biochemically complex than prokaryotes
- Evolutionarily more recent
- Believed to have evolved through endosymbiosis

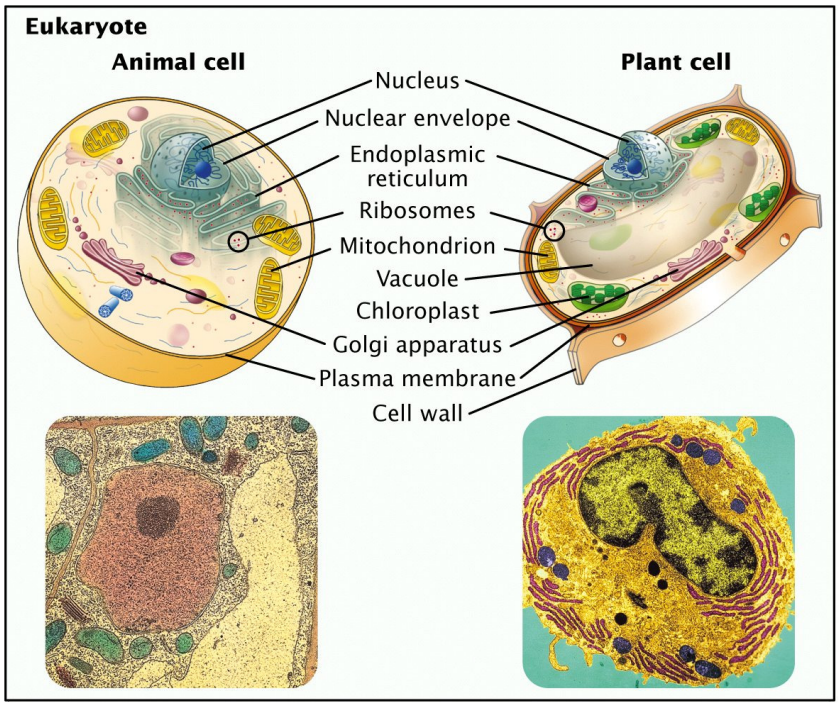


Figure 2: Eukaryotic cell structure

Endosymbiosis

- An ancestral cell engulfed a smaller microbe which continued to survive inside it
- Both host and prey adapt to new environment and become mutually interdependent

1.1.3 Cell functions

Functions of cells provided by organelle.

Nucleus
: Regulate DNA carrying/replication

Endoplasmic reticulum
System of folded membranes used for transport

Ribosomes
Used to produce proteins

Mitochondria
Provide energy through cellular respiration

Vacuole
Used for storage

Chloroplasts
Used to convert light energy into chemical energy for cell

Golgi Apparatus
Packages and transports proteins

	Prokaryotic cells	Eukaryotic cells
Nucleus	Absent	Present
Cell diameter	$1 - 10\mu$m	$10 - 100\mu$m
Genome	One circular module	Multiple linear modules
DNA	Not complexed in eubacteria, some in archaea	Complexed with histomes
DNA quantity	Small	Large
Membrane-bounded organelles	Absent	Present
Cytoskeleton	Absent	Present

Table 1: Comparison of cell types

1.1.4 Phenotype from Genotype

- Characteristics of an organism (phenotype) are determined by the structure and function of its cells (genotype)

1.2 DNA and genes

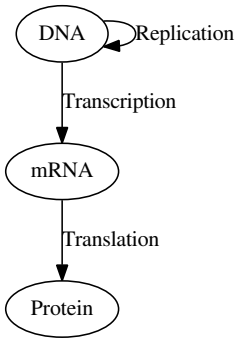


Figure 3: Central dogma of molecular biology

1.2.1 Structure of DNA

- Base
  Bases bind (A-T, C-G) by Hydrogen bonds
  - Pyrimidines
    * Cytosine
    * Thymine
  - Purines
    * Guanine
    * Adenine
- Backbone
  - Phosphate
    Binds ribose sugar
  - Ribose sugar
    Binds phosphate to base

DNA directionality

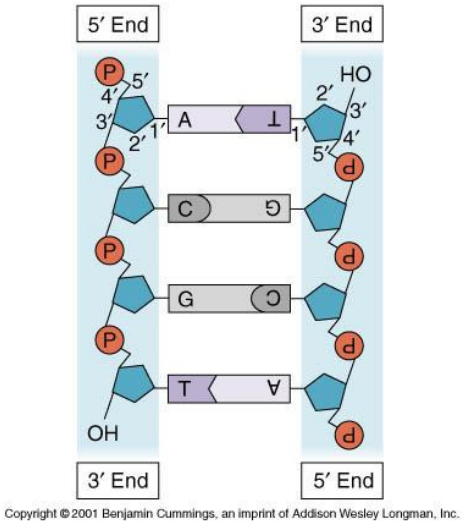


Figure 4: DNA directionality

1.2.2 DNA replication

- Replication is essential for organisms to reproduce
- Without replication, cell division could not occur
- Replication must be high fidelity but have potential for error (to enable evolution)

C-value paradox

The C-value is the size of an organisms genome (defined as the amount of DNA in pico grams in a haploid cell).

C-values not reflective of complexity, i.e. no relation between C-value and number of genes.

1.2.3 Genome structure

Prokaryotic

- Generally small (< 10M bases)
- Single, circular chromosome
- Can have plasmids
  Disposable genetic elements that often carry genes for antibiotic resistance or toxicity
- Information dense
  Little or no space between genes, genes often overlap
- Genes are present in single, uninterrupted units
- Multiple genes may be present in single, uninterrupted units

Eukaryotic

- Usually larger than prokaryotic genomes
- Multiple, linear chromosomes
- Tend to be more information sparse
  Large gaps between genes
- Genes interrupted by non-coding sequence

  Exons coding

  Introns non-coding

1.2.4 DNA packaging

Nuclear DNA is packaged as chromatin, a structure of DNA, RNA and protein.

Purpose of chromatin:

- Packages DNA into a more compact structure
- Reinforces macromolecule to allow mitosis
- Prevents DNA damage
- Regulates gene expression and DNA replication

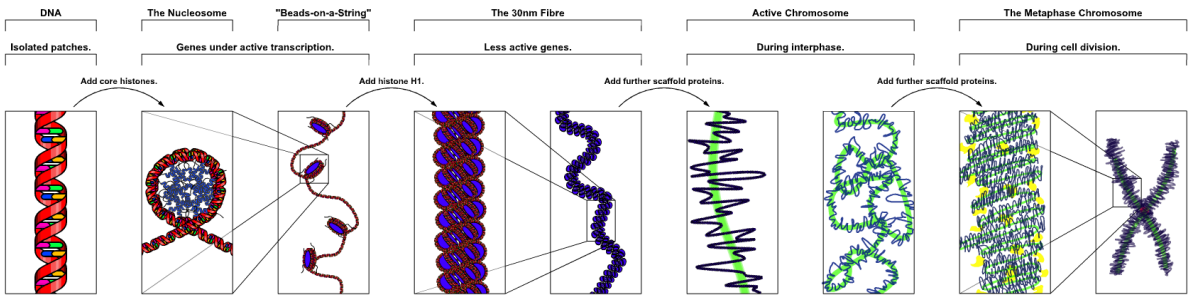


Figure 5: Chromatin structure

1.2.5 Manipulating genomes

- Genetic engineering gives multiple techniques for editing the genome of an organism
- Simplest example is artificial selection to obtain desired traits
- Synthetic biology apples engineering principles to biology

1.2.6 Gene structure

- Genes encode functional RNA modules
- A gene is a functional part of a chromosome
- Every cell contains the same set of genes

1.3 Transcription

Process of turning information stored in DNA into RNA.

Non-template (coding) strand
Contains same base sequence as RNA crated

Template (non-coding) strand
Contains anti-codons of RNA

Promoter
Denotes start of RNA coding region

Coding region
RNA coding

Terminator
DNA coding denoting the end of the coding region

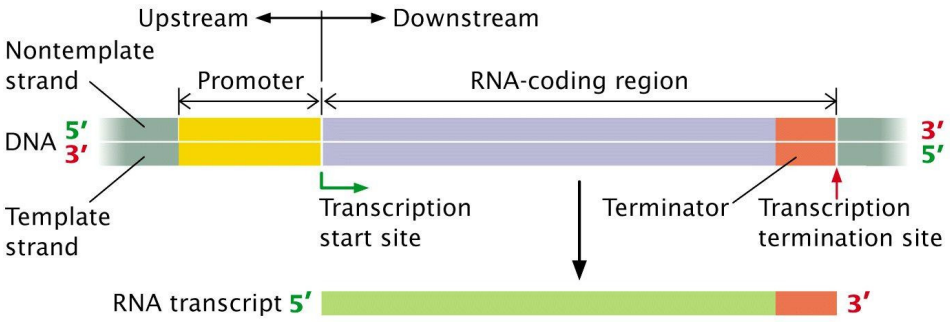


Figure 6: DNA directionality

- The base Thymine (T) in DNA is replaced with Uracil (U) in RNA
- Transcription occurs 5′ to 3′
- RNA polymerase is the enzyme responsible for transcription
- Transcription in eukaryotes is more complex

1.3.1 Transcript processing

Primary transcript must be processed into a mature message.

- Addition of $5'$ cap
  - 7-methlyguanosine
  - Involved in ribosome binding
  - Protective
- $3'$ polyadenylation
  - Addition of tail of adenosine to mRNA
  - Required for nuclear export and stability
- Splicing
  - Primary transcript contains both introns and exons
  - Introns must be removed

1.4 Translation

The process of transforming the information contained in mRNA into an amino acid chain, which is folded into a protein (amino acids joined by peptide bonds).

- Proteins may be structural or have an active role (e.g. enzymes)
- Protein function sis determined by its amino acid sequence and its three dimensional structure

1.4.1 RNA to protein

Nucleotide sequence is read in groups of three (codon). Codons are consecutive and non-overlapping.

Each codon corresponds to one of 20 amino acids.

Special cases:

AUG Methionine indicates start of frame

UAA Frame terminator

UAG Frame terminator

UGA Frame terminator

1.4.2 Amino acid structure

- Common structure to all amino acids
- Side chain defines properties of amino acid

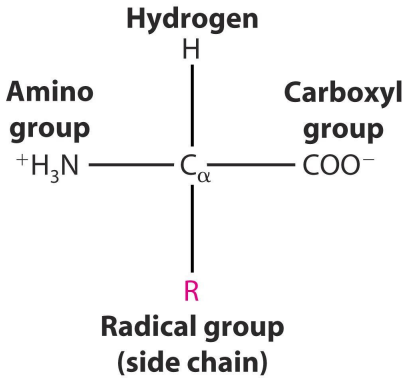


Figure 7: Amino acid structure

Protein is constructed when multiple amino acids are combined into a string by a peptide bond between the carboxyl group of one acid and the amino group of another (giving H_2O as a by-product).

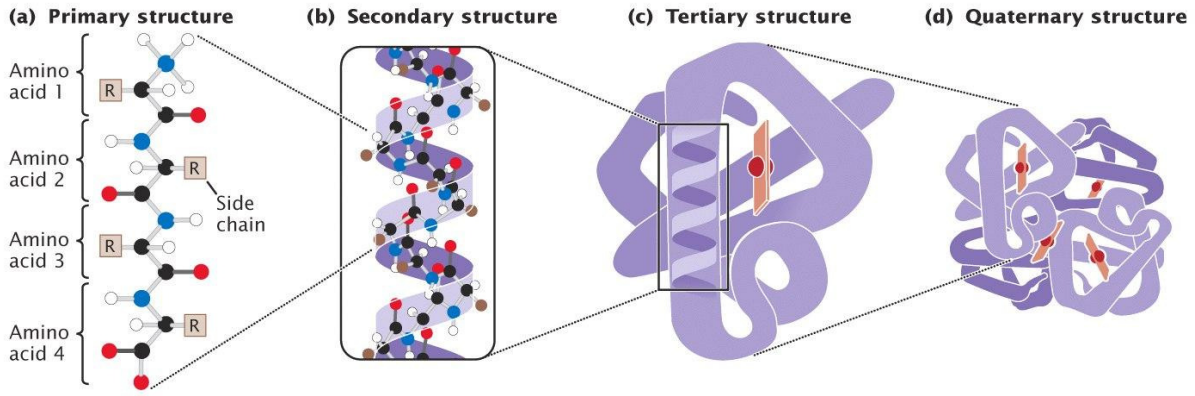


Figure 8: Protein structure

1.4.3 Ribosomes

- Bind to the region of the mRNA that lies just upstream of the AUG codon
- Ribosome Binding Site (RBS) given a consensus sequence which varies with organism

1.4.4 Translation process

0 Binding of amino acid to tRNA

1 Initiation

- Ribosome recognition
  - Prokaryotes: mRNA RBS hydrogen bonds with rRNA
  - Eukaryotes: ribosome attaches to 5′ cap
- $tRNA^{met}$ binds to AUG codon

2 Elongation

3 Termination

- Ribosome pauses at stop codon
- Release factor binds

1.4.5 Control of Gene Expression

- Various levels of control:
  - Transcription
  - Post-transcription
  - Translation
  - Post-translation

2 Genome Sequencing, Assembly and Annotation

2.1 Sequencing

Reasons for sequencing:

- Identify content and structure
- Identify mutations

The process of sequencing produces reads. A read is a single continuous sequence read by the sequencing process.

2.1.1 Sanger sequencing

- Uses DNA polymerase to synthesise DNA

  Enzyme that when given a DNA strand will synthesise the complementary strand
- A mixture of the polymerase, a primer (small strand of complimentary DNA) and deoxynucleotides (DNA nucleotides with fluorescent tag) is made.
- Products then separated. Multiple methods exist but traditionally using gel electrophoresis which allows florescent markers to be seen in four distinct "lanes" corresponding to each DNA nucleotide.
- Accurate but slow

2.1.2 454 pyro-sequencing

- High throughput and lower sample preparation time

2.1.3 Illumina Solexa approach

- Sequencing based on synthesis
- Lower cost per sequenced base but high initial cost

2.1.4 Nanopore

- Feed DNA through a pore in a surface and detect nucleotides as they pass through
- Faster with minimal sample preparation

2.2 Assembly

Sequencing produces several fragments of the full genome, in order to obtain the sequence of the full genome such subsequences must be assembled.

Individual reads are assembled to form contigs. A contig is a group of copied pieces of DNA representing overlapping regions of a chromosome.

Contigs are assembled to give complete sequences.

2.2.1 Shotgun sequencing

- Long sequences are assembled based on overlaps between several subsequences
- Region in full sequence should ideally be covered by several subsequences to ensure good confidence

  For Sanger sequencing 6 fold redundancy is required
- Nucleotide at a given position in the final sequence is given by the most common nucleotide in the aligned reads/contigs
- Assembly can be parameterised in terms of number of mismatches between aligned reads/contigs allowed at a given point in the final sequence
- Computationally intensive method: uses naive string searching which becomes very slow for large genomes and large number of reads

2.2.2 deBruijn assembly

- Split data into a series of small equal length oliomers where oliomers are generated by advancing a sliding window along a sequence one nucleotide at a time
- Build a deBruijn graph of all oliomers where edges represent an alignment between the two joined oliomers
- If successful the graph structure will be a single string containing the aligned sequence, where the sequence forms an Euclidean path through all nodes
- The choice of the length k of the oliomers is critical to the algorithm operating correctly.

  A sub-optimal value of k will cause closed loops in the graph which prevents the correct sequence being obtained

2.2.3 Finishing

After alignment several gaps may be left in the final sequence, leaving the result as a series of contigs.

Can be caused by lack of experimental data or issues with the wet experiment, i.e. AT/GC right regions, homopolymenrs, etc.

Ideal solution is to perform more wet experiments.

2.2.4 Templates

When assembling the sequence of a new organism with known relatives the genomic sequence of the relative can be used to assist positioning of the individual contigs.

This is done by aligning the contigs to the genomic sequence of the relative.

2.3 Annotation

Consensus sequences are analysed to discover features details of which are stored along with the sequence and the region of the sequence it is relevant to.

Typically a time and resource intensive task in the form of a pipeline of individual tools used to identify specific features.

2.3.1 Gene finding

Open reading frames are a region starting with the initialisation codon (`ATG`) and ending in one of three stop codons.

Gene finding can be done using one of two main approaches; comparative genomics or ab initio.

Comparative genomics uses similarity between regions in a sample and other known samples to infer the structure of a gene.

Ab initio methods are based on determining gene structure based solely on the experimental data and known biological theories. External data can be used to improve prediction accuracy.

Prediction of protein coding genes in bacterial genomes is simple and relatively accurate. In eukaryotes predicting intron-exon regions is only around 65% accurate and non-coding exons can cause issues.

3 Evolution

Key relevance is similarity between close evolutionary ancestors which aids prediction of features such as protein function and structure from a sequence.

3.1 Mutations

Point Mutation

When a single nucleotide in the DNA sequence changes.

This can have several different effects on the protein sequence:

Silent
The change in DNA sequence does not change the protein sequence.

e.g. the mutation `TTC` $\rightarrow$ `TTT` gives `Lys` $\rightarrow$ `Lys`

Nonsense
The change in DNA sequence causes a change in protein sequence that does not make sense.

e.g. the mutation `TTC` $\rightarrow$ `ATC` gives `Lys` $\rightarrow$ `STOP`

Missense
The change in DNA sequence causes a change in protein sequence that is valid but may affect the function.

This can be either conservative in which case the new protein has a similar function to the previous or non-conservative in which case the functional change is significant.

Deletion

Where a section of the sequence is removed.

Duplication

Where a section of the sequence is duplicated and inserted after the original section.

Inversion

Where a section of the sequence is reversed in the new sequence.

Insertion

Where a section of the sequence of one chromosome is inserted into the sequence of another chromosome.

Translocation

Where two sections of the sequence of two different chromosomes are swapped.

Frameshift mutations

Mutations where the relative location of coding information is changed due to the insertion, deletion or exchange of a subsequence of a different length to the original.

3.1.1 Gene duplication

- Evolutionary constraints define the tolerance of how much the gene can diverge from its original function after an evolutionary cycle.
- After gene duplication duplicates have reduced evolutionary constraints as the original copy still exists therefore retention of the original function is less important.
- Gene duplication leads to wider divergence in function and evolution of new proteins.
- In bacteria gene duplication can be triggered by external stimuli.

  e.g. temperature, pressure, starvation, etc.
- In eukaryotes gene duplication is vital for evolution. Around $\frac{1}{3}$ of a eukaryotic genome consists of duplicates.
- Duplicate genes with redundant functions can help towards protection against deletion mutations.

3.2 Homology, orthology and parology

Homology
Genes that share a common ancestor.

Orthology
Homologous genes which diverged as a result of speciation.

Parology
Homologous genes within the same genome created as a result of gene duplication.

Xenology
Homologous genes where one gene has been obtained through transfer of genetic material between organisms.

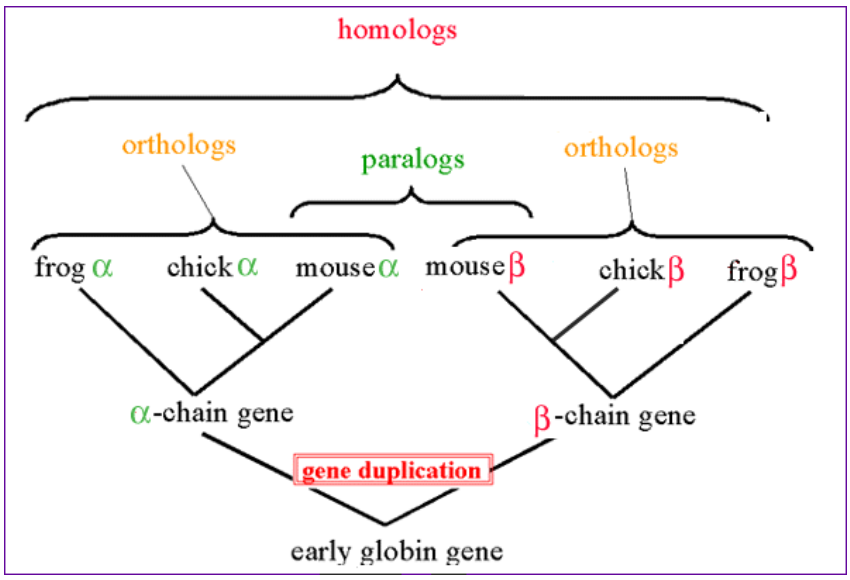


Figure 9: Relationship between homology, orthology and parology

3.3 Homology

- Proteins and genes that have diverged from a common ancestor are homologous.
- Homology is a boolean state, not analogous to sequence similarity.
- Homology cannot be definitively determined from sequence similarity alone, it can however be inferred if the sequences are more than 50% similar.
- Homologous proteins often share properties: same/related function, sites, structure.
- As sequences diverge non-conservative mutations increase and the sequence similarity to the common ancestor decreases while remaining functionally similar, this can make determining homology difficult.

3.4 Orthology

- Homologous genes from different genomes within the same gene family.
- Derived as a result of speciation.
- Often encode proteins that perform the same function in different organisms.

3.5 Parology

- Homologous genes within the same genome within different gene families.
- May perform different functions in each host due to functional change with evolution.

4 Phylogenetics

- The study of evolutionary relationships between groups of organisms.
- Produces hypothesis about evolutionary history of organisms.
- Species (genes) typically shows as evolving as per a tree structure.
- Based on inferring information about evolutionary history from present day knowledge.
  Phylogenetics are hypotheses.
  Changes as new data becomes available.
- Can use molecular data to infer phylogenetic relationships.
- Publish phylogenetics tend to be low quality.

4.1 Workflow

4.1.1 Obtain sequences

- Selection of sequences is dependant on the research.
- Typically use well conserved genes.

4.1.2 Alignment

- Used to provide data from which to create a hypothesis about the relationship between sequences.

4.1.3 Masking

- Separate areas of high similarity from those of low similarity.
- Areas of high similarity infer homology, low similarity cannot be used to infer either way.

4.1.4 Evolutionary model fitting

Multiple methods:

- Unweighted pair group method with arithmetic mean (UPGMA) (simple)
- Maximum likelihood (advanced)
- Bayesian (advanced)

Unweighted pair group method with arithmetic mean

- Uses distance matrix that tabulates the distance between each pair of species.
- Each iteration merge two closest species and recalculate matrix values based on mean of merged species.
- Repeat until all species are grouped.
- Order of grouping gives tree structure.

Example with 5 species: `A`, `B`, `C`, `D`, `E`.

Species	A	B	C	D
B	9	-	-	-
C	8	11	-	-
D	12	15	10	-
E	15	18	13	5

Table 2: UPGMA iteration 1

Merge `D` and `E` (shortest distance 5):

Species	A	B	C
B	9	-	-
C	8	11	-
DE	13.5	16.5	11.5

Table 3: UPGMA iteration 2

Merge `A` and `C` (shortest distance 8):

Species	B	AC
AC	10	-
DE	16.5	12.5

Table 4: UPGMA iteration 3

Reconstruct tree from order of merges:

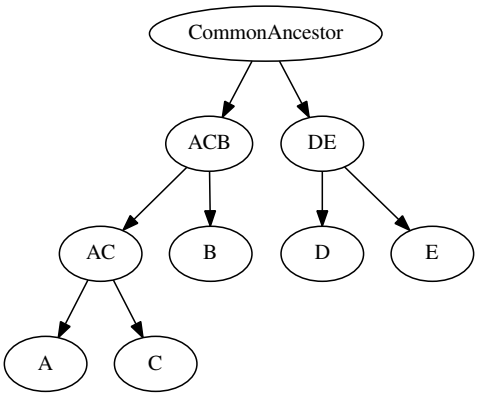


Figure 10: Phylogenetic tree created using UPGMA

4.1.5 Analyse tree

- Tips correspond to modern species.
- Position of fork in tree in terms of depth is reflective of the point in history that the two organisms diverged.

5 Sequence Similarity and Comparison

- Similarity score required for sequence alignment

5.1 String based similarity

- Typically not used for biological sequences
- Certain changes are more likely to occur as a result of evolution

5.1.1 Hamming distance

- Number of positions with mismatching characters in two strings of equal length
- Lower score denotes similar sequences

5.1.2 Levenshtein distance

- Minimum number of edits to change one string into another
- Lower score denotes similar sequences
- A change can be an addition, deletion or substitution

5.2 Sequence Similarity

- Assignment of variable weights to each misalignment
- Additive step scores
- Gap penalty for areas of no alignment

DNA and RNA

- Typically +1 for match and -1 for mismatch
- Complex schemes may take into account that a purine-purine or pyrimidine-pyrimidine mismatch is more common that a purine-pyrimidine mismatch

Proteins

- Learn a matrix of scores based on known amino acid changes during evolution
- Scores derived from frequency of of amino acid substitutions
- Data encoded in PAM matrices

5.2.1 Gap penalty

- Penalise gaps in sequence alignment
- Larger penalty added for opening a gap than extending an already open gap
- Typical penalties for DNA:

  Opening 10
  Extending 0.1

- Typical penalties for proteins:

  Opening 11
  Extending 1

5.2.2 Percent Accepted Mutation (PAM) matrices

- Based on knowledge of evolution
- Probability of a given change can be determined from alignment of homologous sequences
- Relative frequencies of amino acid changes are encoded in the PAM scoring matrix as the probability of mutation to every other amino acid
- Sample data is restricted to sequences sufficiently similar to ensure multiple substitutions have not occurred on the same site
- A PAM matrix measures sequence divergence

  i.e. a 1 PAM substitution matrix is generated from two sequences that are 99% similar
- To assert similarity of more diverse sequences need matrices for much lower sequence similarity
- Can derive PAM matrices for lower similarity by taking powers of existing matrices
- A range of PAM matrices have been generated in this way and the divergence percentage is used in the name of the matrix (e.g. PAM250 denotes expected change of 250%)
- For use in protein comparison the PAM matrix is first normalised and the logarithm taken (known as a log-odds matrix)

  This allows the scores to be added rather than multiplied (as would have had to be done with probabilities)

5.2.3 BLOSSUM matrices

- Multiple alignments of short, continuous (without gaps) sequences are arranged into blocks
- Substitution frequencies for all pairs of amino acids are calculated for each block
- Used in a similar way to PAM matrices to create log-odds scores for amino acid substitution
- Similarity of protein sequences in a block can be adjusted to obtain different similarity matrices
- All BLOSSUM matrices are derived from direct measurements

  i.e. not extrapolated as is done for PAM
- BLOSSUM allows for detection of more distantly related sequences than PAM by reducing the importance of more similar sequences in the block when building the matrix
- BLOSSUM62 (default for many comparison tools) created from blocks where 62% of the amino acids were identical

5.3 Sequence Alignment

- Aligning sequences is useful to determine how similar they are
- Alignment also shows which parts of the sequence are similar
- If two proteins have over 45% identical residue structure then they are likely to have a common or related structure
- If they have over 25% identical residues then they are likely to have similar folding patterns

5.3.1 Exact vs. heuristic

- Exact
  - Guaranteed to find an optimal solution
  - Computationally intensive
  - Good for small numbers of sequences or when algorithm is implemented in hardware
- Heuristic
  - Not guaranteed to find an optimal solution
  - More computationally efficient
  - Better for searching large databases

5.3.2 Global alignment

- Attempts to align all residues in sequence
- Useful when sequences are very similar and of equal length

Exact method: Trajectories

- Trajectory visualised by a 2D matrix with each sequence as indexing row and columns
- Diagonal arrow denotes an alignment between residues
  This can be either a match or mismatch
- A horizontal arrow denotes a gap in the sequence indexing the columns
- A vertical arrow denotes a gap in the sequence indexing the rows

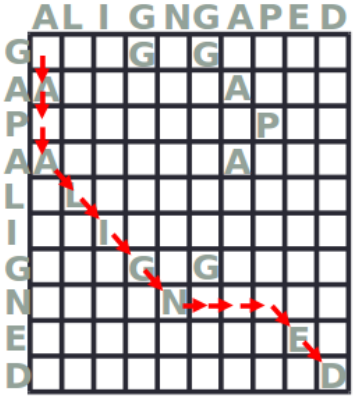


Figure 11: Trajectory example

Global alignment: Needleman Wunsch algorithm

- Computationally intensive
- Can use PAM, BLOSSUM, etc. as a similarity matrix to obtain alignment score
- Construct matrix of dimensions (m, n) where m and n are the sizes of each sequence
- A cell in the matrix (i, j) will contain the optimal score of aligning the first i residues of the first sequence with the first j residues of the second
- Matrix is filled incrementally
- The value of cell (i, j) is determined by the value of cells $(i-1, j)$, $(i, j-1)$ and $(i-1, j-1)$

  The value is the lowest movement score given by the score of the last cell plus the score of the alignment of the current cell (i.e. score of mismatch, gap opening or gap extending)
- An arrow is placed from the cell in the direction of the lowest score
- When all cells are filled cell (m, n) will contain the optimal score of the global alignment
- By following the arrows from cell (m, n) one can obtain the optimal global alignment(s)

  Note that multiple equally scored optimal alignments may be possible, each indicated by branching and joining in the path

Example:

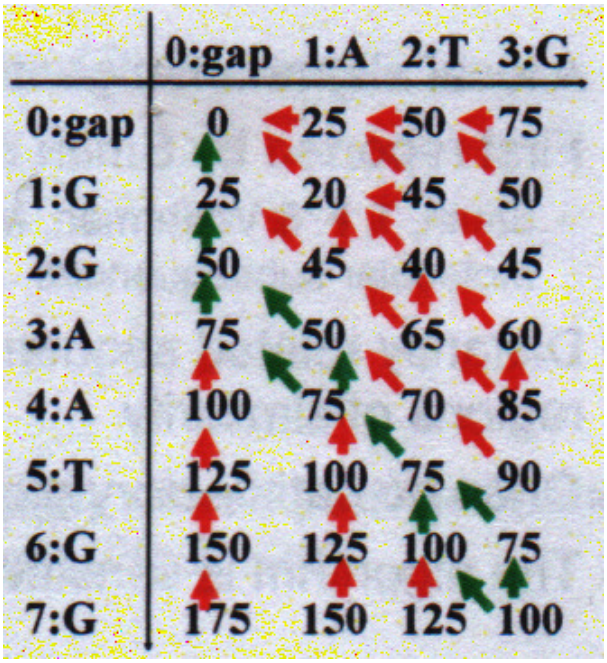


Figure 12: Needleman Wunsch alignment example

Four optimal alignments for figure 12:

- ggaatgg
  ---atg-
- ggaatgg
  ---at-g
- ggaatgg
  --a-t-g
- ggaatgg
  --a-tg-

5.3.3 Local alignment

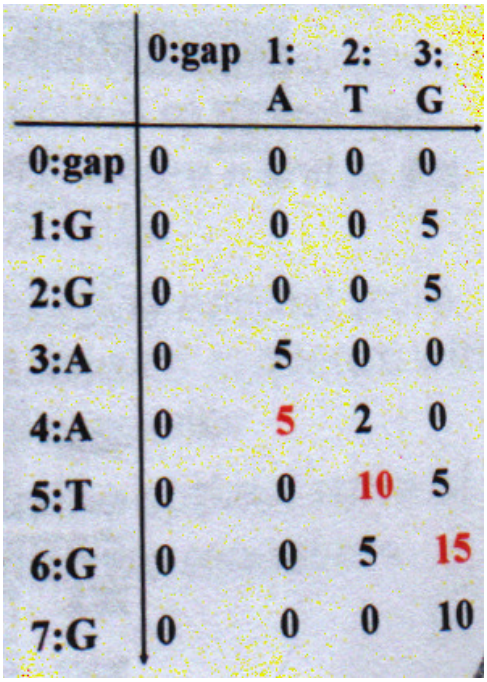
- Aligns subsection of sequences that align best
- Useful when sequences contain sub regions of similarity

Local alignment: Smith Waterman algorithm

- Variation of global alignment method using matrix
- At initialisation all cells are set to zero
- Gaps at the beginning of the sequence are not penalised
- The additional option of ending the sequence is also considered when deciding the score of a given cell
- The optimal score (end of the alignment) can appear anywhere in the matrix

- The start of the alignment can appear anywhere before the end of the alignment in the matrix

Example:



	0:gap	1: A	2: T	3: G
0:gap	0	0	0	0
1:G	0	0	0	5
2:G	0	0	0	5
3:A	0	5	0	0
4:A	0	5	2	0
5:T	0	0	10	5
6:G	0	0	5	15
7:G	0	0	0	10

Figure 13: Smith Waterman alignment example

Scores are given by:

- Match = +5
- Mismatch = -3
- Gap = -5

Value of cell (i, j) is given by:

$$S(i,j) = max \begin{pmatrix} S(i-1,j-1) + mis/matchscore \\ S(i-1,j) + gapscore \\ S(i,j-1) + gapscore \\ 0 \end{pmatrix}$$

Where 0 ends the alignment.

Approximate method: k-tuple

- Take a sample sequence and split it into all possible incremental subsequences of a given size k
  For proteins: $k = 3$
- Search a database for each substring
  Reduces number of candidate strings
- Then extend alignment using sequences of good alignment with substrings

BLAST algorithm

- (Basic Local Alignment Search Tool)
- 100 times faster than exact dynamic programming approaches

Workflow:

1 Split into k letter substrings.

e.g:
```
MATCHES
MAT....
.ATC...
..TCH..
...CHE.
....HES
```

2 Align each sequence to all possible k length subsequences

3 Compute the similarity scores of all alignments

- Discard those that have a score less than the neighbourhood score threshold
- This greatly reduces the number of alignments to be considered

4 Search database for exact matches of each word

5 Ungapped alignments

- If for the same sequence in the database, two non-overlapping hits are found less than a threshold number of residues apart
- In this case the alignment is extended starting with these hits as long as the score is no worse than a given threshold respective of the best score found so far

6 Gapped alignments

- Dynamic programming method is used for gapped alignments

E value

- Describes the number of hits expected when searching a database of a given size
- Decreases exponentially as the score increases
- Essentially describes the amount of background noise

  e.g. $E = 1$ denotes that in a database of the current size, it is expected to find 1 match with a similar score by chance
- Lower E value denotes a higher probability that the sample sequence has matched a homologous protein

  However, almost identical short alignments have a high E value as the length of the sequence is taken into account

5.4 Multiple Sequence Alignment

- Multiple sequences laid out in a grid such that:
  - Relative positions of residues within any one sequence is conserved
  - Similar residues in all sequences are brought into the same vertical register
- Exact approach is too computationally expensive to use
- Can be performed or interpreted by hand/visually by colour coding each residue
- Requires a lot of computational power.
- Often too computationally intensive to use exact methods
  Most commonly used algorithms are heuristic based
- Methods:
  - Progressive alignment strategies (e.g. Clustal, T-Coffee)
  - Iterative methods (e.g. DIALIGN, PRRP)
    Make an initial estimate then iteratively refine it
  - Statistical methods (e.g. HMMER, SAM)
    Generate probabilistic models of sequences, e.g. using a hidden Markov model
  - Based on locally conserved patterns found in same order in sequence

5.4.1 Phylogenetics

- Starting point for phylogenetic analysis
- Can be used to group sequences or subsequences into families
  Given a multiple sequence alignment each alignment column reflects mutations at that site during evolution
- A multiple sequence alignment allows the deduction of the order of appearance of sequences during evolution
- Both global and local alignment used:
  - Protein sequences can be conserved throughout evolutionary change (global)
  - Functional domains of protein sequences may be conserved whilst the sequence diverges (local)

5.4.2 Algorithm: Clustal

- Progressive alignment based
- Very fast
  Can align 10^5 sequences in a couple of hours
- Can be parallelised

Workflow:

1 Compare sequences to obtain a similarity matrix

- Genetic distances between each pair of sequences is computed: $\frac{\#\ mismatched\ positions}{\#\ matched\ positions}$

2 Using similarity matrix, create a tree that relates all sequences

- Computed genetic distances are used to build a phylogenetic tree of all sequences, known as the "guide tree"
- This tree is used to control the order in which sequences are added to the multiple sequence alignment
- Contributions of each sequence to the alignment are weighted by the position of the sequence in the guide tree

3 Perform progressive alignments in which the sequences are aligned in an order determined by the tree

- Most closely related pair(s) are aligned first
- Next closes sequence is then added to alignment until all sequences have been aligned
- Sequences are added to alignment by performing an alignment between the new sequence and the existing alignment

5.4.3 Algorithm: T-Coffee

- Progressive alignment based
- Used within a family of other algorithms for tasks such as accuracy reporting, combining multiple sequence alignments, etc.
- Every possible pair-wise alignment computed using a hidden Markov model which are used to build a library
- A progressive alignment builds up a multiple sequence alignment using pair-wise alignments from the library

  Alignment with highest possible agreement is selected
- Can use any pair-wise alignment method to build library

  Adds flexibility

5.4.4 Algorithm: Muscle

- (MUltiple Sequence Comparison by Log-Expectation)
- Faster for a large multiple sequence alignment
- Relatively complex algorithm
- Two progressive alignment steps followed by an iterative refinement step

Workflow:

1 Progressive alignment 1

- Based on local similarity using k-tuple matching
- (draft progression)

2 Progressive alignment 2

- Based on global alignment
- (improved progression)

3 Iterative refinement

- Alignments are split apart
- Parts are realigned using local conserved regions
- If the alignment is better the changes are saved and the process is repeated

5.4.5 Analysis

Motif

- Region that is conserved between families of proteins
- Typically conserved because they encode part of the function of the protein
- Presence of a motif in a protein sequence can be used to imply function

Pattern

- An amino acid sequence that is related to a motif
- Presence of this pattern denotes a conserved region between protein sequences
- A pattern defines a motif in terms of amino acid sequence

Profile

- An extension of a pattern
- Assigned probabilities of to the occurrence of a particular amino acid at each position of the motif
- Consists of a table of amino acid and gap costs for each position and a set of probabilities for each amino acid at each position
- Used for alignment of more distantly related sequences

5.4.6 Protein Databases

- Several databases have been created which describe motifs in terms of patterns and profiles
- Allows searching for patterns and profiles in a protein query sequence to find possible functions
- The Interpro database combines data from multiple sources to provide a database of protein families and functional sites
- Protein domains typically relate to the conservation of 3D structure
- Multiple sequence alignments can be used to obtain better quality results from searching such databases
  - More sensitive search methods that allow detection of more distant relationships
  - Reducing the number of false reported homologous sequences

PSI-BLAST

- (Position Specific Iterative BLAST)
- More powerful than BLAST for detection of distant relationships
- Starts with a normal BLAST execution
- Derives a pattern from the multiple sequence alignment of the BLAST hits, known as the Position Sensitive Scoring Matrix
  - For each residue in the sequence, compute the distribution of amino acids in the corresponding residues in aligned sequences
    (discard those too similar to the query)
  - Distribution describes the likelihood of each mutation for each residue of the query sequence
  - Also describes which residues are more conserved and which are more susceptible to insertions or deletions
- Uses this pattern information to search the database
- Process is repeated to allow the pattern to be tuned in successive cycles

Hidden Markov Models

- Typically can perform better than PSI-BLAST
- Very good for detecting distant relationships
- Computational structure describing the patterns that define families of homologous sequences
- e.g. HMMER
  - Takes a multiple sequence alignment as input
  - Builds a hidden Markov model form this data
  - Model can be used to query a sequence database to locate homologs

6 Protein Structure Prediction

- Molecules of primary importance to life
- Polypeptide chains created by joining amino acids
- Linear chain folds to produce a 3D structure
- Final 3D structure depends on amino acid composition of the protein

6.1 Amino acids

- All amino acids have a common backbone which is used to form the linear string of amino acids that composes a protein
- The side chain is the unique part of the amino acid and attaches to the backbone by the C_α atom

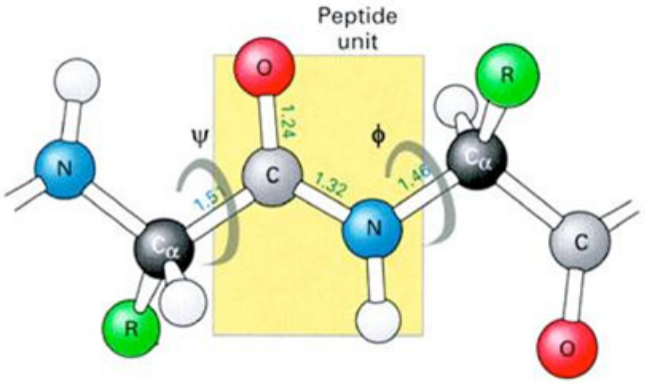


Figure 14: Protein structure

- Different amino acids have different properties
- e.g. Hydrophobic, small, polar, etc.
- Hydrophobicity has the largest effect on the final 3D structure

6.2 Structure

Hierarchical structure:

Primary structure
The amino acid sequence of the protein

Secondary structure
Local interactions between closest residues in the sequence that form one of two structures; α-helices or β-sheets

Tertiary structure
Global interactions of α-helices and β-sheets cause organisation into a stable state to form a complex 3D structure

Super-secondary structure (lesser used)
Recurrent patters of interaction between secondary structures in close proximity

e.g.:

β **hairpin**
An anti-parallel β-sheet arranged in a hairpin shape

$\beta - \alpha - \beta$ **unit**
An α-helix in a strand between two β-sheets, all of which are arranged parallel to each other

Domain (lesser used)
Sub-units within a protein with similar folding stability

- Polypeptide chain is joined such that two angles exist at each side of the peptide unit
  - Ψ, between C_α and N
  - Φ, between C_β and C
  - These angles determine the 3D structure of the protein chain
- An α-helix is the formation of a coil 3.6 residues per turn 5.4 Å wide
- A β-sheet is the formation of one or more straight (parallel) chains
  - If all strands have the same N-C orientation the sheet is known as a parallel sheet
  - If adjacent strands are opposite to each other the sheet is known as an anti-parallel sheet
- Any residues that are not part of an α-helix or β-sheet are said to be in coil state
- Structure of experimentally observed proteins stored in the Protein Data Bank (PDB)

6.2.1 Ramachandran plot

- Plots relationship of Ψ and Φ angles to secondary structure
- Range of observed angles is very constrained and secondary structure types are often clustered
- Raw measured data is often not so clustered

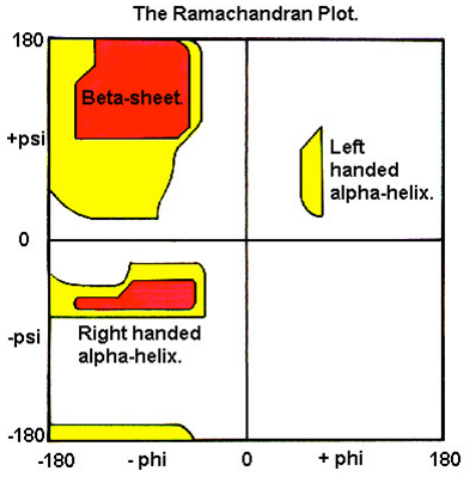


Figure 15: Ramachandran plot

6.2.2 Classification

- Several tertiary structure classifiers exist
- CATH is most up to date
- Uses hierarchical system to catalog proteins:

  1 Class
  - Mainly Alpha
  - Mainly Beta
  - Alpha Beta
  - Few Secondary Structures

  2 Architecture

  3 Fold

  4 Super-family

  5 Domain

6.3 Structure Prediction

- Protein function is dependant on its structure
- Function is determined by which amino acids are exposed to the outside of the protein, which depends on the 3D structure
- Structure determines what substrate the protein can react with
- Protein structure is very difficult (or sometimes near impossible) to measure experimentally
- Protein structure prediction aims to predict the 3D structure of a protein based on its amino acid sequence
- Prediction of 3D structure is typically an optimisation problem
- Prediction of structural aspects of residues (such as secondary structure, contact points, etc.)are typically machine learning problems

6.3.1 Prediction of structural aspects

- Many features are present due to local interactions between amino acids and its close neighbours in the protein chain
- Need to consider more than just the primary structure (sequence)
- Also consider evolutionary information (which residues are conserved)

  Data from Position Specific Scoring Matrices (PSSM) which are created from a multiple sequence alignment

Secondary Structure

- Predicts if a reside belongs to an α-helix, β-sheet or is in coil state
- Uses a window of 15 amino acids (7 either side of the residue being predicted)
- Prediction is typically performed by creating a probability profile of the sequence using a PSSM then by considering each window of the probability profile

Contact Number

- Two residues are said to be in contact if they are less than a threshold distance apart
- The contact number is the number of contacts that a given residue has
- The contact number of each residue gives a vague indication of the density of the protein

Contact Map

- Given two residues determine if they are in contact
- Typically represented as a binary matrix
- Can be used to determine a lot of information about the structure of a protein
  - Basic 3D model
  - Model selection
  - Parameters to energy function of 3D protein structure prediction

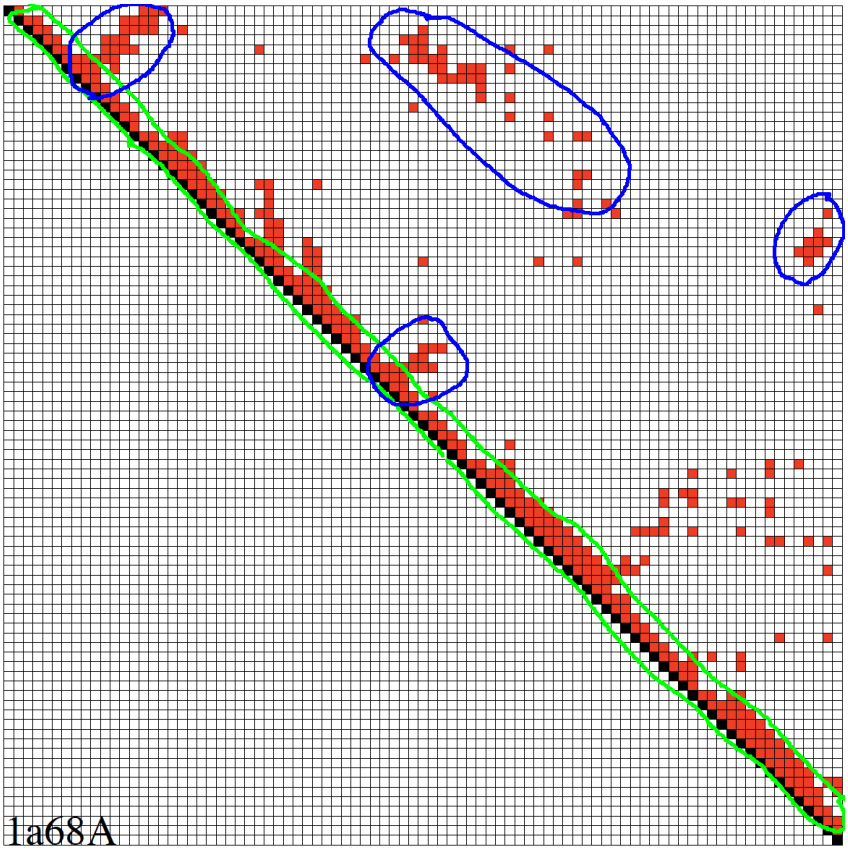


Figure 16: Contact Map

Figure 16 shows a sample contact map. Regions that denote α-helices are indicated in green and those that denote β-sheets are indicated in blue.

Contact map prediction at Newcastle:

- Machine learning approach to contact map prediction

- Use three residue windows:
  - Centred on first residue
  - Centred on second residue
  - Centred on median residue between first and second
- Training data of 2413 proteins with less than 2% real contacts
- Split training set into 50 samples
- BioHEL algorithm used to train 25 rule sets for each of the 50 samples
- Contact is predicted using consensus voting over all 1250 (50×25) rule sets

Other Predictions

- Solvent Accessibility

  Surface area of a residue that is exposed to the environment
- Recursive Convex Hull

  Assign a "layer" to each residue representing their relative proximity to the "centre" of the protein where each layer is a convex hull of points

6.4 3D Structure Prediction

6.4.1 Template based modelling

- Try to identify a template protein
- Adapt the structure of the template to the target protein
- Energy optimisation step aims to find a "at rest" state for amino acid chain in terms of forces acting on atoms

Workflow:

1 Template Identification

- Attempt finding a close homologous sequence, this will have the best change of having a similar structure
- If a template exists but has low sequence similarity it must be found by other means

  Profile based methods
  - Construct 1D representations (profiles) of structures in database
  - Construct profile of target sequence and search for closest matching profile in database
  - Profiles are constructed from structural properties of residues
    - Secondary structure state
    - Solvent accessibility
    - Amino acid properties

  Threading methods
  - Create a catalogue of unique (structural) folds
  - Evaluate how likely it is that the target sequence adopts to each fold and how it adapts (alignment)
  - Choose template (and alignment) that has the lowest estimated energy

2 Alignment

- It is possible that the alignment between the template and target sequences means that atoms are placed in incorrect positions in the 3D structure model
- The offset (alignment) between template and target sequences must therefore be adjusted to correct this
- To correct this information can be used form:
  - Information derived from template structure
  - Predictions about the target

3 Alignment gap correction

- Determine main chain segments of the target that represent the regions containing gaps in alignment
- Stitch them into the main chain of the template to crate an initial model for the target

4 Populate mutations

- Replace the side chains of residues that have been mutated (mismatches in alignment)

5 Atomic position validation

- Inspect the positions of atoms to detect any collisions
- Resolve any serious collisions

6 Refinement by energy optimisation

- Adapt stitched segments to conserved structure
- Adjust side chains to most stable configuration

6.4.2 Ab initio methods

- Without a template sequence ab initio modelling is the only option
- Pure ab initio is costly and ineffective
- Hybrid ab initio/homology methods exist and have better performance
  e.g. fragment assembly

Fragment Assembly

- Most advanced ab initio method
- Break up sequences into subsegments of length 3-9 residues
- Generate structure of subsequences based on library of known fragment structures
- Decoys (candidate structures) are generated from all possible combinations of fragments
- Energy minimisation is applied to all decoys
- Decoys are clustered and the final models are selected form the centres of the largest clusters

6.4.3 I-Tasser method

- Fully automated prediction method

Workflow:

1 Template identification

- MUSTER fold recognition method
- Profile based fold recognition
- For more difficult targets a combination of various methods

2 Structure assembly

- Generate preliminary model with only C_α and side chain positions using template where possible and falling back on ab initio methods
- Two iterations of refinement:
  1 Template based
  2 Cluster models of previous iteration and use centroids as starting point

3 Atomic model construction

- Full models constructed from approximate models produced by cluster centroids
- Backbone is matched to a library of template fragments
- Atomic optimisation performed (focus on Hydrogen bonds)

4 Model selection

- Several full atom models generated from each cluster centroid
- Models are ranked based on:
  - $\frac{\#\, H-bonds}{target\, length}$
  - Comparison score between full atom model and centroid cluster
- Best model is selected

6.5 Assessment of protein structure prediction

- Requirement to compare a predicted protein structure to experimental data
  - Algorithm verification
  - CASP exercise results
- Various methods exist that measure different metrics
- Most popular metric: GDT-TS

6.5.1 GDT-TS

- (Global Distance Test - Total Score)
- Aims to balance rewarding good local similarity and good global similarity
  - A measure that only takes into account local similarity could discard models that only fail on a few amino acids
- Superimposition process is repeated with thresholds of 1, 2, 4 and 8Å
  - Different thresholds used to allow metric to be used with both approximate and high accuracy models

Workflow:

1 All segments of 3, 5 and 7 consecutive amino acids form the model are superimposed to the actual structure

2 Each segment is iteratively extended while the distance between all residue pairs is below a given threshold

3 A final superposition includes the set of segments covering as many residues as possible

Segments do not need to be continuous

6.5.2 GDT-HA

- (Global Distance Test - High Accuracy)
- Uses thresholds of 0.5, 1, 2 and 4Å

7 Network Analysis

7.1 Network

A graph, either directed or undirected, with a set of nodes joined by arcs (directed) or edges (undirected).

7.2 Concepts and Metrics

Degree

- In-degree: number of connections incoming to a node
- Out-degree: number of connections outgoing from a node
- For an undirected graph both are equal

Degree distribution:

Constant All nodes have exactly k links

Random Probability of finding a node with connectivity k is equal to the average connectivity

Scale free No characteristic scale for connectivity of the network

Distance

- Minimum number of links between a pair of nodes

Diameter

- Pajek: longest of shortest paths between all pairs of nodes
- Watts: Average of shortest paths between all pairs of nodes

Density

- Number of edges as a fraction of all possible edges

Cluster coefficient

- Measure of attachment amongst neighbours
- Average density of neighbourhood

Centrality

- Searching for significant nodes based on their connections
- **Degree centrality**
  Number of edges of a node

  Closeness centrality
  High for nodes which have a low shortest path to all other nodes

$$C_H(x) = \sum_{y \neq x} \frac{1}{d(x,y)}$$

  Betweenness centrality
  Fraction of shortest paths of the network of which a node is a member of

$$H_B(v) = \sum_{(s \neq v \neq t) \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

7.3 Kauffman's hypothesis

- Cell types represent attractors in gene expression state
- Number of cell types proportional to root of number of genes
- Genetic regulation can be modelled using a random boolean network
  - Each node has a boolean state
  - Rules define the next state of a node as a function of its current state and the state of its inputs

7.4 Building networks from data

- Many sources of data
- Most data sets are either incomplete, noisy or both
- Data integration is performed to increase quality of network
  - Weight edges by reliability, relevance, etc.
  - Generate high-level probabilistic networks
    (e.g. String DB)
  - Generate networks from experimentally verified data only
    (e.g. BioGRID)

8 Omics Data and Analysis

- General technological field associated with measurement of activity of a cell

8.1 Genome level

8.1.1 Single Nucleotide Polymorphisms (SNP)

- One base pair variation in DNA
- Typically in non-coding region
- When frequent enough in a population, SNP can be linked to a trait (e.g. disease)
- Can be probed in parallel using SNP microarrays

8.1.2 Methylation

- Chemical reaction that blocks transcription of a certain region of a chromosome
- Has a boolean state
- Can be probed using specialised microarrays

8.2 RNA level

8.2.1 RT-PCR

- (Real Time reverse Polymerase Chain Reaction)
- Measures expression of a single predetermined gene
- High accuracy

8.2.2 RNA microarrays

- Measures expression of many genes in parallel
- High noise

8.2.3 RNA sequencing

- Use high throughput technologies to obtain RNA content of a sample
- mRNA converted to cDNA
- cDNA used to generate sequencing library
- Library allows analysis of RNA

8.2.4 RNA-Seq

- Can detect more scarce transcripts than traditional RNA sequencing
- More reliable results

8.3 Protein level

Experimental methods:

- 2D gel electrophoresis
  - Protein mixture separated on gel plate
  - Regions of interest identified using Peptide Mass Fingerprinting
- Liquid chromatography MS/MS
  - Peptide mixture separated on micro column
  - Proteins identified by mass spectroscopy

8.3.1 Peptide Mass Fingerprinting

- Mass spectroscopy of peptides provides list of peaks (mass:charge ratios)
- Peaks compared to databases of peptides
- Identified proteins (with a confidence score)

8.4 Analysis

- Challenge is imbalance between number of samples and number of variables in dataset
- Lack of samples mean only crude analysis can be performed

8.4.1 Normalisation

- Large variation of results due to different experimental conditions/facilities
- Normalisation aims to remove such systematic errors
- Ensures all samples have same statistical distribution

Quantile Normalisation

1 Rank each data point in each sample

2 Take average value for each rank

3 Replace each value with the average for the rank

Gene	Array 1	Array 2	Array 3
A	15	24	13
B	10	11	15
C	20	9	24
D	30	31	28

Table 5: Raw Data

Gene	Array 1	Array 2	Array 3
A	15 (2)	24 (3)	13 (1)
B	10 (1)	11 (2)	15 (2)
C	20 (3)	9 (1)	24 (3)
D	30 (4)	31 (4)	28 (4)

Table 6: Ranks (rank in brackets)

Gene	Array 1	Array 2	Array 3
A	15 (2) [15]	24 (3) [24]	13 (1) [10]
B	10 (1) [10]	11 (2) [15]	15 (2) [15]
C	20 (3) [24]	9 (1) [10]	24 (3) [24]
D	30 (4) [30]	31 (4) [30]	28 (4) [30]

Table 7: Normalised (normalised value in square brackets)

8.4.2 Quality Control

- Step in which obvious anomalous data are removed from the sample data
- Expect all data within the same class to have similar values

8.4.3 Analysis Methods

- Methods can be either supervised or unsupervised
- Unsupervised methods:
  - Principal Component Analysis
  - Hierarchical clustering

  (can also be used for quality control)
- Supervised methods:
  - Univariate (differentially expressed)
  - Multivariate (machine learning)

Principal Component Analysis (PCA)

- Dimension reduction method
- Find a set of vectors that combine to describe the data but are unrelated to each other

Hierarchical clustering

- Each object starts in its own cluster
- Closest two clusters are combined
- Repeat iteratively until all objects are in a single cluster
- Order of merging and distance between pairs of of clusters infer a tree structure

9 Data Standards

- Growing amount of data
- Several data standards for each area of bioinformatics
- Publishing data often requires use of an existing standard
- Standards define minimum data that must be present for publication
- Various fields of bioinformatics use a controlled vocabulary to assist searching and categorisation
- Ontologies used to represent relationships between data sets often using a controlled vocabulary