Defining Sequence Analysis

- **Sequence Analysis** is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

- It includes-

- *Sequencing*:

➢ Sequence Assembly

- *Alignment*:

➢ Searching (in Databases)

ANALYSIS

- Sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences.

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between sequences.

Protein Sequencing

- **Protein sequencing** is a technique to determine the amino acid sequence of a protein, as well as which conformation the protein adopts and the extent to which it is complexed with any non-peptide molecules.

- Methods:

➢ Edman Degradation

➢ Mass Spectroscopy

Principles of Sequence

Alignment

- Alignment is the task of locating "equivalent" regions of two or more sequences to maximize their similarity
- NIKESH NARAYANAN (RED : Mismatches)
- NIGESH NARAYAN- - (gaps)

Sequence Alignment

- ***Sequence Alignment*** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.
- It involves the identification of the **correct location** of **deletions** and **insertions** that have occurred in either of the two lineages since the divergence from a common ancestor.

Sequence Alignment:-

- This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences.
- We usually compare sequences in order to check evolutionary relationship and prediction of protein structure and function.

- It is an important first step toward structural and functional analysis of newly determined sequences.
- As new biological sequences are being generated at exponential rate, sequence comparison is becoming increasingly important to draw functional and evolutionary inference

TYPES

- On the basis of number of comparing sequencing strand, it is of two types:

➢ Pair-wise Alignment

➢ Multiple Sequence Alignment

Pairwise Sequence Alignment

- Sequence comparison lies at the heart of bioinformatics analysis.
- It is an important first step towards structural and functional analysis of newly determined sequences.
- As new biological sequences are being generated at exponential rates, sequence comparison is becoming increasingly important to draw functional and evolutionary inference of a new protein with proteins already existing in the database.

- The most fundamental process in this type of comparison is **sequence alignment.**
- This is the process by which sequences are compared by searching for common character patterns and establishing residue correspondence among related sequences.
- **Pairwise sequence alignment** is the process of aligning two sequences and is the basis of database similarity searching and multiple sequence alignment

- Pairwise sequence alignment is the fundamental component of many bioinformatics applications.
- Pairwise sequence alignment is used to identify regions of similarity that may indicate functional, structural and or/ evolutionary relationships between two biological sequence e.g say: (protein or nucleic acid)

- It is extremely useful in structural, functional, and evolutionary analyses of sequences. Pairwise sequence alignment provides inference for the relatedness of two sequences.

Pairwise alignment:-

- Pairwise sequence alignment methods are used to find the best-matching local or global alignments of two query sequences.
- Pairwise alignments can only be used between two sequences at a time.
- It is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Pair-wise sequence alignments

```
A: C A T - T C A - C
   |   |     | |   |
B: C - T C G C A G C
```

- DNA and proteins are products of evolution.
- For example, active site residues of an enzyme family tend to be conserved because they are responsible for catalytic functions.

- Therefore, by comparing sequences through alignment, patterns of conservation and variation can be identified. The degree of sequence conservation in the alignment reveals evolutionary relatedness of different sequences, whereas the variation between sequences reflects the changes that have occurred during evolution in the form of substitutions, insertions, and deletions.

- ***Identifying the evolutionary relationships between sequences helps to characterize the function of unknown sequences.***
- *When a sequence alignment reveals significant similarity among a group of sequences, they can be considered as belonging to the same family*

- Therefore, **sequence alignment** can be used as basis for prediction of structure and function of uncharacterized sequences.

- Sequence alignment provides inference for the relatedness of two sequences under study.

- If the two sequences share significant similarity, that the two sequences must have derived from a common evolutionary origin.

- When a sequence alignment is generated correctly, it reflects the evolutionary relationship of the two sequences: regions that are aligned but not identical represent residue substitutions; regions where residues from one sequence correspond to nothing in the other represent insertions or deletions that have taken place on one of the sequences during evolution.

Multiple Sequence Alignment

- A natural extension of pairwise alignment is multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences.

- Related sequences are identified through the database similarity searching.
- As the process generates multiple matching sequence pairs, it is often necessary to convert the numerous pairwise alignments into a single alignment, which arranges sequences in such a way that evolutionarily equivalent positions across all sequences are matched.

- There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments.

- Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.
- Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.
- *However, the amount of computing time and memory it requires increases exponentially as the number of sequences increases.* As a consequence, full dynamic programming cannot be applied for datasets of more than ten sequences. In practice, ***heuristic approaches*** are most often used.
- Multiple sequence alignment is to arrange sequences in such a way that a maximum number of residues from each sequence are matched up according to a particular scoring function.
- They are more computationally complex than pairwise.

Multiple Sequence Alignment

- **Multiple sequence alignment** (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.

➢ **ClustalW, PROBCONS, MUSCLE**

```
P69905  (HBB_HUMAN)  MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS  53
P68871  (HBB_HUMAN)  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN  58
P02144  (MYG_HUMAN)  -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS  59
                      : *:  :   *   ****         * * *::   * *   *  *  *

P69905  (HBB_HUMAN)  AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH  113
P68871  (HBB_HUMAN)  PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH  118
P02144  (MYG_HUMAN)  EDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSK  119
                       :* **  *  *:   :         :  *:: *  *  :    : ::   ::  *  :

P69905  (HBB_HUMAN)  LPAEFTPAVHASLDKFLASVSTVLTSKYR------ 142
P68871  (HBB_HUMAN)  FGKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
P02144  (MYG_HUMAN)  HPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
                        :*    : : :*        :: :*:
```

- A natural extension of pairwise alignment is multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences.
- There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can. For example, it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by comparing only two sequences.

Multiple Sequence Alignment

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY

- When two sequences are descended from a common evolutionary origin, they are said to have a ***homologous relationship*** or share ***homology***. A related but different term is ***sequence similarity*****,** which is the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity.
- To be clear**,** ***sequence homology*** is an inference or a conclusion about a common ancestral relationship drawn from **sequence similarity** comparison when the two sequences share a high enough degree of similarity.
- On the other hand, *similarity* is a direct result of observation from the sequence alignment.
- Sequence similarity can be quantified using percentages; homology is a qualitative statement. For example, one may say that two sequences share 40% similarity but It is incorrect to say that the two sequences share 40% homology. They are either homologous or non homologous.
- Generally, if the sequence similarity level is high enough, a common evolutionary relationship can be inferred.

Similarity versus Homology*

- Similarity refers to the likeness or % identity between 2 sequences
- Similarity means sharing a statistically significant number of bases or amino acids
- Similarity does not imply homology

- Homology refers to shared ancestry
- Two sequences are homologous if they are derived from a common ancestral sequence
- Homology usually implies similarity

Types of homology

- **Homologues** (**homologous genes**) are genes that derive from a common ancestor-gene
- **Orthologues** (orthologs) are homologous genes in different species
- **Paralogues** (paralogs) are homologous genes in one species that derive from gene duplication

When one gene is duplicated, the duplication event results in two paralogous genes (paralogues)

Studies of paralogs have found that one paralogue of a pair often retains the ancestral gene's function, while the other paralogue is free to evolve and adopt new functions

Similarity versus Homology*

- Similarity can be quantified
- It is correct to say that two sequences are X% identical
- It is correct to say that two sequences have a similarity score of Z
- It is generally incorrect to say that two sequences are X% *similar*

Homologues & All That*

- Homologue (or Homolog)
  - Protein/gene that shares a common ancestor and which has good sequence and/or structure similarity to another (general term)
  - **Homology**: genes that derive from a common ancestor- these gene are called ***homologs***
- Paralogue (or Paralog)
  - A homologue which arose through gene duplication in the same species/chromosome
  - **Paralogous** genes are homologous genes in *one* organism that derive from **gene duplication**
  - **Gene duplication**: one gene is duplicated in multiple copies that are therefore free to evolve and assume new functions
- Orthologue (or Ortholog)
  - A homologue which arose through speciation (found in different species)
  - **Orthologous** genes are homologous genes in ***different*** organisms

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY

- Another set of related terms for sequence comparison are **sequence similarity** and **sequence identity.**
- Sequence similarity and sequence identity are synonymous for nucleotide sequences.
- For protein sequences, however, the two concepts are very different. In a protein sequence alignment, *sequence identity* refers to the percentage of matches of the same amino acid residues between two aligned sequences.
- *Similarity* refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.

Percent Identity as a Measure for Quantifying Sequence Similarity

- Identity is the number of identical bases or amino acids matched between two aligned sequences

- Percent identity is obtained by dividing this number by the total length of the aligned sequences and multiplying by 100

- Sequence similarity is a measure of an empirical relationship between sequences. Its common objective is establishing the likelihood for sequence homology i.e chance that sequences has evolved from a common ancestor.

- Sequence identity is the amount of characters which match exactly between two different sequences. A similarity score is therefore aimed to approximate the evolutionary distance between a pair of nucleotide or protein sequences.

- The overall goal of pairwise sequence alignment is to find the best pairing of two sequences, such that there is maximum correspondence among residues.
- To achieve this goal, one sequence needs to be shifted relative to the other to find the position where maximum matches are found.
- There are two different alignment strategies that are often used: **global alignment** and **local alignment.**

Types of Alignment

Based on Completeness

- Global
- Local

Based on Numbers

- Pair wise alignment
- Multiple sequence Alignment

Global Alignment and Local Alignment

- In ***global alignment,*** two sequences to be aligned are assumed to be generally similar over their entire length. Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences. You take entirety of both sequences into consideration when finding alignment.

- ***Local alignment,*** on the other hand, does not assume that the two sequences in question have similarity over the entire length. It only finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions. When you take small portion into account.

- This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences.
- The two sequences to be aligned can be of different lengths. This approach is more appropriate for aligning divergent biological sequences containing only modules that are similar, which are referred to as ***domains*** or ***motifs***.

```
seq1   EARDF-NQYYSSIKRSGSIQ
          .  :  .:::::::::.  .  .
seq2   LPKLFIDQYYSSIKRTMG-H
```

global sequence alignment

```
seq1   NQYYSSIKRS
       .:::::::::.
seq2   DQYYSSIKRT
```

local sequence alignment

Global vs. Local Alignments

- Global alignment algorithms start at the beginning of two sequences and add gaps to each until the end of one is reached.

- Local alignment algorithms finds the region (or regions) of highest similarity between two sequences and build the alignment outward from there.

Global Alignment and Local Alignment

- In *global alignment*, two sequences to be aligned are assumed to be generally similar over their entire length.

- Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences.
- *Local alignment*, on the other hand, does not assume that the two sequences in question have similarity over the entire length.
- It only finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions.
- The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods.

Algorithms

- An ***algorithm*** is a sequence of instructions that one must perform in order to solve a well-formulated ***problem***
- *First you must identify exactly what the problem is!*
- A ***problem*** describes a class of computational tasks. A problem for ***instance*** is one particular input from that task

Alignment Algorithms

- Very short or very similar sequences can be aligned by hand. However, most interesting problems require alignment of lengthy numerous numerous sequences that cannot be aligned by human effort.
- Instead, human knowledge is applied in constructing algorithms to produce high quality sequence alignments.
- Alignment algorithms, both **global** and **local**, are fundamentally similar and only differ in the optimization strategy used in aligning similar residues.
- Both types of algorithms can be based on one of the three methods:
- the ***dot matrix method, the dynamic programming method,*** and the ***word method.*** The word method, which is used in fast database similarity searching .
- The dot matrix method is useful in visually identifying similar regions, but lacks the sophistication of the other two methods.

Dot Matrix Method

- The most basic sequence alignment method is the dot matrix method, also known as the ***dot plot method***. It is a graphical way of comparing two sequences in a two dimensional matrix. In a dot matrix, two sequences to be compared are written in the horizontal and vertical axes of the matrix.

- The comparison is done by scanning each residue of one sequence for similarity with all residues in the other sequence. If a residue match is found, a dot is placed within the graph.
- Otherwise, the matrix positions are left blank. When the two sequences have substantial regions of similarity, many dots line up to form contiguous diagonal lines, which reveal the sequence alignment.
- If there are ***interruptions*** in the middle of a diagonal line, they indicate ***insertions*** or ***deletions***. The dot matrix method gives a direct visual statement of the relationship between two sequences and helps easy identification of the regions of greatest similarities

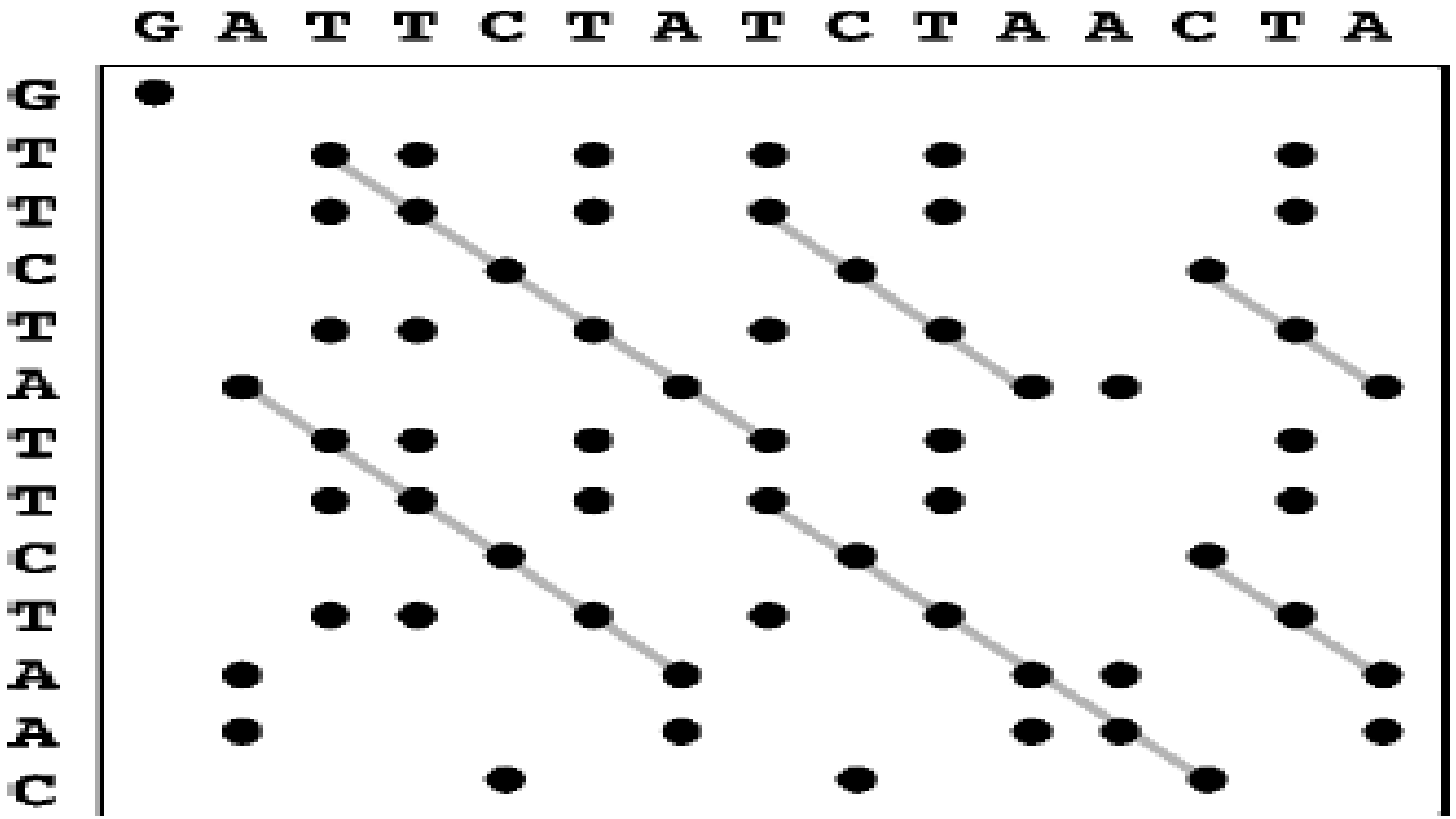


Figure 1: Example of comparing two sequences using dot plots. Lines linking the dots in diagonals indicate sequence alignment. Diagonal lines above or below the main diagonal represent internal repeats of either sequence.

- **Dynamic programming** is an exhaustive and quantitative method to find optimal alignments. This method effectively works in three steps.

- It first produces a sequence versus sequence matrix.
- The second step is to accumulate scores in the matrix.
- The last step is to trace back through the matrix in reverse order to identify the highest scoring path.

This scoring step involves the use of scoring matrices and gap penalties.

Dynamic Programming Method

- Dynamic programming is a method that determines optimal alignment by matching two sequences for all possible pairs of characters between the two sequences.
- It is fundamentally similar to the dot matrix method in that it also creates a two dimensional alignment grid.
- However, it finds alignment ***in a more quantitative way*** by converting a dot matrix into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

Database Similarity Searching

- A main application of pairwise alignment is ***retrieving biological sequences in databases*** based on similarity.
- Thus, database similarity searching is ***pairwise alignment on a large scale.*** However, the dynamic programming method is slow and impractical to use in most cases.
- Special search methods are needed to speed up the computational process of sequence comparison.

- There are ***unique requirements*** for implementing algorithms for sequence database searching. The first criterion is ***sensitivity,*** which refers to the ability to find as many correct hits as possible. It is measured by the extent of inclusion of correctly identified sequence members of the same family. These correct hits are considered "true positives" in the database searching exercise.

- The second criterion is ***selectivity,*** also called ***specificity,*** which refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered "**false positives.**"
- The third criterion is ***speed,*** which is the time it takes to get results from database searches.
- Depending on the size of the database, speed sometimes can be a ***primary concern***. Ideally, one wants to have the greatest sensitivity, selectivity, and speed in database searches.

- In ***database searching,*** as well as in many other areas in bioinformatics, are two fundamental types of algorithms. One is the ***exhaustive type,*** which uses a rigorous algorithm to find the best or exact solution for a particular problem by examining all mathematical combinations.
- Dynamic programming is an example of the exhaustive method and is computationally very intensive.
- Another is the ***heuristic type,*** which is a computational strategy to find an empirical or near optimal solution by using rules of thumb.
- Essentially, this type of algorithms take shortcuts by reducing the search space according to some criteria. It is often used because of the need for obtaining results within a realistic time frame without significantly sacrificing the accuracy of the computational output.

HEURISTIC DATABASE SEARCHING

- Searching a large database using the dynamic programming methods, such as the ***Smith–Waterman algorithm,*** although accurate and reliable, is too slow and impractical when computational resources are limited
- Thus, **speed** of searching became an important issue. To speed up the comparison, **heuristic methods** have to be used.
- The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming.
- Currently, there are two major heuristic algorithms for performing database searches: **BLAST** and **FASTA.**

Profiles and Hidden Markov Models

- One of the applications of multiple sequence alignments in identifying related sequences in databases is by construction of position-specific scoring matrices (PSSMs), profiles, and hidden Markov models (HMMs). These are ***statistical models*** that reflect the frequency information of amino acid or nucleotide residues in a multiple alignment.

- The purpose of establishing the mathematical models is to allow partial matches with a query sequence so they can be used to detect more distant members of the same sequence family, resulting in an increased sensitivity of database searches.

PROFILES

- Actual multiple sequence alignments often contain gaps of varying lengths. When gap penalty information is included in the matrix construction, a profile is created. In other words, a profile is a PSSM with penalty information regarding insertions and deletions for a sequence family.

- However, in the literature, *profile* is often used interchangeably with PSSM, even though the two terms in fact have subtle but significant differences. **Position-specific scoring matrices**
- PSSMs, profiles, and HMMs are statistical models that represent the consensus of a sequence family. Because they allow partial matches, they are more sensitive in detecting remote homologs than regular sequence alignment methods.

- A PSSM by definition is a scoring table derived from ungapped multiple sequence alignment.

- A profile is similar to PSSM, but also includes probability information for gaps derived from gapped multiple alignment. An HMM is similar to profiles but differentiates insertions from deletions in handling gaps.

- The ***probability calculation*** in HMMs is more complex than in profiles. It involves traveling through a special architecture of various observed and hidden states to describe a gapped multiple sequence alignment.

- As a result of flexible handling of gaps, ***HMM is more sensitive than profiles*** in detecting remote sequence homologs.

- All three types of models require training because the statistical parameters have to be determined according to alignment of sequence families.

MARKOV MODEL AND HIDDEN MARKOV MODEL

Markov Model

- A more efficient way of computing matching scores between a sequence and a sequence profile is through the use of HMMs, which are statistical models originally developed for use in speech recognition.

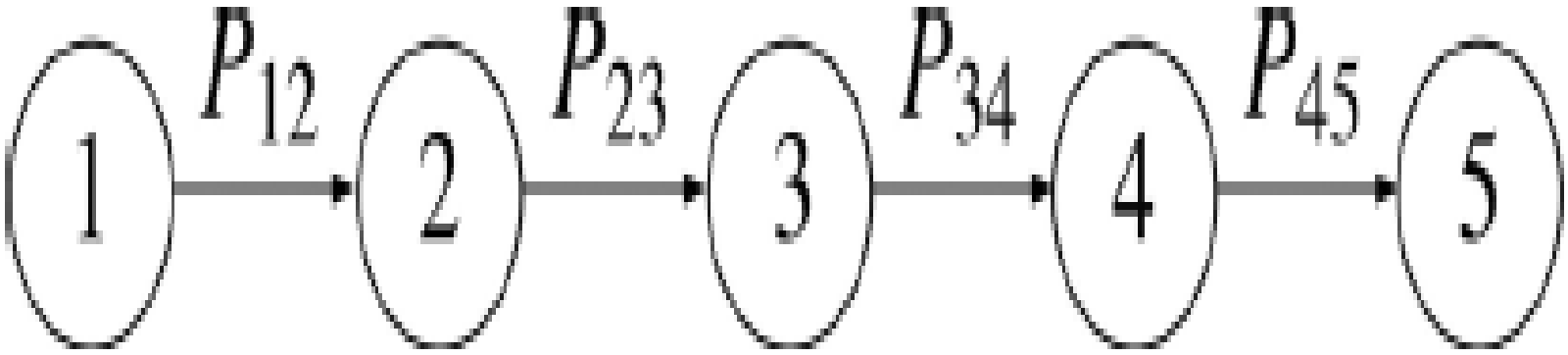
- This statistical tool was subsequently found to be ideal for describing sequence alignments. To understand HMMs, it is important to have some general knowledge of Markov models.

- A Markov model, also known as ***Markov chain,*** describes a sequence of events that occur one after another in a chain. Each event determines the probability of the next event.

- A Markov chain can be considered as a process that moves in one direction from one state to the next with a certain probability, which is known as ***transition probability.***

- A good example of a Markov model is the *signal change of traffic lights in which the state of the current signal depends on the state of the previous signal* (e.g., green light switches on after red light, which switches on after yellow light).

- ***Biological sequences written as strings of letters*** can be described by Markov chains as well; each letter representing a state is linked together with transitional probability values.



A simple representation of a markov chain which consist of a linear events or states (numbered) linked by transition probability values between events (states)

- An **HMM** combines two or more Markov chains

- In an HMM, as in a Markov chain, the probability going from one state to another state is the ***transition probability.*** The probability value associated with each symbol in each state is called ***emission probability.***

Applications

- An advantage of HMMs over profiles is that the probability modeling in HMMs has more predictive power. This is because an HMM is able to differentiate between insertion and deletion states, whereas in profile calculation, a single gap penalty score that is often subjectively determined represents either an insertion or deletion.
- Because the handling of insertions and deletions is a major problem in recognizing highly divergent sequences, HMMs are therefore more robust in describing subtle patterns of a sequence family than standard profile analysis.
- HMMs are very useful in many aspects of bioinformatics. Although an HMM has to be trained based on multiple sequence alignment, once it is trained, it can in turn be used for the construction of multiple alignment of related sequences. HMMs can be used for database searching to detect distant sequence homologs.
- HMMs are also used in protein family classification through motif and pattern identification