Bioinformatics

- Human Genome Project

What are databases

- A database is an organised collection of structured information or data, typically stored in a computer system.
- Consists of basic units called records or entries
- Protein entries → records
- Protein properties → fields

Three types of database Management system :

(i) Flat-file indexing system → can be much useful by sorting and ordering
→ PDB used flat-file indexing system
→ Used in System Retrival System (SRS)
→ drawbacks : ① (indexing) → the index must be consistant
② storage.

(ii) Relational DBMS → collection of tables

(iii) Object-oriented DBMS

Relational DBMS

— 2-D format storing
— unique columns in the short table act as identifier → these are called primary key
[candidate key → when more than 1 unique columns → out of these 1 is primary key and others alternate key]
- primary key becomes foreign key for another table.

drawback : Scalability

③ Object Oriented DBMS
↳ in the form of Objects as used in Object oriented programming.

Database Vs Databank Vs Data Warehousing

Biological database classified as

① Primary (archival) → stores info without manipulation
↳ Raw Data
↳ Mostly from research scientist

② Secondary database. → only data of specific category
(curated)
eg: mutans.

Composite Databases (Swiss Prot + TrEMBL) → Uni Prot KB

Reasons to search database → ① Contain homologous sequence or not.
② Repeat elements or regulatory seq. in non-coding DNA strand

DNS Vs Protein searches → There are very different DNA sequences that code for similar protein sequences.

- DNA → 4 letters → 25%
Protein → 20 letters

→ ① degeneracy
② Randomness

Three data retrieval systems in molecular biology → ① SRS
② Entrez
③ DBGET

What can be discovered about a gene by database search

→ Evolutionary information
→ Genomic information
→ Structural information
→ Expression info.
→ Functional info.

Biological Databases

(i) Nucleic Acid → GenBank
(ii) Protein Sequence → UniProtKB
(iii) Structure → PDB

(i) GenBank (primary database)
↳ First DNA Sequence Database
Gene Sequence Database (GSDB)
1979, NCBI
ENA → European Nucleotide Archive
DDBJ → DNA Databank of Japan
~ 19.8 trillion

* Submission of data to GenBank

(i) Online web server → BankIt
(ii) Stand alone → Sequin
(iii) Batch Submitter → Tbl2asn
Table2asn
Tbl2asn

Types of sequences which can be deposited ?

(i) Whole cell genomes
(ii) EST → Expressed sequence tags cDNA
(iii) GSS → Genome sequence survey
(iv) Shotgun Sequence

What can not be deposited ?

(i) < 200 nt
(ii) mixture of [DNA + RNA Sequence]
(iii) Primers

Metabase

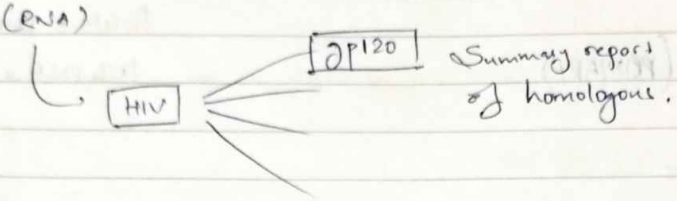
↳ a database of database
↳ BIOGRAPH
↳ NCBI is not a Metabase, it's an organisation

Bac, VIR, INV, PLT, ...

GenBank ⟶ Genomes

Retrovirus ⟵ VIRAL Genomes

(RNA)



Viruses, Eukaryotes
(Homo sapiens)
↓
??
Past Infection

Genome sizes

(00) UniProt KB → Protein Sequences database (PRIMARY)

DNA ⟶ translate into protein

UniProt KB → Swiss Prot
→ trEMBL

(B) Proteomics

(00) 90%. Archival 25%, 50%, 90%.

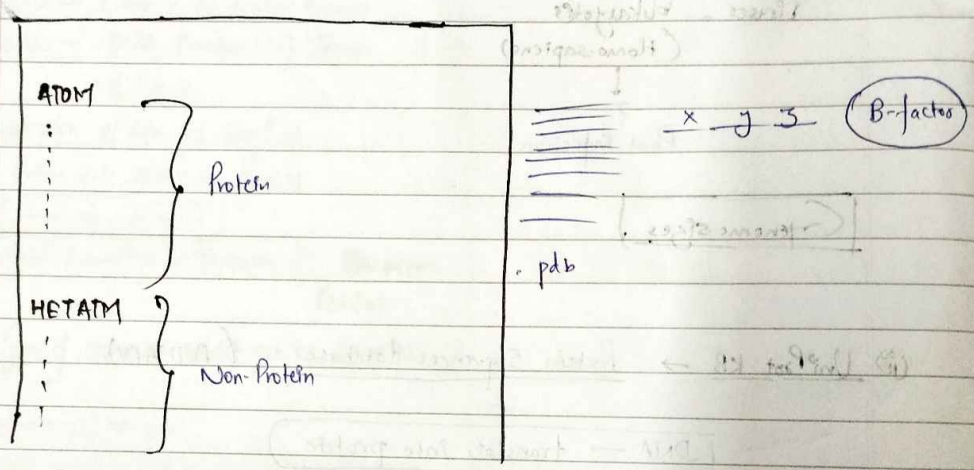
(3) PDB : Structure Database (Textfile)
→ experimentally determined structures

NDB, CSD
Protein,
Protein + Complexes,
DNA, RNA, Sugars.

~ 2,00,000 1978 models

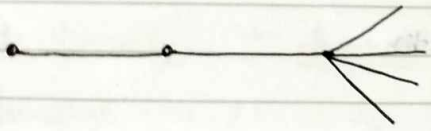
Alphafold (PRIMARY)

Flat-file dbms



Homology — Richard Owen (1843)

"Same Organ in different animals under every variety of form and function"



Homology

- Orthology
- Paralogy
- Xenology

Homoplasy

↳ The counterpart of homology is usually considered to be homoplasy.

↳ A homoplasious trait is a similarity among organisms that was not inherited from the common ancestor of these organisms.

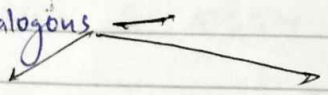
How to relate homology with 'gene'

Homology →
1. Organism level
2. Trait - Specific
3. Gene - specific

Ag gene present in two diff. species is called Orthology

Any gene present within the same species that has arisen through the gene duplication events. As a result of duplication, it can have different function while sharing a common ancestor → Paralogy → gene duplication level.
Relationship b/w two gene

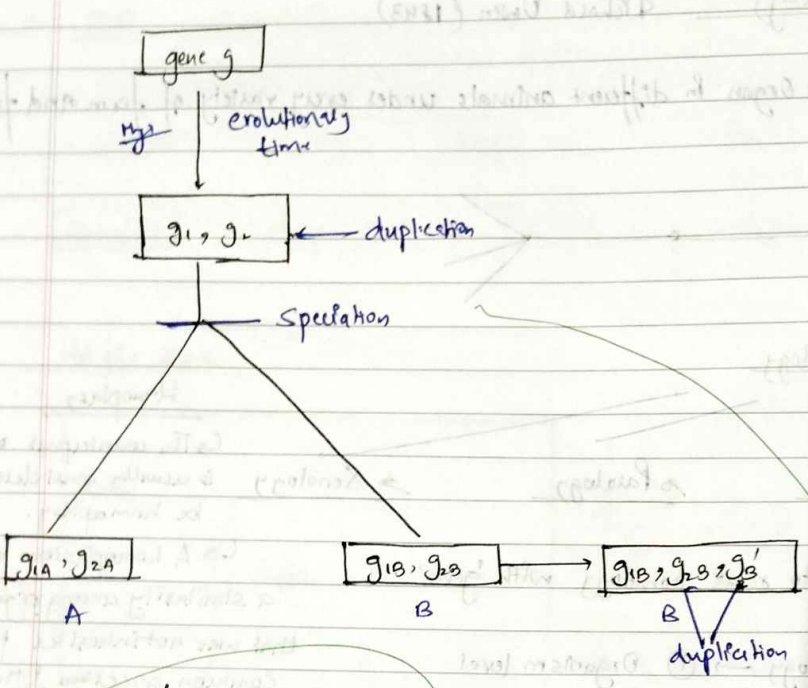
Analogous



Xenology: Xenology is the formation of different sequences due to Horizontal gene transfer ~~species~~ b/w diff. species. HGT is the transfer of genetic material from one org. to another through plasmid exchange or viral infection

Inparalogy →

Outparalogy → follows speciation after duplication in different species.



g_{1A}, g'_{2B} ⟷ orthology

g_{1A} and g_{1B} ⟷ O

g_{2A} and g_{2B} ⟷ O

g_{1A} and g_{2A} ⟷ P

g_{1B} and g_{2B} ⟷ P

Paralogy

① Ortho (Syn) paralog → after speciation

② Out (Allo) paralog → Before Speciation

③ Ohno paralog ⟷ G_1 → Genes that are result of a duplication of a whole genome.

G_1 = $G_{1.1}$ $G_{1.2}$

proteins that have similar active sites but different backbone sequence

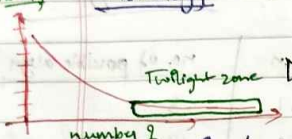
Convergent Evolution : It refers to the phenomenon in which the physical traits of the two diff. species similar but they do not have any common ancestor. The similarity b/w the species is due to the adaption to similar environment or functional requirements.

Analogy
due to
→ Convergent evolution
→ parallel evolution
→ Reversal evolution

Pseudo-Homology → Shows similarity but diff. ancestors.

& Homology

Homology → similarity (%) of genes



Doolittle theory (Doolittle's Rules of Thumb)

g_1 & g_2 ≥ 30 → homologous
≤ 15-30 → may/may not be homologous
≤ 15% → not homologous

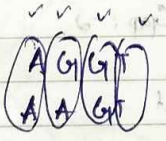
if two genes g_1 & g_2 are similar ⇏ homologous.
if two genes g_1 & g_2 are homologous ⇒ similar.

} Similarity Vs Homology

Comparison (pattern 1 and pattern 2)

DNA : A/C/G/T

① Alignment



Similarity = 3/4
similarity : 4/4 →

- Match
- Mismatch
- Gap
- Global alignment
- local align...

S_1 : GGG S_2 : ATGGG

GGG – –
ATGGG

⇒ | – – GGG |
| A T GGG | → global alignment

| GGG |
| GGG | → local alignment → max of no. matches without a gap.

Pair wise sequence alignment.

- fission yeast (Schizosaccharomyces cerevisiae) & humans (cdk2 gene) had a common ancestor 1 billion years ago.

S1 : G T A T T A G C A

S2 : G T A C C

Cyclin-dependent kinase-CDK2 (a clock gene which controls the cell cycle.)

T	A	G	C
T	A	C	C

→ local alignment.

Highly conserved region of two sequences

G	T	A	T	T	A	G	C	A
G	T	A	-	-	-	C	C	-

→ covered the whole length of the sequences. → Global Alignment

No. of possible alignment : n and m

n = 9

m = 5

without gap = n−m+1

with gap = n^m if n ≠ m

m,n	no. of possible align
1,1	3
2,2	13
3,3	63
4,4	321
5,5	1683
6,6	8989

Dots

A	G	G	C	A
A	A	G	C	-

+2 +1 +2 +2 −2 = 5

A	-	G	G	C	A
A	A	G	-	C	-

+2 −2 +2 −2 +2 −5 = 0

Scoring scheme Assume that M : +2

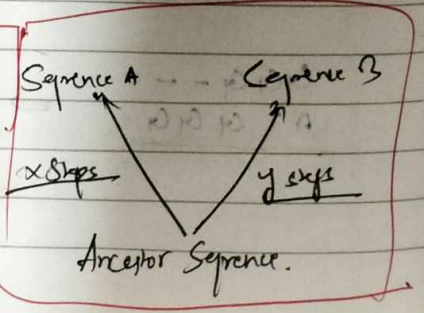
Mis : +1

Gap : −2

∴ Higher score → Better Alignment

Optimal Alignment : $\max_{\ell} S_\ell(x,y)$

$d(S_1,S_2)$ = max alignment of $S_1 \& S_2$ score (alignment)



PIK-CDK2

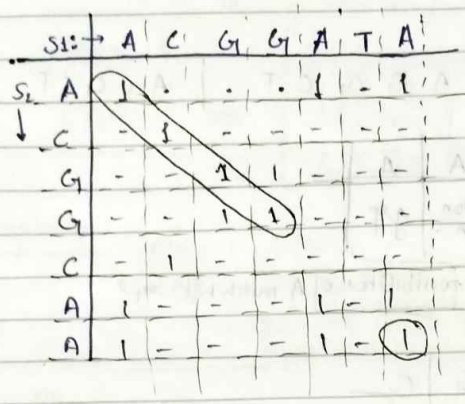
- Sequence
- about
- Gn
- Gn

PIK-CDK2

Significance of sepence alignmen

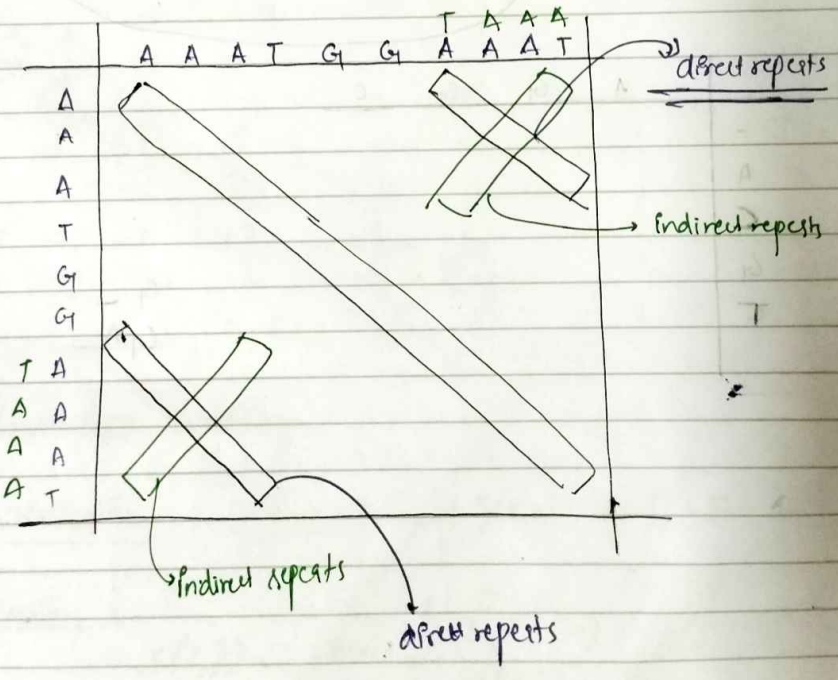
- Sepence alignment is useful for discovering functional, structural and evolutionary information about the biological sepences.
- Can locate subsepences in DNA which can be useful to identify regulatory elements
- Can locate DNA sepences that might overlap: DNA sequencing

Methods : (i) Dot Plot Method



A C G G - - - A
A C G G - - - A

Cons
- As the length of sequences increases, matrix size increases.
- database search



(GN) Global Alignment : Needleman – Wunsch Algorithm

(LS) Local Alignment : Smith – Waterman Algorithm

2. Dynamic Programming (DP method) ⟶ best method to alignment.

A G G C T | A C G C T

Seq1	A	A	A
Seq2	A	Non A	gap

3 possibilities of A match with Seq.2.

Step1	G	G	G
Step2	G	non G	gap

$S_{max} = S_1(A/A) + S_2(G/\text{non } G) +$

	–	A	G	G	C
–					
A					
G					
G					
T					

G –
G A

Global Alignment (Needle Man Wunch Algorithm)

(i) Initialization

$$S(0,0) = 0$$

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + w(x_i, y_j) \\ S(i-1, j) + \text{Gap-penalty} \\ S(i, j-1) + \text{Gap-penalty} \end{cases}$$

Time complexity

Space : $O(mn)$

Time : $O(mn)$

- filling the matrix $O(mn)$
- Backtrace $O(m+n)$

M : +1 , MM : -1 , Gap : -3

	-	G	A	A	G	A
-	0	-3	-6	-9	-12	-15
G	-3	-1	-2	-5	-8	-11
A	-6	-2	2	-1	-4	-2
A	-9	-5	-1	3	0	-3
A	-12	-8	-4	0	2	-1
T	-15	-11	-7	-3	-1	1
T	-18	-14	-10	-6	-4	-2

0+1 = 1
-6
-6

-2-1-4

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + \delta_{x,y} \\ S(i, j-1) + \delta_{gap} \\ S(i-1, j) + \delta_{gap} \end{cases}$$

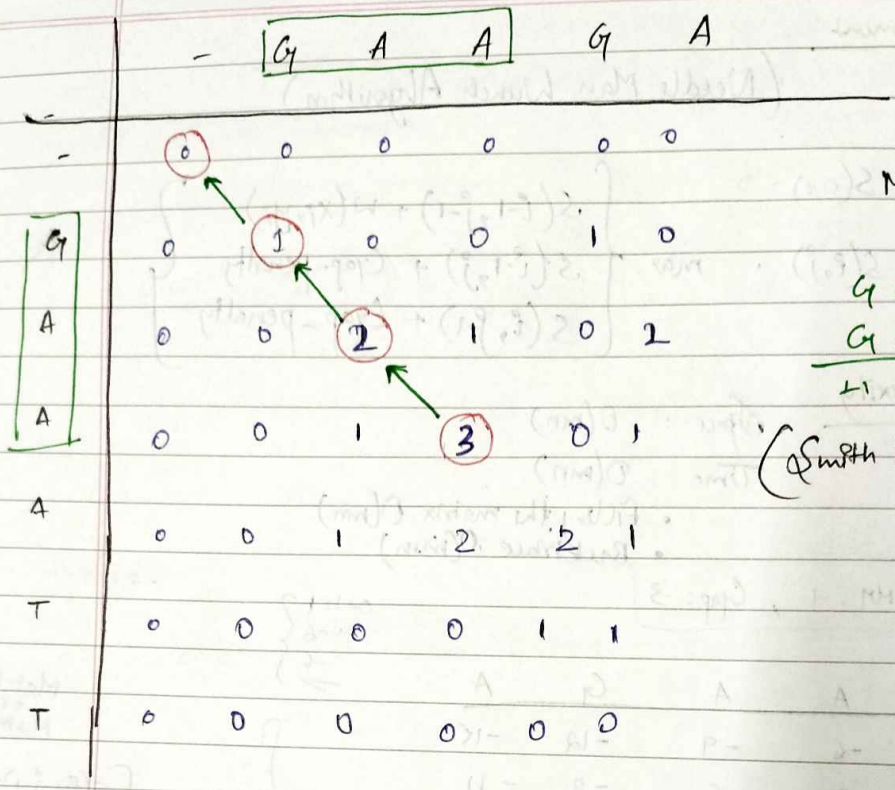
Match(M) or Mismatch(MM)

Local Alignment

(i) Initialization $S(0,0) = 0$, $S(i,0) = S(0,j) = 0 \quad \forall\, i, j$

(2) Iteration

$$S(i,j) = \max \begin{cases} 0 \\ S(i-1, j-1) + \delta_{x,y} \\ S(i-1, j) + \delta_{gap} \\ S(i, j-1) + \delta_{gap} \end{cases}$$



Max Value (1)

G A A
G A A
+1 +1 +1 : (3)

(Smith Waterman)

Overlapping

Similar to Global Alignment except gaps penalties deserve immunity

1. Initialization

$$S(0,0) = 0 \quad , \quad S(i,0) = S(0,j) = 0$$

2) Iteration

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + S(i,j) \\ S(i, j-1) + S_{gap} \\ S(i-1, j) + S_{gap} \end{cases}$$

{ maximum value either in last row or last column }

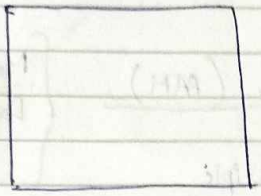
3) Trace back

Maximum value either in last or last column

18/08/23

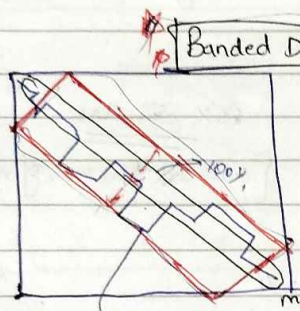
10^{10} operation/s

Disadvantages : time and space complexity.



time taken : $O(mn)$
Tracing back : $O(m+n) \leq O(10^3)$

$m \leq 10^3$
$n \leq 10^3$



$S_1 : m$, $S_2 : n$
$m \simeq n$

$Sim(S_1, S_2) \geq 90\%$.
$Sim(S_1, S_2) = 100\%$.

Time : $O(kn) = O(n)$ → Only a diagonal band of the dynamic programming matrix is computed and filled.
Reduces time & memory requirement.

If the diagnol band consists of k diagonals (width k), then dynamic programming taken $O(kn)$, much faster than $O(n^2)$ of standard DP

= AGGT	Match : +2	+3	+5
AACGT	M-M : -1	-2	+4
-4 +2 -1 +2 +2 ←	Gap : -4	-1	-6
-1 +3 -2 +3 +3			

Hamming Distance

A M A N
A M I N

$(\frac{3}{4}) \times 100$

① → Hamming Distance

Levenshtein Distance

A M A I N (Relationship)
A M A I N

Substitution / Addition / deletion
1
to make similar.

6
2<1

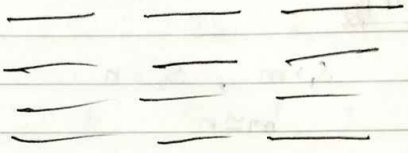
Substitution (or Scoring) Matrices

PAM / BLOSUM

(i) PAM (or Percent Accepted mutation (PAM))

{Margaret Dayhoff Pn 1978}

34 families of protein → Multiple Sequence Alignment.



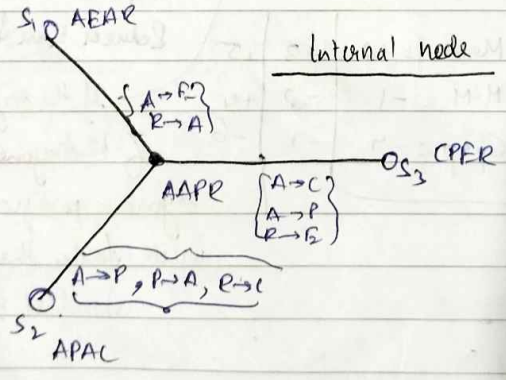
85% Similar

{counted the no. of mutation}

A E A R
A P A C
C P E R

A → A $(M_{A \to A}) = 4$

$M_{AC} = 2$

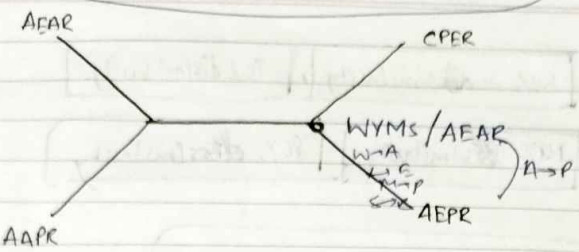


$(P^1)^N = P^N$

Rooted tree →

Unrooted tree →

B Cell receptor
↳ Antigen Binding Receptor



- Minimum no. of mutation
- Parsimonous tree.

A_{xy}

$$P_{xy} = m_x \left(\frac{A_{xy}}{\sum_{h \neq x} A_{xh}} \right) \quad x \neq y$$

→ for non-diagonal values of matrix

Diagonal value of matrix

$$P_{xx} = 1 - m_x$$

m_x : mutability of x

$$m_x = \frac{1}{L \cdot p_x \cdot z} \sum_{d \neq x} A_{xd}$$

L: length of alignment

p_x : background prob. or a priori prob.

z : scaling factor.

21/08/23

PAM2 : PAM1 × PAM1 ⇒ 2% accepted mutation

PAMn : PAM1 × PAM(n-1) : (PAM1)^n

PAM 1 → 10% dis.
PAM 7 → 15% dis
PAM 30 → 30% dis

PAM 159 → 30% dissimilarity ; 70% dissimilarity.

PAM 350 → 14% dissimilarity ; 86% dissimilarity

1990, Henikoff and Henikoff , BLOSUM

n Blocks (500 BLOCKS)

BLOcks SUbstitution Matrix.

q_{ij} = no of times i changed to j for $i, j \in AA$

p_{ij} is the probability that a mutation occurs between amino acid i and amino acid j.

$$p_{ij} = \frac{q_{ij}}{\sum_{j=1}^{20} \sum_{i=1}^{20} q_{iq}}$$

Marginal Prob p_i :

$$\frac{\sum_{j=1}^{20} q_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{20} q_{iz}}$$

Log Odd ratio : S_{ij}

$$S_{ij} = 2 \log_2 \frac{p_{ij}}{p_i \times p_j}$$

$P[x,y] = P[x] \times P[y]$ ⟷ x, y → independent

$P[x|y] = P\left(\frac{x}{y}\right) \cdot P(y)$ ⟶ x, y → dependent

P C V A P

P C A A C

C P V A P

P A C A V

	A	C	P	V	
A	6×2	3	1	2	18
C	3	1×2	7	3	15
P	1	7	4×1	2	18
V	2	3	2	1×1	9
					60

$P_A = \frac{18}{60}$ $P_C = \frac{15}{60}$, $P_P = \frac{18}{60}$, $P_V = \frac{9}{60}$

$P_{ij}^{\circ} = \frac{9 \cdot 9^{\circ}}{60}$

$S_{ij} = 2 \log_2 \frac{P_{ij}}{P_i \times P_j}$

$S_{AA} = 2 \log_2 \frac{P_{AA}}{P_A \times P_A}$

$S_{AA} = 2 \log_2 \frac{12/60}{18/60 \times 18/60}$

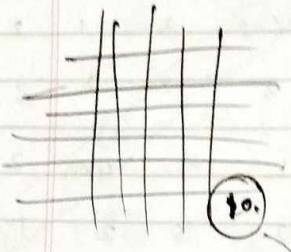
BLOSUM45 : 45% simi ≈ PAM100

BLOSUM60 : 60% simi

BLOSUM80 : 80% simi

↓ divergence

(higher index are appropriate for more conserved proteins



$2 \log_e x$

$2^{(10/2)} : 2^5 : 32$ ~~means~~ the matrix is 32% more efficient for alignments.

Higher number ⓘ better alignment.

standard in protein database searches.

BLOSUM62, 2008 Nature Biotechnology

BLOSUM62 (corrected)
RBLUM62 (corrected) } still BLOSUM62 performs better than RBLOSUM62

With PAM matrices (Percentage difference)

↳ The score indicates the percentage of substitution per position ⇒ higher index are app. for more distant proteins

With BLOSUM matrices (percentage similarity)

↳ The score indicates the percentage of conservation ⇒ higher index are appropriate for more conserved proteins.

Scoring & Matrices / Mismatch

Gaps

Total gap cost = $g_{open\ cost} + (n-1) g_{ext\ cost}$ ... } n of gaps

A P A A C P
A P - - - P

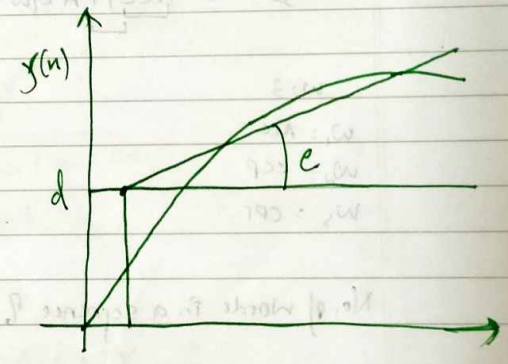
open gap — internal gaps

A P A A C C P C
- - - C C P C

End gaps -1<<-25

$\gamma(n) = d + (n-1)e$

- d: gap open
- e: gap extend



BLOSUM 80 — PAM 1 — BLOSUM 62 — PAM 120 — BLOSUM 45 — PAM 250

Less Divergent ⟵ ⟶ More divergent

22/08/23

Pairwise Sequence Alignment

- Methods of Sequence Alignment
  - Dot Plot Method.
  - D.P.
  - Substitution Matrices.
  - Heuristic Methods.

Pattern Search

BLAST – Basic local alignment Search Tool, Altschul, 1990

(i) Let us divide query sequence into small words

S1 = ACCPTAGWY

W = 3

W_1 = ACC

W_2 = CCP

W_3 = CPT

No. of words in a sequence of length L = L-W+1

(ii) Assign the score for each words.

W_1 = ACC . (5+3+3) = 11

→ probability of A matching to A
C " C
C " C

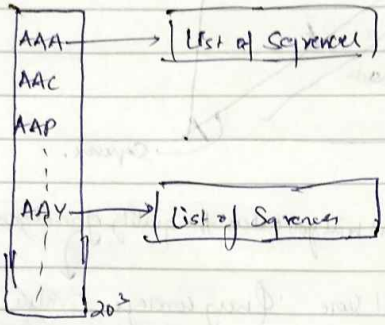
W_2 = CCP = 3+3+4 = 10

⋮ ⋮ ⋮

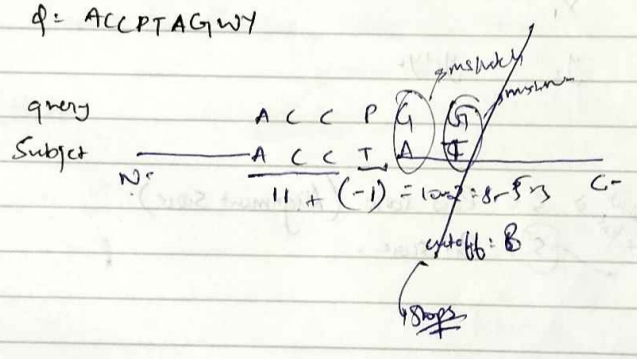
Let us decide a cut-off score : 10

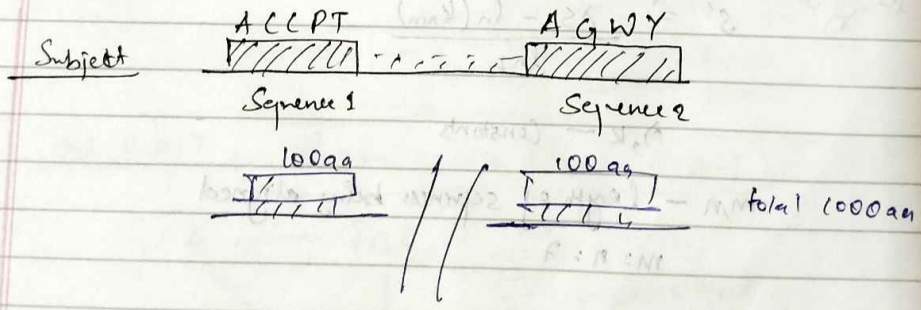
(iii) Scan the database sequence using these high scoring words.

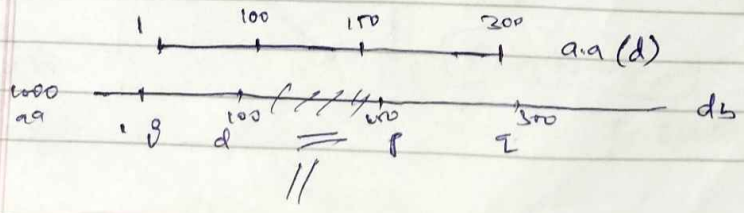
Table of words Vs Sequences in DB.

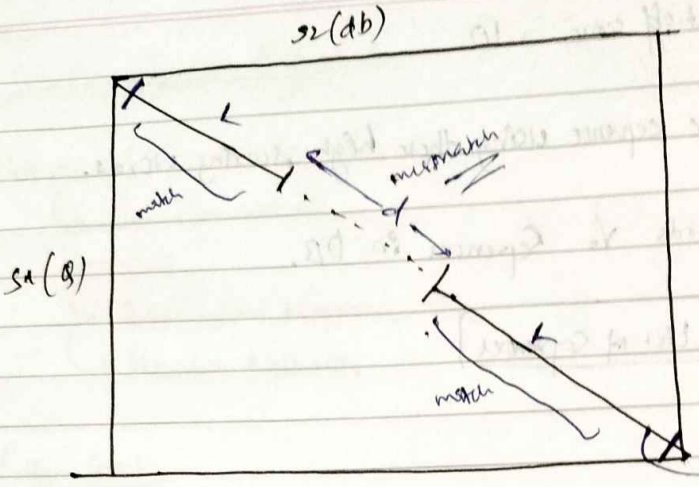


Q: ACCPTAGWY







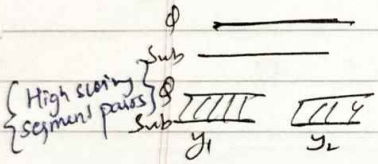


d = 10 aa

How will you assess the quality of an alignment → Sequence Identity

BLAST

Max Score . Total Score . Query coverage . E-value . SI



x_1

y_2 y_1+y_2

If is the required size of a sequence database in which the current match would be found by chance.

S = Raw score (Alignment score)

S' : Bit score.

$$S' = \frac{\lambda S - \ln(K m n)}{\ln 2} \quad \ldots$$

$\lambda, K \rightarrow$ Constants

m, n – length of sequences being aligned

m = n = 7

$P(x \geq s) \quad \frac{\text{P-Value}}{} \cdot \frac{1}{2^{s'}}$

Prob that we get a score similar to 's' by just random alignment.

E-value = N * P-value

$\text{E-value} = \frac{N}{2^{s'}} \quad (N = n \times m)$

Prob. of getting a better alignment by just random Alignment.

query length = n
m = database length

Lower the E-Value, better the alignment m = database length

BLAST → blastn (two nucleotides), blastp (two proteins), ~~blastn~~ blastn (Que = Protein & Data = Nucleic Acid), blastx (translated → Protein), tblastx (In accu → 6 protein)

PSI-blast, PHI-blast. . . .

translate & into Protein

Q: DNA or Nucleic Acid Sq.
Database: Protein

Protein → Search

tblastx → t → DNA translated Protein.
x → DNA translated Protein

PSI-BLAST → Position specific Iterative BLAST

A → {B} → {C} → Remote Homolgy

QY → characterised.

PSI-Blast → Pattern Hit Initiated

for detection of Remote Homology

28/08/23

Comparison of Sequences
↳ Alignment

① Whole genome sequence comparison : Duplication / Translocations / Inversion / deletion.

② Sequences obtained from NGS : Some regions of the genome are deposited to the sequencer

③ Non Coding Regions

④

Time Complexity

29/08/23

Dynamic Programming

S_i – S_j S_i / – S_j S_j } → pair seqn

S_1, S_2, S_3

S_i
S_j
S_k

Multiple Seqn alignment

S_i	S_j	–	S_i	–	S_i
–	S_j	–	S_j	S_j	–
–	–	S_k	–	S_k	S_k

7 alignment

$$S(S_i, S_j, S_k) = (\max) \begin{cases} S_{i-1}, S_{j-1}, S_{k-1} + \delta_{i,j,k} \\ \vdots \\ \vdots \\ \vdots \end{cases}$$

{ no. of seqrences = k
size of seqrence = n }

Time Complex : $O\left((2^k - 1) * (n^k)\right)$

Heuristic Methods

(i) Star Alignment

S_1 : ACT → start Sequence

S_2 : TCT

S_3 : CT

S_4 : CCT

△ Aligning S_2, S_3, S_4 with S_1

(ii)

(iii)

(i) One of the Sequences from multiple sequence alignment

(S_1, S_2), (S_1, S_3), (S_1, S_4)

ACT	ACT	ACT
TCT	-CT	C-CT

⇒

-ACT	S_1
-TCT	S_2
--CT	S_3
C-CT	S_4

different

(S_1, S_4), (S_1, S_2), (S_1, S_3)

ACT	ACT	ACT
CCT	TCT	-CT

⇒

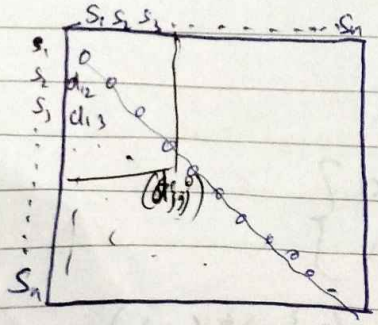
ACT	S_1
CCT	S_2
TCT	S_3
-CT	S_4

How to fix this order problem

Progressive Method

(i) Calculate a distance matrix b/w Sequences

d =



$d_{ij} = \frac{S_{ij}}{L_{ij}}$

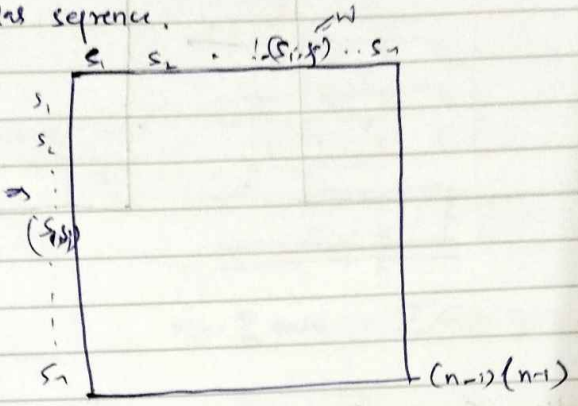
S_{ij} = no. of substitution

L_{ij} = length of the alignment

gaps are not counted.

(ii) To find the most similar sequence.

(iii) Merge the most similar sequence.



(i) PGMA (Pair Group Method - Arithmetic Mean)

$$d(w,x) = \frac{d(u,x) + d(v,x)}{2}$$

$u = S_p$
$v = S_q$

(ii) WPGMA

$$d(w,x) = w_1 d(u,x) + w_2 d(v,x)$$

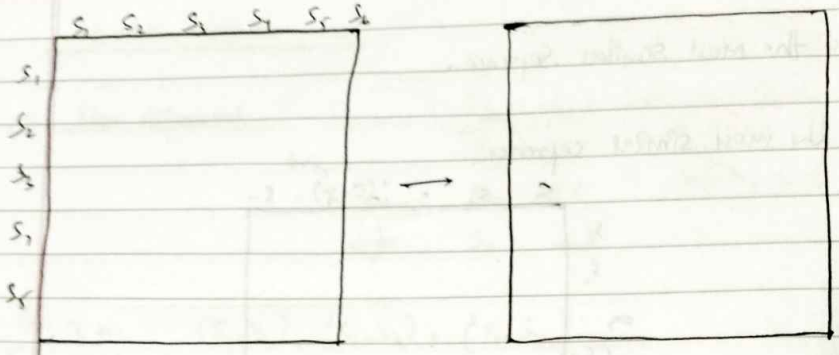
(iii) Unweighted PGMA (UPGMA)

$$m(u) = \text{no. of nodes under } u$$

$$d(w,x) = a(u)d(u,x) + b(v)d(v,x)$$

$$a(u) = \frac{m(u)}{m(u)+m(v)}$$

$$b(x) = \frac{m(v)}{m(u)+m(v)}$$



Multiple Sequence Alignment

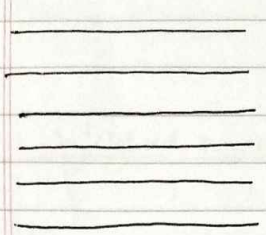
↳ How to align multiple sequences.

(a) Dynamic Programming and its limitation → $O(n^2)$

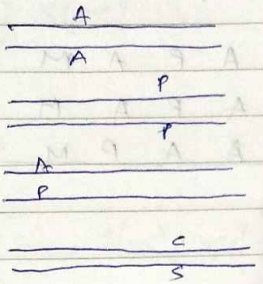
(b) Heuristic method → Progressive method of MSA (clustalW)

T-COFFEE (Evaluate the alignment)

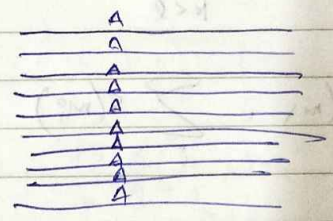
AA = x_1
AP = x_2



$\hookrightarrow \frac{N(N-1)}{2}$



No. of pairs : $\sum x_i$



Score = $\frac{\#\text{ of conserved pairs}}{\text{total no. of pairs}}$

~~Score = # of~~

Ex:

S_1 : APAM
S_2 : PAPA
S_3 : APPA

(S_1S_2) :

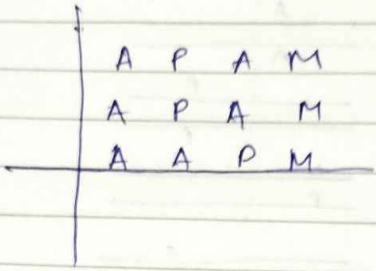
	A	P	M
A	2	3	
P	3		
M			

(S_1S_3) :

(S_2S_3)

How to Score the MSA ?

① Longest Multiple Column (LMC)



② Sum of Pairs

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

$$S(m) = \sum_i S(m_i)$$

$S(m_i)$: Score at the i^{th} position

$s(m_i^k, m_i^l)$: Score b/w pairs of two sequence k & l

$S(m_1)$: $S(A,A) + S(A,A) + S(A,A)$
$= 3\,S(A,A)$

$S(m_2)$: $S(P,P) + S(P,A) + S(P,A)$
⋮

$S(m)$: $S(m_1) + S(m_2) + S(m_3) + S(m_4) \cdots n$

Normalised sum of pairs :

$$S(m_i) = \frac{1}{N} \sum_i s(m_i)$$

Weighted Normalised sum of pairs

$$S(m_i) = \sum_{k \in A} w_{k,a} \, s(m_i^k, m_i^a)$$

MSA, : 100
MSA$_2$: 80

Star Consensus Sum of pairs

$$S(m_i) = \sum_k \delta(m_i, m_c^k)$$

→ Consensus sequence.

MSA
A P A M M
P A P A M
A P A M M

consensus sequence
A P A M M

$$S(m_i) = S(A,A) + S(A,P) + S(A,A)$$

How to find Consensus sequence

A R A F
A A P F
S P S A

Consensus sequence → { A P A/P/S F

Entropy Measurement (Calculation)

Shannon's Entropy : $-\sum_{x \in \text{letters}} \sum_{i=1}^{L} p_x \log_2 p_x$

A	C	T	G
A	C	G	G
A	T	G	C
A	C	T	A

$E_1 = -\left\{ p_A \log_2 p_A + 0 + 0 + 0 \right\} = \left\{ -4 \times (1 \times \log_2 1) \right.$

$E_2 = -\left\{ 3 \times (p_x \log_2 p_x) + (p_T \log_2 p_T) \right\}$

E. $E_3 = -\left[2 \times (p_A \log_2 p_A) + 2 \times (p_T \log_2 p_T) \right]$

$-\left\{ 3 \times \left(\frac{3}{4} \log \frac{3}{4}\right) + \frac{1}{4} \log_2 \frac{1}{4} \right\}$

06/09/23

Motifs
- Structural motifs X
- Sequence motifs → conserved ; short sequence segment having same functions.

Signature / Patterns

① DNA : TATAAT (RE), GAATTC (TF), CACGTG

② RNA : RBS (AGGAGGU), Splicing of pre-mRNA

③ Protein : PKKKRKV (NLS), X-C-X-C-X-H-X-H (Zinc finger Motifs)

① Genes, Patterns → Pattern Matching,

② Genes, do not know the pattern → Discover the pattern.

③ Patterns donot know the set of genes → Classification

How to find motifs?

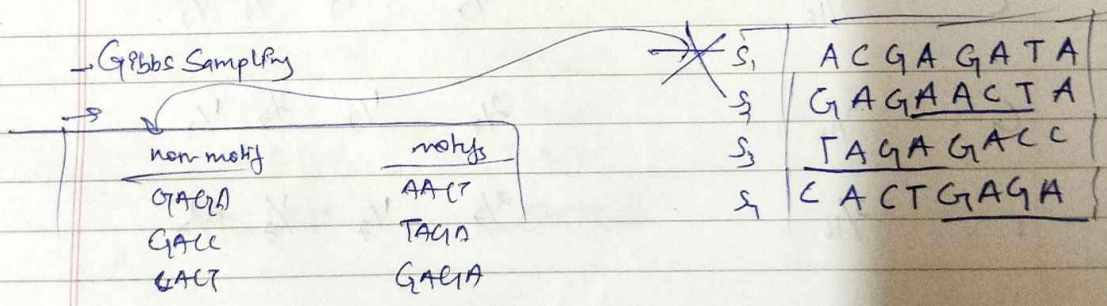
① Enumeration (word-based) methods:

② Probabilistic methods

③ Nature-Inspired Algorithms

Probabilistic Methods –

– Gibbs Sampling



$$p_{c,0} = \frac{n_{c,0} + k_c}{(N-1)(L-W)+k_b}$$

$n_{c,0}$: # of times 'c' has appeared in the non-motif region

k_c, k_b – pseudo-counts

$k_c = 1$, $k_b = 4$ (DNA)
$= 20$ (AA)

L: length of sequence
N: No. of sequence
W: motif length

$$p_{c,k} = \frac{n_{c,k} + k_c}{(l_w)(N-1) + k_b}$$

$$p_{a,k} = \frac{n_{a,k} + k_c}{(l_w)(n-1) + k_b}$$

① Non-Motif

G A G A
G A C C
C A C T

k=0

A : 5/16

C : 5/16

G : 4/16

T : 2/16

Motif

A A C T
T A G A
G A G A

K=1 K=2 K=3 K=4

2/7 4/7 1/7 3/7

1/7 1/7 2/7 1/7

2/7 1/7 3/7 1/7

2/7 1/7 1/7 2/7

S1 : ACGAGATA

motif	non-motif

S1 : ACGA GATA

	Motif	Non-Motif	Weakness (motif / non-motif)
ACGA	$\frac{2}{7}\times\frac{1}{7}\times\frac{2}{7}\times\frac{3}{7}$	$\frac{5}{16}\times\frac{6}{16}\times\frac{5}{16}\times\frac{5}{16}$	0.78
CGAG	$\frac{1}{7}\times\frac{1}{7}\times\frac{1}{7}\times\frac{1}{7}$		= 0.068
GAGA	$\frac{2}{7}\times\frac{4}{7}\times\frac{3}{7}\times\frac{3}{7}$	$\frac{4}{16}\times\frac{7}{16}\times\frac{5}{16}\times\frac{5}{16}$	= 4.91
AGAT	$\frac{2}{7}\times\frac{1}{7}\times\frac{1}{7}\times\frac{2}{7}$	$\frac{5}{16}\times\frac{4}{16}\times\frac{5}{16}\times\frac{2}{16}$	= 0.64
GATA	$\frac{2}{7}\times\frac{4}{7}\times\frac{1}{7}\times\frac{3}{7}$	$\frac{4}{16}\times\frac{5}{16}\times\frac{2}{16}\times\frac{5}{16}$	= 3.21

Non-motif (removing S_1) : CACT GAGA

Non-motif	motif
CACT	GAGA
CACC	TAGA
GAGA	AACT

A

C

G

T

Artificial Bee Colony Algorithm

	1	2	3	4	5	6	7	8
S_1	A	C	G	A	G	A	T	A
S_2	G	A	G	A	A	C	T	A
S_3	T	A	G	G	A	G	A	C
S_4	C	A	C	T	G	A	G	A

Position :

4	3	1	5

A	G	A	T
G	A	A	C
T	A	G	A
G	A	G	A

P(G A A A) : . .

1	2	3	4

A	C	G	A
A	G	A	A
G	A	G	A
T	G	A	G

P (A G G A) :

Score $= \sum \frac{\max(b_i)}{w \times N}$

no. word length

$\therefore \frac{2/4 + 3/4 + 2/4 + 2/4}{4 \times 4}$

How to find Motifs ??

How to represent motifs

① Consensus pattern

T A T A A T

A T T A A T

C A T A A G

G A T A A T

C A T A A T

IUPAC

A and G : R
C and T : Y
A and T : W
G and C : S
A and C : M
G and T : K
A, C and G : not T : V
(not T)
(not U)

C, G & T : not A : B

A, C, G, T : N → Any

A, G & T : not C : D

② Regular Expression

= [ACGT] - [AT] - T - A - A - [GT]

= [ACGT] - {CG} - T - A - A - {AC}

= [ACGT] - {CG} - T - A_2 - {AC}

↳ This can also have a range

A(n,m) → maximum no of times
min no. of times

A(n,·) , A(·,m)

eg → A(2,6) A(·,2)

2 to ∞ −∞ to 2

(3) Position Weight Matrix

Count Matrix

	1	2	3	4	5	6
A	1	3	0	4	4	0
C	1	0	0	0	0	0
G	1	0	0	0	0	1
T	1	1	4	0	0	2

frequency Matrix

$$\begin{bmatrix} 1/4 & 3/4 & 0 & 1 & 1 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1 & 0 & 0 \end{bmatrix}$$

$$f_{ij} = \frac{n_{ij}}{\sum_{r=1}^{t} n_{rj}}$$

Position Weight Matrix

$$W_{ij} = \ln\left(\frac{f_{ij}}{p_i}\right)$$

p_i: prior or background prob. of letter (A, C, G, T)

$$W_{ij}' = \ln\left(\frac{f_{ij}'}{p_i}\right)$$

$$f_{ij}' = \frac{n_{ij} + p_i \times k}{\sum n_{ij} + k}$$

k: pseudo Constant

→ Most Probable Motif (Psi-Blast)

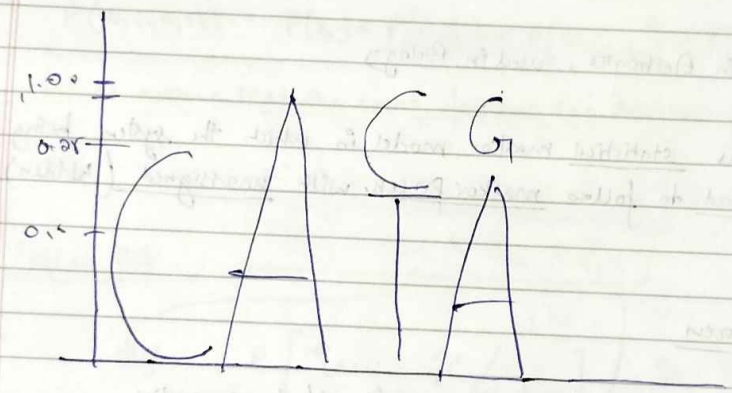
C A C A : .

0.93 + 1.25 + 1.25 + 0.93

PWM

	1	2	3	4
A	0.43	1.25	0.95	1.3
C	0.93	-0.5	1.25	0.55
G	-0.58	-1.3	-0.75	0.25
T	-1.25	0.6	0.60	- -0.5

(4) Web logo / Sequence logo



PSSM ⟷ how to measure the quality of PSSM ??

↳ Position-Specific Scoring Matrix

Information content of PSSM :

$$I_{ij} = f_{ij} \log_2 \left(\frac{f_{ij}}{p_i} \right)$$

$$\text{Imatrix} = \sum_{i} \sum_{j} I_{ij}$$ better quality

Hidden Markov Model

$P(AAAA) = (1/4)^4$
$P(AATA) = (1/4)^4$
$P(AATA) = (1/4)^4$
⋮ (44)
$P(TTTT) = (1/4)^4$

DNA

$P(AAAA) = (1/20)^4$
⋮
$(20)^4$
$P(YYYY) = (1/20)^4$

Protein

- Implement in Electronics, used in Biology

→ Defined as statistical markov model in which the system being modeled is assumed to follow markov process with unassigned (hidden) states.

Markov process

└ A process which ~~follows~~ satisfies the Markov properties.

Markov Property

└ If anyone is able to predict the future given only the present

India Vs Pakistan 1000 bet

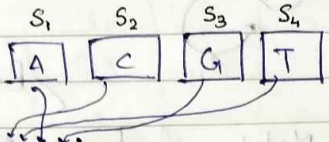
INR 1000

$X_0 = 1000$ ₹ , $X_1 = 1500$, $X_2 = 2000$, $X_3 = 1500$ $X_4 = $ ____

X_4: 1000 → Y_1, 2000 → Y_2

X_4 is predicted given X_3.

↳ This is called Markov property.



P(CTAG) :

$P(S_2 S_4 S_1 S_3) = P[S_2] \times P[S_4/S_2] \times P[S_1/S_2 S_4] \times P[S_3/S_1, S_2, S_4]$

Let us assume that the event follows the Markov property.

: $P[S_2] \times P[S_4/S_2] \times P[S_1/S_4] \times P[S_3/S_1]$

$P[S_2] \times a_{24} \times a_{41} \times a_{13}$

Let us Define

$a_{ij} = P[q_{t+1} = S_j / q_t = S_i]$ q_t : Represents a state at time t.

$\forall i, j$

Transition Prob

$a_{ij} \geq 0 \quad \forall i, j$

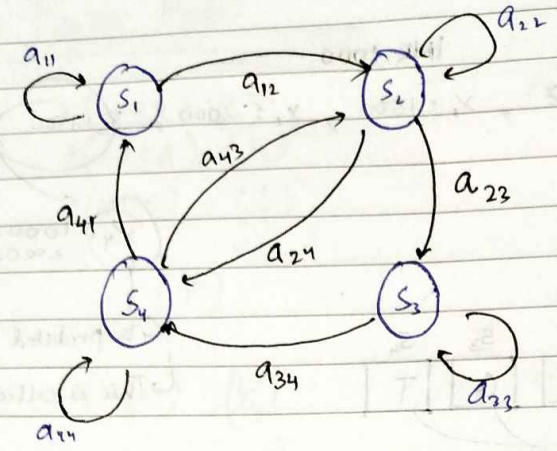
$\sum a_{ij} = 1$

Initial state Property

$\pi_i = P[q_1 = S_i] \quad \forall i$

$\sum \pi_i = 1 \quad 1 \leq i \leq N$

$= \pi_2 \times a_{24} \times a_{41} \times a_{13}$



Every node is connected.

Non-Ergodic Model Vs Ergodic Model

↳ If any one of the node is not directly connected then it is called non-ergodic model.

① No. of state = $\{S_1, S_2 \ldots S_N\}$

② Transition prob = a_{ij}

③ Initial State Prob = π_i

Example: Weather Prediction

S_1 : Sunny
S_2 : Rainy
S_3 : Cloudy

$O = S_2 S_3 S_1 S_1 S_2 S_3 S_2$

$P[O] = P[S_3 S_3 S_1 S_1 S_1 S_2 S_3]$

$= P[S_3] \times P[S_3/S_3] \times P[S_1/S_3] \times P[S_1/S_1] \times P[S_1/S_1] \times P[S_2/S_1] \times P[S_3/S_2]$

$= \pi_3 \times a_{33} \times a_{31} \times a_{11} \times a_{12} \times a_{23} \times a_{32}$

Given

	S_1	S_2	S_3
S_1	0.4	0.3	0.3
S_2	0.2	0.6	0.2
S_3	0.1	0.7	0.2

$P[O] = 1 \times 0.2 \times 0.1 \times 0.4 \times 0.3 \times 0.2 \times 0.7$

Ques What is the prob. same weather is Cont. for d days

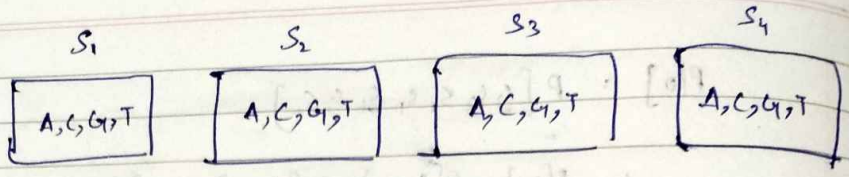
$= \pi_i \times (a_{ii})^{d-1} (1 - a_{ii}) = p_i(d)$

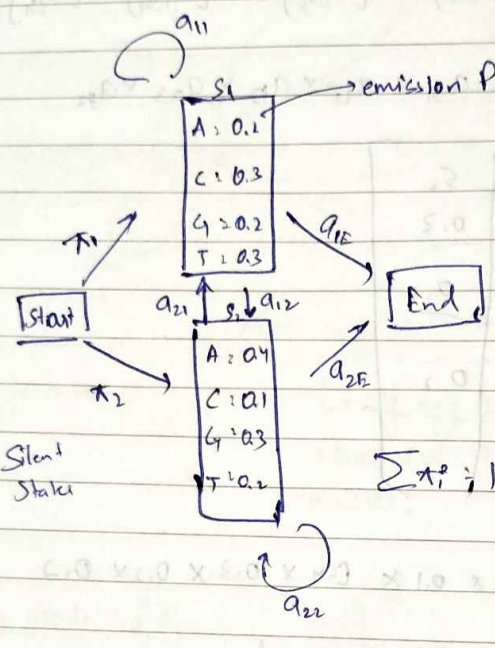
$\bar{d}_i = \sum_{d=1}^{\infty} d \cdot p_i(d) = \sum_{d=1}^{\infty} d_i (a_{ii})^{d-1} (1 - a_{ii}) = \frac{1}{1 - a_{ii}}$

$\bar{d}_1 = \frac{1}{1 - 0.4}$

$\bar{d}_2 = \frac{1}{1 - 0.6}$

$\bar{d}_3 = \frac{1}{1 - 0.2}$





* States are hidden

$a_{11} + a_{12} + a_{1E} = 1$

$a_{22} + a_{21} + a_{2E} = 1$

$\sum \pi_i = 1$

Observation $L = 1$,

$= \{0, 1, E\}_{\text{or } 3}, \{0, 2, E\}, \{$

Observation $L = 2$

$= \{0, 1, 1, 3\}, \{0, 1, 2, 3\}, \{0, 2, 1, 3\}, \{0, 2, 2, 3\}$

Possible Paths $\rightarrow 2^L \rightarrow N^L$