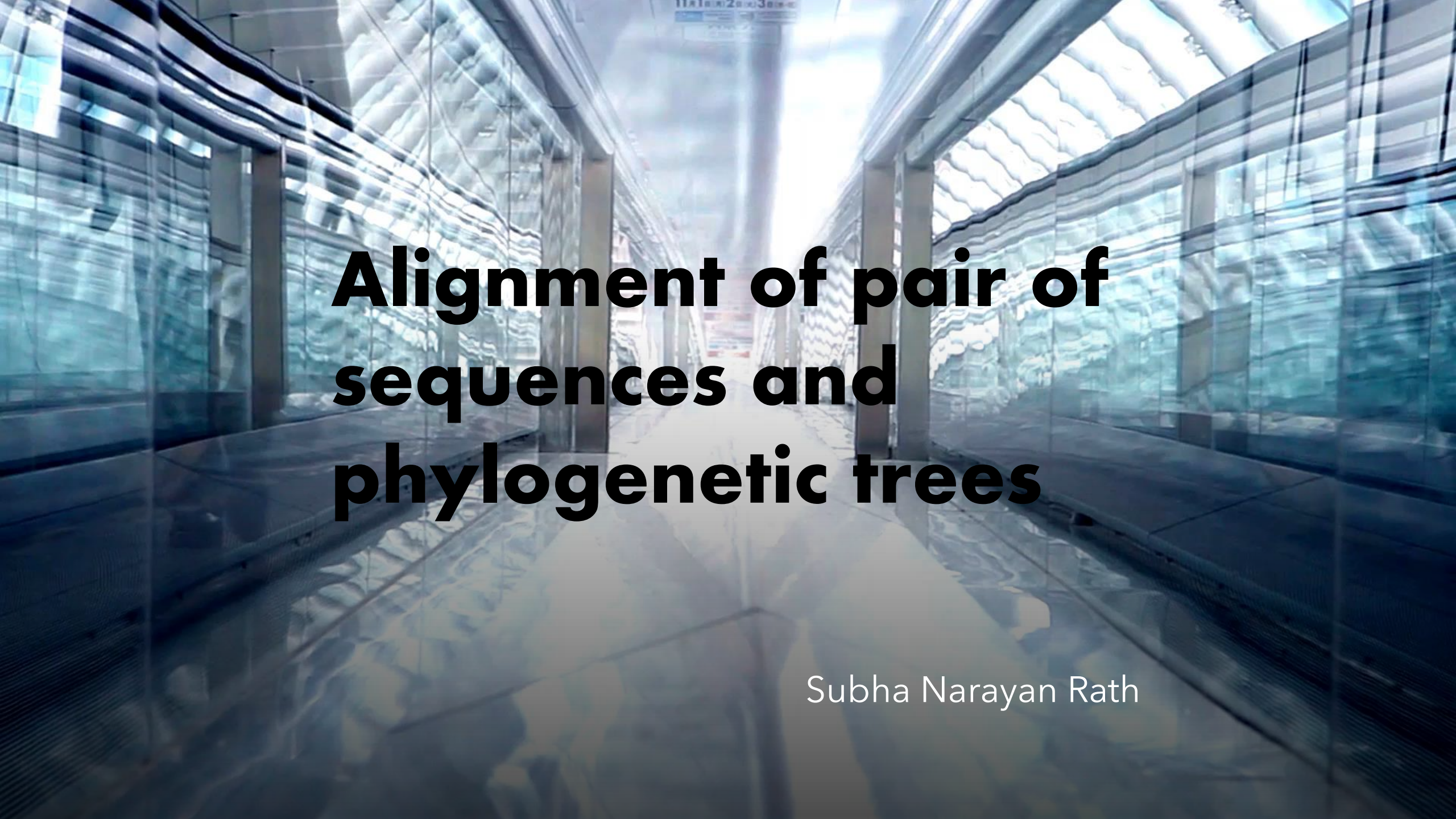
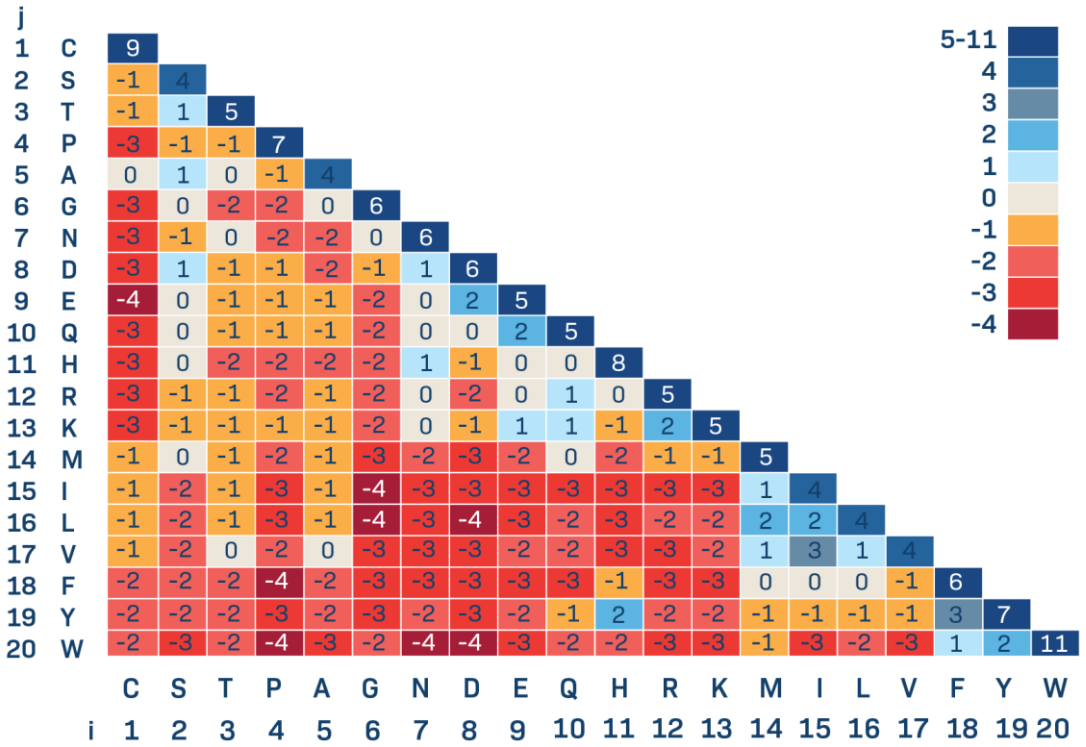


Alignment of pair of sequences and phylogenetic trees

Subha Narayan Rath

BLOSUM matrix of Henikoff and Henikoff

- Using much larger amount of data available now
- Means BLOcks SUbstittuion Matrix and based on BLOCKS database (representing known protein families) of aligned protein sequences
- From family of closely related proteins alignable without gaps… they calculated the ratio of number of observed pairs of amino acids at any position to the number expected from overall amino acid frequencies
- They have sequence identities higher than a threshold e.g. BLOSUM 62% is commonly used where the matrix built using sequences no more than 62% similarity



Scoring insertions and deletions (substitution matrix) or gap weighting

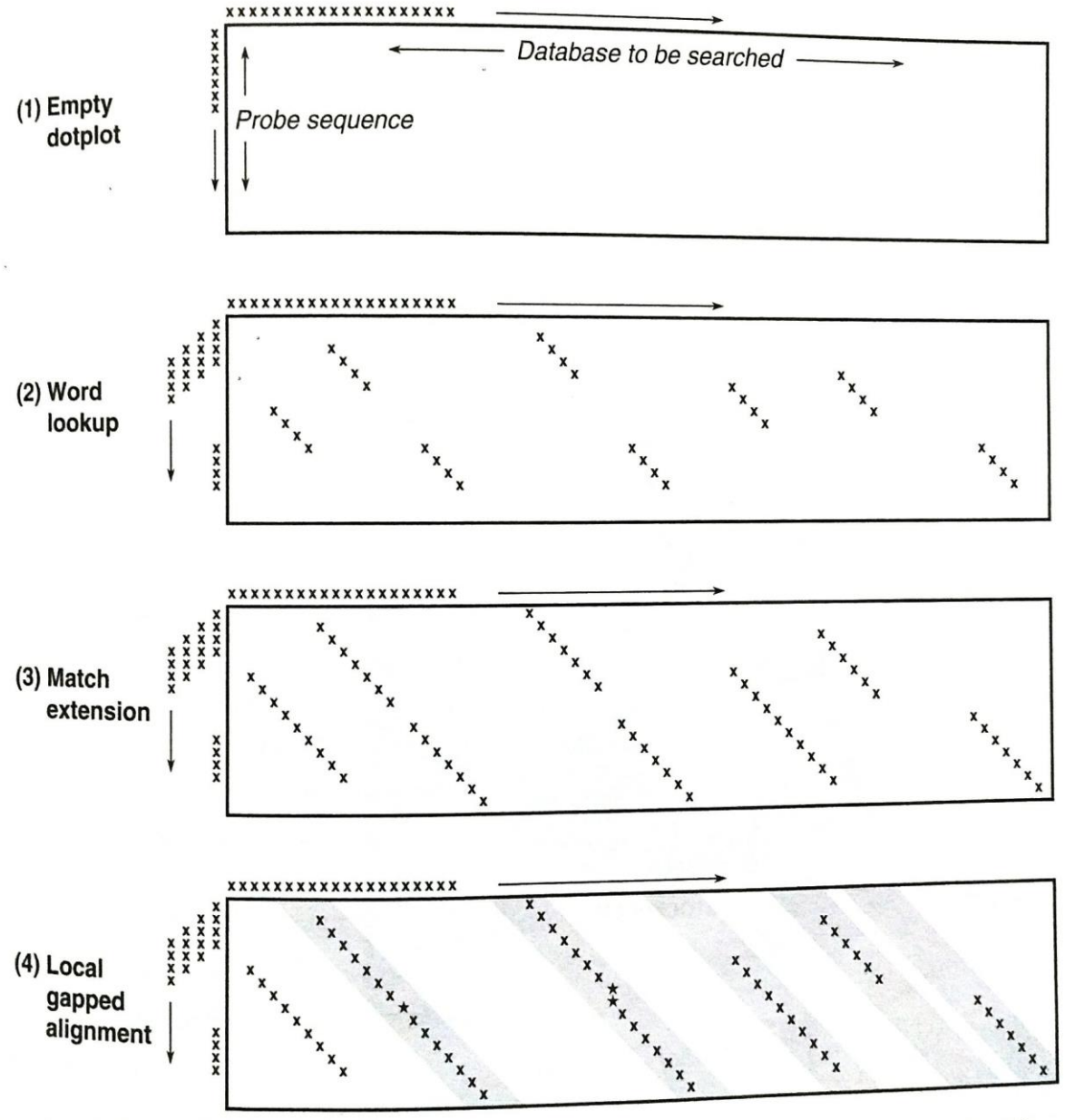
- In addition to substituion matrix: there is a way of gap weighting too
- Aligning DNA sequnces: CLUSTAL-W is recommended
- Alignign Protein sequences: BLOSUM 62 is recommended

	Matrix/ protocol recommended	Gap initiation	Gap extension	Match	Mismatch
DNA	CLUSTAL-W	10	0.1	1	0
Protein	BLOSUM62	11	1	Matrix	Matrix

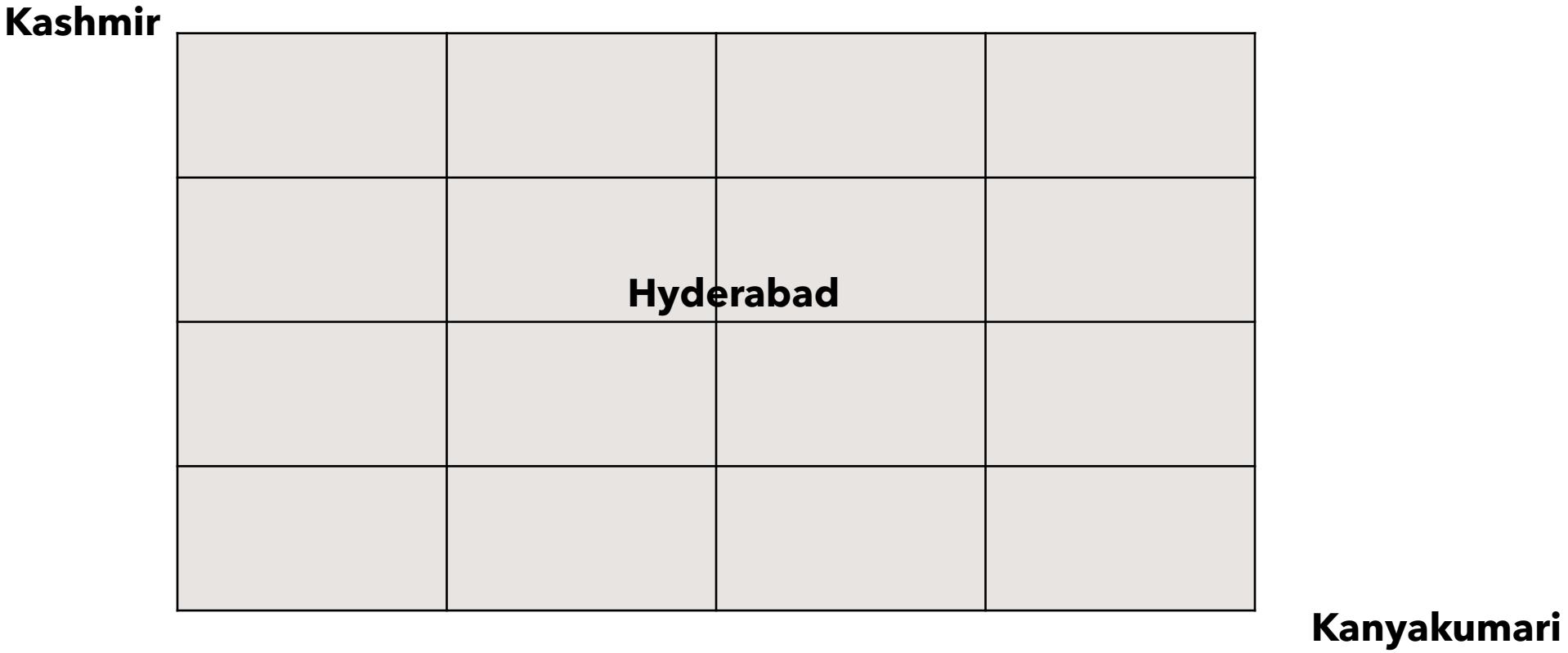
Computing the alignment of two sequences

- An algorithm used for this: dynamic programming and very imp for molecular biology
- Guarantee: to give an optimal global alignment
- Problem1: many alignment may give the same optimal score
- Problem 2: technical: the time required to align two sequences is proportional to n * m, as it is the size of edit matrix that must be filled in
- Variations of the dynamic programming method:
- 1. entire sequence to entire sequence: global match
- 2. region of one sequence to entire other sequence: local match
- 3. region of one to region of another: motif match
- Typical approximation approach would take a small integer k > all instances of each k-tuple of residues in the probe sequence that is found in database sequences

Dynamic programming: for BLAST search



How to travel from start to finish by passing through hyderabad



6+6 only!!
However, detailed algorithm can be read from the book!!

Multiple sequence alignment and database searching

- Searching a database for homologues of known protein is a central theme of bioinformatics
- 3 imp methods are there
- A. Profiles
- B. PSI-BLAST
- C. Hidden Markov models (HMM)
- THE goal is to find high sensitivity or high specificity sequences to find.

A. Profiles

- It express the patterns inherent in a MSA of a set of homologous sequences
- They help in following:
- A. greater accuracy in alignment of distantly related sequences
- B. set of residues that are highly conserved are likely to be part of active site and give clues to function
- C. Identification of other homologous sequences
- D. set of residues which are of little conservation are in surface loops and used for vaccine design
- E. Most structure prediction methods rely on the profiles

Matching MSA of thioredoxins from 25-30 position...

Residue number	Number of occurrences																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
25	1									2								13		
26			16																	
27					16															
28																7	1		5	3
29	16																			
30			1	4									2			1	7	1		

What is score of VDFSAE??

Amino acid colors

Side Chain Chemistry - Amino Acids
Aromatic - Phenylalanine (F), Tryptophan (W), Tyrosine (Y)
Acidic (negatively charged) - Aspartate (D), Glutamate (E)
Basic (positively charged) - Arginine (R), Histidine (H), Lysine (K)
Nonpolar (aliphatic) - Alanine (A), Glycine (G), Isoleucine (I), Leucine (L), Methionine (M), Proline (P), Valine (V)
Polar (neutral) - Cysteine (C), Asparagine (N), Glutamine (Q), Serine (S), Threonine (T)

B. PSI-BLAST

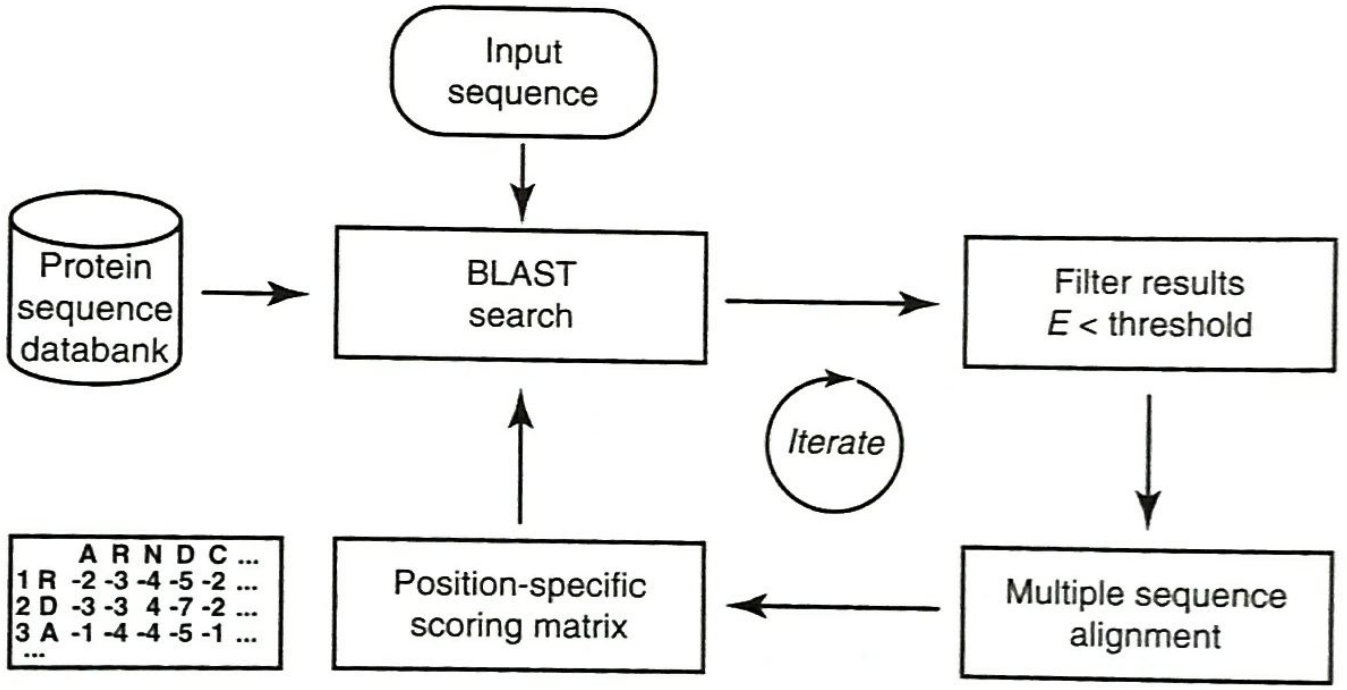
- IT is a program that searches the data bank for sequences similar to a query sequence
- It derives pattern information from a multiple sequence alignment of initial hits
- And reprobes the database using the pattern
- Then it repeats the process, fine tuning the pattern in successive cycles
- Very powerful: in picking distant relationships

E is usually 0.005;

program will make position-specific-scoring matrix

the matrix can be used as an alternative to input sequence and substitution matrix in a BLAST search

the procedure usually converge



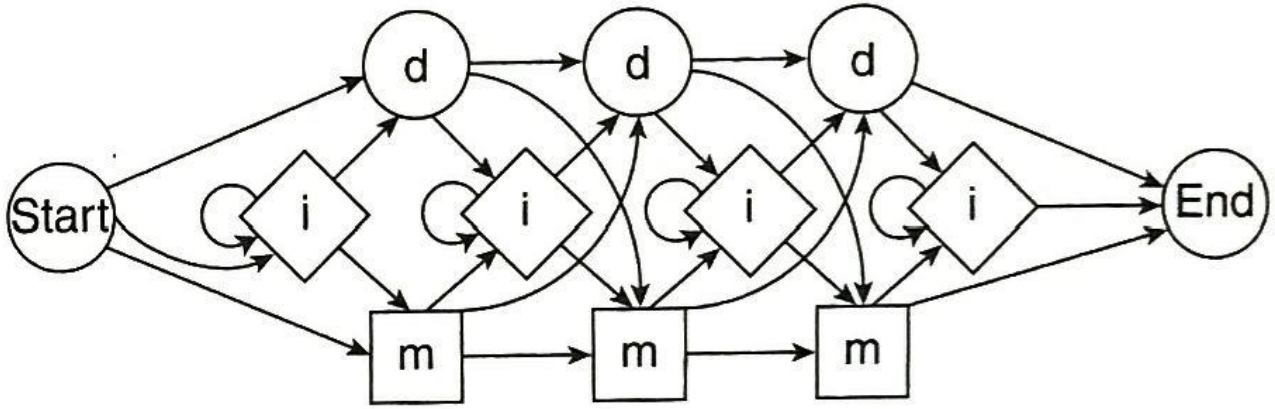
C. Hidden Markov models

- It is a computational structure for describing subtle patterns that define families of homologous sequences
- 1. distant relative prediction
- 2. prediction of protein folding patterns

- HMMs are more general than profiles:
- 1. possibility of introducing gaps into the generated sequence with position specific gap penalties
- 2. HMMs carry out the alignment and the assignment of probabilities together

HMM..output sequence from succession of match state and insert state

- EACH residue position to a MSA, HMM contains a match state (m), delete state (d)
- Insert state (I) appear between residue positions and at the beginning and at the end
- Proabibility of each match state is position-dependent and it emits a match
- Delete state skips a column and starts a gap opening and another delete state from previous delete state causes gap extension
- Insert state: causes new residue that does not correspond to a position in the alignment table appears in the emitted sequence
- Traverse the network without m or d state at each position is not possible



HMM can detect distant homologous

- Only the current state influences the choice of the successor
- The system has no memory of the history
- The succession of amino acids emitted causes the output and visible
- The state sequence that generates the characters remains internal to the system, that is hidden
- By probability distribution associated with the individual states the system models the patterns inherent in a family of sequences