Bioinformatics

Scoring Matrices

David Gilbert

Bioinformatics Research Centre

www.brc.dcs.gla.ac.uk

Department of Computing Science, University of Glasgow

Scoring Matrices

- Learning Objectives
  - To explain the requirement for a scoring system reflecting possible biological relationships
  - To describe the development of PAM scoring matrices
  - To describe the development of BLOSUM scoring matrices

Scoring Matrices

- Database search to identify homologous sequences based on similarity scores
- Ignore position of symbols when scoring
- Similarity scores are additive over positions on each sequence to enable DP
- Scores for each possible pairing, e.g. proteins composed of 20 amino acids, 20 x 20 scoring matrix

Scoring Matrices

- Scoring matrix should reflect
  - Degree of biological relationship between the amino-acids or nucleotides
  - The probability that two AA's occur in homologous positions in sequences that share a common ancestor
    - Or that one sequence is the ancestor of the other
- Scoring schemes based on physico-chemical properties also proposed

Scoring Matrices

- Use of Identity
  - Unequal AA's score zero, equal AA's score 1. Overall score can then be normalised by length of sequences to provide percentage identity
- Use of Genetic Code
  - How many mutations required in NA's to transform one AA to another
    - Phe (Codes UUU & UUC) to Asn (AAU, AAC)
- Use of AA Classification
  - Similarity based on properties such as charge, acidic/basic, hydrophobicity, etc

Scoring Matrices

- Scoring matrices should be developed from experimental data
  - Reflecting the kind of relationships occurring in nature

- Point Accepted Mutation (PAM) matrices
  - Dayhoff (1978)
  - Estimated substitution probabilities
  - Using known mutational (substitution) histories

Scoring Matrices

- Dayhoff employed 71 groups of near homologous sequences (>85% identity)
- For each group a phylogenetic tree constructed
- Mutations accepted by species are estimated
  - New AA must have similar functional characteristics to one replaced
  - Requires strong physico-chemical similarity
  - Dependent on how critical position of AA is to protein
- Employs time intervals based on number of mutations per residue

Scoring Matrices

Overall Dayhoff Procedure:-

- Divide set of sequences into groups of similar sequences – multiple alignment for each group
- Construct phylogenetic tree for each group
- Define evolutionary model to explain evolution
- Construct substitution matrices
  - The substitution matrix for an evolutionary time interval *t* gives for each pair of AA (*a, b*) an estimate for the probability of *a* to mutate to *b* in a time interval *t*.

Scoring Matrices

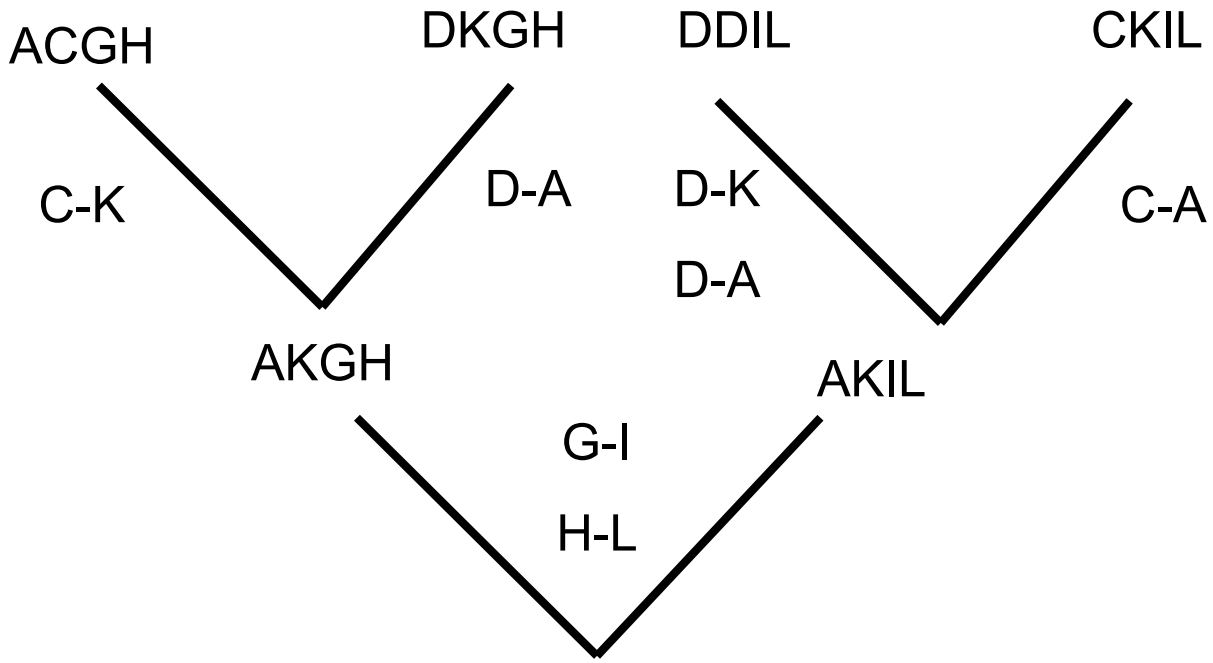
- Evolutionary Model
  - *Assumptions : The probability of a mutation in one position of a sequence is only dependent on which AA is in the position*
  - Independent of position and neighbour AA's
  - Independent of previous mutations in the position

- No need to consider position of AA's in sequence

- Biological clock – rate of mutations constant over time
  - Time of evolution measured by number of mutations observed in given number of AA's. 1-PAM = one accepted mutation per 100 residues

Scoring Matrices

- Calculating Substitution Matrix – count number of accepted mutations



	A	C	D	G	H	I	K	L
A		1	2					
C	1						1	
D	2						1	
G						1		
H								1
I				1				
K		1	1					
L					1			

Scoring Matrices

- Once all accepted mutations identified calculate
  - The number of *a* to *b* or *b* to *a* mutations from table – denoted as f_{ab}
  - The total number of mutations in which *a* takes part – denoted as $f_a = \Sigma_{b \neq a} f_{ab}$
  - The total number of mutations $f = \Sigma_a f_a$ (each mutation counted twice)

- Calculate relative occurrence of AA's
  - p_a *where* $\Sigma_a p_a = 1$

Scoring Matrices

- Calculate the relative mutability for each AA
  - Measure of probability that *a* will mutate in the evolutionary time being considered
- Mutability depends on f_a
  - As f_a *increases so should mutability* m_a *; AA occurring in many mutations indicates high mutability*
  - *As* p_a *increases mutability should decrease ; many occurrences of AA indicate many mutations due to frequent occurrence of AA*
- Mutability can be defined as $m_a = K f_a / p_a$ where K is a constant

Scoring Matrices

- Probability that an arbitrary mutation contains *a*
  - $2f_a / f$
- Probability that an arbitrary mutation is from *a*
  - f_a / f
- For 100 AA's there are $100p_a$ occurrences of *a*
- Probability to select *a* $1/100p_a$
- Probability of any of *a* to mutate
  - $m_a = (1/100p_a) \times (f_a / f)$
- Probability that *a* mutates in 1 PAM time unit defined by m_a

Scoring Matrices

- Probability that *a* mutates to *b* given that *a* mutates is f_{ab} / f_a
- Probability that *a* mutates to *b* in time *t = 1 PAM*
  - $M_{ab} = m_a f_{ab} / f_a$ *when* $a \neq b$

Log-odds PAM 250 matrix

```
X=0
C 12
S  0  2
T -2  1  3
P -3  1  0  6
A -2  1  1  1  2
G -3  1  0 -1  1  5
N -4  1  0 -1  0  0  2
D -5  0  0 -1  0  1  2  4
E -5  0  0 -1  0  0  1  3  4
Q -5 -1 -1  0  0 -1  1  2  2  4
H -3 -1 -1  0 -1 -2  2  1  1  3  6
R -4  0 -1  0 -2 -3  0 -1 -1  1  2  6
K -5  0  0 -1 -1 -2  1  0  0  1  0  3  5
M -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2  0  0  6
I -2 -1  0 -2 -1 -3 -2 -2 -2 -2 -2 -2 -2  2 5
L -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3  4 2  6
V -2 -1  0 -1  0 -1 -2 -2 -2 -2 -2 -2 -2  2 4  2  4
F -4 -3 -3 -5 -4 -5 -4 -6 -5 -5 -2 -4 -5  0 1  2 -1  9
W  0 -3 -3 -5 -3 -5 -2 -4 -4 -4  0 -4 -4 -2 -1 -1 -2  7 10
Y -8 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3  2 -3 -4 -5 -2 -6  0  0 17
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  W  Y
```

Dayhoff mutation matrix (1978) - summary

- Point Accepted Mutation (PAM)
- Dayhof matrices derived from sequences 85% identical
- Evolutionary distance of 1 PAM = probability of 1 point mutation per 100 residues
- Likelihood (*odds*) ratio for residues *a* and *b* :
  Probability a-b is a mutation / probability a-b is chance
- PAM matrices contain *log-odds* figures
  val > 0 : likely mutation
  val = 0 : random mutation
  vak < 0 : unlikely mutation
- 250 PAM : similarity scores equivalent to 20% identity
- low PAM - good for finding short, strong local similarities
  high PAM = long weak similarities

Scoring Matrices

- What about longer evolutionary times ?
- Consider two mutation periods 2-PAM
  - *a* is mutated to *b* in first period and unchanged in second
    - *Probability is* $M_{ab} M_{bb}$
  - *a* is unchanged in first period but mutated to *b* in the second
    - *Probability is* $M_{aa} M_{ab}$
  - *a* is mutated to *c* in the first which is mutated to *b* in the second
    - *Probability is* $M_{ac} M_{cb}$
- Final probability for *a* to be replaced with *b*
  - $M^2_{ab} = M_{ab} M_{bb} + M_{aa} M_{ab} + \Sigma_{c \neq a,b} M_{ac} M_{cb} = \Sigma_{c} M_{ac} M_{cb}$

Scoring Matrices

- Simple definition of matrix multiplication
  - $M^2_{ab} = \Sigma_c M_{ac} M_{cb}$
  - $M^3_{ab} = \Sigma_c M^2_{ac} M_{cb}$ *etc*

- Typically M^{40} M^{120} M^{160} M^{250} *are used in scoring*

- *Low values find short local alignments, High values find longer and weaker alignments*

- Two AA's can be opposite in alignment not as a results of homology but by pure chance

- Need to use odds-ratio $O_{ab} = M_{ab} / P_b$ *(Use of Log)*
  - $O_{ab} > 1$: *b* replaces *a* more often in bologically related sequences than in random sequences where *b* occurs with probability P_b
  - $O_{ab} < 1$: *b* replaces *a* less often in bologically related sequences than in random sequences where b occurs with probability P_b

BLOSUM Scoring Matrices

- PAM matrices derived from sequences with at least 85% identity
- Alignments usually performed on sequences with less similarity
- Henikoff & Henikoff (1992) develop scoring system based on more diverse sequences
- BLOSUM – BLOcks SUbstitution Matrix
- Blocks defined as ungapped regions of aligned AA's from related proteins
- Employed > 2000 blocks to derive scoring matrix

BLOSUM Scoring Matrices

- Statistics of occurrence of AA pairs obtained
- As with PAM frequency of co-occurrence of AA pairs and individual AA's employed to derive Odds ratio
- BLOSUM matrices for different evolutionary distances
  - Unlike PAM cannot derive direct from original matrix
  - Scoring Matrices derived from Blocks with differing levels of identity

BLOSUM Scoring Matrices

- Overall procedure to develop a BLOSUM X matrix
  - Collect a set of multiple alignments
  - Find the Blocks (no gaps)
  - Group segments of Blocks with X% identity
  - Count the occurrence of all pairs of AA's
  - Employ these counts to obtain odds ratio (log)

- Most common BLOSUM matrices are 45, 62 & 80

Scoring Matrices

- Differences between PAM & BLOSUM
  - PAM based on predictions of mutations when proteins diverge from common ancestor – explicit evolutionary model
  - BLOSUM based on common regions (BLOCKS) in protein families
- BLOSUM better designed to find conserved domains
- BLOSUM - Much larger data set used than for the PAM matrix
- BLOSUM matrices with small percentage correspond to PAM with large evolutionary distances
  - BLOSUM64 is roughly equivalent to PAM 120