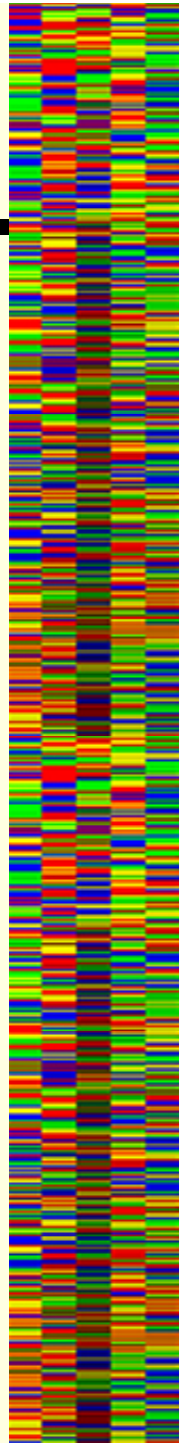
# Genomics

## 8-12 September

| | | | | | |
|---|---|---|---|---|---|
| 6 | M 8 | MG | Scoring Matrices | | Ch 3 and Ch 4 |
| 7 | W 10 | MG | Pairwise Alignment | | |
| 8 | F 12 | MG | Pairwise Alignment | Hw2 | |

## Reading: Mount - ch 3 and 4



# Genomics

## *Sequence database searching - BLAST*

### *blastp*

- compares an amino acid query sequence against a protein sequence database

### *blastn*

- compares a nucleotide query sequence against a nucleotide sequence database

### *blastx*

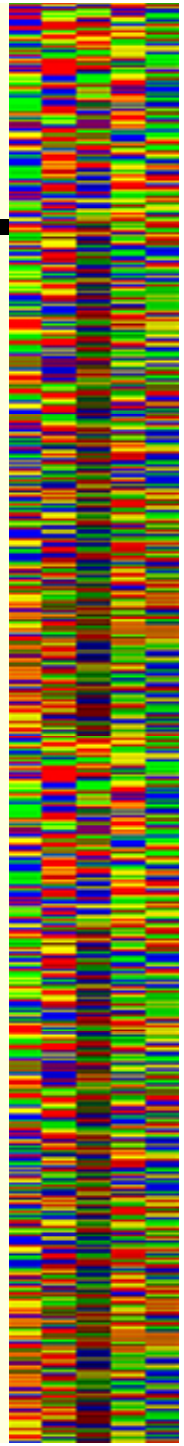
- compares a nucleotide query sequence translated in all reading frames against a protein sequence database, e.g., unknown genomic sequence

### *tblastn*

- compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames, e.g., trying to find exons in genomic DNA

### *tblastx*

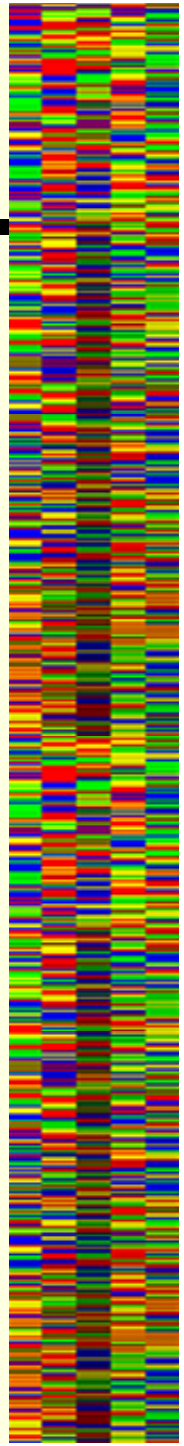
- compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database, e.g. two unknown genes where you don't know the exons



# Genomics

## *Sequence database search - Filtering*

- ***Purpose – remove parts of query that are likely to obscure correct results***
  - repeats
  - low-entropy sequences – regions of low compositional complexity – regions that only have one or two sequence characters
- ***BLASTN uses dust program***
  - low complexity sequences (similar to seg)
  - SINES,LINES, known repeated sequences, vector sequences?
- ***BLASTP uses seg program***
  - Altschul, S. F., M. S. Boguski, W. Gish, J. C. Wootton (1994). Issues in searching molecular sequence databases. Nat Genet 6: 119-129.
  - Wootton, J. C. and S. Federhen (1993). Statistics of local complexity in amino acid sequences and sequence databases. Computers in Chemistry 17:149-163.
  - Wootton, J. C. and S. Federhen (1996). Analysis of compositionally biased regions in sequence databases. Methods in Enzymology 266: 554-571.
- ***Filtered regions are marked with Xs in output, and are not included in search or alignment so a perfect match may not have 100% identity***
  - filtered regions don't count as matches!



## Sequence database search - Filtering

- Calculate compositional complexity, K

$$K = 1 / L \log_N( L!/\Pi n_i!)$$

where L is the window length and N is the alphabet size (4 for DNA, 20 for protein) and $n_i$ is the count of each kind of character

- For GAGAGAGA, $n_G=4$ $n_A=4$ $n_T=n_C=0$

$K = 1/ 8 \log_4(8!/4!4!0!0!) = 1/8 \log_4( 40320 / 576 ) = 1 / 8 \log_4 ( 70 ) = 1 / 8 x3.06$
$K = 0.041$

- Note that only composition, not order is important
- More examples in book

# Genomics

## Sequence database search - Filtering

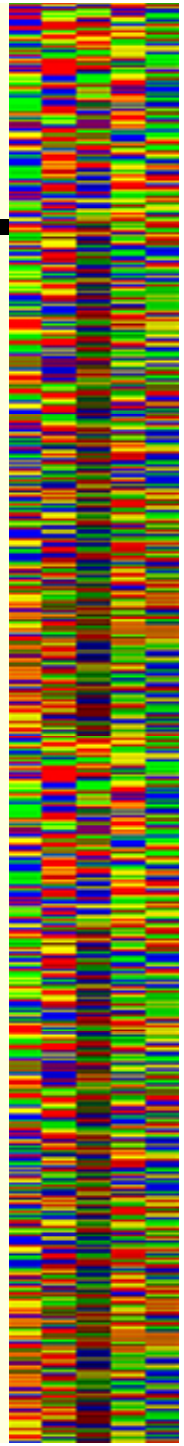
- ***Seg filtering of human MTG8***
- ***putative transcription factor in** acute myeloid leukemia*

*Human MGT8a protein*

| Low-complexity segments | | High-complexity segments |
|---|---|---|
| | 1-24 | MPDRTEKHSTMPDSPVDVKTQSRL |
| tpptmpppptt | 25-35 | |
| | 36-60 | QGAPRTSSFTPTTLTNGTSHSPTAL |
| ngapsppngfsngpsssssssslanqqlpp | 61-89 | |
| | 90-258 | ACGARQLSKLKRFLTTLQQFGNDISPEIGE<br>RVRTLVLGLVNSTLTIEEFHSKLQEATNFP<br>LRPFVIPFLKANLPLLQRELLHCARLAKQN<br>PAQYLAQHEQLLLDASTTSPVDSSELLLDV<br>NENGKRRTPDRTKENGFDREPLHSEHPSKR<br>PCTISPGQRYSPNNGLSYQ |
| pnglphptpppp | 259-270 | |
| | 271-377 | QHYRLDDMAIAHHYRDSYRHPSHRDLRDRN<br>RPMGLHGTRQEEMIDHRLTDREWAEEWKHL<br>DHLLNCIMDMVEKTRRSLTVLRRCQEADRE<br>ELNYWIRRYSDAEDLKK |
| gggsssshs | 378-386 | |
| | 387-554 | RQQSPVNPDPVALDAHREFLHRPASGYVPE<br>EIWKKAEEAVNEVKRQAMTELQKAVSEAER<br>KAHDMITTERAKMERTVAEAKRQAAEDALA<br>VINQQEDSSESCWNCGRKASETCSGCNTAR<br>YCGSFCQHKDWEKHHHICGQTLQAQQQGDT<br>PAVSSSVTPNSGAGSPMD |
| tppaatprsttpgtpstiettp | 555-576 | |
| | 577-577 | R |

# *Scoring Systems*

- ***Simple matching of identical characters (identities) is usually sufficient for DNA.***
  - Scoring systems including transition/transversion are possible and sometimes used
- ***Proteins residues have complex patterns of similarity based on the chemistry of the amino acid side chains.***
- ***Protein comparisons generally use a 20 x 20 comparison table to describe the similarity of the amino acid residues.***
- ***Two major scoring systems for proteins:***
  - PAM or Dayhoff (MDM78, PAM-250, PAM-125)
  - BLOSUM series



# *Scoring Systems*

## *Simple Systems - unclassified situations*

- ***The simplest scoring systems score some kind of property, hydrophobicity is a good example.***

| | | | |
|---|---|---|---|
| ***R*** | ***-4.5*** | ***S*** | ***-0.8*** |
| ***K*** | ***-3.9*** | ***T*** | ***-0.7*** |
| ***D*** | ***-3.5*** | ***G*** | ***-0.4*** |
| ***Q*** | ***-3.5*** | ***A*** | ***1.8*** |
| ***N*** | ***-3.5*** | ***M*** | ***1.9*** |
| ***E*** | ***-3.5*** | ***C*** | ***2.5*** |
| ***H*** | ***-3.2*** | ***F*** | ***2.8*** |
| ***P*** | ***-1.6*** | ***L*** | ***3.8*** |
| ***Y*** | ***-1.3*** | ***V*** | ***4.2*** |
| ***W*** | ***-0.9*** | ***I*** | ***4.5*** |

- ***To tell how surprising, you must have an idea of what kinds of scores you expect to see in sequences. If you are looking for transmembrane sequences, you must know what scores are typical of non-transmembrane sequences***

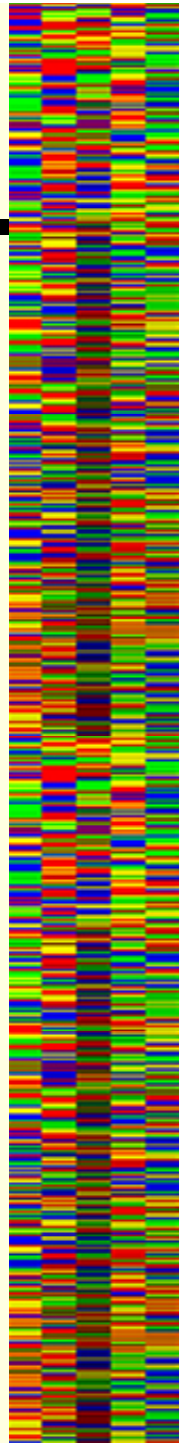
# Scoring Systems

## Simple systems - classified situations

- *Since we know that runs of hydrophobic amino acid residues are common in proteins due to stretches of beta-strand that cross the protein core, we may want to include this information*
- *We can use the frequencies that we observe a specific sequence in transmembrane regions and non-transmembrane regions to make a better discriminator*
- *For instance, consider the three residue sequence IVI and the sequence IAG. Say we count these up and find the following*

| *sequence* | *#tm* | *#non-tm* |
|---|---|---|
| *IAG* | *3/1000* | *4/1000* |
| *IVI* | *12/1000* | *1/1000* |

- *Clearly the IVI sequence makes a better discriminator.*



# Scoring Systems

## Simple Systems - classified situations

- *A simple way to capture our qualitative sense that IVI is more discriminating that IAG is to look at the ratio of the frequency of each sequence in the TM and non-TM classes:*

  $R_{IAG} = F_{IAG,TM} / F_{AIG,NTM} = 0.003/0.004 = 0.75$
  $R_{IVI} = F_{IVI,TM} / F_{IVI,NTM} = 0.012/0.001 = 12.0$

- *If you looked at another sequence, e.g., AST and found*

  $$R_{AST} = F_{AST,TM} / F_{AST,NTM} = 0.001/0.012 = 0.008$$

  *you have the beginning of a system to discriminate TM and non-TM regions.*
- *However, note that if you add up these scores, the scores favoring TM regions will dominate those favoring non-TM regions even though the degree of difference in occurrence is similar as for IVI and AST.*
- *This can be simply adjusted for by taking the log of the ratio, i.e.,*

  $\ln R_{IVI} = 2.48$ $\ln R_{AST} = -2.48$

- *This is a log-odds scoring system.*

# Scoring Systems

## Alignments

- *For alignments, we want a scoring system that discriminates between sequences that represent homology, and unrelated sequences.*
- *A log-odds scoring system is suitable*
- *For alignments we are interested in the frequencies of the amino acid pairs in alignments of homologous proteins vs what we expect to see in unrelated proteins*

```
52 LSDGEWQLVLNVWGKVEADIPGHGQEVL 79
   || .:   |   |||| |.    | | |
 2 LSPADKTNVKAAWGKVGAHAGEYGAEAL 29
```

- *That is L:L, S:S, D:P, G:A*
- *We can count their frequencies is a set of classified examples of true homologous sequence pairs and unrelated sequence pairs to construct a log-odds scoring system*

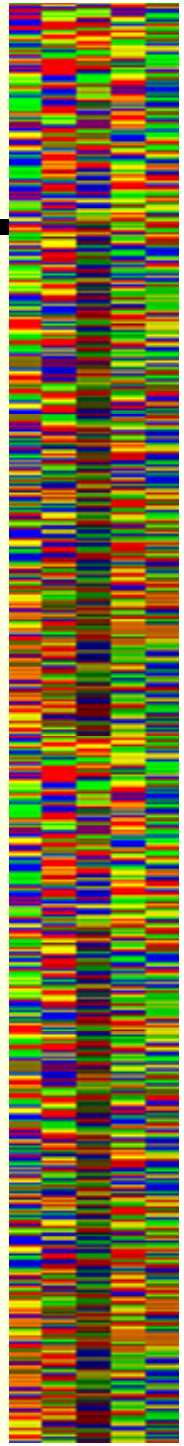
# Scoring Systems

## Log-Odds Scoring

- *Log-Odds scoring systems compare two models, a foreground model, Q, and a background model, P. The background model is often a random sequence model.*
- *The score for comparing two residues, $S_{ij}$, is the log of the ratio of the foreground and background probabilities, i.e., how many times more likely than chance you are to see the residues in the matched in the foreground model.*

$$S_{ij} = ln( q_{ij} / p_{ij} )$$

- *Where $q_{ij}$ and $p_{ij}$ are the scores for comparing residues i and j in the foreground and background models, respectively*



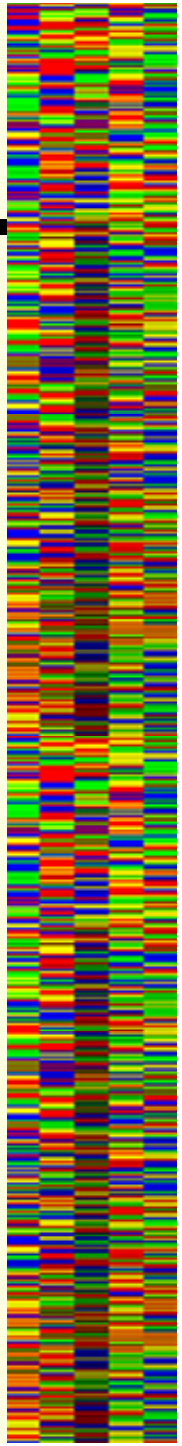
# Scoring Systems

## Log-odds scoring

- *A log-odds scoring system evaluates the relative probabilities of a match representing true homology versus the chance that a match occurs at random, i.e. the relative probability of two models*

$$s_{ij} = \ln( q_{ij} / p_i \; p_j )$$

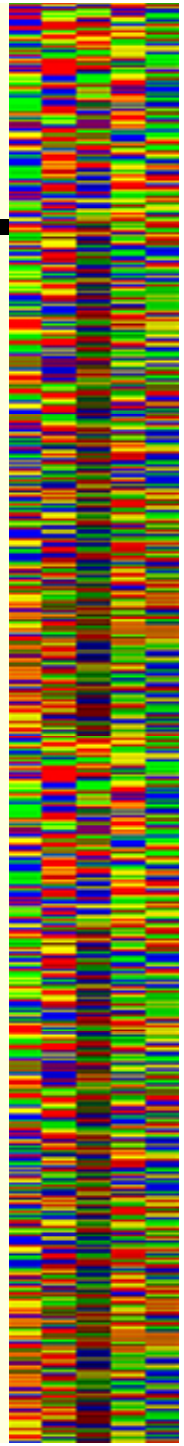
- *Normally, one multiplies probabilities - since these are log probabilities you get the overall probability by adding them up*
- *When added up over a matching segment, you get the probability that the segment represents homology relative to the probability that it represents a random match, i.e. how much more likely than chance is it that the matching segment represents homology*



# Scoring Systems

## PAM vs BLOSUM

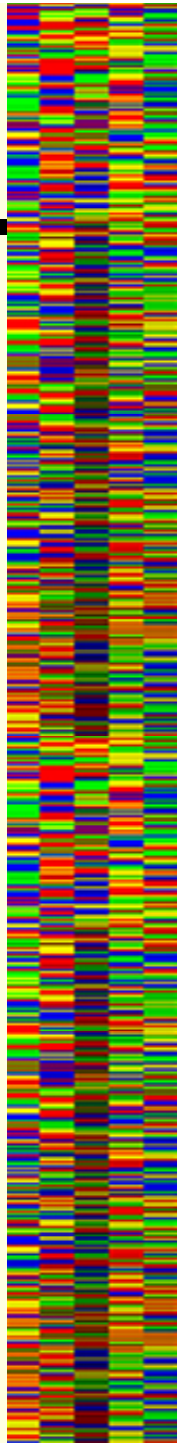
- ***There are two popular scoring models for protein sequences***
- ***PAM (Percent Accepted Mutation)***
  - Based on explicit evolutionary model
  - Represents a specific evolutionary distance
  - Ranges from identical to completely random
- ***BLOSUM (BLOcks SUbstitution Matrix)***
  - Based on empirical frequencies
  - Always a blend of distances as seen in the database and PROSITE
  - Narrower range than PAM matrix



# Scoring Systems

## PAM Matrices

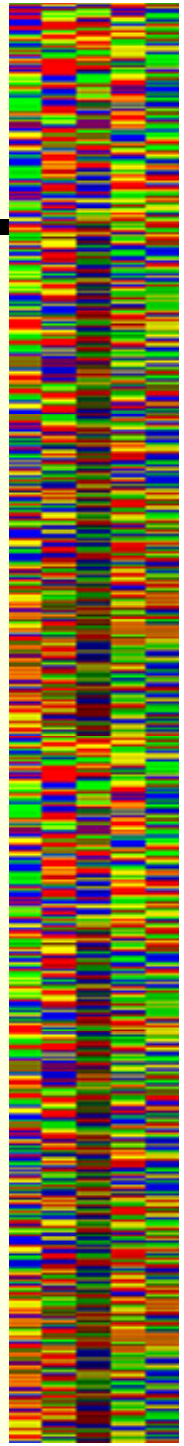
- ***PAM means "percent accepted mutations"***
  - accepted means fixed in the population and is therefore a more complex process than simply mutation
- ***PAM-1 therefore is a scoring system for sequences in which 1% of the residues have undergone mutation***
- ***PAM-250 represents 250% mutation, i.e., an average of 2.5 accepted mutation per residue - a very distant relationship***
- ***PAM tries to model what happens at long evolutionary distances based on a simple Markov model derived from closely related sequences.***



# Scoring Systems

## PAM Matrices

- ***PAM model of evolution was originally proposed by Margaret Dayhoff and co-workers in the 60's***
- ***This work has proven to be very enduring, it is still difficult to do better than this group did in the sixties***
- ***Approach:***
  - Carefully examine the kinds of mutations that occur in closely related protein sequence, *i.e.*, at short evolutionary times
  - Extrapolate these differences to greater mutational distance/longer times



# Scoring Systems

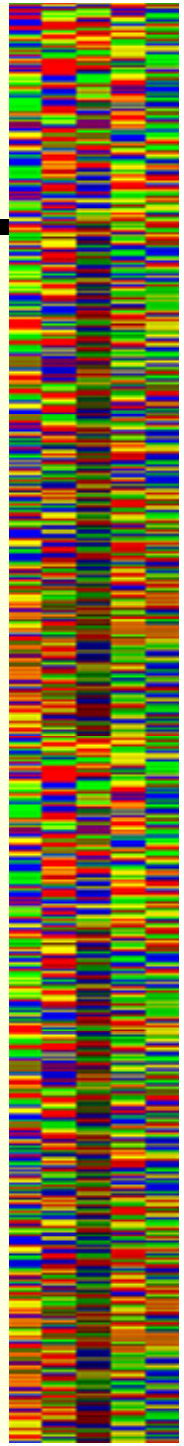
## PAM Matrices

- ***Accepted point mutations - tabulate actual mutations by looking at proteins that are sufficiently closely related that there is no ambiguity in alignment***
  - Sequences no more than 15% different so that changes can be thought of as a single evolutionary step
  - Consider a tree to correctly count changes
  - 1572 changes in 71 families
- ***Mutation probability matrix - probability that residue in column j will be replaced by residue in row i after some amount of evolution***

  $M_{jj} = 1 - \lambda m_j$ $\quad M_{ij} = \lambda m_j A_{ij} / \Sigma A_{ij}$

  $m_i$ ***= mutability of residue i (probability of mutating)***
  $A_{ij}$ ***= number of accepted point mutations***
  $\lambda$ ***= proportionality constant***





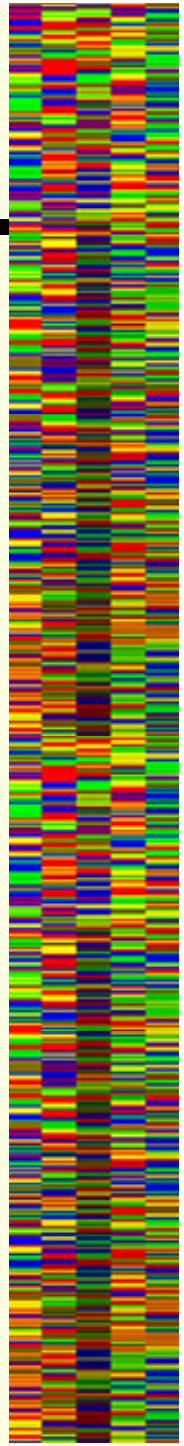
# Scoring Systems

## PAM Matrices

- ***Relatedness odds matrix***
  - log-odds form of the mutation probability matrix.
  - Remember that a log-odds matrix compares two models, in this case a model of relationship by homology (the mutation probability matrix) and relationship by chance

$$R_{ij} = M_{ij} / f_j$$

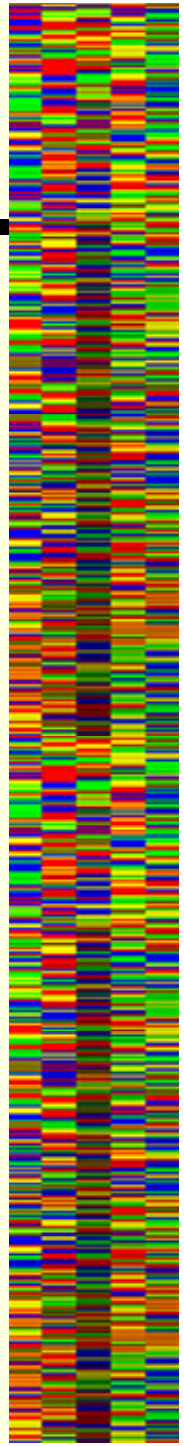
  - Where $f_i$ is the frequency of amino acid residue *i* at random



# Scoring Systems

## PAM Matrices

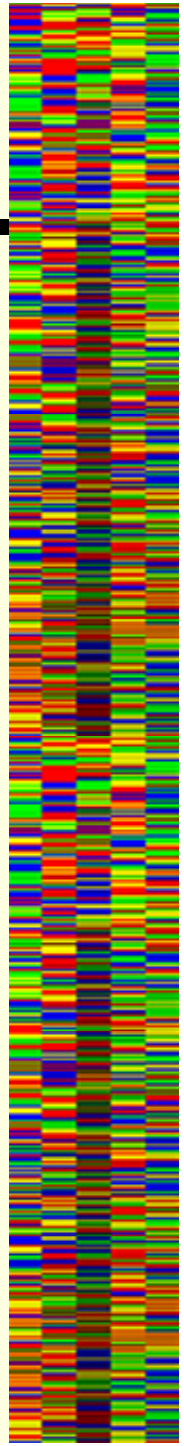
- ***Problems with the PAM approach:***
  - Not all positions are the same
    - Evolutionary rates vary greatly within a sequence
    - Each position has a unique three dimensional environment
    - Environment changes over evolutionary time
  - The most mutable positions were inadvertently selected as the basis for the calculation
    - proteins change more rapidly at the least constrained positions and most slowly at buried “core” positions



# Scoring Systems

## *BLOSUM (BLOcks SUbstitution Matrix)*

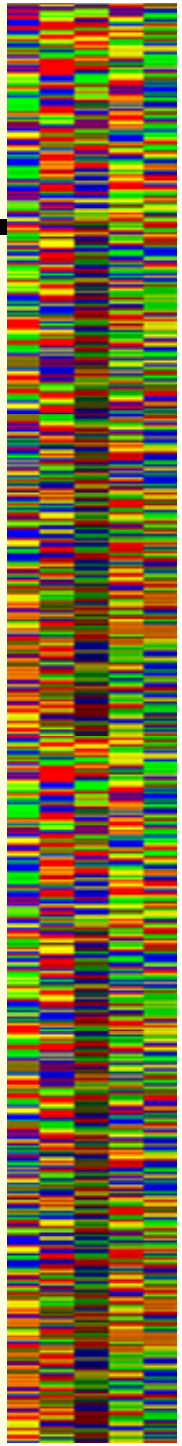
- ***Based on PROSITE signatures***
  - Signatures are short expressions like C-X-X-C-X-X-X-C
- ***Locally align each feature to get "blocks“***
- ***Blocks are locally conserved regions, i.e., more constrained regions likely to be related to structure/function***
- ***Blocks contain sequences at all different evolutionary distances and may be highly biased (e.g. many identical sequences)***



# Scoring Systems

## BLOSUM Matrices

- ***Dealing with bias and distance***
  - Cluster all sequences with less than X% identities
    - Clustered sequences count as 1 sequence
    - if X is 100% it simply removes identical sequences
    - if X < 100% it reduces the weight on closely related sequences
- ***Calculate substitution frequencies and log-odds matrix***
- ***This gives a BLOSUM X table***
  - BLOSUM 62 - sequences greater than 62% identical are clustered
  - BLOSUM 80 - sequences greater than 80% identical are clustered



# Scoring Systems

## BLOSUM Matrices

- *METHYLTRANSFERASE BI*

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

*Unweighted (BLOSUM 100) count of transitions for column 1, total $(n^2 - n)/2$ transitions*

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 4$ $c_{FV} = 2$

$c_{II} = 0$ $c_{IL} = 4$ $c_{IV} = 2$

$c_{LL} = 6$ $c_{LV} = 8$

$c_{VV} = 1$

# Scoring Systems

***BLOSUM Matrices***

***Unweighted (BLOSUM 100) count ($c_{ij}$) of transitions for column 1***

| N=28 | F | I | L | V |
|---|---|---|---|---|
| F | 0 | 1 | 4 | 2 |
| I | 1 | 0 | 4 | 2 |
| L | 4 | 4 | 6 | 8 |
| V | 2 | 2 | 8 | 1 |

***N = 28 transitions, $f_{ij} = c_{ij} / N$***

| | F | I | L | V |
|---|---|---|---|---|
| **F** | 0.00 | 0.04 | 0.14 | 0.07 |
| **I** | 0.04 | 0.00 | 0.14 | 0.07 |
| **L** | 0.14 | 0.14 | 0.21 | 0.29 |
| **V** | 0.07 | 0.07 | 0.29 | 0.04 |

***Log-Odds $s_{ij} = \log_2( f_{ij} / p_i p_j )$ -***
***Background frequencies, $p_i$, from database***

- ***$p_F = 0.0397$***
- ***pI = 0.0529***
- ***$p_L = 0.0917$***
- ***$p_V = 0.0649$***

| | F | I | L | V |
|---|---|---|---|---|
| F | 0.00 | 4.09 | 5.29 | 4.79 |
| I | 4.09 | 0.00 | 4.88 | 4.38 |
| L | 5.29 | 4.88 | 4.67 | 5.59 |
| V | 4.79 | 4.38 | 5.59 | 3.08 |