Profiles and Multiple Alignments

COMP 571

Luay Nakhleh, Rice University

Outline

* Profiles and sequence logos
* Profile hidden Markov models
* Aligning profiles
* Multiple sequence alignment by gradual sequence addition

Profiles and Sequence Logos

Sequence Families

* Functional biological sequences typically come in families
* Sequences in a family have diverged during evolution, but normally maintain the same or a related function
* Thus, identifying that a sequence belongs to a family tells about its function

Profiles

* Consensus modeling of the general properties of the family
* Built from a given multiple alignment (assumed to be correct)

Sequences from a Globin Family

```
Helix                 AAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus            Ls....  v a W kv .  .   g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE            FFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGGG     HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus    v.   f  l . ..  ....    f   . aa. k. .     l sky
```

Alignment of 7 globins

The 8 alpha helices are shown as A-H above the alignment

Ungapped Score Matrices

* A natural probabilistic model for a conserved region would be to specify independent probabilities $e_i(a)$ of observing amino acid a in position i
* The probability of a new sequence x according to this model is

$$\mathbf{P}(x|M) = \prod_{i=1}^{L} e_i(x_i)$$

Log-odds Ratio

* We are interested in the ratio of the probability to the probability of x under the random model

$$S = \sum_{i=1}^{L} \log \frac{e_i(x_i)}{q_{x_i}}$$

Position specific score matrix (PSSM)

Non-probabilistic Profiles

* Gribskov, McLachlan, and Eisenberg 1987
* No underlying probabilistic model, but rather assigned position specific scores for each match state and gap penalty
* The score for each consensus position is set to the average of the standard substitution scores from all the residues in the corresponding multiple sequence alignment column

Non-probabilistic Profiles

```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                 ***  *****
```

The score for residue 'a' in column 1

$$\tfrac{5}{7}s(\mathrm{V},a)+\tfrac{1}{7}s(\mathrm{F},a)+\tfrac{1}{7}s(\mathrm{I},a)$$

s(a,b) : standard substitution matrix

Non-probabilistic Profiles

* They also set gap penalties for each column using a heuristic equation that decreases the cost of a gap according to the length of the longest gap observed in the multiple alignment spanning the column

Representing a Profile as a Logo

* The score parameters of a PSSM are useful for obtaining alignments, but do not easily show the residue preferences or conservation at particular positions.
* This residue information is of interest because it is suggestive of the key functional sites of the protein family.

Representing a Profile as a Logo

* A suitable graphical representation would make the identification of these key residues easier.
* One solution to this problem uses information theory, and produces diagrams that are called logos.

Representing a Profile as a Logo

* In any PSSM column u, residue type a will occur with a frequency $f_{u,a}$.
* The entropy in that is position is defined by

$$H_u = -\sum_a f_{u,a} \log_2 f_{u,a}$$

Representing a Profile as a Logo

* The maximum value of H_U occurs if all residues are present with equal frequency, in which case H_U takes the value $\log_2(20)$ for proteins.

Representing a Profile as a Logo

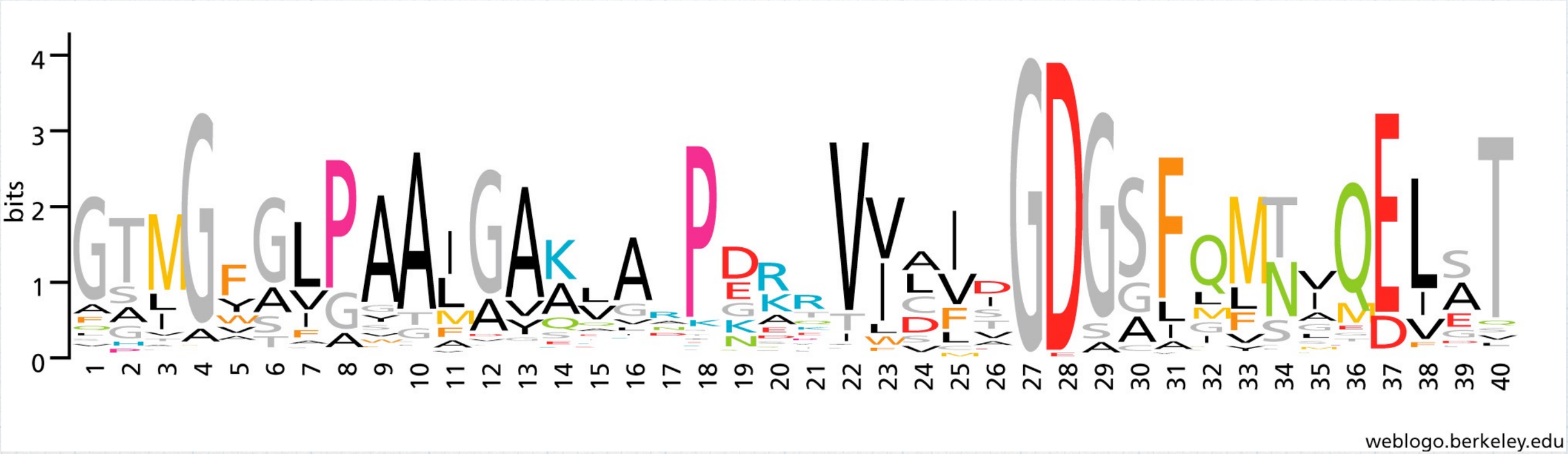
* The information present in the pattern at position u is given by

$$I_u = \log_2 20 - H_u$$

Representing a Profile as a Logo

* If the contribution of a residue is defined as $f_{u,a}I_u$, then a logo can be produced where at every position the residues are represented by their one-letter code, with each letter having a height proportional to its contribution.

Representing a Profile as a Logo



Profile HMMs

Problem with the Approach

* If we had an alignment with 100 sequences, all with a cysteine (C), at some position, the probability distribution for that column for an "average" profile would be exactly the same as would be derived from a single sequence
* Doesn't correspond to our expectation that the likelihood of a cysteine should go up as we see more confirming examples

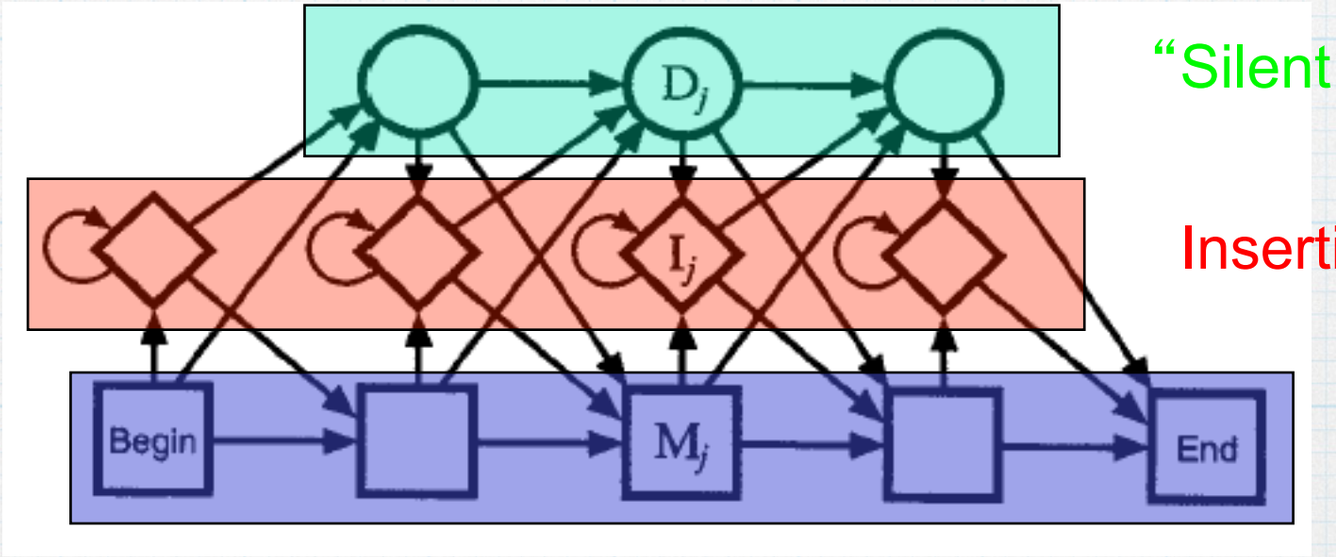
Similar Problem with Gaps

```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                 ***  *****
```

Scores for a deletion in columns 2 and 4 would be set to the same value

More reasonable to set the probability of a new gap opening to be higher in column 4

Adding Indels to Obtain a Profile HMM



Profile HMMs generalize pairwise alignment

Deriving Profile HMMs from Multiple Alignments

* Essentially, we want to build a model representing the consensus sequence for a family, rather than the sequence of any particular member
* Non-probabilistic profiles and profile HMMs

Basic Profile HMM Parameterization

* A profile HMM defines a probability distribution over the whole space of sequences
* The aim of parameterization is to make this distribution peak around members of the family
* Parameters: probabilities and the length of the model

Model Length

* A simple rule that works well in practice is that columns that are more than half gap characters should be modeled by inserts

Probability Values

$$a_{k\ell} = \frac{A_{k\ell}}{\sum_{\ell'} A_{k\ell'}} \qquad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')}$$

k, ℓ : indices over states

$a_{k\ell}, e_k(a)$: transition and emission probabilities

$A_{k\ell}, E_k(a)$: transition and emission frequencies

Problem with the Approach

* Transitions and emissions that don't appear in the training data set would acquire zero probability (would never be allowed)
* Solution: add pseudo-counts to the observed frequencies
* Simples pseudo-count is Laplace's rule: add one to each frequency

Example

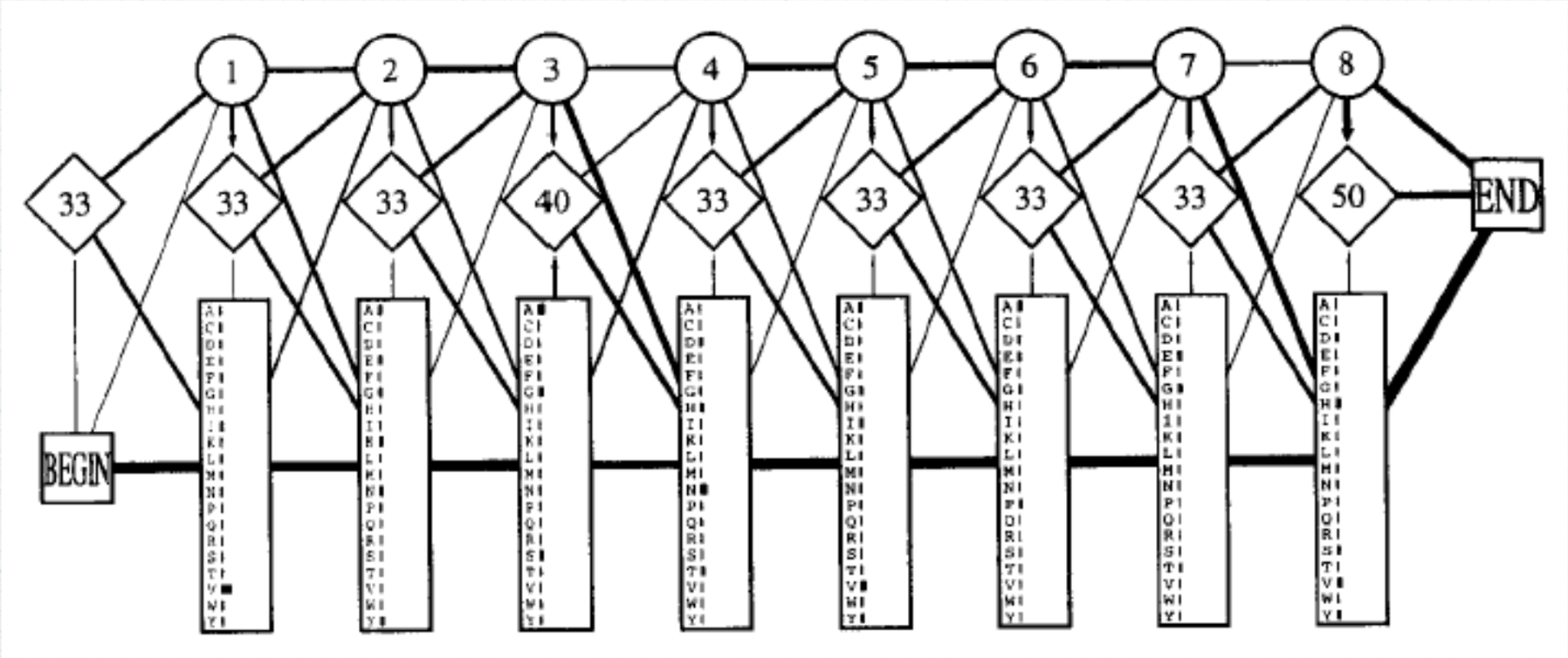
```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                  ***  *****
```

$e_{M_1}(\mathtt{V}) = 6/27, e_{M_1}(\mathtt{I}) = e_{M_1}(\mathtt{F}) = 2/27$

$e_{M_1}(a) = 1/27$ for all residue types a other than V, I, F

$a_{M_1M_2} = 7/10, a_{M_1D_2} = 2/10$ and $a_{M_1I_1} = 1/10$

Example: Full Profile HMM



Searching with Profile HMMs

* One of the main purposes of developing profile HMMs is to use them to detect potential membership in a family
* We can either use Viterbi algorithm to get the most probable alignment or the forward algorithm to calculate the full probability of the sequence summed over all possible paths

Viterbi Algorithm

$$
\begin{aligned}
V_j^{\mathrm{M}}(i) &= \log \frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_{j-1}^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_{j-1}\mathrm{M}_j}; \end{cases} \\
V_j^{\mathrm{I}}(i) &= \log \frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_j^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_j\mathrm{I}_j}, \\ V_j^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_j\mathrm{I}_j}, \\ V_j^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_j\mathrm{I}_j}; \end{cases} \\
V_j^{\mathrm{D}}(i) &= \max \begin{cases} V_{j-1}^{\mathrm{M}}(i) + \log a_{\mathrm{M}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{I}}(i) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{D}}(i) + \log a_{\mathrm{D}_{j-1}\mathrm{D}_j}. \end{cases}
\end{aligned}
$$

Forward Algorithm

$$
\begin{aligned}
F_j^{\mathrm{M}}(i) &= \log \frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \log\Big[a_{\mathrm{M}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{M}}(i-1)\big) \\
&\qquad + a_{\mathrm{I}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{I}}(i-1)\big) + a_{\mathrm{D}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{D}}(i-1)\big)\Big]; \\
F_j^{\mathrm{I}}(i) &= \log \frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \log\Big[a_{\mathrm{M}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{M}}(i-1)\big) \\
&\qquad + a_{\mathrm{I}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{I}}(i-1)\big) + a_{\mathrm{D}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{D}}(i-1)\big)\Big]; \\
F_j^{\mathrm{D}}(i) &= \log\Big[a_{\mathrm{M}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{M}}(i)\big) + a_{\mathrm{I}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{I}}(i)\big) \\
&\qquad + a_{\mathrm{D}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{D}}(i)\big)\Big].
\end{aligned}
$$

Aligning Profiles

* Aligning two PSSMs or profile HMMs can be effective at identifying remote homologs and evolutionary links between protein families.

Comparing Two PSSMs by Alignment

* The alignment of two PSSMs cannot proceed by a standard alignment technique.
* Consider the alignment of two columns, one from each PSSM.
* As neither represents a residue, but just scores, there is no straightforward way of using them together to generate a score for use in an alignment algorithm.

Comparing Two PSSMs by Alignment

* The solution to this problem is to use measures of the similarity between the scores in the two columns.

Comparing Two PSSMs by Alignment

* The program LAMA (Local Alignment of Multiple Alignments) solves one of the easiest formulations of this problem, not allowing any gaps in the alignment of the PSSMs.

Comparing Two PSSMs by Alignment

* Consider two PSSMs A and B that consist of elements $m^A_{u,a}$ and $m^B_{v,a}$ for residue type a in columns u and v, respectively.
* LAMA uses the Pearson correlation coefficient r_{A_u,B_v} defined as

$$r_{A_u,B_v} = \frac{\sum_a (m^A_{u,a} - \bar{m}^A_u)(m^B_{v,a} - \bar{m}^B_v)}{\sqrt{\sum_a (m^A_{u,a} - \bar{m}^A_u)^2 \sum_a (m^B_{v,a} - \bar{m}^B_v)^2}}$$

Comparing Two PSSMs by Alignment

* The correlation value ranges from 1 (identical columns) to -1.
* The score of aligning two PSSMs is defined as the sum of the Pearson correlation coefficients for all aligned columns.

Comparing Two PSSMs by Alignment

* As no gaps are permitted in aligning two PSSMs, all possible alignments can readily be scored by simply sliding one PSSM along the other, allowing for overlaps at either end of each PSSM.
* The highest-scoring alignment is then taken as the best alignment of the two families.

Comparing Two PSSMs by Alignment

* Assessing the significance of a given score:
  * The columns of the PSSMs are shuffled many times, recording the possible alignment scores at each time, and then the z-score is computed.

Comparing Two PSSMs by Alignment

* Once a significance alignment has been detected, a plot of the correlation coefficient values can help to identify the columns for which the families have similar residues.

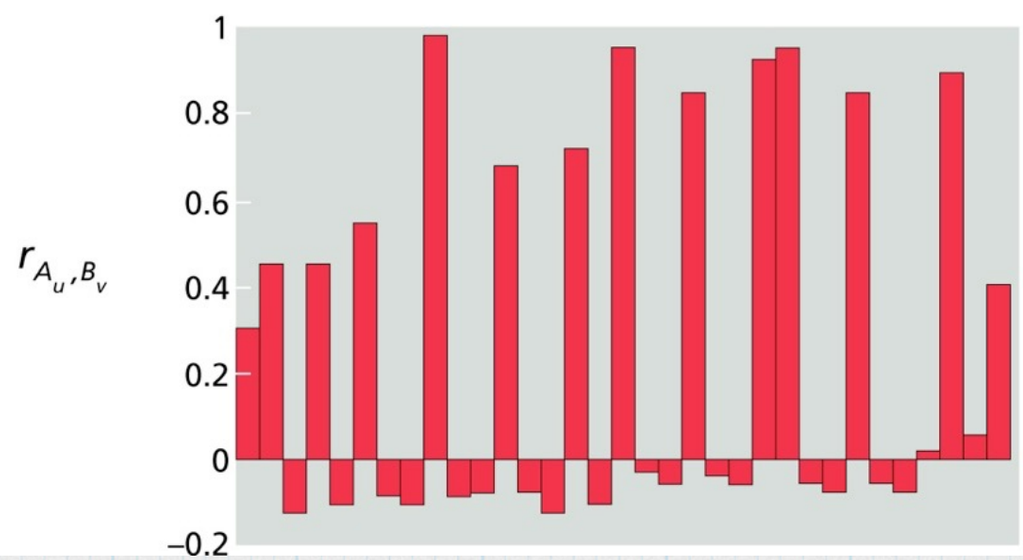
Comparing Two PSSMs by Alignment

(A)

```
OXDA_FUSSO   319   LDDETWIVHNYGHSGWGYQGSYGCAENVVQLVD   351
OXDD_BOVIN   294   DSRRLPVVHHYGHGSGGIAMHWGTALEATRLVN   326
OXDA_HUMAN   299   GPSNTEVIHNYGHGGYGLTIHWGCALEAAKLFG   331
OXDA_MOUSE   297   GSSSAEVIHNYGHGGYGLTIHWGCAMEAANLFG   329
OXDA_PIG     299   GSSNTEVIHNYGHGGYGLTIHWGCALEVAKLFG   331
OXDA_RABIT   299   GPSKTEVIHNYGHGGYGLTIHWGCALEAAKLFG   331

DHSA_BACSU   229   GEFIQIHPTAIPGDDKLRLMSESARGEGGRVWT   261
DHSA_ECOLI   234   QDMEMWQFHPTGIAGAGVLVTEGCRGEGGYLLN   266
FRDA_WOLSU   249   GNMEAVQFHPTPLFPSGILLTEGCRGDGGILRR   281
DHSA_BOVIN   289   QDLEFVQFHPTGIYGAGCLITEGCRGEGGILIN   321
DHSA_RICPR   238   QDMEFVQFHPTGIYGAGCLITEGARGEGGYLVN   270
DHSA_YEAST   279   QDLEFVQFHPSGIYGSGCLITEGARGEGGFLVN   311
FRDA_ECOLI   224   RDMEFVQYHPTGLPGSGILMTEGCRGEGGILVN   256
FRDA_PROVU   225   RDMEFVQYHPTGLPGSGILMTEGCRGEGGILVN   257
```

(B)



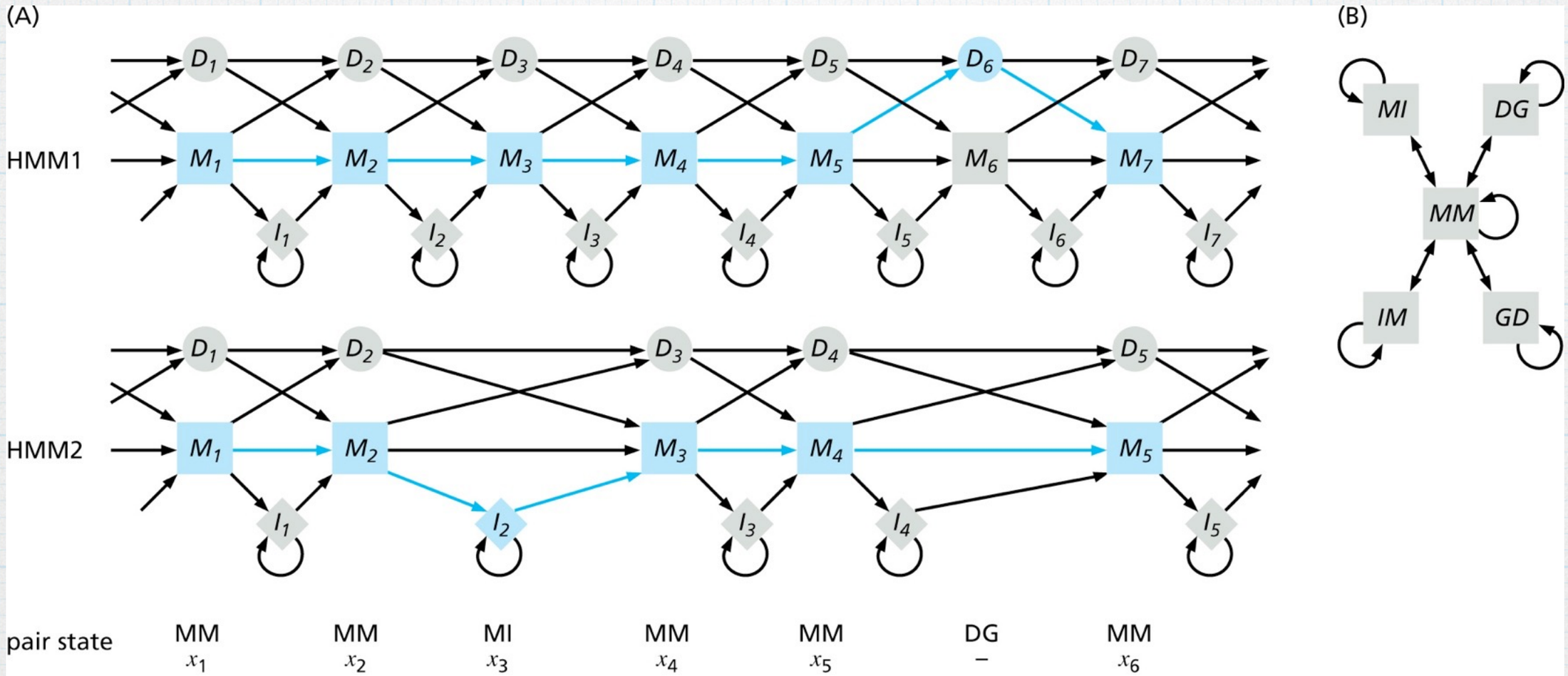
Aligning Profile HMMs

* One way to align two alignments is to turn one into a profile HMM, and then modify Viterbi's algorithm to find the most probable set of paths, which together emit the other alignment (this is the basis for method COACH: COmparison of Alignments by Constructing HMMs).

Aligning Profile HMMs

* The HHsearch method aligns two profile HMMs and is designed to identify very remote homologs.
* It uses a variant of the Viterbi algorithm to find the alignment of the two HMMs that has the best log-odds score.

Aligning Profile HMMs



Multiple Sequence Alignment by Gradual Sequence Addition

* Multiple alignments are more powerful for comparing similar sequences than profiles because they align all the sequences together, rather than using a generalized representation of the sequence family.

* One way of building a multiple alignment is simply to superpose each of the pairwise alignments.
* However, this method is unlikely to give the optimal multiple alignment.

* The pairwise dynamic programming algorithms we described can be modified to optimally align more than two sequences.

* However, this approach is computationally inefficient, and is infeasible in practice.

* As a result, many alternative multiple alignment methods have been proposed, which are not guaranteed to find the optimal alignment but can nevertheless find good alignments.

Progressive Alignment

* The majority of heuristic methods create a multiple alignment by gradually building it up, adding sequences one at a time.
* This is often referred to as <u>progressive alignment</u>.

Progressive Alignment

* The order in which sequences are added to the alignment makes a big difference in the quality of the produced alignment.

Progressive Alignment

* One technique to determine a "good" order:
  * Compute pairwise similarity
  * Build a phylogenetic tree
  * Use the tree to guide the multiple alignment

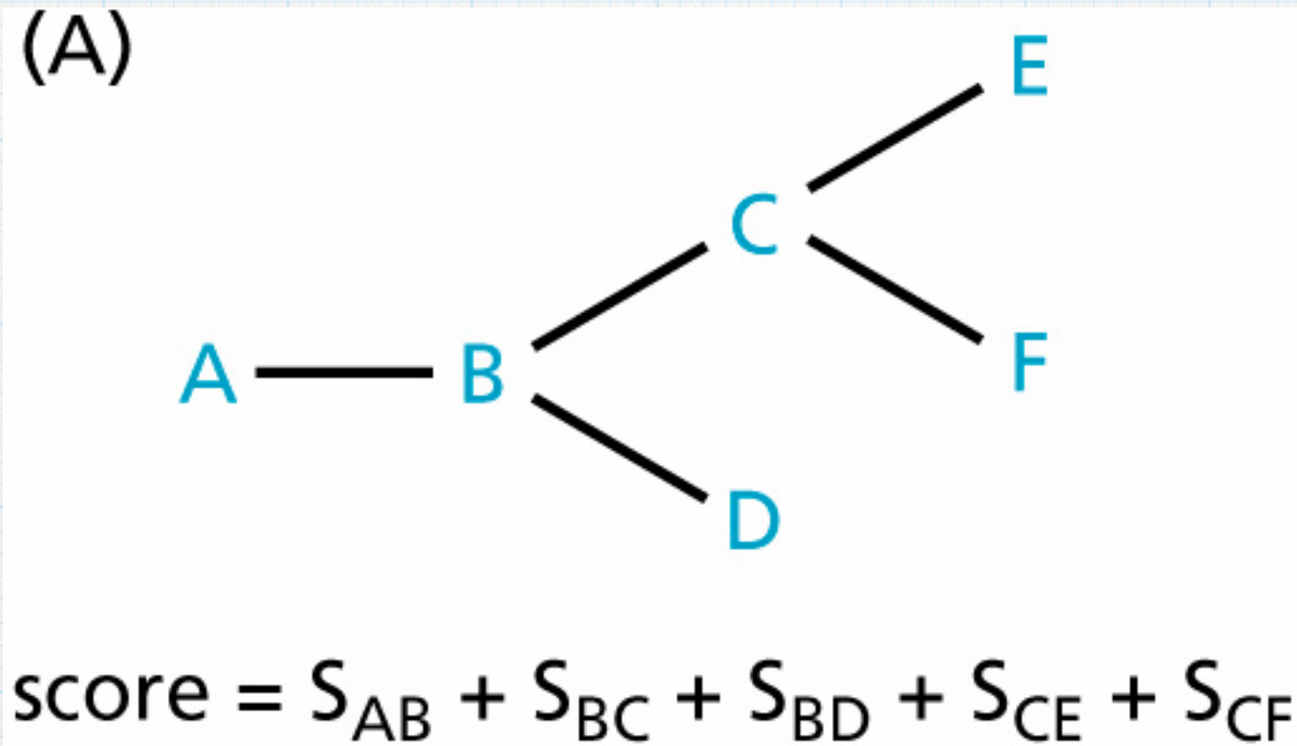
Progressive Alignment

* For example, ClustalW and T-Coffee perform Needleman-Wunsch global alignment for every pair of sequences, and from these alignments obtain the measure used in constructing the guide tree.

Progressive Alignment

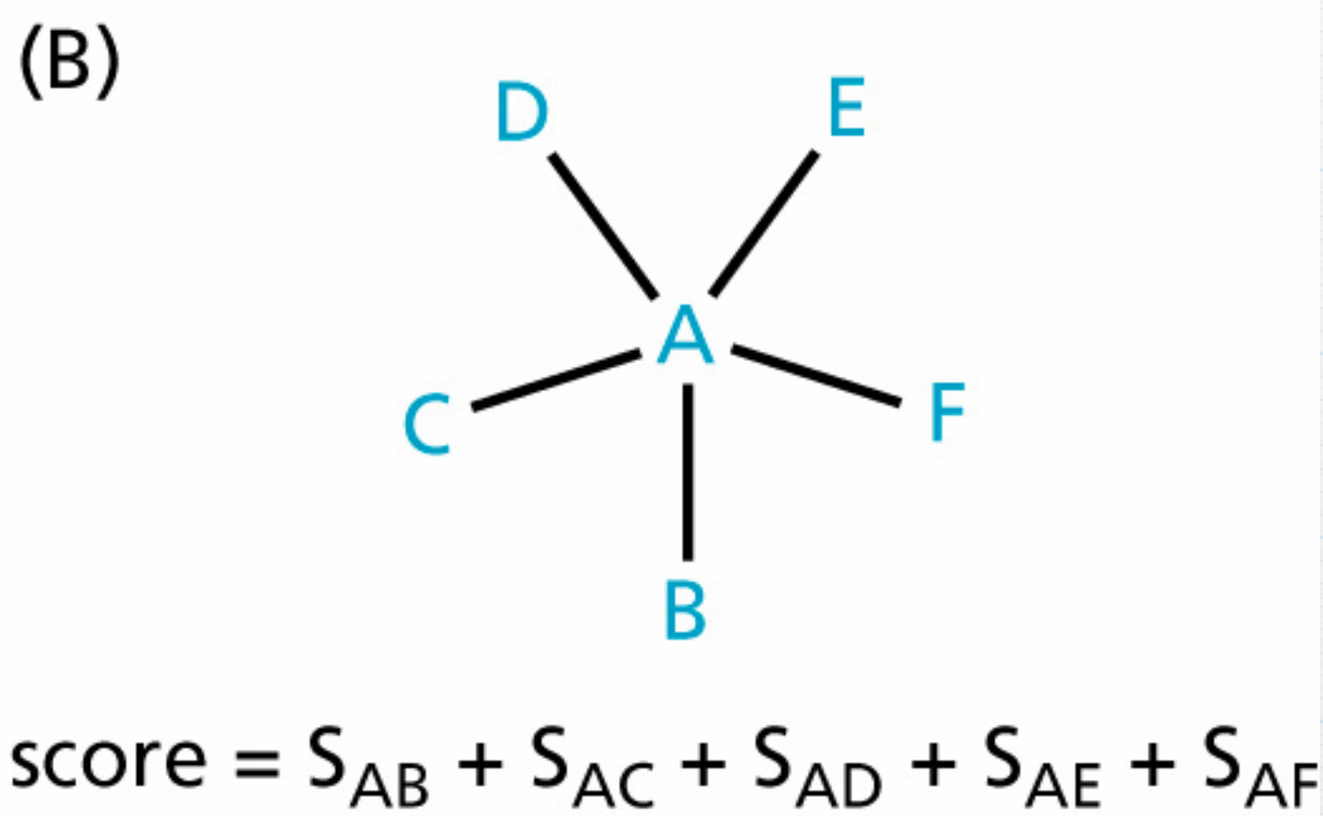
* When the number of sequences is very large, pairwise global alignments of all sequence pairs can take a very long time.
* Methods such as MUSCLE and MAFFT use approximation techniques to quantify pairwise (dis)similarities.

Scoring Schemes for Multiple Alignments



Ideally!

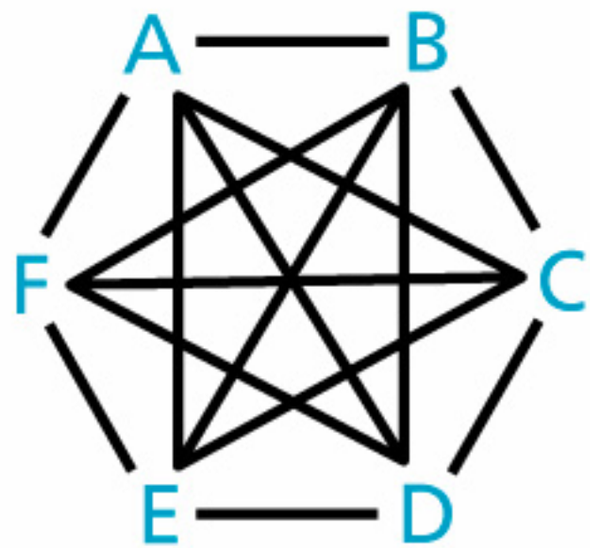
Scoring Schemes for Multiple Alignments



Star

Scoring Schemes for Multiple Alignments

(C)



$$\text{score} = S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF} + S_{BC} + S_{BD} + S_{BE} + S_{BF} + S_{CD} + S_{CE} + S_{CF} + S_{DE} + S_{DF} + S_{EF}$$

Sum-of-pairs (SP)

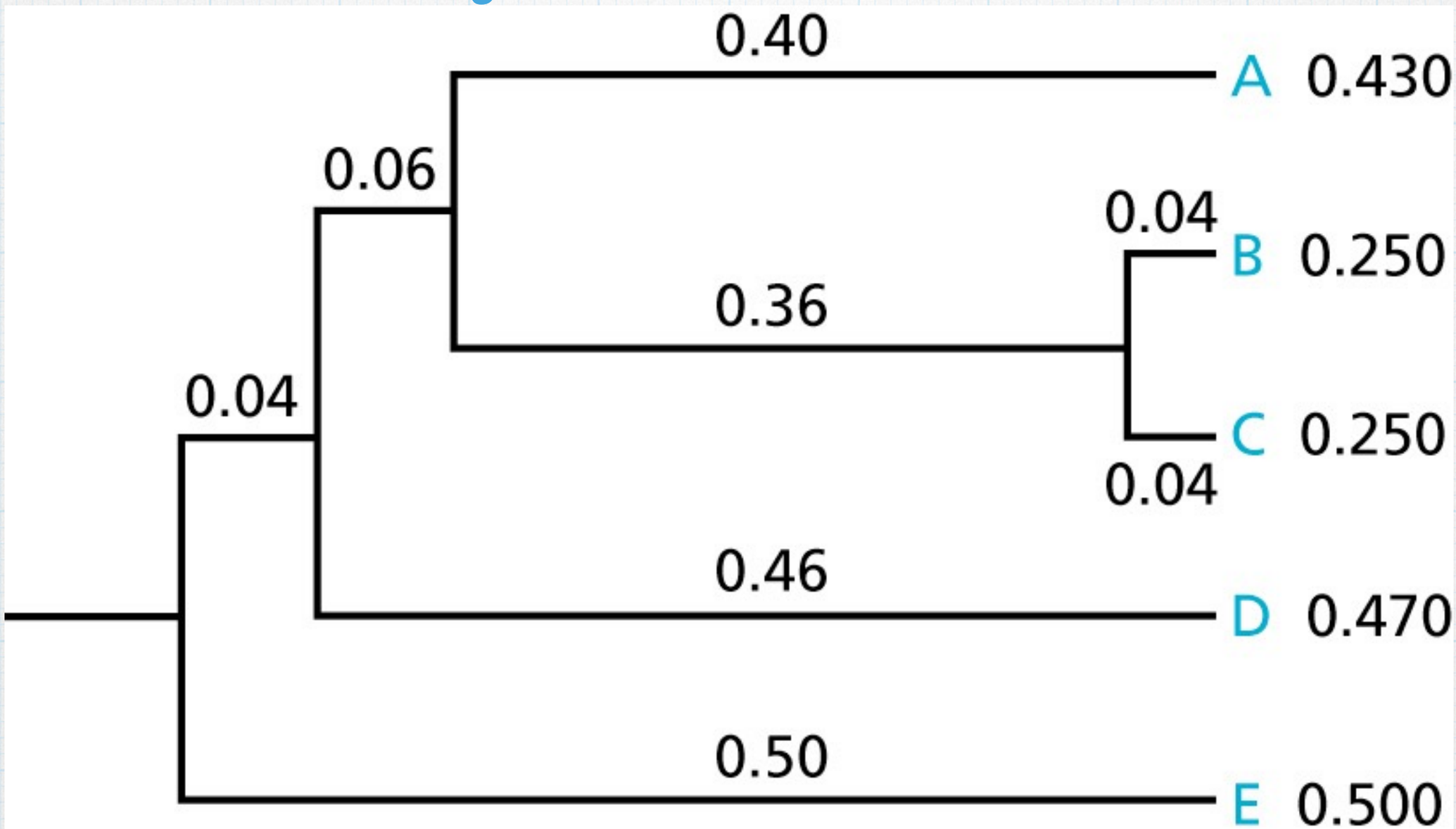
Scoring Schemes for Multiple Alignments

* When SP is used, all sequences should not be regarded as equally independent or useful, and should be weighted to take account of this.
* For example, two identical sequences give exactly the same information as just one of them, whereas two very different sequences give significantly more information than either of them individually.

Scoring Schemes for Multiple Alignments

* One way to weight a sequence is to use the sum of branch lengths from the sequence at the leaf to the root of the guide tree, with each branch length being divided by the number of leaves "under" it (this is the weighting scheme used in ClustalW).

Scoring Schemes for Multiple Alignments



Questions?