BioE/MCB/PMB C146/246, Spring 2005
Problem Set 5

1. MSA is able to run quickly in some cases if it can set a high lower bound on the alignment score. The lower bound is based on the score of a fast, heuristic multiple sequence alignment. The highest possible score for any sum-of-pairs score is obtained if each pairwise alignment is the optimal pairwise alignment (usually not the actual result of an m.s.a.). So, the difference between the score of the bad m.s.a. and the sum of all but one *optimal* pairwise alignments is the lower bound for the remaining pairwise alignment in the actual m.s.a. This pairwise minimum is used to find the set of all position pairs in those two sequences that contribute to alignments scoring above the minimum.

If some aspect of the sequences makes the pairwise minimum a lot lower than than the pairwise score for those sequences in the full dynamic programming, the ”volume” of the dynamic programming matrix which must be computed is much larger and MSA will be much slower. The pairwise minimum could be low if the original heuristic m.s.a. score is too low, or if the sum of optimal pairwise alignments is too high.

2. Some alignment programs put no cost on adding a gap in the newest sequence in an alignment if there is a gap in the existing alignment. It has positive effects, since gaps are more likely to occur in the same column than scattered in different places in each sequence, but it also means that each gap is locked into the alignment forever, and if the original decision to add a gap was incorrect, it can have a negative effect on the alignment.

Feng-Doolittle progressive alignment was the first algorithm to describe the ”once a gap, always a gap” principle. Other progressive alignment programs tend to have the same feature, such as PILEUP. ClustalW doesn’t force gaps to occur in the same positions, but it has biased gap penalties that tend to mean gaps are stuck where they first appeared.

Iterative alignment methods were designed to avoid locking in aspects of early alignments when adding new sequences, so programs like MultAlign, IterAlign, and PRPP don’t have the problem. MSA compares all sequences at once, rather than adding new sequences to an existing alignment. SAGA is a progressive alignment, but it avoids the problem by having a randomization step in which gaps can be moved. New methods like MUSCLE and ProbCons use iterative refinement.

3. Three of ClustalW’s heuristics: Substitution matrix variations: ClustalW uses different weight matrices for different alignments within a multiple sequence alignment. Closely-related sequences are aligned with weight matrices based on similar sequences, and distant sequences use matrices that include more divergent sequences. While the choice of weight matrix is less important for closely-related sequences, the hope is that a good choice of matrix based on similarity will allow ClustalW to align distant sequences in the ”twilight zone” of low identity.

Gap penalties: ClustalW changes its gap penalties in a number of ways. The overall gap penalties are affected by the lengths of the sequences being aligned, their similarity, and the scoring matrix. Also, the gap penalties at a given position are modifed based on the adjacent residue, the presence of a gap in the same place in other sequences (more favorable), the presence of a gap in a nearby position (less favorable),

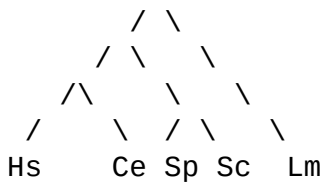
and whether the position is in a stretch of hydrophilic residues. The heuristics help create alignments with gaps concentrated in the same area in different sequences, rather than spread out along the sequences, which makes sense biologically. If there was an insertion or deletion in an ancestral sequence, it is likely to be found in most of the descendants; insertions and deletions should not be viewed as independent events for each sequence being added to the alignment. The gap penalty modifications also attempt to model the likelihood of seeing an insertion or deletion in regions of a protein based on the biology of that region.

Sequence weighting: ClustalW weights the sequences based on a rough guide tree, with the weight as the distance of each sequence from the root. The weights are used as multiplication factors when scoring the alignments. The result is that groups of closely-related sequences are down-weighted so that they don't contribute an unreasonable amount to the alignment; more information is provided by distantly-related sequences.

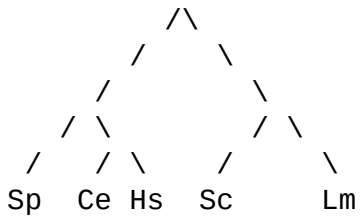
Weights are also used in the opposite manner to add a closely related sequence into an alignment. The most closely related sequences in the existing alignment are given more weight, so that the new sequence will be more likely to use the same gap locations as its relative. If this heuristic were not used, there would be no difference between near and distant relatives, and gap locations would be less accurate.

4.A. ClustalW constructs a guide tree before doing a multiple sequence alignment by doing pairwise alignments between all sequences. Distances from the pairwise alignments are used to build a tree by neighbor-joining. Then the sequences closest together in the tree are aligned, the next-closest sequence is aligned to the existing alignment, and so on. The sequences are weighted in the alignment based on their distance, so that closely-related sequences don't contribute too much to the end alignment. After the alignment, ClustalW can generate a corrected tree which can be used for bootstrapping.

B. Alignment attached. The output tree does not agree with the species tree. The species tree is:



ClustalW gave approximately (not to scale):



So S. cerevisiae is in the wrong place.

5. Pairwise alignments (global, no end gap penalty):

```
DEFI_APIME vs DEFI_AESCY
Score: 53

--G-QVNDSACAANCLSL-GKAGGHCE---KVGCICR
GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCY

DEFI_APIME vs DEFA_ZOPAT
Score: 100

--G-QVNDSACAANCLSLGKAGGHC-EK-VGCICR
IAGTKLNSAACGAHCLALGRRGGYCNSKSV-CVCR

DEFI_APIME vs SAPC_SARPE
Score: 84

GQVNDSACAANCLSLGKAGGHCE-KVG-CICR
G-VQHSACALHCVFRGNRGGYCTGK-GICVCR
```

Master-slave alignment:

```
DEFI_APIME/53-82     --G-QVNDSACAANCLSL-GKAGGHC-E---K-VG-CICR
DEFI_AESCY/1-37      GFGCPLDQMQCHRHCQTITGRSGGYC-SGPLK-LT-CTCY
DEFA_ZOPAT/10-43     IAGTKLNSAACGAHCLAL-GRRGGYCNS---KSV--CVCR
SAPC_SARPE/10-39     --G--VQHSACALHCVFR-GNRGGYC-TG--K--GICVCR
```

Pfam alignment:

```
DEFI_APIME/53-82     G...QVNDSACAANCLSLG.KAGGHCE...KVGCICR
DEFI_AESCY/1-37      GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCY
DEFA_ZOPAT/10-43     IAGTKLNSAACGAHCLALG.RRGGYCNS..KSVCVCR
SAPC_SARPE/10-39     G....VQHSACALHCVFRG.NRGGYCTG..KGICVCR
```

Note that the master-slave alignment inserts more gaps near the end of the multiple sequence alignment. This behavior is a consequence of the method for combining the individual pairwise alignments: a gap in the master sequence in one pairwise alignment is propagated to the other sequences. Other alignment methods, such as Pfam, consider the cost of opening a gap in the other sequences combined, not just in pairwise alignments.

6. A. The entropy score for a column i is $S(m_i) = -\sum_a c_{ia} \log_2 p_{ia}$.

Ignoring any residues aligned with gaps, the possibilities are that all three amino acids in a column are the same; two are the same and the third differs; or all three are different. In the first case, the score is $-3 \log_2 1 = 0$. In the second, the score is $-2 \log_2 \frac{2}{3} - \log_2 \frac{1}{3} = 2.7549$. In the last case, the score is $-3 \log_2 \frac{1}{3} = 4.7549$.

So, the score for the whole alignment is $18 \times 4.7549 + 40 \times 2.7549 = \mathbf{195.8}$.

B. To compute the sum of pairs score for a single column, we need to enumerate all possible pairs among the sequences (there are 3 pairwise combinations for 3 sequences) and add together the three substitution scores from BLOSUM62. We then add the scores for all columns.

Ignoring any residues aligned with gaps, the score for the alignment is **1002**.

7. A. ProbCons uses "consistency" to avoid making some of the errors made in progressive alignments, instead of only repairing them in a later refinement step. To align sequences x and y, ProbCons uses pairwise alignments with intermediate sequences z to determine what residues in x and y should align. A similar concept is used in T_COFFEE.

B. **ProbCons:** Computes a matrix of pairwise alignments showing the probability of each residue in x being aligned with each residue in y, constructs a guide tree based on these probabilities, do a progressive alignment according to the order of the guide tree, and then iteratively refine by partitioning the alignment and re-aligning.
MUSCLE: Computes a distance matrix between all pairs of sequences based on counting k-mer matches, builds a tree from this distance matrix, does a progressive alignment based on the tree, builds a new distance matrix from the pairwise relationships implied by that multiple alignment, builds a new tree, then does another progressive alignment. To refine the progressive alignment, it iteratively splits the tree, computes the two sub-alignments and aligns to each other, then compares the result to the beginning alignment.

```
CLUSTAL W (1.82) multiple sequence alignment


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        MSLPLKT-----IVHLVKPFACTARFSARYPIHVIVVAVLLSAAAYLSVT
Schizosaccharomyces_pombe       MIYKLAARYPIQVIAIVGILVSMAYFSFLEALTQEDFPVLIRALKRFGIL
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    -------------MLSRLFRMHGLFVASHPWEVIVGTVTLTICMMS--M


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        QSYLNEWK----------LDSNQYSTYLSIKPDEL-------FEKCTHYY
Schizosaccharomyces_pombe       DGFPNTRLPNEMILKLSSVQGEDASVWEQIPAAELGGEGFVDFDITQWYY
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    NMFT--------------GNNKICGWN--------------YECPKFEE


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        RSPVSDTWKLLSSKEAADIYTPFHYY-LSTISFQSKDNSTTLPSLDDVIY
Schizosaccharomyces_pombe       PANAKVDVAQLVEPYRNDCIFHDASG-ACHFFFKEVGNWTVSSIALPSNL
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    DVLSSDIIILTITRCIAILYIYFQFQNLRQLGSKYILGIAGLFTIFSSFV


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        SVDHTRYLLSEEPKIPTELVSENGTKWRLRNNSNFILDLHNIYRNMVKQF
Schizosaccharomyces_pombe       ANPPIDYFLDSSSTVIQRILPAIR-----EHGISWSWLLQLIARTWMNTL
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    FSTVVIHFLDKELTGLNEALPFFLLLIDLSRASTLAKFALSSN-SQDEVR


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        SNKTSEFDQFDLFIILAAYLTLFYTLCCLFNDMRKIGSKFWLSFSALSNS
Schizosaccharomyces_pombe       K-IASQASKTELLIVGTAYACMLISIVSLYLKMRRLGSKFWLFFSVLLST
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    ENIARGMAILGPTFTLDALVECLVIGVGTMSGVRQLEIMCCFGCMSVLAN


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        ACALYLSLYTTHSLLKKPASLLSLVIGLPFIVVIIGFKHKVRLAAFSLQK
Schizosaccharomyces_pombe       LFSVQFAMTLVRASGVR-ISLVSLIESLPFLINVVALDKAAELTRQVITR
Caenorhabditis_elegans          --------------------------------------------------
Homo_sapiens                    YFVFMTFFPACVSLVLELSRESREGRPIWQLSHFARVLEEEENKPNPVTQ


Leishmania_major                --------------------------------------------------
Saccharomyces_cerevisiae        FHRISIDKKITVSNIIYEAMFQEGAYLIRDYLFYISSFIGCAIYARHLPG
```

```
Schizosaccharomyces_pombe      CS--VSDSHSPMHEDIAKACRNAAPPILRHFSFG---IVVLAIFSYCNFG
Caenorhabditis_elegans         ------------------------------------------------
Homo_sapiens                   RVKMIMSLGLVLVHAHSRWIADPSPQNSTADTSKVSLGLDENVSKRIEPS

Leishmania_major               ------------------------------------------------
Saccharomyces_cerevisiae       LVNFCILSTFMLVFDLLLSATFYSAILSMKLEINIIHRSTVIRQTLEEDG
Schizosaccharomyces_pombe      IKQFFLFAAVM-IYDLLLLFSFFVAILTLKLEMRRYNAKDDVRKVLIEEG
Caenorhabditis_elegans         ---------------------------------MVADKRKLLDFLQSCD
Homo_sapiens                   VSLWQFYLSKMISMDIEQVITLSLALLLAVKYIFFEQTETESTLSLKNPI

Leishmania_major               ------------------------------------------------
Saccharomyces_cerevisiae       VVPTTADIIYKDETASEPHFLRS-NVAIILGKASVIGLLLLINLYVFTDK
Schizosaccharomyces_pombe      LSESTARHVADGNDSSATTSAGSRYFKVRYGTKIILFIFIAFNLFELCSI
Caenorhabditis_elegans         ISDEASRKIEN--------------------------FICENYDQKEK
Homo_sapiens                   TSPVVTQKKVPDN-----------------------------CCRREPMLV

Leishmania_major               ------------------------------------------------
Saccharomyces_cerevisiae       LNATILNTVYFDSTIYSLP--NFINYKDIGNLSN----------------
Schizosaccharomyces_pombe      PFKHYAATSAAAARLIPLVRSQYPDFKSQRLLDDGVFDDVLSAISSMSNI
Caenorhabditis_elegans         PKRKPLFSVGEDD-------------------------------------
Homo_sapiens                   RNNQKCDSVEEETG------------------------------------

Leishmania_major               ------------------------------------------------
Saccharomyces_cerevisiae       -QVIISVLPKQYYTPLKKYHQIEDSVLLIIDSVSNAIRDQFISKLLFFAF
Schizosaccharomyces_pombe      ESPSVRLLPAVFYGAELSSTSFLSTIHSFINNWSHYISASFLSKWIVCAL
Caenorhabditis_elegans         ----------------DSEEYVSPIKETCETGTQCKR------------
Homo_sapiens                   ----------------INRERKVEVIKPLVAETDTPNRATFVVGNSSLLD

Leishmania_major               ----------------------------------MRRSLLLACSAAK--
Saccharomyces_cerevisiae       AVSISINVYLLNAAKIHTGYMNFQPQSNKIDDLVVQQKSATIEFSETRSM
Schizosaccharomyces_pombe      SLSIAVNVFLLNAARLNS--IKEEPEKKVVEKVVEVVKYIPSSNSSSIDD
Caenorhabditis_elegans         ----------------------DVEYPEIESGNRSLDEISREWKECK--
Homo_sapiens                   TSSVLVTQ--------------EPEIELPREPRPNEECLQILGNAEKG-
                                                                              .

Leishmania_major               ------------------------------------------------GE
Saccharomyces_cerevisiae       PASSGLETPVTAKDIIISEEIQNNECVYALSSQDEPIRPLSNLVELMEKE
Schizosaccharomyces_pombe      IQKDE------------------------IAQESVVRSLEECITLYNNG
Caenorhabditis_elegans         ------------------------------------------------
Homo_sapiens                   ------------------------------------------------
```

```
Leishmania_major                SWASMSDTEIMKQVENKKIAFHGLEQALAPDYDRAIAIRREIVKKKICPS
Saccharomyces_cerevisiae        QLKNMNNTEVSNLVVNGKLPLYSLEKKLE-DTTRAVLVRRKALS-TLAES
Schizosaccharomyces_pombe       QISTLNDEEVVQLTLAKKIPLYALERVLK-DVTRAVVIRRTVVSRSSRTK
Caenorhabditis_elegans          ---DVTPGEAVRLLRRGQAKSRELESRFP--AEQAIPIRRTFINKKFEN-
Homo_sapiens                    -AKFLSDAEIIQLVNAKHIPAYKLETLME-THERGVSIRRQLLSKKLSEP
                                .   :.  *  .    :    **  : .   :.: :**  :. .

Leishmania_major                PAATHPLERVPYKNYDWSSVVGQSCENILGYVPVPVGLAGPLLLDGK-EV
Saccharomyces_cerevisiae        PILVS--EKLPFRNYDYDRVFGACCENVIGYMPIPVGVIGPLIIDGT-SY
Schizosaccharomyces_pombe       TLESS---NCPVYHYDYSRVLNACCENVIGYMPLPLGVAGPLIIDGK-PF
Caenorhabditis_elegans          ---------LPYLGYDYTLATECCCENVIGYTPVPVGVAGPLTLNGTSEI
Homo_sapiens                    SSLQY----LPYRDYNYSLVMGACCENVIGYMPIPVGVAGPLCLDEK-EF
                                .    . .  *   *::  .   .***::** *:*:*: *** :: .

Leishmania_major                ALPMATTEGALVASAHRGARAINLSGGCRTAVLKEGMTRAPVVEVNSFD-
Saccharomyces_cerevisiae        HIPMATTEGCLVASAMRGCKAINAGGGATTVLTKDGMTRGPVVRFPTLIR
Schizosaccharomyces_pombe       YIPMATTEGALVASTMRGCKAINAGGGAVTVLTRDQMSRGPCVAFPDLTR
Caenorhabditis_elegans          YVPMATTEGALIASTNRGMNVIRAAGGVETSIFNSGMTRAPVVKFPTARD
Homo_sapiens                    QVPMATTEGCLVASTNRGCRAIGLGGGASSRVLADGMTRGPVVRLPRACD
                                 :*******.*:**: ** ..*  .**  : :  . *:*.* * .

Leishmania_major                -EAITVIKFCEERFDVLREAFESTTRFGKLLSIKCAMAGRQVHLRFSAFT
Saccharomyces_cerevisiae        SGACKIWLDSEEGQNSIKKAFNSTSRFARLQHIQTCLAGDLLFMRFRTTT
Schizosaccharomyces_pombe       AGRAKIWLDSPEGQEVMKKAFNSTSRFARLQHIKTALAGTRLFIRFCTST
Caenorhabditis_elegans          AVSMKRWLEHPENQDRARQEFQSCSRFAKLKSIDITIDGNLAYLRFDAHT
Homo_sapiens                    SAEVKAWLETSEGFAVIKEAFDSTSRFARLQKLHTSIAGRNLYIRFQSRS
                                    .      *     :: *:* :**.:*  :.  : *   .:** : :

Leishmania_major                GDAMGMNMITKGCDKALQVLQQHI--PSVRVLTLSGNFCTDKKPSALNWV
Saccharomyces_cerevisiae        GDAMGMNMISKGVEYSLKQMVEEYGWEDMEVVSVSGNYCTDKKPAAINWI
Schizosaccharomyces_pombe       GDAMGMNMISKGVEHALVVMSNDAGFDDMQVISVSGNYCTDKKPAAINWI
Caenorhabditis_elegans          GDAMGMNMISKSCDSTMRFLMENF--PEMTVLALSGNLCVDKKAAAKNWT
Homo_sapiens                    GDAMGMNMISKGTEKALSKLHEYF--PEMQILAVSGNYCTDKKPAAINWI
                                *********:*. : ::  : :     .: :::*** *.***.:* **

Leishmania_major                EGRGKSVVAEAVIKRDVVESVLKCTVDSVVSLNVTKNLRGSALAGSIGGF
Saccharomyces_cerevisiae        EGRGKSVVAEATIPGDVVKSVLKSDVSALVELNISKNLVGSAMAGSVGGF
Schizosaccharomyces_pombe       DGRGKSVIAEAIIPGDAVKSVLKTTVEDLVKLNVDKNLIGSAMAGSVGGF
Caenorhabditis_elegans          EGRGRSVVAECLIPREVVTKTLRTTPEQLAYLTTTKLHIGSSRAGAVGGS
Homo_sapiens                    EGRGKSVVCEAVIPAKVVREVLKTTTEAMIEVNINKNLVGSAMAGSIGGY
                                :***:**:.*. *  ..* ..*:   . :  :.  *   **: **::**

Leishmania_major                NAHAANIVAALYIATGQDPAQVVESATCMTTVDK---AGEDLVISLMMPS
Saccharomyces_cerevisiae        NAHAANLVTALFLALGQDPAQNVESSNCITLMKE---VDGDLRISVSMPS
Schizosaccharomyces_pombe       NAHAANIVTAVYLATGQDPAQNVESSNCITLMDN---VDGNLQLSVSMPS
Caenorhabditis_elegans          NAHAANIVAAIFIATGQDAAQVVSSSMCSTRMEVT--ADKNLYVSCTLPC
Homo_sapiens                    NAHAANIVTAIYIACGQDAAQNVGSSNCITLMEASGPTNEDLYISCTMPS
```

```
                                ******:*:*:::* ***.** * *: * * :.    .. :* :*  :*.

Leishmania_major                IEVGAVGGGTGLSSQRAMLELMGCAGSNKEDPGAHSRQIARVVAGAVICG
Saccharomyces_cerevisiae        IEVGTIGGGTVLEPQGAMLDLLGVRGPHPTEPGANARQLARIIACAVLAG
Schizosaccharomyces_pombe       IEVGTIGGGTVLEPQGAMLDLLGVRGAHMTSPGDNSRQLARVVAAAVMAG
Caenorhabditis_elegans          VEVGTVGGGTILAPQRACLESLGCAGPNKEQPGQNAERLAEVIAATVLAG
Homo_sapiens                    IEIGTVGGGTNLLPQQACLQMLGVQGACKDNPGENARQLARIVCGTVMAG
                                :*:*::**** * .* * *: :*  *.   .** ::.::*.::. :*:.*

Leishmania_major                ELSLLSGLAAGHLLSAHMKLNRKP------PTP-----------------
Saccharomyces_cerevisiae        ELSLCSALAAGHLVQSHMTHNRKTNKANELPQPSNKGPPCKTSALL----
Schizosaccharomyces_pombe       ELSLCSALASGHLVKSHIGLNRSALNTPAMDSSAKKPATDALKSVNSRVP
Caenorhabditis_elegans          ELSLMAALTTNELVSSHMKLNRSKQQLYADDSGKATHFEKEVEKAGSLLS
Homo_sapiens                    ELSLMAALAAGHLVKSHMIHNRSKINLQDLQGACTKKTA-----------
                                **** :.*::..*:.:*:  **.

Leishmania_major                ----------------------
Saccharomyces_cerevisiae        ----------------------
Schizosaccharomyces_pombe       GR--------------------
Caenorhabditis_elegans          GKSGNIKLKRLPQDVVQCSNIL
Homo_sapiens                    ----------------------
```