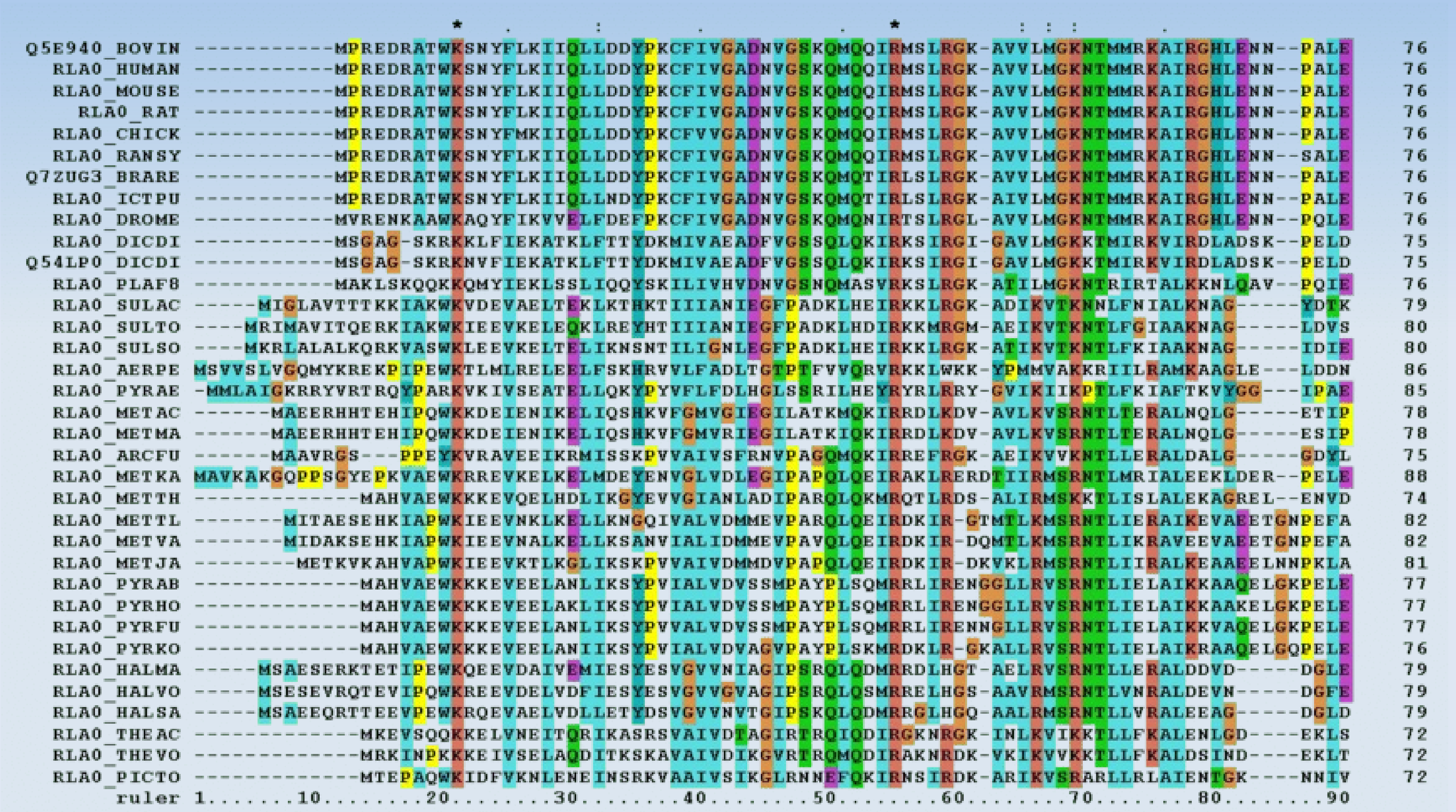
Topic 8: Multiple Sequence Alignment



```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE 76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE 76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE 76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD 75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD 75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE 76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK 79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS 80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE 80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN 86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE 85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP 78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP 78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL 75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE 88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD 74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA 82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA 82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA 81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE 77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE 77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE 77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE 76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE 79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE 79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD 79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS 72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT 72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV 72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

First 90 aa of a protein multiple sequence alignment of the acidic ribosomal protein P0 (L10E) from several organisms – ClustalX. (*Wikipedia*)

Multiple Sequence Alignment

The pair-wise sequence alignment problem:

- we have 2 strings (sequences)
- find the ”best match” between the two
- Usually local alignment but can do global

Multiple (3+) Sequence Alignment:

→ usually related by evolution

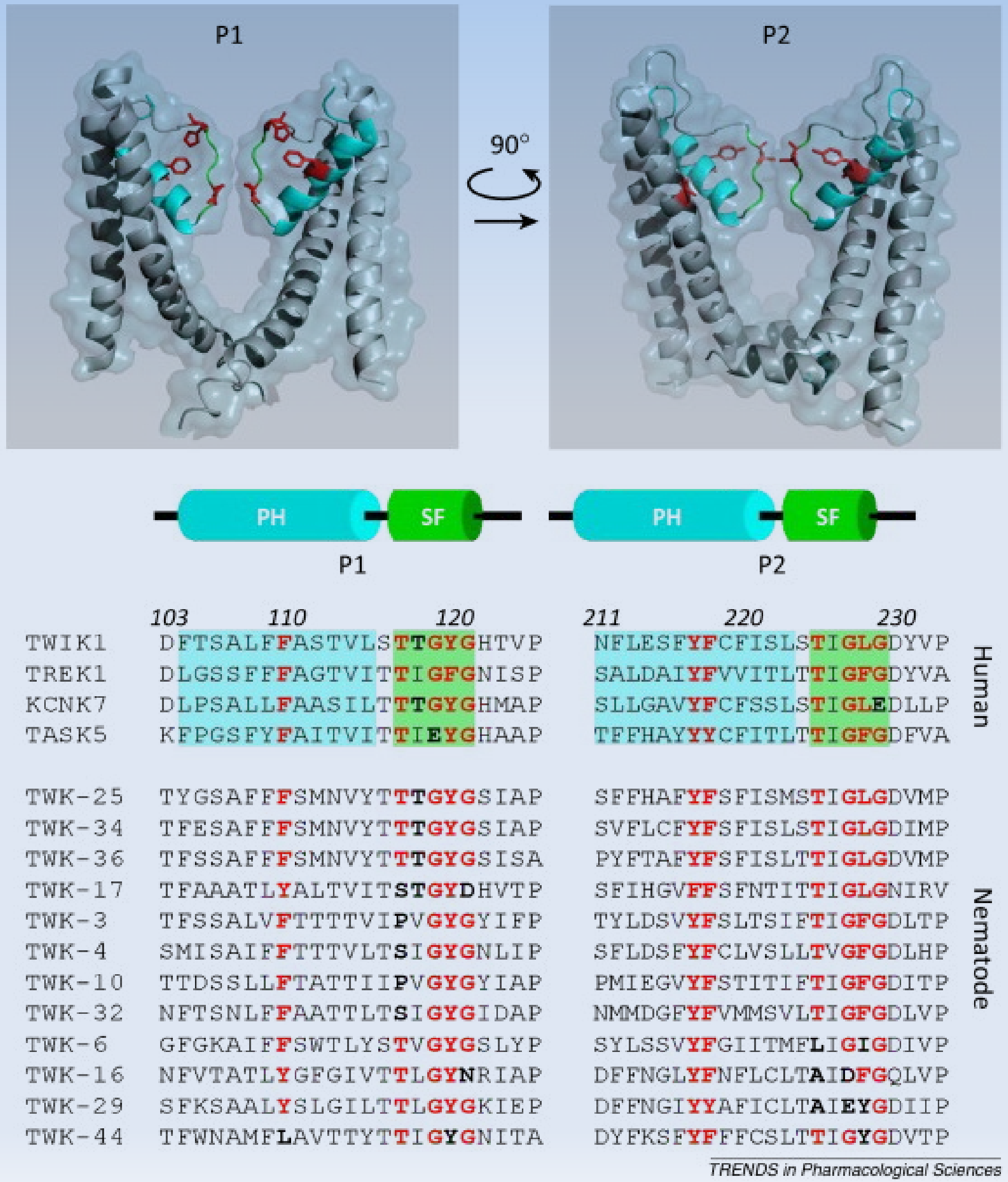
→ what is common between the sequences is usually conserved protein function

almost always global alignment – whole genes

Multiple Sequence Alignment

Example:

Selectivity for ions is a critical function preserved over a long evolutionary period



```
            103       110         120       211      220         230
TWIK1       DFTSALFFASTVLSTTGYGHTVP       NFLESFYFCFISLSTIGLGDYVP
TREK1       DLGSSFFFAGTVITTIGFGNISP       SALDAIYFVVITLTTIGFGDYVA
KCNK7       DLPSALLFAASILTTTGYGHMAP       SLLGAVYFCFSSLSTIGLEDLLP
TASK5       KFPGSFYFAITVITTIEYGHAAP       TFFHAYYYCFITLTTIGFGDFVA

TWK-25      TYGSAFFFSMNVYTTTGYGSIAP       SFFHAFYFSFISMSTIGLGDVMP
TWK-34      TFESAFFFSMNVYTTTGYGSIAP       SVFLCFYFSFISLSTIGLGDIMP
TWK-36      TFSSAFFFSMNVYTTTGYGSISA       PYFTAFYFSFISLTTIGLGDVMP
TWK-17      TFAAATLYALTVITSTGYDHVTP       SFIHGVFFSFNTITTIGLGNIRV
TWK-3       TFSSALVFTTTTVIPVGYGYIFP       TYLDSVYFSLTSIFTIGFGDLTP
TWK-4       SMISAIFFTTTVLTSIGYGNLIP       SFLDSFYFCLVSLLTVGFGDLHP
TWK-10      TTDSSLLFTATTIIPVGYGYIAP       PMIEGVYFSTITIFTIGFGDITP
TWK-32      NFTSNLFFAATTLTSIGYGIDAP       NMMDGFYFVMMSVLTIGFGDLVP
TWK-6       GFGKAIFFSWTLYSTVGYGSLYP       SYLSSVYFGIITMFLIGIGDIVP
TWK-16      NFVTATLYGFGIVTTLGYNRIAP       DFFNGLYFNFLCLTAIDFGQLVP
TWK-29      SFKSAALYSLGILTTLGYGKIEP       DFFNGIYYAFICLTAIEYGDIIP
TWK-44      TFWNAMFLAVTTYTTIGYGNITA       DYFKSFYFFFCSLTTIGYGDVTP
```

TRENDS in Pharmacological Sciences

Multiple Sequence Alignment

Nucleotide Example:

```
TACGG_G
TAC_GTG
AA_GGTG
AACAG_A
```

Protein MSA vs. Nucleotide MSA

If you want to compare a number of genes, concentrate on the protein sequences rather than nucleotide sequences

Protein MSAs are more informative

More likely to be accurate (20 aa vs. 4 nucs)

Can translate back to multiple nucleotide sequences after doing protein MSA.

Multiple Sequence Alignment

The ”best” pair-wise sequence alignment is easy: with a scoring matrix and gap penalties → find the **highest scoring** alignment.

Multiple Sequence Alignment: no simple criteria.

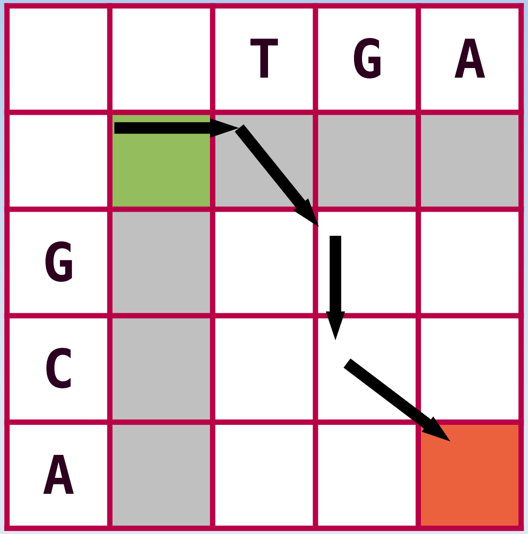
How do we score an alignment of N sequences?

- Sum of pairwise alignments? Depends on N
- Average?

Both are used

Multiple Sequence Alignment – How?

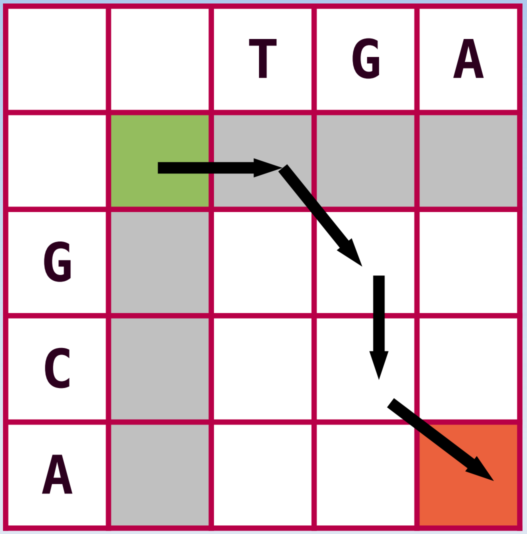
We looked at Dynamic Programming for Pairwise Sequence Alignment



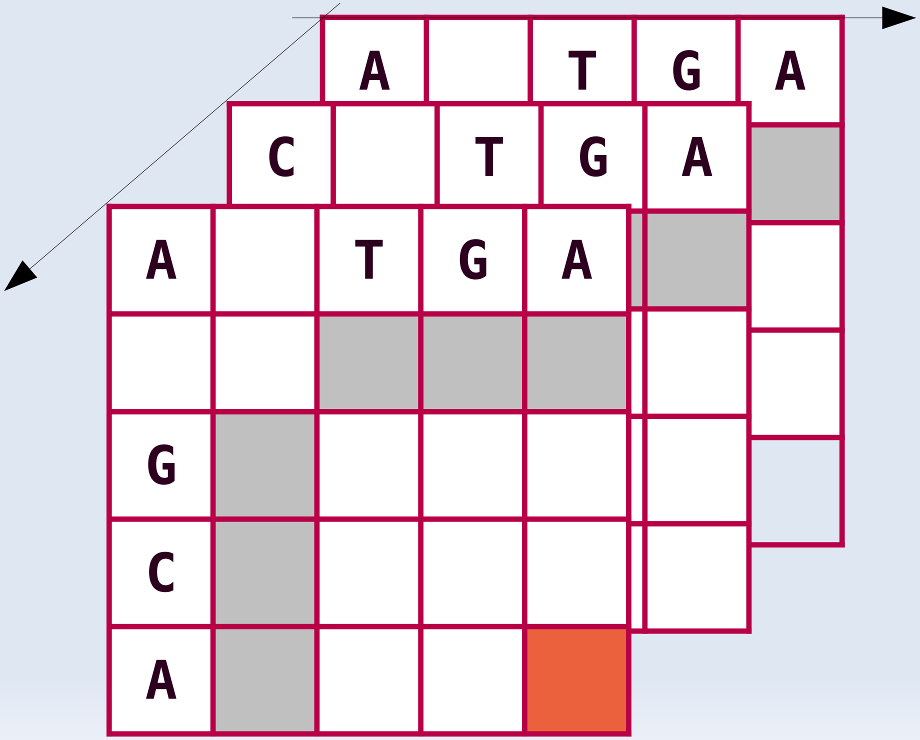
Can we do DP for 3 sequences?

Multiple Sequence Alignment – How?

DP Pairwise Sequence Alignment



What about DP for 3 sequences?



Multiple Sequence Alignment – How?

A **Fully** Dynamic Programming approach to multiple sequence alignment will work.

But: very expensive to compute

For M sequences of length N, time complexity is N^M

→ for protein sequences of length 500, programs that use a fully DP approach are limited to ~10 sequences on a fast computer

We may want to do MANY sequences, each 1000s of nucleotides or amino acids long!

Progressive Multiple Sequence Alignment

Most MSA algorithms use the **Progressive** Alignment approach

First, do M(M-1)/2 pair-wise alignments (using DP) → scored using protein scoring matrices, gaps

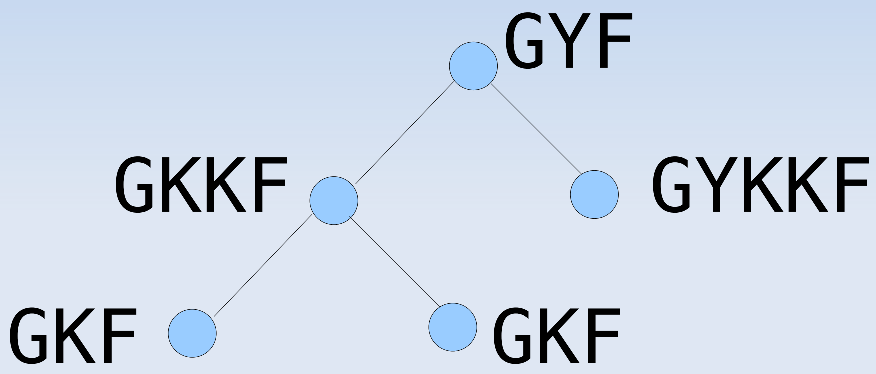
Get the two most related sequences - highest scoring pair

Then progressively add next highest pair, etc. to build up the MSA

→ MSA depends on the best pairwise alignments

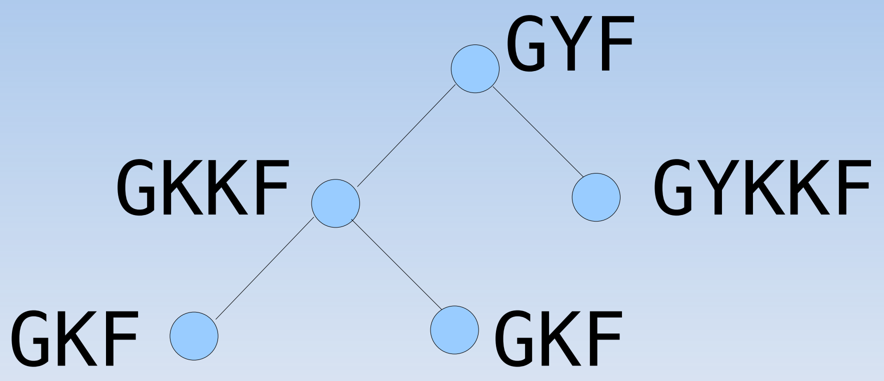
Progressive MSA

To build up the MSA progressively, we use a Tree:



Progressive MSA

An MSA gives us M(M-1)/2 induced pairwise alignments



GYF

GKKF

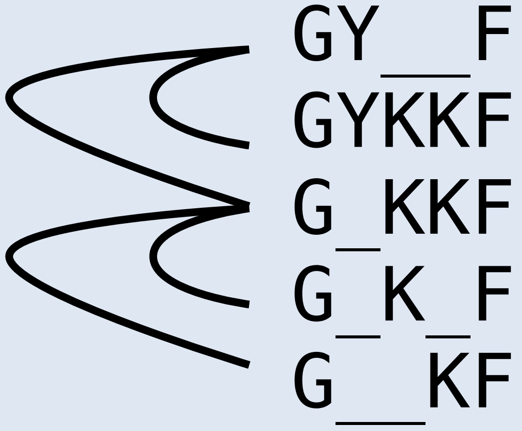
GYKKF

GKF

GKF

Neighbors in the tree (e.g. GY--F and GYKKF) should have optimal ”induced” pairwise alignments.

Non-neighbors (e.g. GY--F and G-K-F) can have less than optimal alignments.



```
GY__F
GYKKF
G_KKF
G_K_F
G__KF
```

Advantages, Disadvantages

Advantages:

Progressive MSAs are usually fast (there is a range)

Alignments are generally of high quality

Disadvantages:

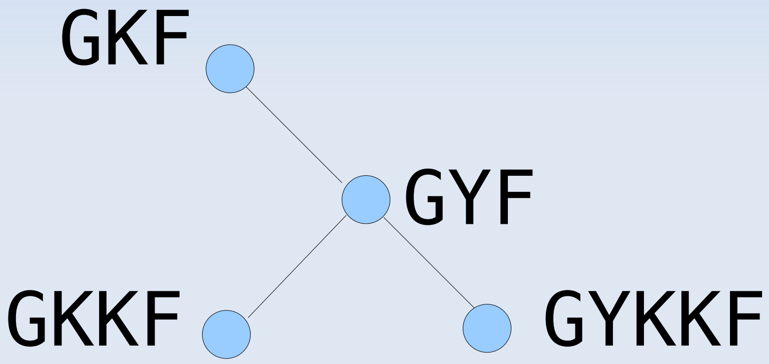
Most progressive MSA methods ”fix” early alignments and do not ”reconsider” later

If initial alignments of MSA are made on distantly related sequences, there may be errors

Compare: Clustal Omega allows guide tree iterations.

Center Tree Progressive MSA

The Center Tree approach starts with the sequence that has the minimum total ”distance” from all other sequences – this goes in the center of a tree.

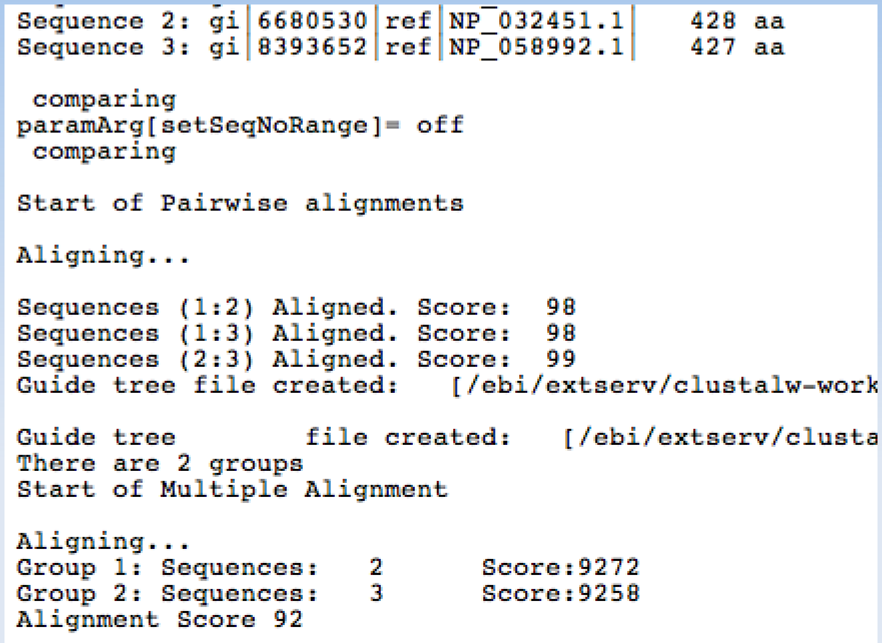


Create a pair-wise alignment of closest sequences

Add sequences to alignment in order of distance from center.

Clustal MSA

1) Compute all pairwise alignments



```
Sequence 2: gi|6680530|ref|NP_032451.1|     428 aa
Sequence 3: gi|8393652|ref|NP_058992.1|     427 aa

 comparing
paramArg[setSeqNoRange]= off
 comparing

Start of Pairwise alignments

Aligning...

Sequences (1:2) Aligned. Score:  98
Sequences (1:3) Aligned. Score:  98
Sequences (2:3) Aligned. Score:  99
Guide tree file created:   [/ebi/extserv/clustalw-work

Guide tree        file created:   [/ebi/extserv/clusta
There are 2 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:9272
Group 2: Sequences:   3      Score:9258
Alignment Score 92
```

2) Sort the alignments in order of scores

3) Create a ”guide tree” from sorted pairwise alignments: add sequences as long as no cycles → minimum spanning tree (also used in computer network routing, statistical clustering)

4) Brings in sequences in order of scores.

Clustal Omega output

1) Pairwise alignment scores

2) The MSA

Trees that estimate phylogeny:

3) Cladogram – branches of equal length (no information on evolutionary time)

4) Phylogram – branches of unequal length (length proportional to evolutionary change)

MSAs based on k-mers

An approach similar to the way Blast works

1) List all k-letter words in each sequence
2) Find best matches
3) Extend matches

How good is an MSA?

Not easy to tell – ultimately, we have to look at biological implications of an MSA

One way to check MSA algorithms is to use a ”benchmark” of accepted MSAs:

BaliBASE - this is BaliBASE 4

BaliBASE 2

OXBench

These use knowledge of protein structure and other information.

Purpose of MSAs

1. Identify conserved regions of proteins, find patterns and protein domains

2. MSAs help with predicting protein secondary structure and performing phylogenetic analysis (Lecture 9) → evolutionary relationship.

3. MSAs can be used to generate Position-Specific Scoring Matrix for sequence search (e.g. PSI-Blast)

Uses of MSAs – motifs, profiles

1. Sequence similarity usually implies a similarity in biological function
2. Similar biological function is less likely to imply sequence similarity

One use of MSAs is to find protein families or motifs – see Prosite, Pfam, PSI-Blast

The idea is to find patterns in the sequences of proteins with similar function. #2 makes this hard.

Pairwise vs. Multiple Sequences

Pairwise	Multiple
Compares two sequences	Compares three+ sequences
DNA, RNA, or Protein	Protein usually but DNA and RNA possible too
Can use local or global alignment	Usually uses many global pairwise alignments
Goals: find similar subsequences	Goals: find similar protein structure, phylogenetic or evolutionary relationship