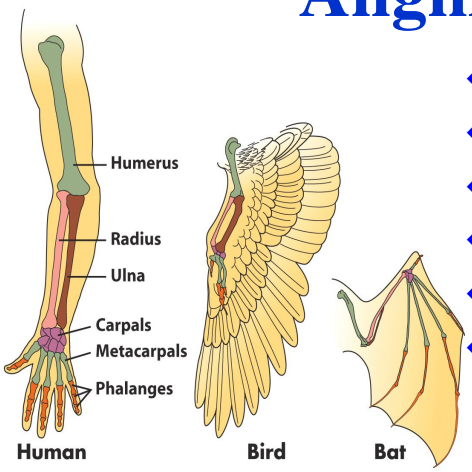
Introduction to Bioinformatics

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

Multiple Sequence Alignment



- Progressive Alignment
- Iterative Pairwise
- Guide Tree
- ClustalW
- Co-linearity
- Multiple Sequence Alignment Editors

```
                    *       20        *       40        *       60        *       80
Wombat     : AAAGTTAATGAGTGGTTATCCAGAAGTAGTGACATTTTAGCCTCTGATAACTCCAACGGTAGGAGCCATGAGCAGAGCGCAGA :  83
Opossum    : AAAGTTAATGAGTGGTTATTCAGAAGTAATGACGTTTTAGCCCCAGATTACTCAAGTGTTAGGAGCCATGAACAGAATGCAGA :  83
Armadillo  : AAAGTTAACGAGTGGTTTTCCAGAGGTGATGACATATTAACTTCTGATGACTCACACGATAGGGGGTCTGAATTAAATGCAGA :  83
Sloth      : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGACATACTAACTTCTGATGACTCACACAATGGGGGGTCTGAATCAAATGCAGA :  83
Dugong     : AAAGTTAATGAGTGGTTTTTCAGAAGTGATGGCCTG---------GATGACTTGCATGATAAGGGGTCTGAGTCAAATGCAGA :  74
Hyrax      : AAAGTTAATGAGTGGTTTTCCAGAAGTGACAACCTA---------AGTGATTCACCTAGTGAGGGGTCTGAATTAAATGGAAA :  74
Aardvark   : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGGCCTG---------GATGGCTCACATGATGAAGGGTCTGAATCAAATGCAGA :  74
Tenrec     : AAGGTTAACGAGTGGTTTTCCAAAAGCCACGGCCTG---------GGTGACTCTCGCGATGGGCGGCCTGAGTCAGGCGCAGA :  74
Rhinoceros : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGAAATATTAACTTCTGATGACTCACATGATGGGGGGCCTGAATCAAATACTGA :  83
Pig        : AAAGTTAATGAGTGGTTTTCTAGAAGCGATGAAATGTTAACTTCTGACGACTCACAGGACAGGAGGTCTGAATCAAATACTGG :  83
Hedgehog   : AAAGTGAATGAATGGCTTTCCAGAAGTGATGAACTGTTAACTTCTGATGACTCATATGATAAGGGATCTAAATCAAAAACTGA :  83
Human      : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGAACTGTTAGGTTCTGATGACTCACATGATGGGGAGTCTGAATCAAATGCCAA :  83
Rat        : AAAGTGAATGAGTGGTTTTCCAGAACTGGTGAAATGTTAACTTCTGACAATGCATCTGACAGGAGGCCTGCGTCAAATGCAGA :  83
Hare       : AAAGTTAACGAGTGGTCTCCAGAAGTAATGAAATGTTAACTCCTGATGACTCACTTGACCGGCGGTCTGAATCAAATGCCAA :  83
               AAaGTtAAtGAgTGGtTtTccAgAagt atga  T          gatgactca   gat  g gg cTga t aaatgc ga

                      *      100        *      120        *      140        *
Wombat     : GGTGCCTAGTGCCTTAGAAGATGGGCATCCAGATACCGCAGAGGGAAATTCTAGCGTTTCTGAGAAGACTGAC : 156
Opossum    : GGCAACCAATGCTTTAGAATATGGGCATGTAGAGACA---GATGGAAATTCTAGCATTTCTGAAAAGACTGAT : 153
Armadillo  : AGTAGCTGGTGCATTGAAAGTT------TCAAAAGAAGTAGATGAATATTCTAGTTTTTCAGAGAAGATAGAC : 150
Sloth      : AGTAGTTGGTGCATTGAAAGTT------CCAAATGAAGTAGATGGATATTCTGGTTCTTCAGAGAAGATAGAC : 150
Dugong     : AGTAGCTGGTGCTTTAGAAGTT------CCAGAAGAAGTACATGGATATTCTAGTTCTTCAGAGAAAATAGAC : 141
Hyrax      : AGTGGCTGGTCCAGTAAAACTT------CCAGGTGAAGTACATAGATATTCTAGTTTTCCAGAGAACATAGAT : 141
Aardvark   : AATAGGTGGTGCATTAGAAGTT------TCAAATGAAGTACATAGTTACTCTGGTTCTTCAGAGAAAATAGAC : 141
Tenrec     : CGTAGCTGTAGCCTTCGAAGTT------CCAGACGAAGCATGTGAATCTTATAGTTCTCCAGAGAAAACAGAC : 141
Rhinoceros : AGTAGCTGGTGCAGTAGAAGTT------CAAAATGAAGTAGATGGATATTCTGGTTCTTCAGAGAAAATAGGC : 150
Pig        : GGTAGCTGGTGCAGCAGAGGTT------CCAAATGAAGCAGATGGACATTTGGGTTCTTCAGAGAAAATAGAC : 150
Hedgehog   : AGTAACTGTAACAACAGAAGTT------CCAAATGCAATAGATAGRTTTTTTGGTTCTTCAGAGAAAATAAAC : 150
Human      : AGTAGCTGATGTATTGGACGTT------CTAAATGAGGTAGATGAATATTCTGGTTCTTCAGAGAAAATAGAC : 150
Rat        : AGCTGCTGTTGTGTTAGAAGTT------TCAAATGAAGTGGATGGATGTTTCAGTTCTTCAAAGAAAATAGAC : 150
Hare       : AGTGGCTGGTGCATTAGAAGTC------CCAAAGGAGGTAGATGGATATTCTGGTTCTACAGAGAAAATAGAC : 150
               gt gctg tgc  t gAagtt       cA a gaag a atggatatT t Gtt TtCagAgAA Atagac
```

Part of the alignment of the DNA sequences of the BRCA1 gene

From "Bioinformatics and Molecular Evolution" by Paul Higgs and Teresa Attwood

Aligning BRCA1 Sequences

```
                          *           *           *           *           *
Wombat      : KVNEWLSRSSDILASDNSNGRSHEQSAEVPSALEDGHPDTAEGNSSVSEKTD : 52
Opossum     : KVNEWLFRSNDVLAPDYSSVRSHEQNAEATNALEYGHVET-DGNSSISEKTD : 51
Armadillo   : KVNEWFSRGDDILTSDDSHDRGSELNAEVAGALKV--SKEVDEYSSFSEKID : 50
Sloth       : KVNEWFSRSDDILTSDDSHNGGSESNAEVVGALKV--PNEVDGYSGSSEKID : 50
Dugong      : KVNEWFFRSDGL---DDLHDKGSESNAEVAGALEV--PEEVHGYSSSSEKID : 47
Hyrax       : KVNEWFSRSDNL---SDSPSEGSELNGKVAGPVKL--PGEVHRYSSFPENID : 47
Aardvark    : KVNEWFSRSDGL---DGSHDEGSESNAEIGGALEV--SNEVHSYSGSSEKID : 47
Tenrec      : KVNEWFSKSHGL---GDSRDGRPESGADVAVAFEV--PDEACESYSSPEKTD : 47
Rhinoceros  : KVNEWFSRSDEILTSDDSHDGGPESNTEVAGAVEV--QNEVDGYSGSSEKIG : 50
Pig         : KVNEWFSRSDEMLTSDDSQDRRSESNTGVAGAAEV--PNEADGHLGSSEKID : 50
Hedgehog    : KVNEWLSRSDELLTSDDSYDKGSKSKTEVTVTTEV--PNAIDXFFGSSEKIN : 50
Human       : KVNEWFSRSDELLGSDDSHDGESESNAKVADVLDV--LNEVDEYSGSSEKID : 50
Rat         : KVNEWFSRTGEMLTSDNASDRRPASNAEAAVVLEV--SNEVDGCFSSSKKID : 50
Hare        : KVNEWFSRSNEMLTPDDSLDRRSESNAKVAGALEV--PKEVDGYSGSTEKID : 50
                KVNEWfs4    6    d s      e n              e         eki
```

Alignment of BRCA1 protein sequences for the same region on the gene

From "Bioinformatics and Molecular Evolution" by Paul Higgs and Teresa Attwood

What is Multiple Alignment

Most simple extension of pairwise alignment

Given:

- Set of sequences
- Match matrix
- Gap penalties

Find:

Alignment of sequences such that an optimal score is achieved.

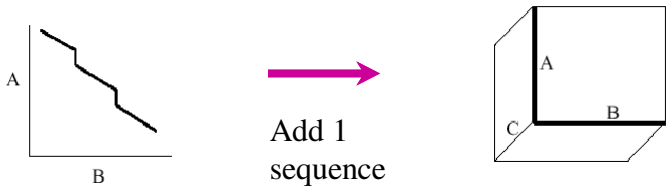
Uses of Multiple Alignment

A good **alignment** is critical for further analysis

- Determine the **relationships** between a group of sequences
- Determine the **conserved** regions
- **Evolutionary Analysis**
  - Determine the phylogenetic relationships and evolution
- **Structural Analysis**
  - Determine the overall structure of the proteins

Alignment Difficulties

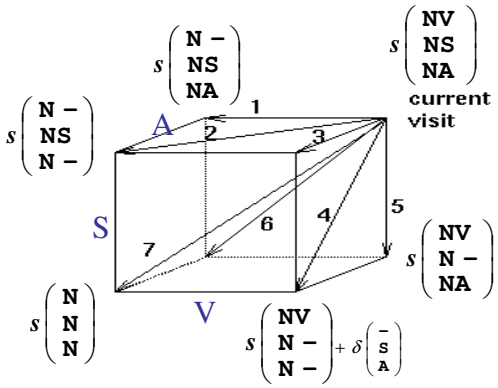
We cannot use sequence comparison algorithms (dynamic programming) and just add more sequences.



With each sequence, another dimension is added that we need to search for the optimal alignment.

Hyperlattice Computation

$k=3 \quad 2^k - 1 = 7$



$$s\begin{pmatrix}NV\\NS\\NA\end{pmatrix} = \max \left\{ \begin{array}{l} s\begin{pmatrix}N\\N\\N\end{pmatrix} + \delta\begin{pmatrix}V\\S\\A\end{pmatrix} \\ s\begin{pmatrix}NV\\N-\\N-\end{pmatrix} + \delta\begin{pmatrix}-\\S\\A\end{pmatrix} \\ s\begin{pmatrix}N-\\NS\\N-\end{pmatrix} + \delta\begin{pmatrix}V\\-\\A\end{pmatrix} \\ s\begin{pmatrix}N-\\N-\\NA\end{pmatrix} + \delta\begin{pmatrix}V\\S\\-\end{pmatrix} \\ s\begin{pmatrix}N-\\NS\\NA\end{pmatrix} + \delta\begin{pmatrix}V\\-\\-\end{pmatrix} \\ s\begin{pmatrix}NV\\N-\\NA\end{pmatrix} + \delta\begin{pmatrix}-\\S\\-\end{pmatrix} \\ s\begin{pmatrix}NV\\NS\\N-\end{pmatrix} + \delta\begin{pmatrix}-\\-\\A\end{pmatrix} \end{array} \right.$$

MSA: Exact vs. Heuristic

- The **exact algorithm**
  - traverses the entire search space
  - finds overall measure of alignment quality and tries to maximize this quality.
- The operation is computationally intensive.
- The largest computers can only optimally align a few sequences (7-8).
- Therefore, we have to use **heuristics**; i.e., faster algorithms, if we want to align many sequences.

Heuristic Algorithms

- Based on a **progressive pairwise** alignment approach
  - ClustalW (**Clust**er **Al**ignment)
  - PileUp (GCG)
  - MACAW
- Builds a global alignment based on **local alignments**
- Builds local multiple alignments
- Based on **Hidden Markov Models**
- Based on **Genetic algorithms**.

Progressive Strategies for MSA

- A common strategy to the MSA problem is to **progressively align** pairs of sequences.
  - A starting pair of sequences is selected and aligned
  - Each subsequent sequence is aligned to the previous alignment.
- **Progressive alignment** is a greedy algorithm.

Iterative Pairwise Alignment

- The greedy algorithm:

  align some pair

  while not done

  pick an unaligned string "near" some aligned one(s)

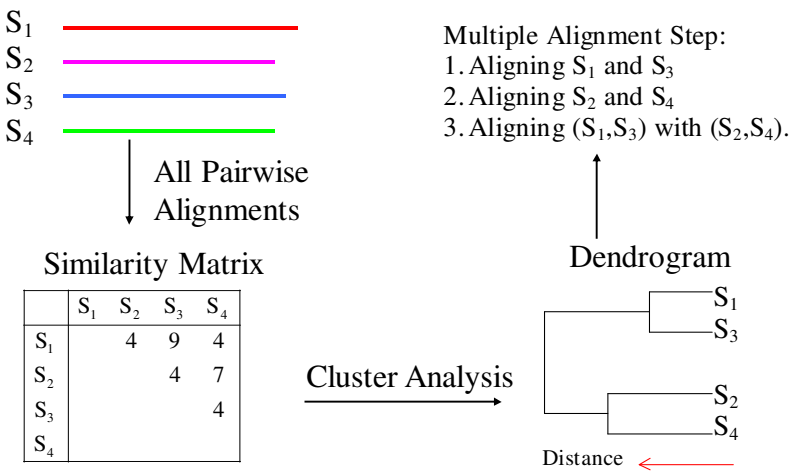
  align with the previously aligned group

- There are many variants to the algorithm.

ClustalW: Package for MSA

- **ClustalW** [the **W** is from **W**eighted] is a software package for the MSA problem.
- Different weights are given to sequences and parameters in different parts of the alignment to and create an alignment that makes sense biologically.
- **Scalable Gap Penalties** for protein profile alignments
  - A gap opening next to a conserved hydrophobic residue can be penalized more heavily than a gap opening next to a hydrophilic residue.
  - A gap opening very close to another gap can be penalized more heavily than an isolated gap.

Steps of ClustalW



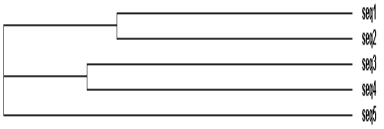
	S_1	S_2	S_3	S_4
S_1		4	9	4
S_2			4	7
S_3				4
S_4				

ClustalW: An Example

```
CLUSTAL W (1.82) multiple sequence alignment


seq3            FEGGILVEAL 10
seq4            FDG-ILVQAV 9
seq5            YEGGAVVQAL 10
seq1            YDG-GAVEAL 9
seq2            YDG-G--EAL 7
                ::*    :*:
```

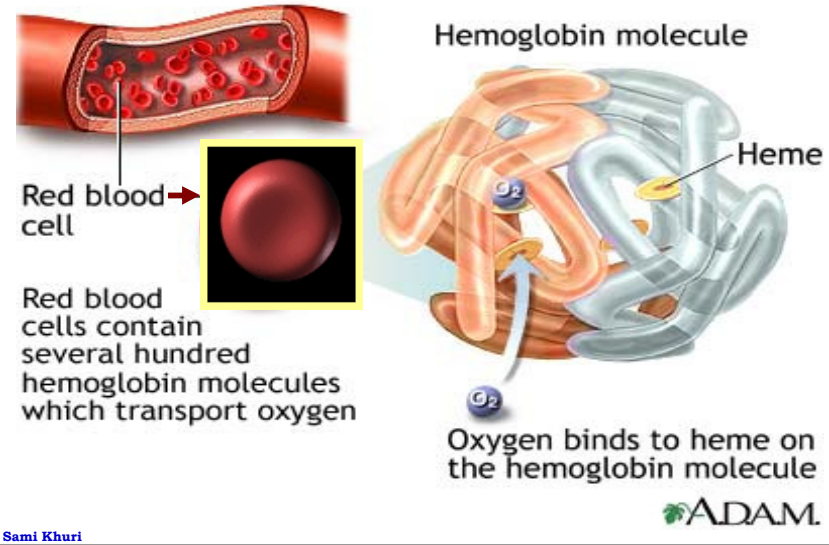
* = identity
: = strongly conserved
. = weakly conserved

By using the same five sequences and aligning them with CLUSTALW, we get the illustrated results.

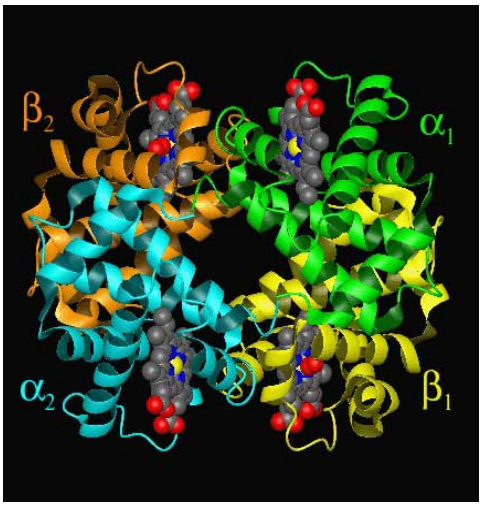


```
HBA_HUMAN       ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS- 49
HBA_HORSE       ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS- 49
HBA_CHICK       --------MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHF-DLS- 50
HBB_HUMAN       --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST 50
HBB_BOSMU       ---------MLTAEEKAAVTAFWGKVK--VDEVGGEALGRLLVVYPWTQRFFESFGDLSS 49
HBB_HORSE       --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN 50
HBB_MACGI       --------VHLTAEEKNAITSLWGKVA--IEQTGGEALGRLLIVYPWTSRFFDHFGDLSN 50
MYG_PHYCA       ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT 51
GLB5_PETMA      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT 60
LGB2_LUPLU      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE 52
                        *:  :   :   :  .           :  .:   * :   *  :   .

HBA_HUMAN       ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL 100
HBA_HORSE       ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL 100
HBA_CHICK       ----HGSAQIKGHGKKVVAALIEAANHIDD-----IAGTLSKLSDLHAHKLRVDPVNFKL 101
HBB_HUMAN       PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL 105
HBB_BOSMU       ADAVMNNPKVKAHGKKVLDSFSNGMKHLDD-----LKGTFAALSELHGDKLHVDPENFKL 104
HBB_HORSE       PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL 105
HBB_MACGI       AKAVMANPKVLAHGAKVLVAFGDAIKNLDN-----LKGTFAKLSELHCDKLHVDPENFKL 105
MYG_PHYCA       EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF 106
GLB5_PETMA      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV 118
LGB2_LUPLU      VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV 109
                      . .:  *.  :   .                   :  *. *.  .  :    : .

HBA_HUMAN       LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------ 141
HBA_HORSE       LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------ 141
HBA_CHICK       LGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR------ 142
HBB_HUMAN       LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------ 146
HBB_BOSMU       LGNVLVVVLARHFGKEFTPVLQADFQKVVVGVANALAHRYH------ 145
HBB_HORSE       LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------ 146
HBB_MACGI       LGNIIVICLAEHFGKEFTIDTQVAWQKLVAGVANALAHKYH------ 146
MYG_PHYCA       ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG 153
GLB5_PETMA      LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY------- 149
LGB2_LUPLU      VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA--- 153
                :   :   :           .         .   :
```

Red Blood Cells



Hemoglobin Structure



- 4 subunits: 2 alpha and 2 beta (red and gold)
- Each subunit holds a heme group
- Each heme group contains an iron atom, which is responsible for the binding of oxygen

Practical Considerations

- When to use Clustal?
- Can be used to align any group of protein or nucleic acid sequences that are related to each other over their entire lengths.
- Clustal is optimized to align sets of sequences that are entirely co-linear, i.e. sequences that have the same protein domains, in the same order.

When Not To Use Clustal

- Sequences do not share common ancestry.
- Sequences are partially related.
- Sequences include short non overlapping fragments.

Structural Alignment

- What you really want to do is "align regions of similar function".
- These are the areas that are evolutionarily conserved. (Folds, domains, disulfide bonds)

- **Problem**
  - The computer does not know anything about the structure or function of the proteins.
- **Solution**
  - Use computer alignment as a first step, then manually adjust the alignment to account for regions of structural similarity.

MSA Editors

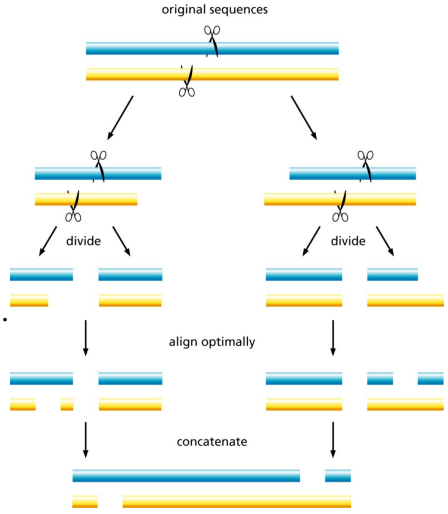
- Once the multiple alignment is produced, it may be necessary to edit the sequence manually to obtain a more reasonable or expected alignment.
- Some of the considerations for an editor:
  - the use of colors to aid in the visual representation of the alignment,
  - the capability of recognizing the alignment format,
  - the ability of using the mouse to add, delete, or move sequences, thus allowing for an adequate windows interface.

Divide and Conquer Alignment

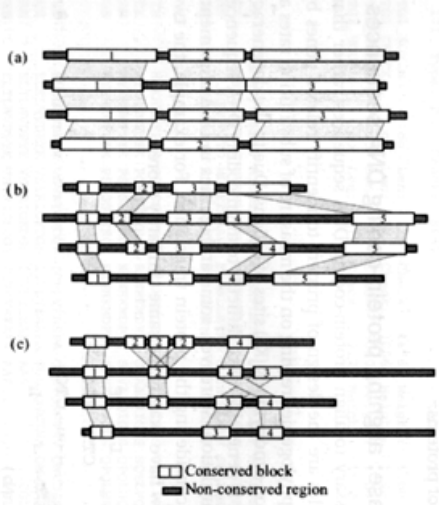
Use the **divide and conquer** technique to perform multiple sequence alignment.

MSA with the Divide & Conquer Method by J. Stoye Gene 211(2), GC45-GC56, 1998. (Gene-COMBIS).

http://bibiserv.techfak.uni-bielefeld.de/dca/algorithm/



Appropriate Approaches



	Comment	Appropriate approach
(a)	Sequences are related over their entire length.	Progressive global alignment method (e.g. CLUSTAL W).
(b)	Sequences share conserved blocks, separated by non-conserved regions containing large indels. Blocks are consistent (i.e. in the same order) but not necessarily uniform (i.e. some blocks may be missing in some sequences).	Block-based global alignment method (e.g. DIALIGN, ITERALIGN). Compare alignments produced by different programs (including progressive methods).
(c)	Sequences contain a non-consistent set of conserved blocks (i.e. some blocks are duplicated or occur in a different order along sequences).	Motif-based local alignment method (e.g. MEME). Compare alignments produced by different programs.

Figure 4 Choice of multiple alignment methods according of the nature of the sequence set.