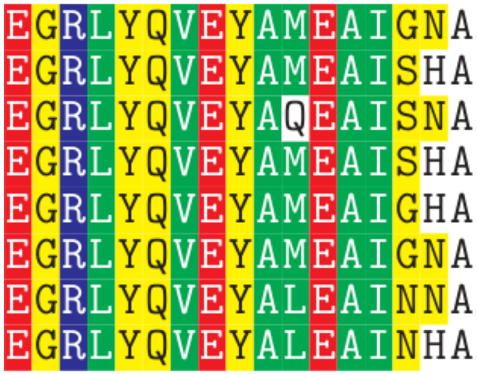
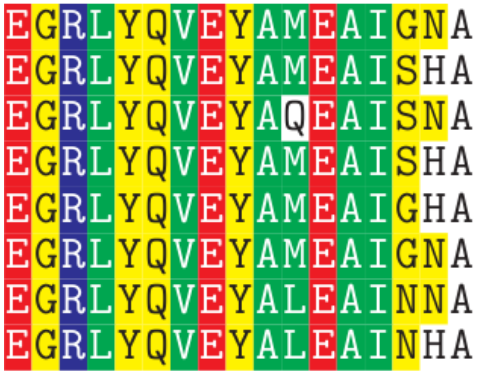
Bioinformatics

Four
Multiple Sequence Alignment

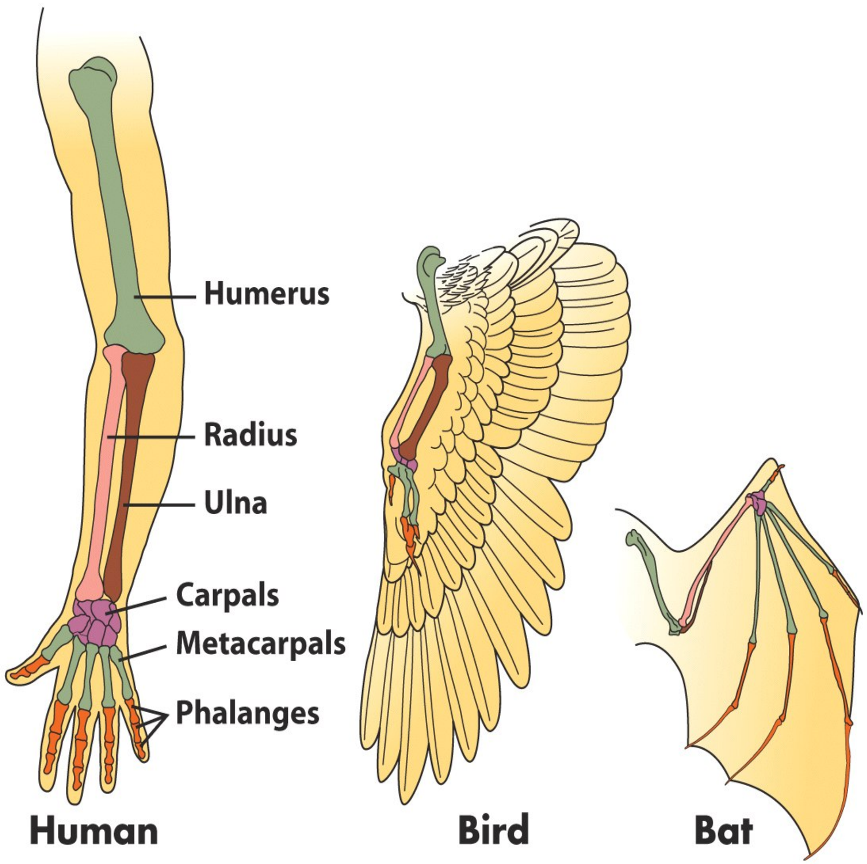


Wendy Lee
Dept of Computer Science
San José State University

Biology/CS/SE 123A
Fall 2014



Multiple Sequence Alignment



- Progressive Alignment
- Guide Tree
- ClustalW
- TCoffee
- Muscle
- MAFFT

```
                           *        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       : AAAGTTAACGAGTGGTTCTCCAGAAGTAATGAAATGTTAACTCCTGATGACTCACTTGACCGGCGGTCTGAATCAAATGCCAA :  83
```

```
                       *       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------CCAAATGCAATAGATAGRTTTTTGGTTCTTCAGAGAAAATAAAC : 150
Human      : AGTAGCTGATGTATTGGACGTT------CTAAATGAGGTAGATGAATATTCTGGTTCTTCAGAGAAAATAGAC : 150
Rat        : AGCTGCTGTTGTGTTAGAAGTT------TCAAATGAAGTGGATGGATGTTTCAGTTCTTCAAAGAAAATAGAC : 150
Hare       : AGTGGCTGGTGCATTAGAAGTC------CCAAAGGAGGTAGATGGATATTCTGGTTCTACAGAGAAAATAGAC : 150
```

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

Aligning Kinases: An Example

```
p110β        SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT
p110δ        TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT
p110α        TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT
p110γ        TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT
p110_dicti   TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT
cAMP-kinase  QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA
```

Multiple sequence alignment between a cAMP-kinase and 5 PI-3 kinases. Green indicates total conservation (identical residues), while blue indicates physicochemically conserved residues (belonging to the same partition of amino acids).

Pairwise vs. Multiple Alignment

```
p110α         TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI
cAMP-kinase   QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE
```

```
p110β         SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT
p110δ         TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT
p110α         TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT
p110γ         TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT
p110_dicti    TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT
cAMP-kinase   QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA
```

Top Figure: The pairwise alignment of the two homologous kinases does not align the important active-site residues and the DFG motif (in green).
Bottom Figure: The multiple sequence alignment of 5 homologous kinases forces the best-conserved regions to be matched.

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

Uses of Multiple Alignment

From a good **alignment**, one can

• Infer phylogenetic relationships; evolution of organisms.

• Elucidate biological facts about proteins: most conserved regions are usually biologically significant.

• Formulate and test hypothese about protein 3-D structure (based on conserved regions).

• Formulate and test hypotheses about protein function (see which regions of a gene, or its derived protein, are susceptible to mutaton & which can have one residue replaced by another without changing the function)

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.

Step One of ClustalW: Pairwise Alignments

1) Perform pairwise alignments of all sequences Compare each sequence with each other calculate a **distance matrix**.

	A	B	C
A	-		
B	.87	-	
C	.59	.60	-

Distance Matrix

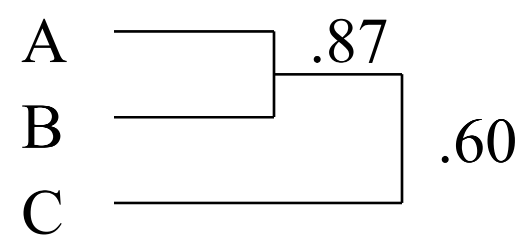
Distance = Number of exact matches divided by the sequence length (ignoring gaps).

Note that .87 means 87% identical.

Step Two of ClustalW: Create Guide Tree

2) Use the results of the Distance Matrix to create a **Guide Tree** to help determine in what order the sequences are aligned.

	A	B	C
A	-		
B	.87	-	
C	.59	.60	-



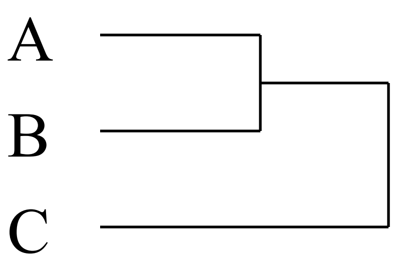
Guide Tree

The **Guide Tree**, or Dendrogram has no phylogenetic meaning.

It cannot be used to show evolutionary relationships.

Step Three of ClustalW: Progressive Alignment

3) Use the **Guide Tree** to align the sequences



- Align A and B first
- Then add sequence C to the previous alignment

Align the most closely related sequences first, then add in the most distantly related ones and align them to the existing alignment, inserting gaps if necessary.

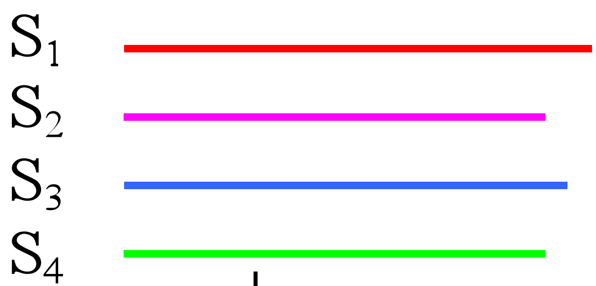
Multiple Alignment Problems

- Does the quality of the **guide tree** matter?
  - Not for very closely related sequences, but perhaps for distantly related ones.
- **Local minimum** problem
  - If the initial alignments have a problem, they cannot be removed during subsequent steps.

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



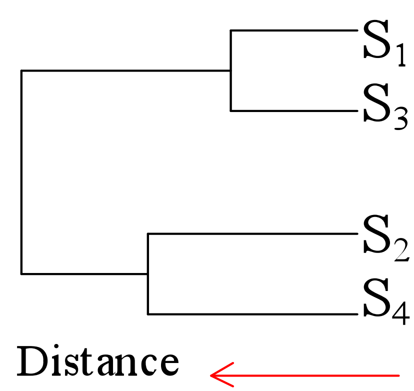
All Pairwise Alignments

Similarity Matrix

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

Cluster Analysis

Dendrogram



Multiple Alignment Step:

1. Aligning S_1 and S_3
2. Aligning S_2 and S_4
3. Aligning (S_1,S_3) with (S_2,S_4).

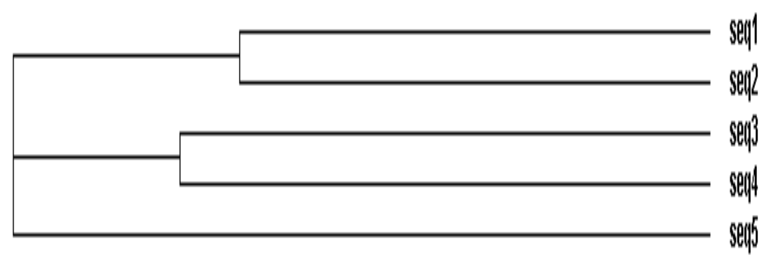
ClustalW: An Example

```
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
                ::*    :*:
```

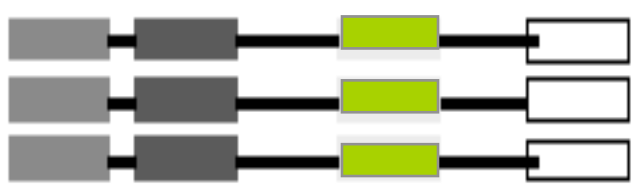
* = identity
: = strongly conserved
. = weakly conserved

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



Practical Considerations

- When to use ClustalW?
- 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 ClustalW

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

Alignment Problems

- Final result sometimes depends on the **order** that sequences were analyzed.
- **Gaps** can make alignment unrealistically long.
- Sequences of **different lengths** can cause problems.
- **Non-conserved** regions can dilute conserved areas.
  - Only need to align the shared domain.
  - So trim away any excess sequence and realign.

Clustal Omega

Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega

Fabian Sievers[1,8], Andreas Wilm[2,8], David Dineen[1], Toby J Gibson[3], Kevin Karplus[4], Weizhong Li[5], Rodrigo Lopez[5], Hamish McWilliam[5], Michael Remmert[6], Johannes Söding[6], Julie D Thompson[7] and Desmond G Higgins[1,*]

In this paper, we describe a new program called Clustal Omega, which can align virtually any number of protein sequences quickly and that delivers accurate alignments. The accuracy of the package on smaller test cases is similar to that of the high-quality aligners. On larger data sets, Clustal Omega outperforms other packages in terms of execution time and quality. Clustal Omega also has powerful features for adding sequences to and exploiting information in existing alignments, making use of the vast amount of precomputed information in public databases like Pfam.

DNA or Protein Alignment

- If we are comparing two or more sequences, is it better to align the **DNA**, or **Protein**?

  It depends on what we want to compare.
  - If **protein function**, then look at the amino acids
  - If **genetic changes**, then look at the DNA

- The **initial mutations** take place at the DNA level, but the **evolutionary pressure** occurs at the protein level.

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.

Alternatives to CLUSTALW (I)

- **Clustal Omega**
- **TCoffee**: A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures.
  - Good for distantly related sequences too.
  - www.tcoffee.org
- MUSCLE: **Mu**ltiple **S**equence **C**omparison by **L**og-**E**xpectation
  - www.drive5.com/muscle

Alternatives to CLUSTALW (II)

- **MAFFT**: **M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform.
  - A good balance between accuracy and speed.
  - http://mafft.cbrc.jp/alignment/software/
- **PRRN**: A web-based multiple sequence alignment package.
  - http://www.genome.jp/tools/prrn/

Alternatives to CLUSTALW (III)

University of Göttingen | Faculty of Biology | Inst. of Microbiology and Genetics | Dep. of Bioinformatics

DIALIGN [home]

Dialign 2.2.1 - Welcome

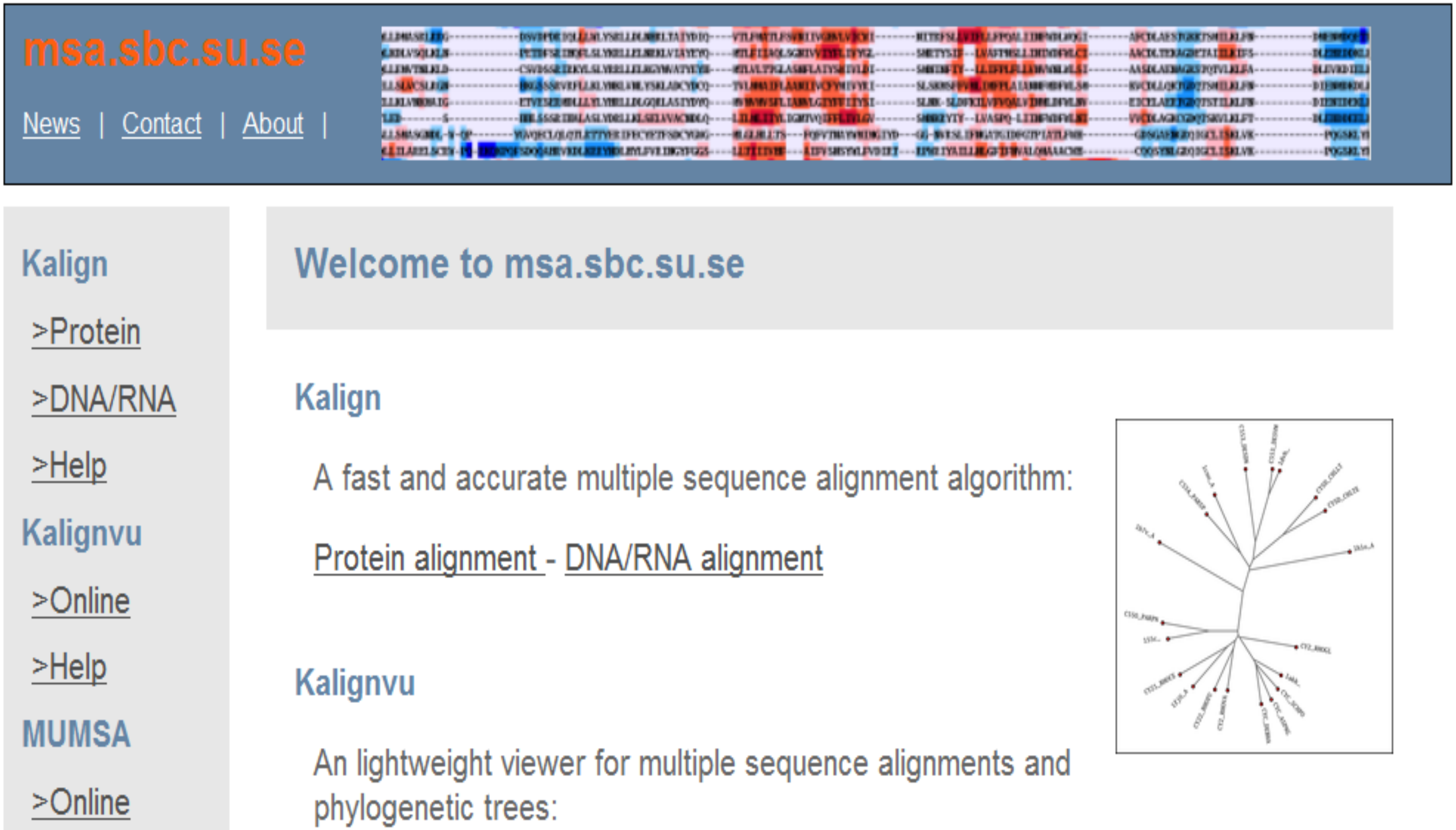
This is the new home page of the DIALIGN multiple-alignment program at *Göttingen Bioinformatics Compute Server (GOBICS)*

If you use *DIALIGN*, please cite this paper:

L. Al Ait, Z. Yamak, B. Morgenstern (2013)
DIALIGN at GOBICS - multiple sequence alignment using various sources of external information
Nuc. Acids Research 41, W3-W7

http://dialign.gobics.de

Alternatives to CLUSTALW (IV)



http://msa.sbc.su.se/cgi-bin/msa.cgi

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.

MSA Editor and Formatter Programs

- Multiple Sequence Alignment programs:
  - CINEMA (Color Interactive Editor for Multiple Alignments)
  - GDE (Genetic Data Environment)
  - GeneDoc
  - MACAW
- Multiple Sequence Alignment programs:
  - Boxshade
  - CLUSTALX