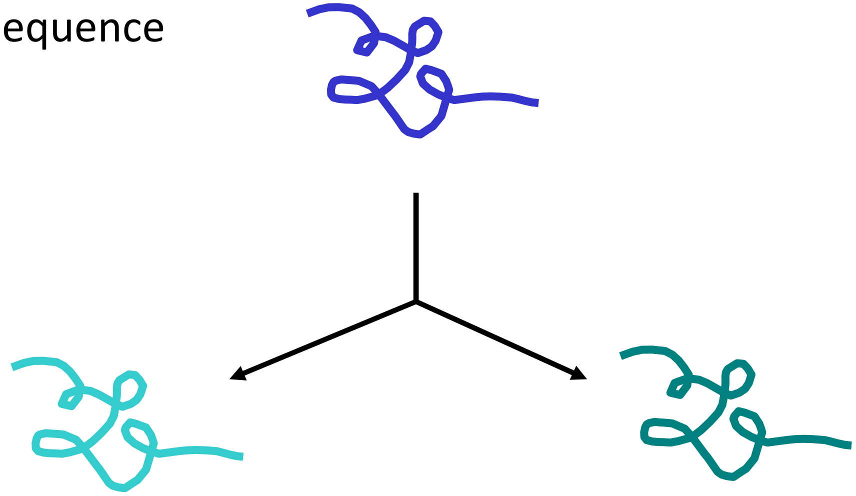
Gene Sequence Analysis

Lecture 3: **Multiple Alignment**

02/06/2024
Phuc Loi Luu, PhD
Apdated from Dr. Morgan Langille

Importance of Similarity

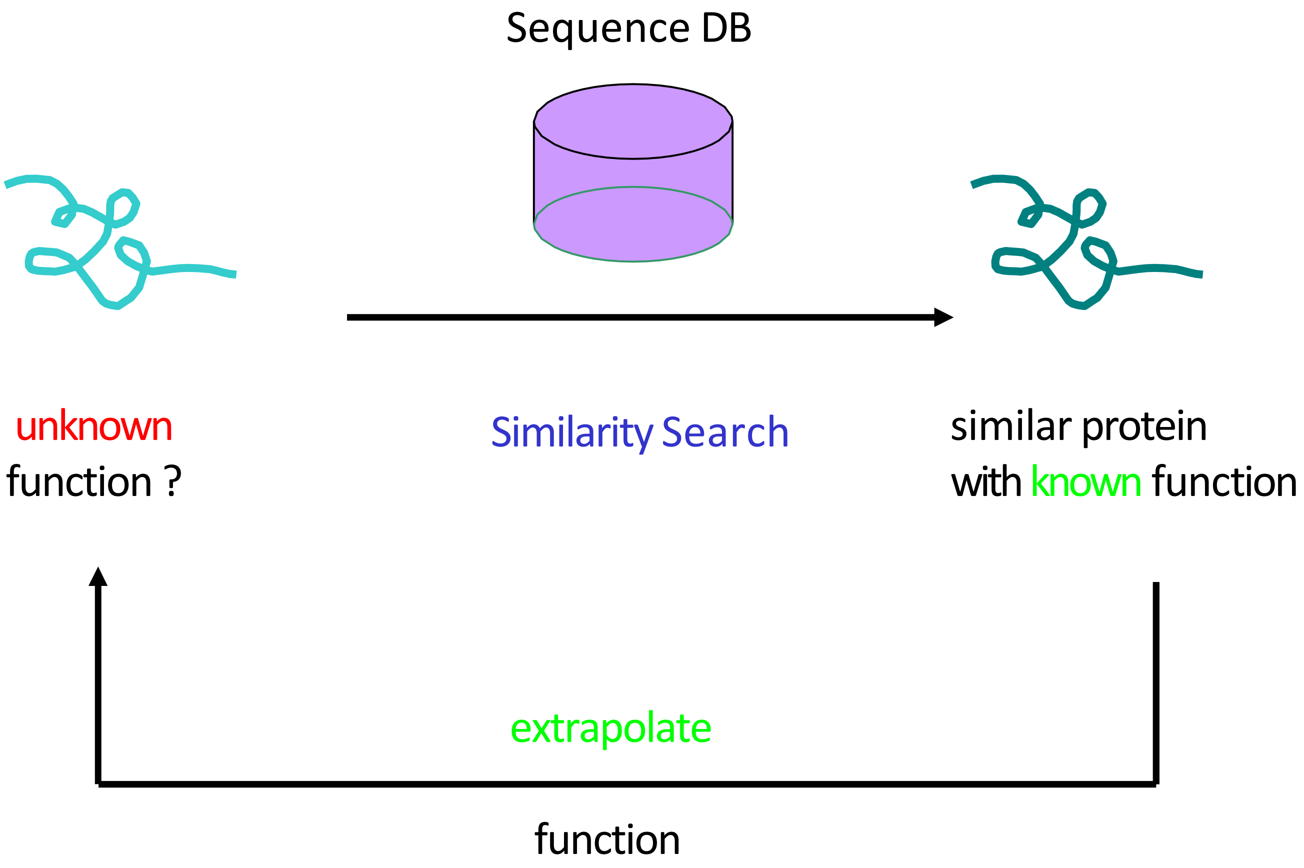


Ancestral protein/gene sequence

Similar (homologous) protein/gene sequences

Similar sequences: probably have the same ancestor, share the same structure, and have a similar biological function

Importance of Similarity



Sequence DB

unknown
function ?

Similarity Search

similar protein
with known function

extrapolate

function

Importance of Similarity

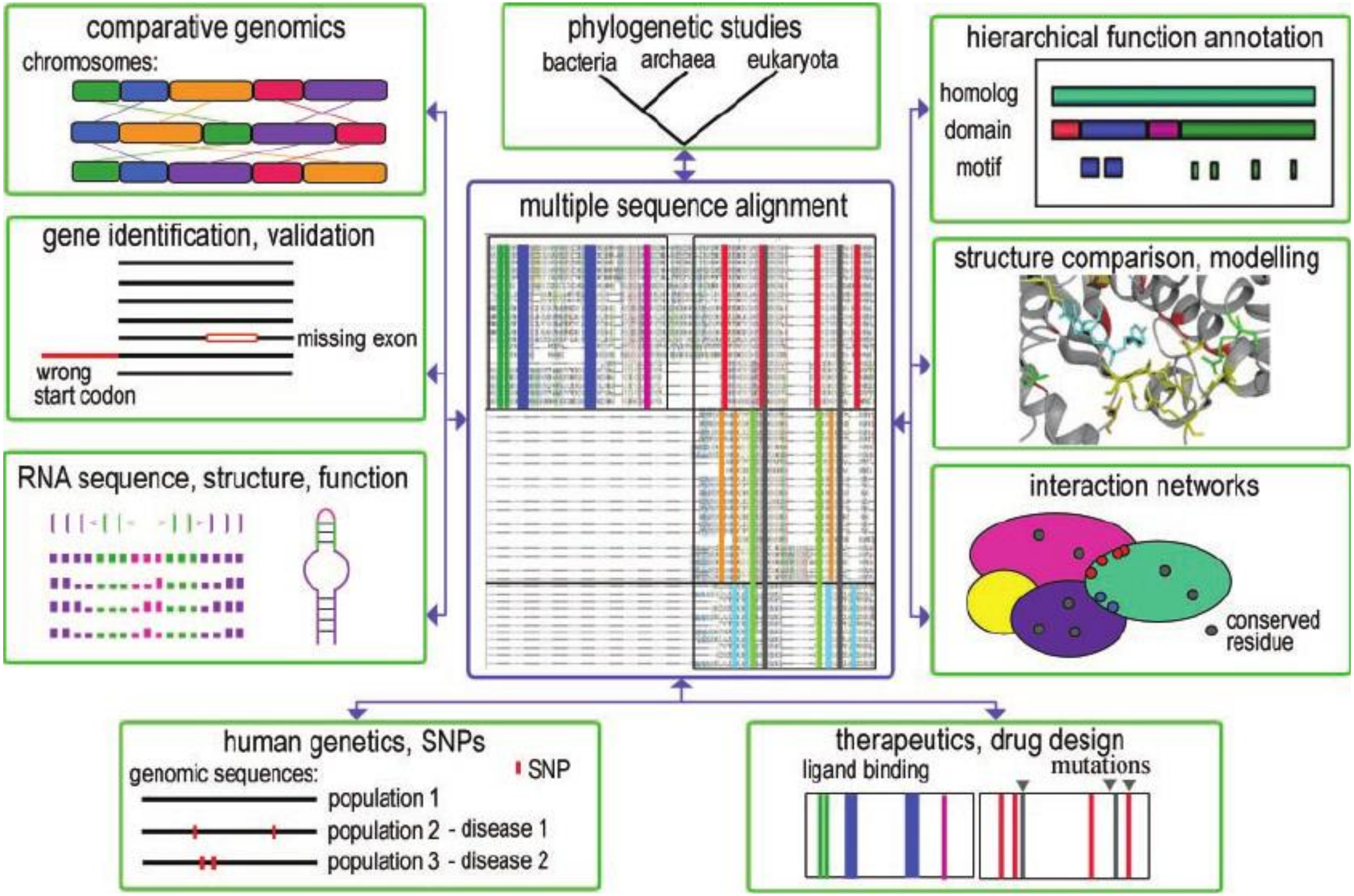
Rule-of-thumb:
If your sequences are more than **100 amino acids** long (or 100 nucleotides long) you can considered them as homologues if **25%** of the **aa** are identical (**70%** of **nucleotide** for DNA). Below this value you enter the twilight zone.

Twilight zone = protein sequence similarity between ~0-20% identity:
is not statistically significant, i.e. could have arisen by chance.

Beware:

- E-value (*Expectation value*)
- length of the segments similar between the two sequences
- The number of insertions/deletions

Examples of molecular biology applications (shown in green boxes) that rely on multiple sequence alignments



10.1093/nar/gki735

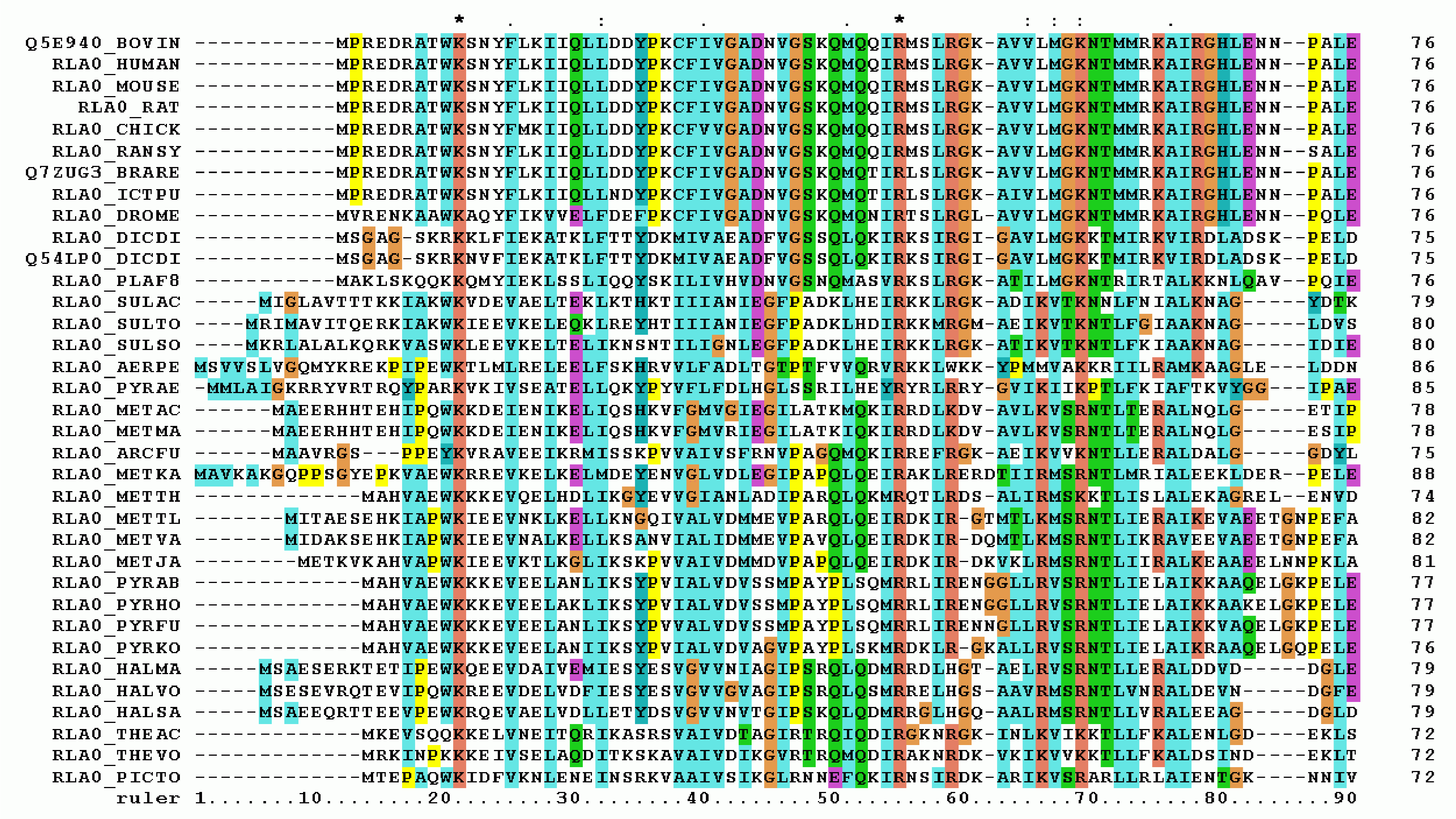
Outline

- What is a multiple alignment?
- Why do we need a multiple alignment?
- Characters of evolution
- How to create multiple alignments?
- Editing Alignments
- Viewing Alignments

What is a multiple alignment?

- Simply an alignment of more than 2 sequences
- Sequences are aligned globally (end to end)
- Multiple Sequence Alignment (MSA) programs try to insert gaps in the sequences so that *homologous characters* are aligned

What is a multiple alignment?



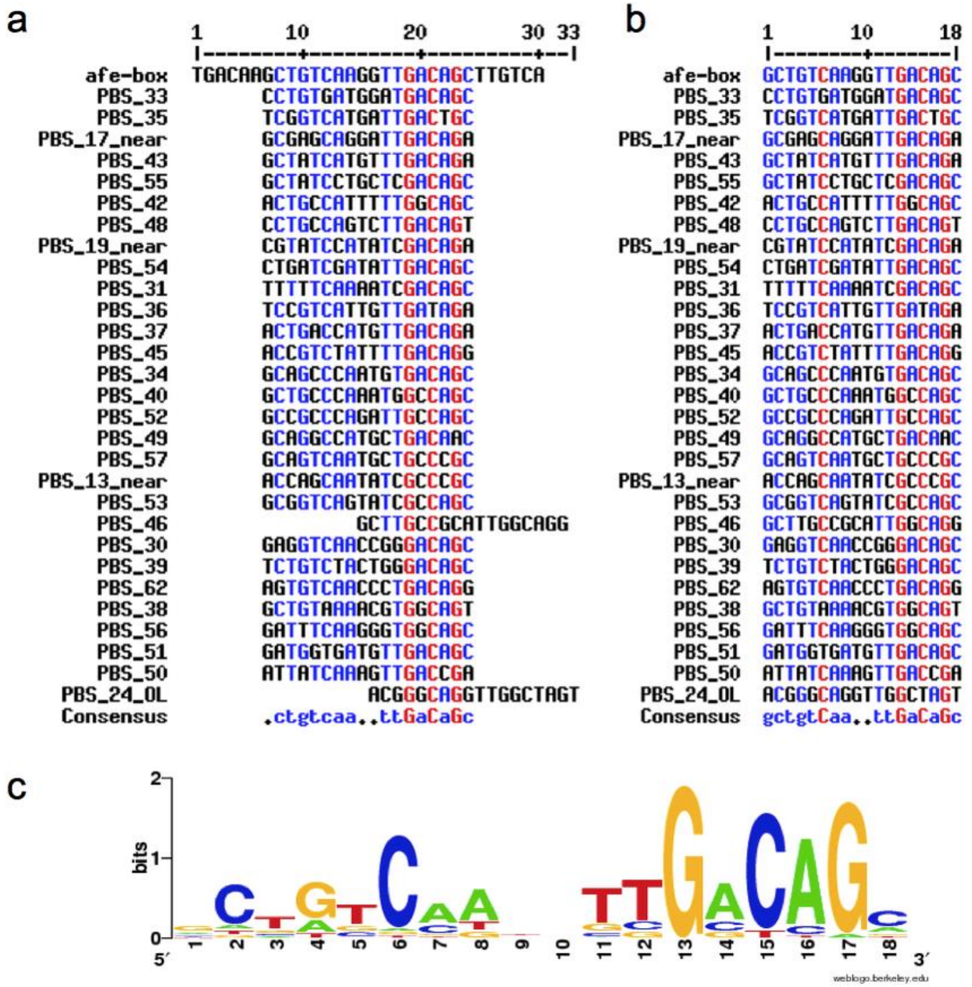
```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE     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
```

https://en.wikipedia.org/wiki/Multiple_sequence_alignment

What to do with a MSA?

- To look for sites of interest/conservation within a gene (motifs, binding sites, etc.)
- To build phylogenetic trees
- To identify positive selection

- To look for sites of interest/conservation within a gene (motifs, binding sites, etc.)

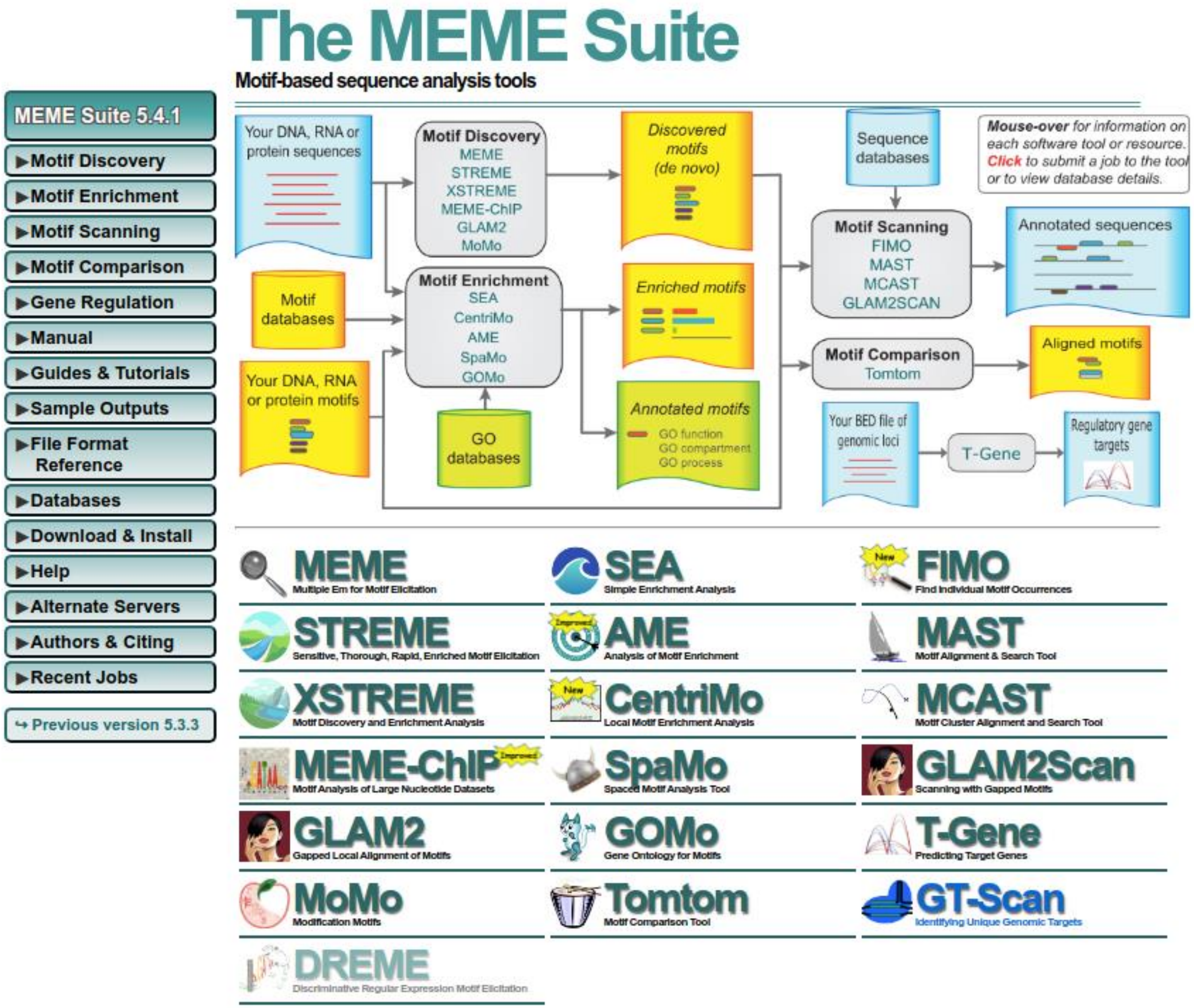


```
            1        10        20        30 33
            |--------+---------+---------+--|
    afe-box TGACAAGCTGTCAAGGTTGACAGCTTGTCA
     PBS_33       CCTGTGATGGATGACAGC
     PBS_35       TCGGTCATGATTGACTGC
PBS_17_near       GCGAGCAGGATTGACAGA
     PBS_43       GCTATCATGTTTGACAGA
     PBS_55       GCTATCCTGCTCGACAGC
     PBS_42       ACTGCCATTTTTGGCAGC
     PBS_48       CCTGCCAGTCTTGACAGT
PBS_19_near       CGTATCCATATCGACAGA
     PBS_54       CTGATCGATATTGACAGC
     PBS_31       TTTTTCAAAATCGACAGC
     PBS_36       TCCGTCATTGTTGATAGA
     PBS_37       ACTGACCATGTTGACAGA
     PBS_45       ACCGTCTATTTTGACAGG
     PBS_34       GCAGCCCAATGTGACAGC
     PBS_40       GCTGCCCAAATGGCCAGC
     PBS_52       GCCGCCCAGATTGCCAGC
     PBS_49       GCAGGCCATGCTGACAAC
     PBS_57       GCAGTCAATGCTGCCCGC
PBS_13_near       ACCAGCAATATCGCCCGC
     PBS_53       GCGGTCAGTATCGCCAGC
     PBS_46               GCTTGCCGCATTGGCAGG
     PBS_30       GAGGTCAACCGGGACAGC
     PBS_39       TCTGTCTACTGGGACAGC
     PBS_62       AGTGTCAACCCTGACAGG
     PBS_38       GCTGTAAAACGTGGCAGT
     PBS_56       GATTTCAAGGGTGGCAGC
     PBS_51       GATGGTGATGTTGACAGC
     PBS_50       ATTATCAAAGTTGACCGA
  PBS_24_OL                ACGGGCAGGTTGGCTAGT
  Consensus       .ctgtcaa..ttGaCaGc
```

```
            1        10      18
            |--------+-------|
    afe-box GCTGTCAAGGTTGACAGC
     PBS_33 CCTGTGATGGATGACAGC
     PBS_35 TCGGTCATGATTGACTGC
PBS_17_near GCGAGCAGGATTGACAGA
     PBS_43 GCTATCATGTTTGACAGA
     PBS_55 GCTATCCTGCTCGACAGC
     PBS_42 ACTGCCATTTTTGGCAGC
     PBS_48 CCTGCCAGTCTTGACAGT
PBS_19_near CGTATCCATATCGACAGA
     PBS_54 CTGATCGATATTGACAGC
     PBS_31 TTTTTCAAAATCGACAGC
     PBS_36 TCCGTCATTGTTGATAGA
     PBS_37 ACTGACCATGTTGACAGA
     PBS_45 ACCGTCTATTTTGACAGG
     PBS_34 GCAGCCCAATGTGACAGC
     PBS_40 GCTGCCCAAATGGCCAGC
     PBS_52 GCCGCCCAGATTGCCAGC
     PBS_49 GCAGGCCATGCTGACAAC
     PBS_57 GCAGTCAATGCTGCCCGC
PBS_13_near ACCAGCAATATCGCCCGC
     PBS_53 GCGGTCAGTATCGCCAGC
     PBS_46 GCTTGCCGCATTGGCAGG
     PBS_30 GAGGTCAACCGGGACAGC
     PBS_39 TCTGTCTACTGGGACAGC
     PBS_62 AGTGTCAACCCTGACAGG
     PBS_38 GCTGTAAAACGTGGCAGT
     PBS_56 GATTTCAAGGGTGGCAGC
     PBS_51 GATGGTGATGTTGACAGC
     PBS_50 ATTATCAAAGTTGACCGA
  PBS_24_OL ACGGGCAGGTTGGCTAGT
  Consensus gctgtCaa..ttGaCaGc
```

(a) Alignment including the complete version of the afe-box (30 nucleotides); **(b)** Alignment including the central palindromic module of the afe-box (18 nucleotides). PBS_24_OL and PBS_46 align with better scoring to the right half of the complete afe-box. Numbers correspond to PBSs listed in Supplementary Table 1;
(c) Sequence logo file created from the sequence alignment shown in b.

- To look for sites of interest/conservation within a gene (motifs, binding sites, etc.)



The MEME Suite

Motif-based sequence analysis tools

Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.

MEME – Multiple Em for Motif Elicitation

SEA – Simple Enrichment Analysis

FIMO (New) – Find Individual Motif Occurrences

STREME – Sensitive, Thorough, Rapid, Enriched Motif Elicitation

AME (Improved) – Analysis of Motif Enrichment

MAST – Motif Alignment & Search Tool

XSTREME – Motif Discovery and Enrichment Analysis

CentriMo (New) – Local Motif Enrichment Analysis

MCAST – Motif Cluster Alignment and Search Tool

MEME-ChIP (Improved) – Motif Analysis of Large Nucleotide Datasets

SpaMo – Spaced Motif Analysis Tool

GLAM2Scan – Scanning with Gapped Motifs

GLAM2 – Gapped Local Alignment of Motifs

GOMo – Gene Ontology for Motifs

T-Gene – Predicting Target Genes

MoMo – Modification Motifs

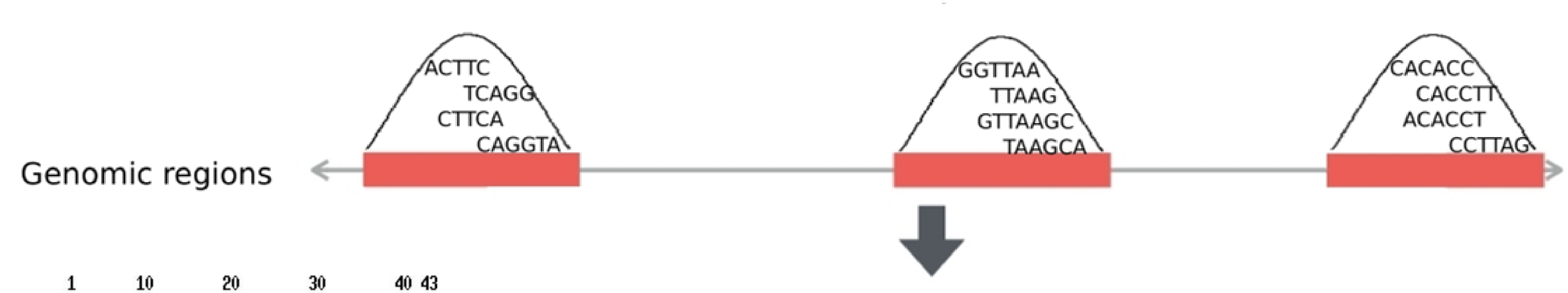
Tomtom – Motif Comparison Tool

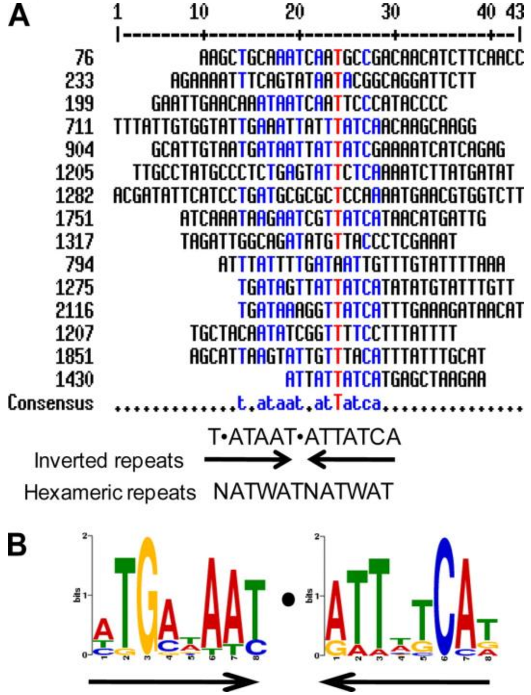
GT-Scan – Identifying Unique Genomic Targets

DREME – Discriminative Regular Expression Motif Elicitation

https://meme-suite.org/meme/index.html

- To look for sites of interest/conservation within a gene (motifs, binding sites, etc.)



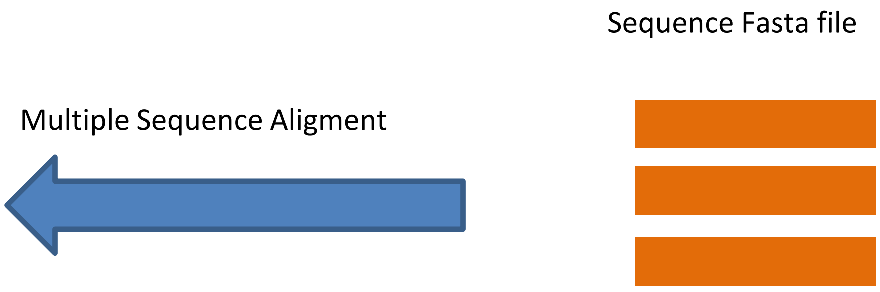


```
          1       10        20        30        40 43
          |--------+---------+---------+---------+--|
  76               AAGCTGCAAATCAATGCCGACAACATCTTCAACC
 233            AGAAAATTTCAGTATAATACGGCAGGATTCTT
 199          GAATTGAACAAATAATCAATTCCCATACCCC
 711      TTTATTGTGGTATTGAAATTATTTATCAACAAGCAAGG
 904          GCATTGTAATGATAATTATTATCGAAAATCATCAGAG
1205        TTGCCTATGCCCTCTGAGTATTCTCAAAATCTTATGATAT
1282      ACGATATTCATCCTGATGCGCGCTCCAAAATGAACGTGGTCTT
1751             ATCAAATAAGAATCGTTATCATAACATGATTG
1317             TAGATTGGCAGATATGTTACCCTCGAAAT
 794                 ATTTATTTTGATAATTGTTTGTATTTTAAA
1275                   TGATAGTTATTATCATATATGTATTTGTT
2116                   TGATAAAGGTTATCATTTGAAAGATAACAT
1207              TGCTACAATATCGGTTTTCCTTTATTTT
1851              AGCATTAAGTATTGTTTACATTTATTTGCAT
1430                        ATTATTATCATGAGCTAAGAA
Consensus ..............t.ataat.atTatca...............
```

T•ATAAT•ATTATCA

Inverted repeats

Hexameric repeats NATWATNATWAT

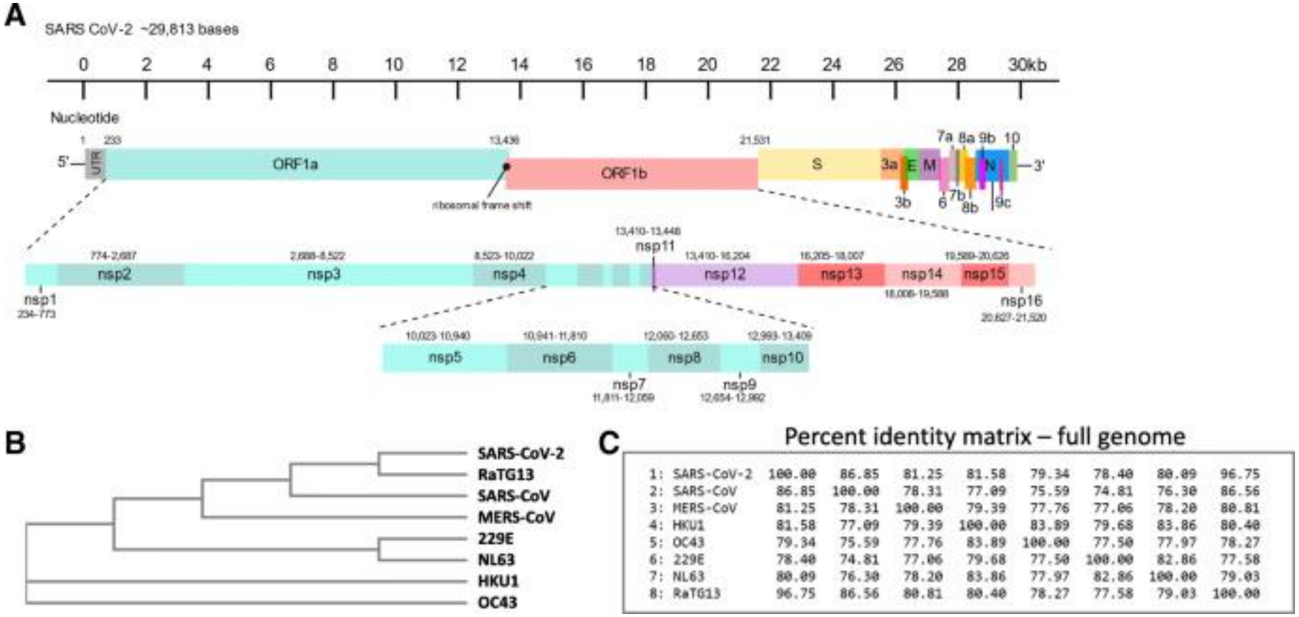


Sequence Fasta file

Multiple Sequence Aligment

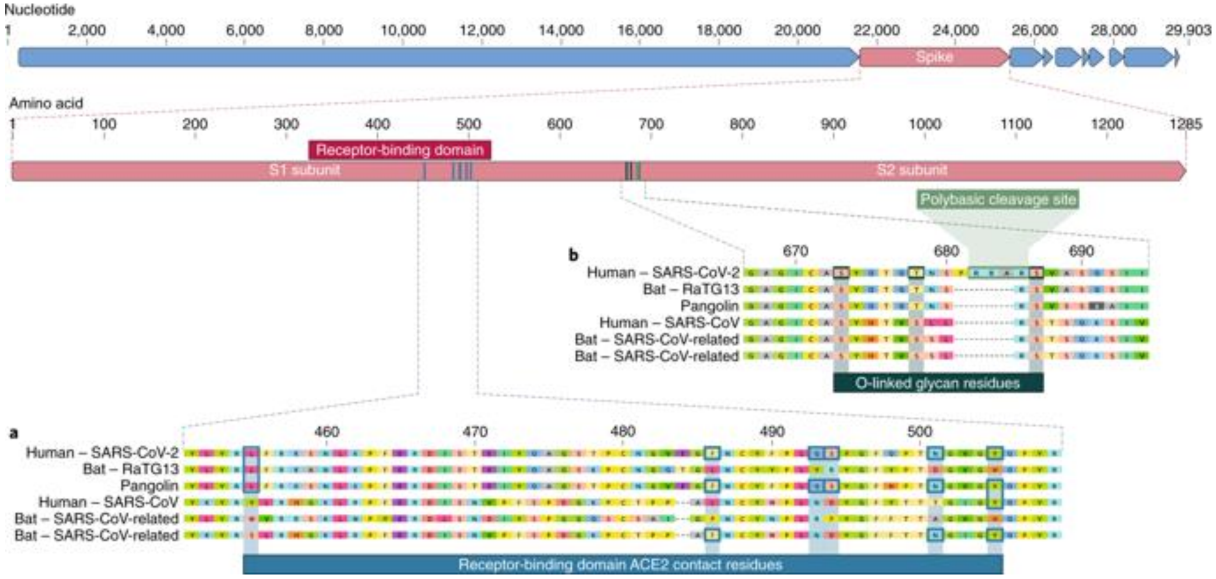
https://meme-suite.org/meme/index.html

Reconstruct phylogenetic trees



Percent identity matrix – full genome

```
1: SARS-CoV-2  100.00   86.85   81.25   81.58   79.34   78.40   80.09   96.75
2: SARS-CoV     86.85  100.00   78.31   77.09   75.59   74.81   76.30   86.56
3: MERS-CoV     81.25   78.31  100.00   79.39   77.76   77.06   78.20   80.81
4: HKU1         81.58   77.09   79.39  100.00   83.89   79.68   83.86   80.40
5: OC43         79.34   75.59   77.76   83.89  100.00   77.50   77.97   78.27
6: 229E         78.40   74.81   77.06   79.68   77.50  100.00   82.86   77.58
7: NL63         80.09   76.30   78.20   83.86   77.97   82.86  100.00   79.03
8: RaTG13       96.75   86.56   80.81   80.40   78.27   77.58   79.03  100.00
```



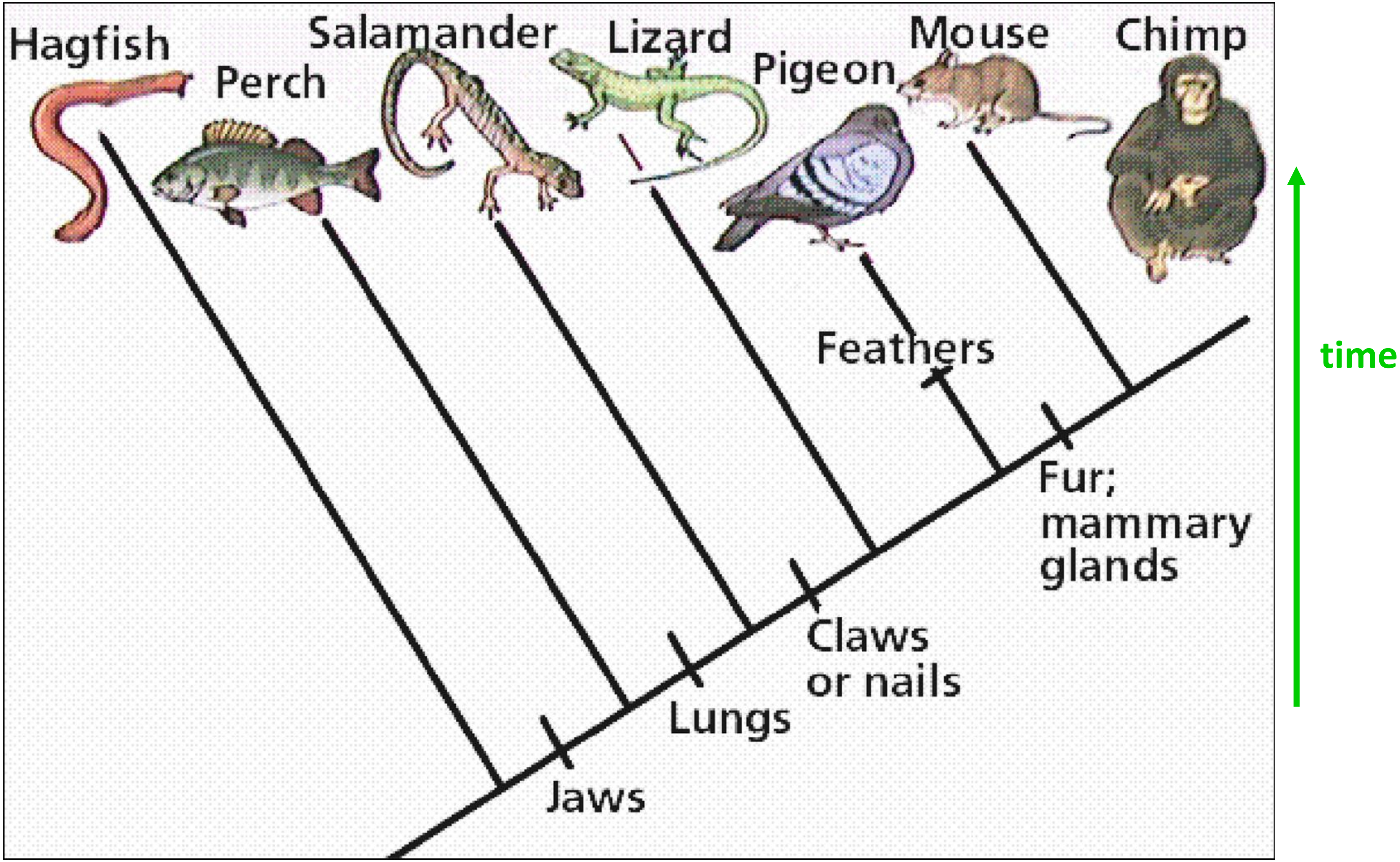
10.1007/s40142-020-00197-5

Characters

- Heritable changes in features (morphology, DNA sequence etc…)

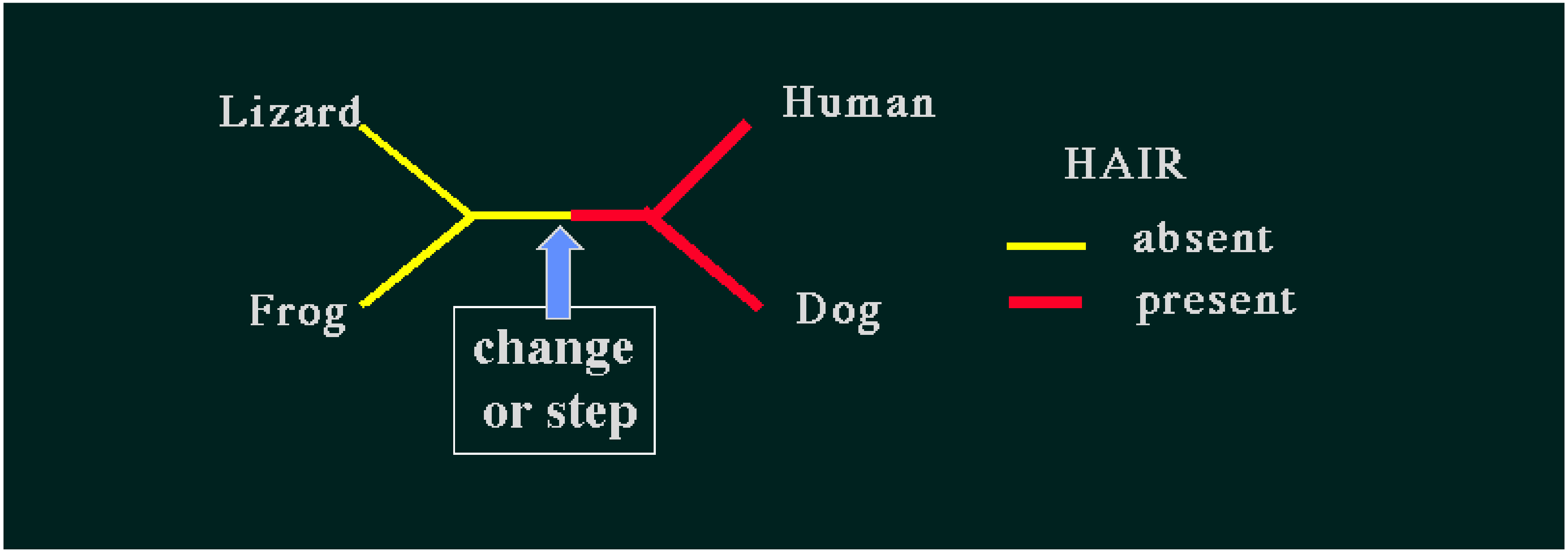
- The more similar characters you have, the more related you are

Evolution and characters



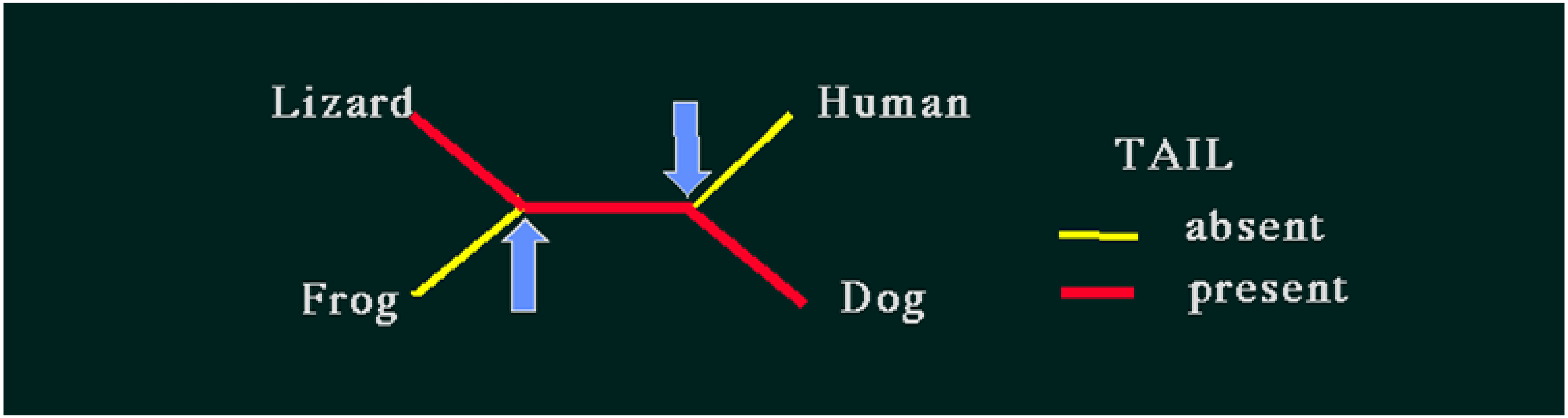
A Unique Character: Hair for Mammals

- Hair evolved only once and is “unreversed”
- Presence of hair is strong indication that organism is a mammal



Homoplasy: The formation of tails

- Tails evolved independently in the ancestors of frogs and humans
- Presence of a tail is no useful conclusions



Classification according to characters – more characters can be good

	Colour	**Skin**	**Cost**
Beef	red	no	$$$
Duck	red	yes	$$$
Pork	white	no	$$
Chicken	white	yes	$
Tofu	white	sometimes	$

Is Chicken most similar to Tofu?

Classification according to characters – increasing the number of characters

	Colour	Skin	Cost	Legs	Feathers	Hair
Beef	red	no	$$$	four	no	yes
Duck	red	yes	$$$	two	yes	no
Pork	white	no	$$	four	no	yes
Chicken	white	yes	$	two	yes	no
Tofu	white	sometimes	$	none	no	no

Evolution and characters – the importance of comparing characters with common origins (homologous)

bioinformatics

bioinformatics

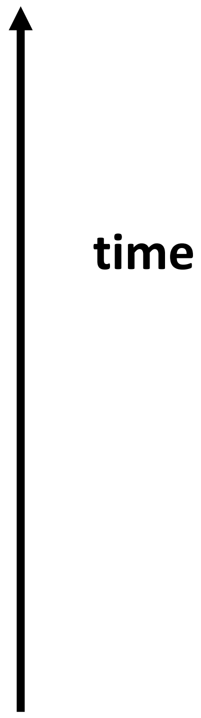
bioinformatios

oinformatios

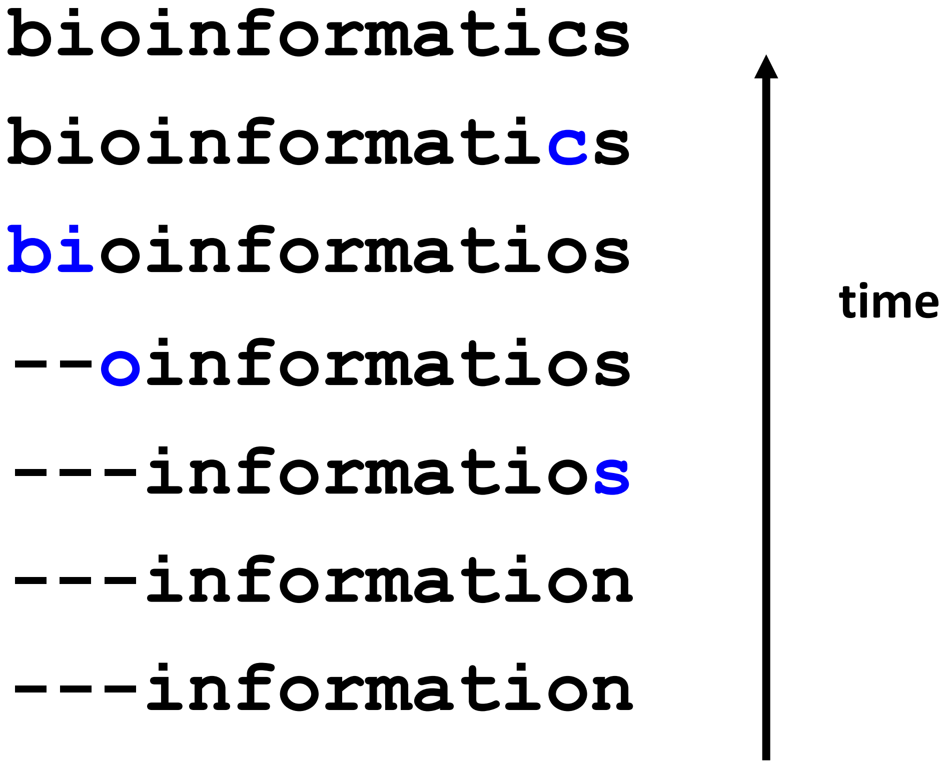
informatios

information

information



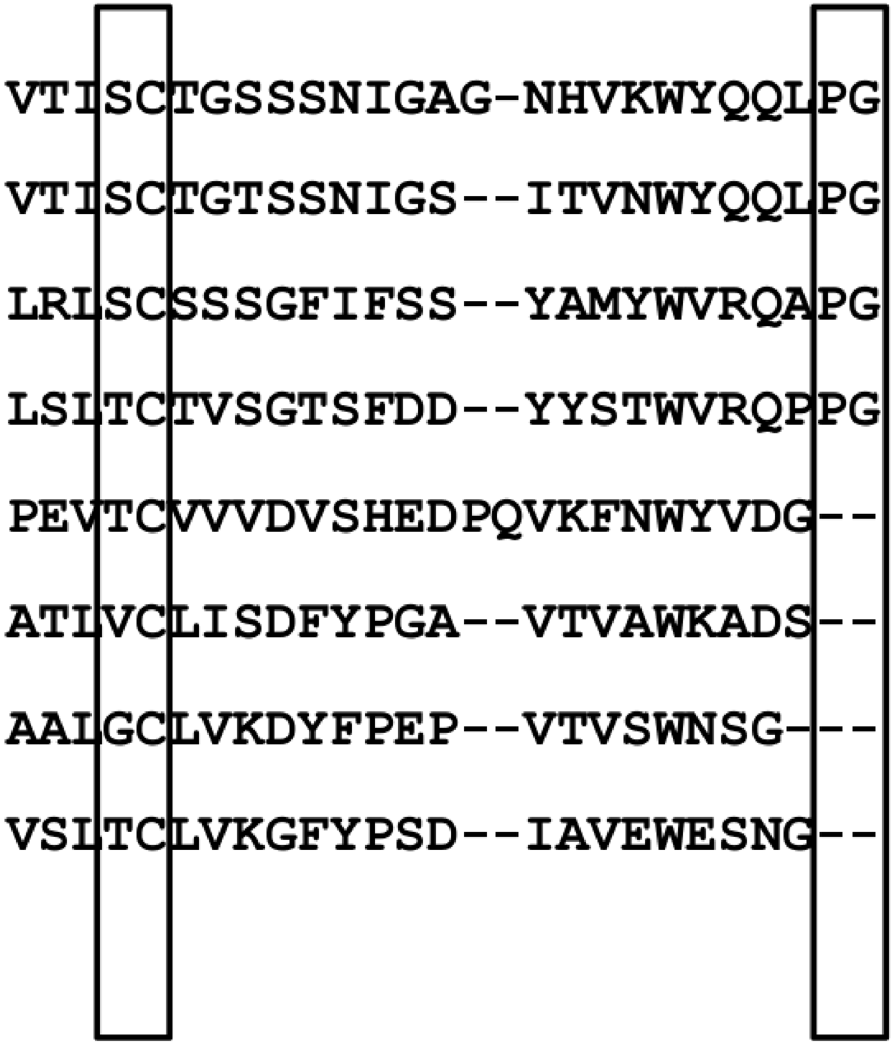
Evolution and characters



```
bioinformatics
bioinformatics
bioinformatios
--oinformatios
---informatios
---information
---information
```

- Gaps represent non-homologous positions in the sequence.
- They reflect the occurrence of insertions/deletions or other rearrangements during the evolutionary process.

Multiple Sequence Alignment



```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

The sole purpose of multiple sequence alignments is to place *homologous positions* of *homologous sequences* into the *same column*.

Multiple sequence alignments and phylogenetic analysis

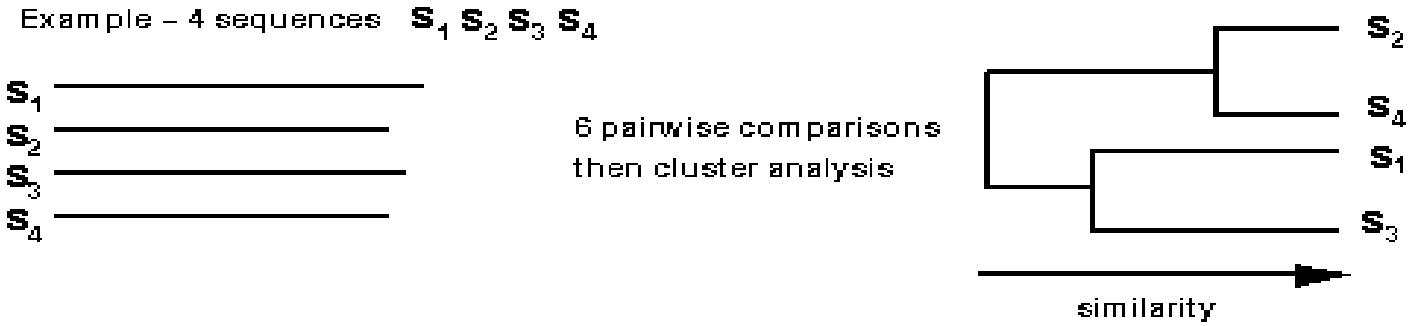
- *First step in any phylogenetic analysis*

- Phylogenetic analysis only as good as the alignment

in --> out!

Steps in Multiple Alignment

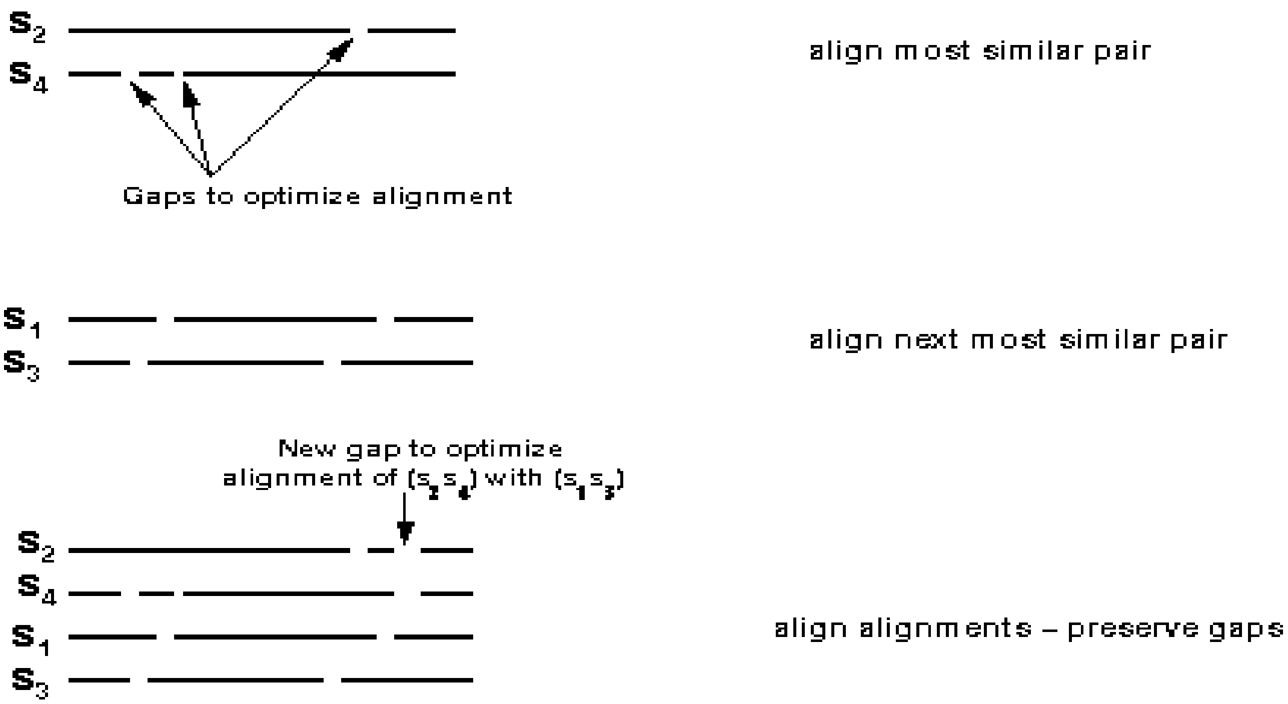
(A) Pairwise Alignment



Example – 4 sequences s_1 s_2 s_3 s_4

6 pairwise comparisons then cluster analysis

(B) Multiple alignment following the tree from A



align most similar pair

Gaps to optimize alignment

align next most similar pair

New gap to optimize alignment of (s_2s_4) with (s_1s_3)

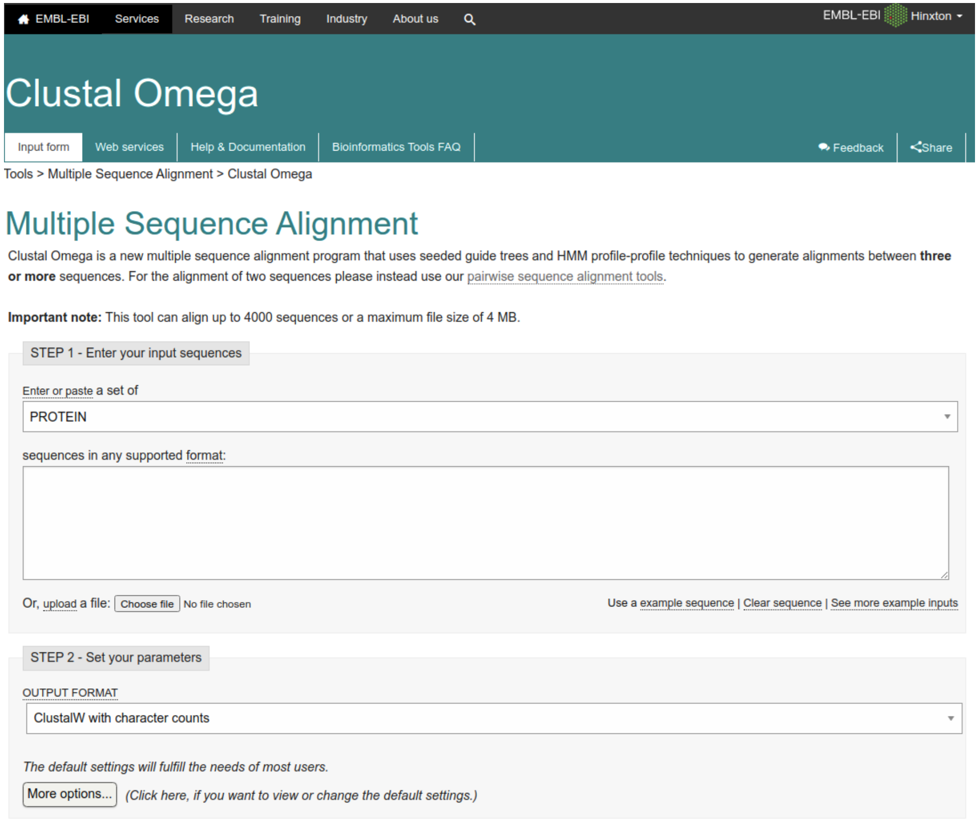
align alignments – preserve gaps

Creating a MSA

- Clustal
  - Been around for ever and widely used
  - ClustalW (command line)
  - ClustalX (GUI)
  - Also available on many web servers
  - http://www.clustal.org/clustal2/
- Muscle
  - Faster and maybe more accurate than Clustal
  - Command line only
  - There are web servers http://www.ebi.ac.uk/Tools/msa/muscle/
- T-Coffee
  - Most accurate, but also the slowest
  - Also has special variations for RNA, protein structure, etc.
  - http://tcoffee.crg.cat/

Need something faster

- Clustal Omega: HMM Based
- http://www.clustal.org/omega/



Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, upload a file: Choose file No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

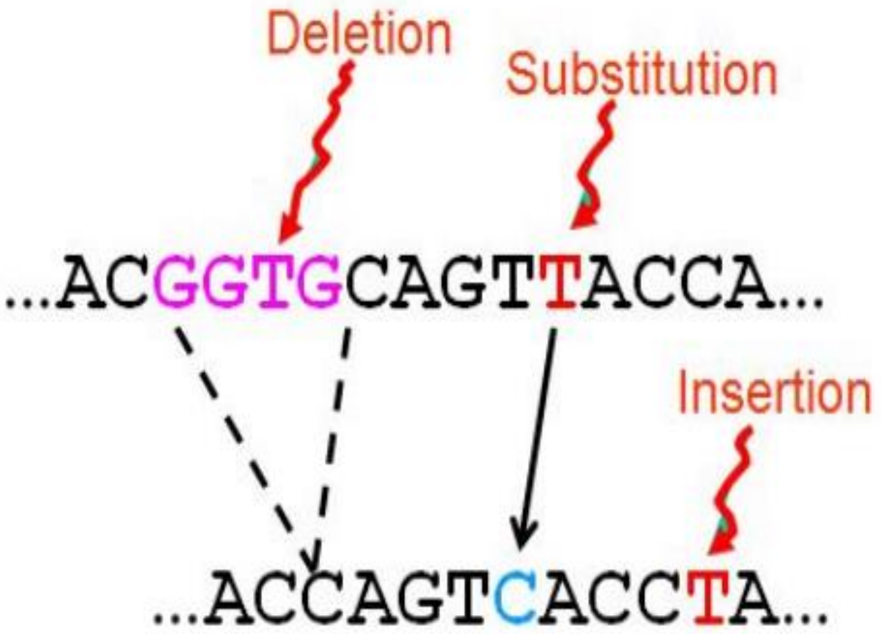
ClustalW with character counts

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

Editing Alignments

- A MSA is rarely perfect
- Downstream tools will presume columns are homologous
- Remove unreliably aligned regions for phylogenetic analysis



```
...ACGGTGCAGTTACC-A...
...AC----CAGTCACCTA...
```

Editing Alignments

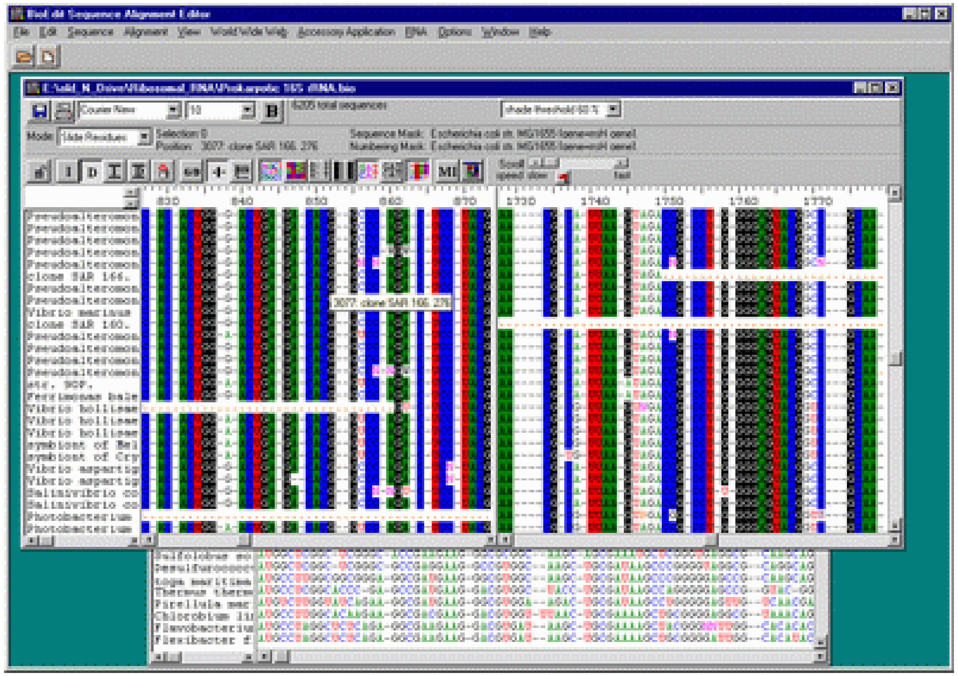
- A MSA is rarely perfect
- Downstream tools will presume columns are homologous
- Remove unreliably aligned regions for phylogenetic analysis

```
ILPITSPSKEGYESGKAPDEFSSGG
ILPEH--IKDDGELGAAPHSFSTAG
VLPLD-----S--AGRPADSFSAAG
VLPVDR-------DGQARDEYT-VG
VLPVDN-------KGEARDEYT-VG
LLPYDD-------QGRPQDDYSRAG
GIVSRSG---SNFDGEPKDSYGKVG
```

Delete?

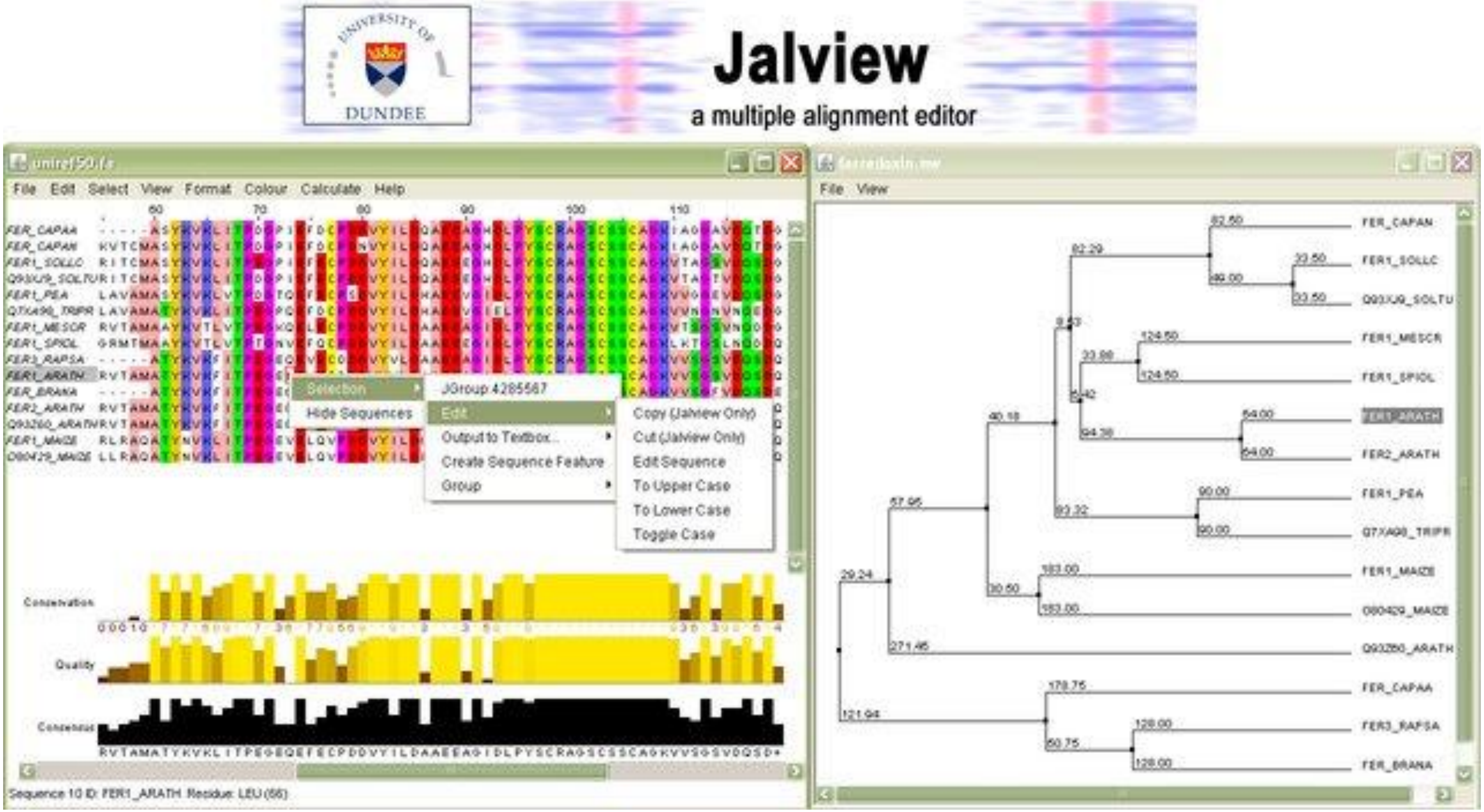
BioEdit: a biological sequence alignment editor

- An intuitive multiple document interface with convenient features
- Several sequence manipulation and analysis options and links to external analysis



https://bioedit.software.informer.com/7.2/

JalView: Manual

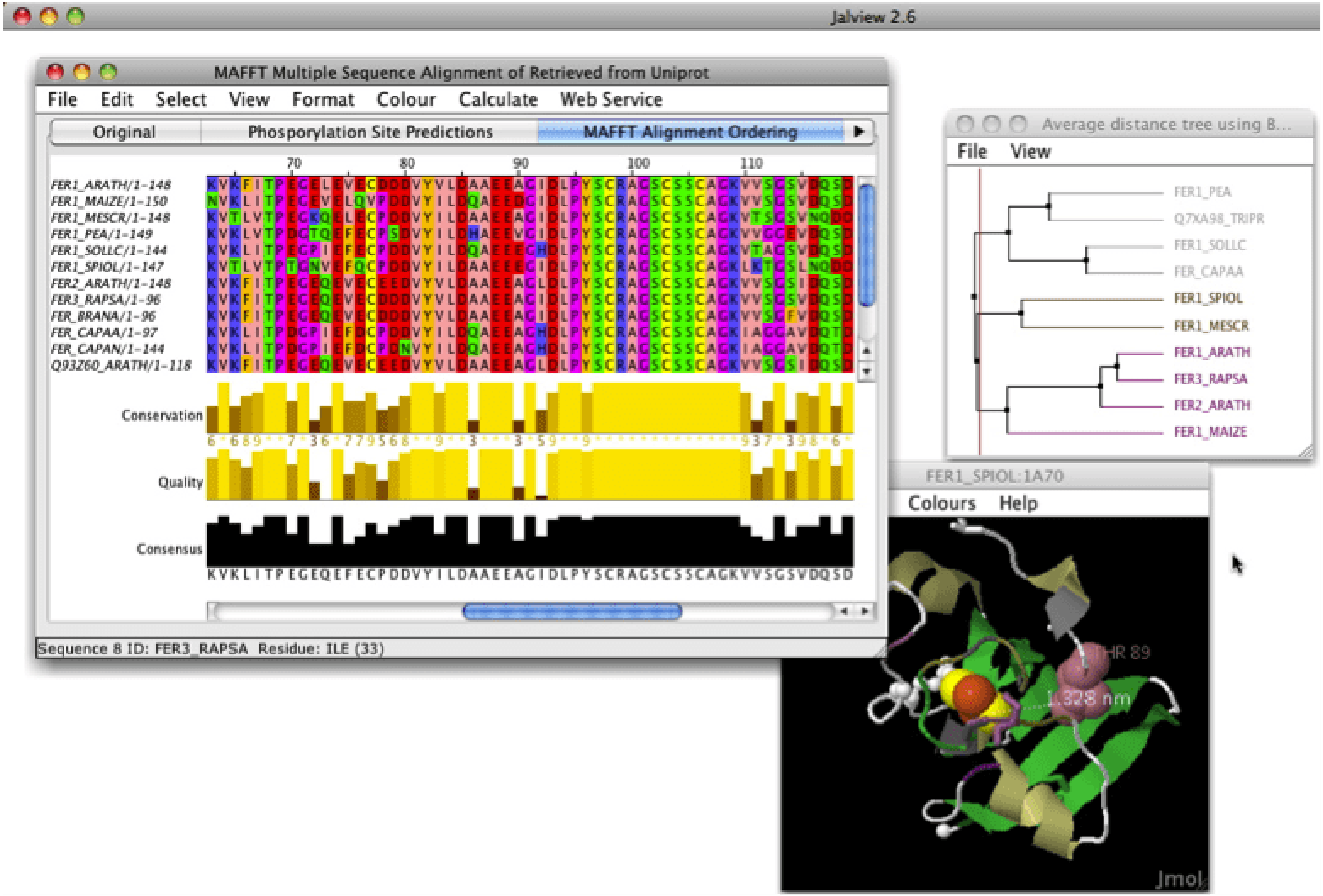


Sequence 10 ID: FER1_ARATH Residue: LEU (56)

http://www.pdg.cnb.uam.es/cursos/CNB2013_seq/UsingJalView.html

http://www.jalview.org/

JalView: Manual



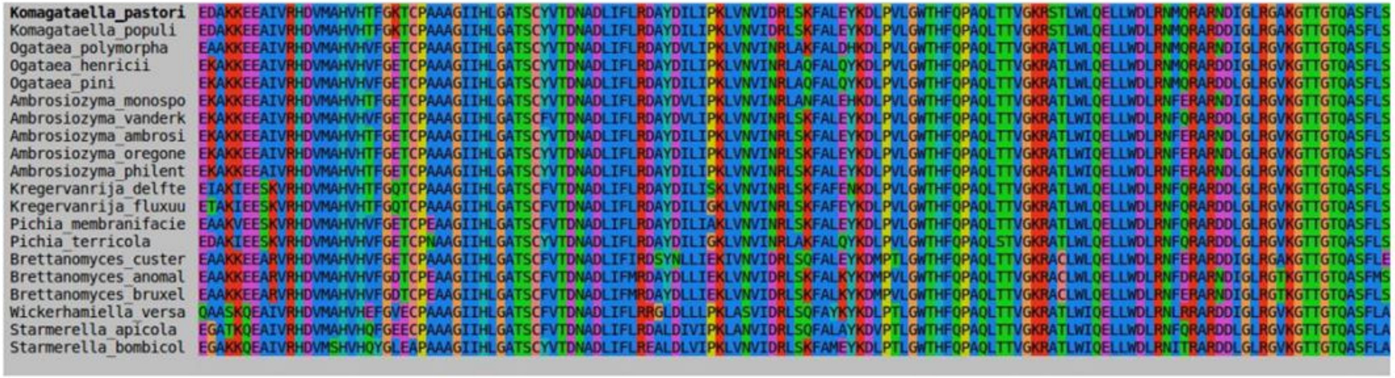
Sequence 8 ID: FER3_RAPSA Residue: ILE (33)

http://www.pdg.cnb.uam.es/cursos/CNB2013_seq/UsingJalView.html

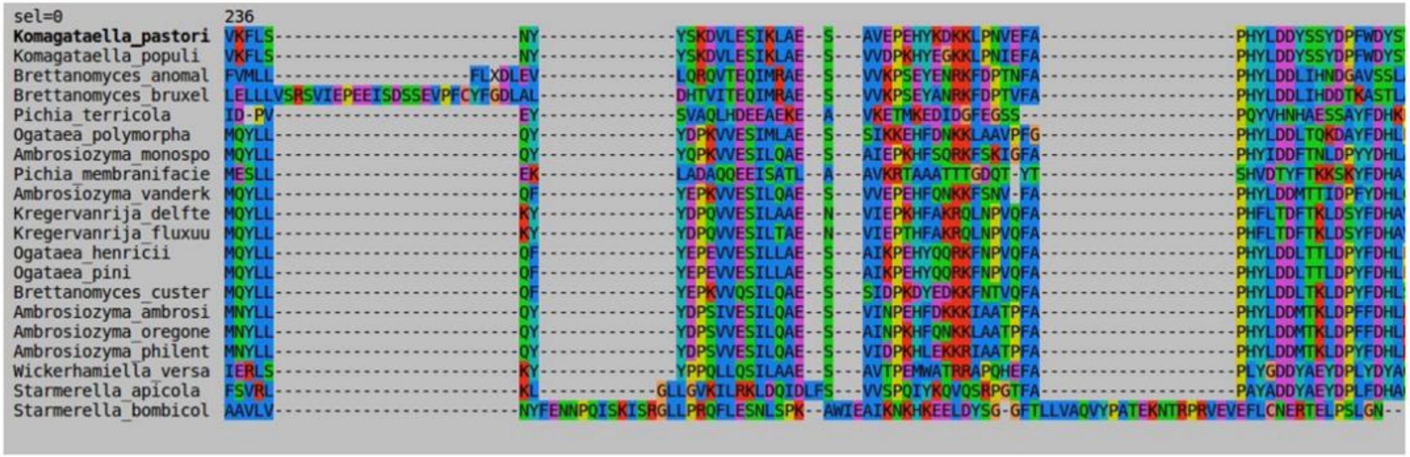
http://www.jalview.org/

GBlocks: Automatic

Conserved



Noisy

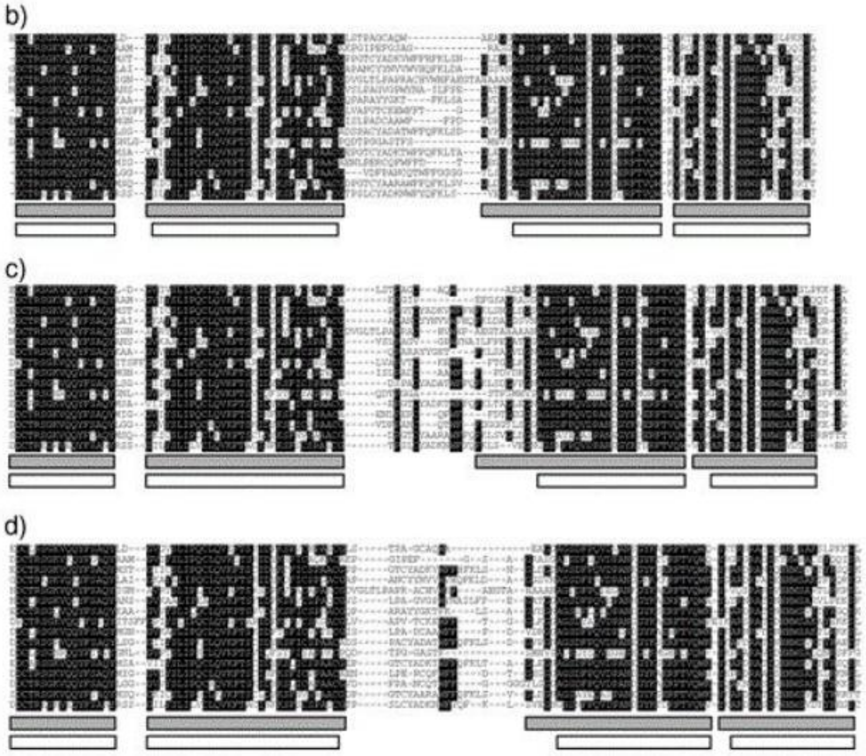


http://molevol.cmima.csic.es/castresana/Gblocks.html

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

The first one: GBLOCKS (http://molevol.cmima.csic.es/castresana/Gblocks.html)

Gblocks selects blocks in a similar way as it is usually done by hand but following a reproducible set of conditions. The selected blocks must fulfill certain requirements with respect to the lack of large segments of contiguous nonconserved positions, lack of gap positions and high conservation of flanking positions, making the final alignment more suitable for phylogenetic analysis.



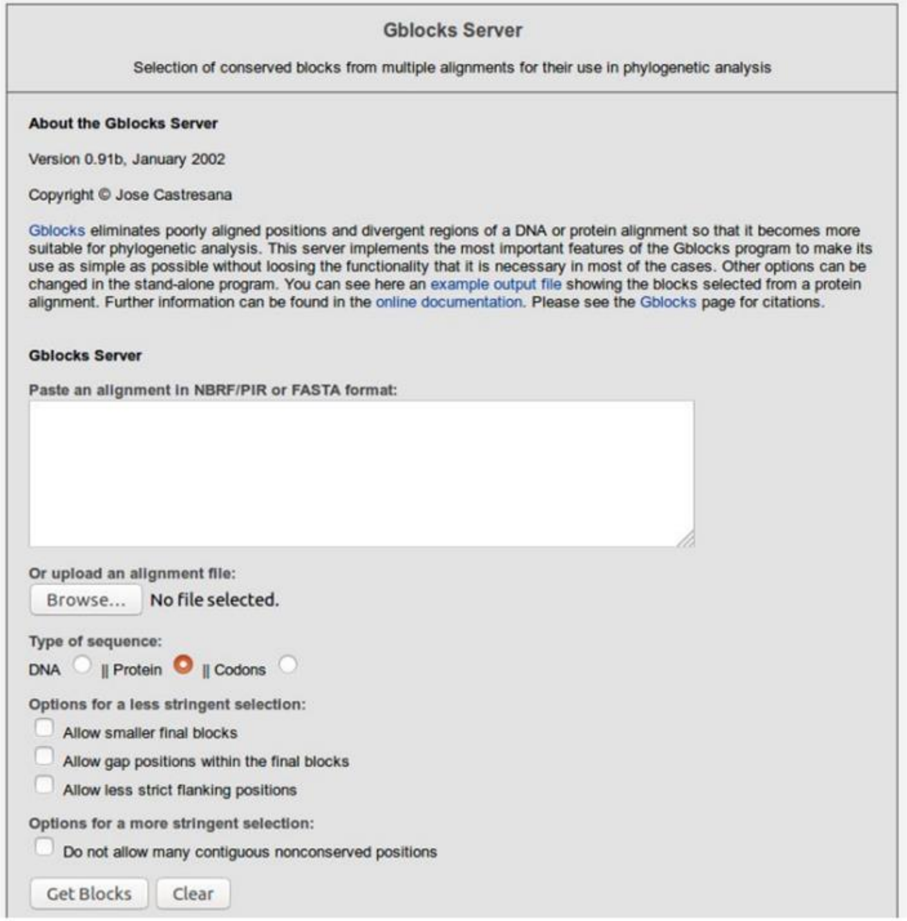
The white and grey blocks under the alignments represent the parts of the alignment that Gblocks would keep using a more relaxed and a more stringent approach.

http://molevol.cmima.csic.es/castresana/Gblocks.html

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

GBlocks: Automatic

How to run Gblocks: website



Gblocks Server

Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis

About the Gblocks Server

Version 0.91b, January 2002

Copyright © Jose Castresana

Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment so that it becomes more suitable for phylogenetic analysis. This server implements the most important features of the Gblocks program to make its use as simple as possible without loosing the functionality that it is necessary in most of the cases. Other options can be changed in the stand-alone program. You can see here an example output file showing the blocks selected from a protein alignment. Further information can be found in the online documentation. Please see the Gblocks page for citations.

Gblocks Server

Paste an alignment in NBRF/PIR or FASTA format:

Or upload an alignment file:

No file selected.

Type of sequence:

DNA ○ || Protein ● || Codons ○

Options for a less stringent selection:

□ Allow smaller final blocks

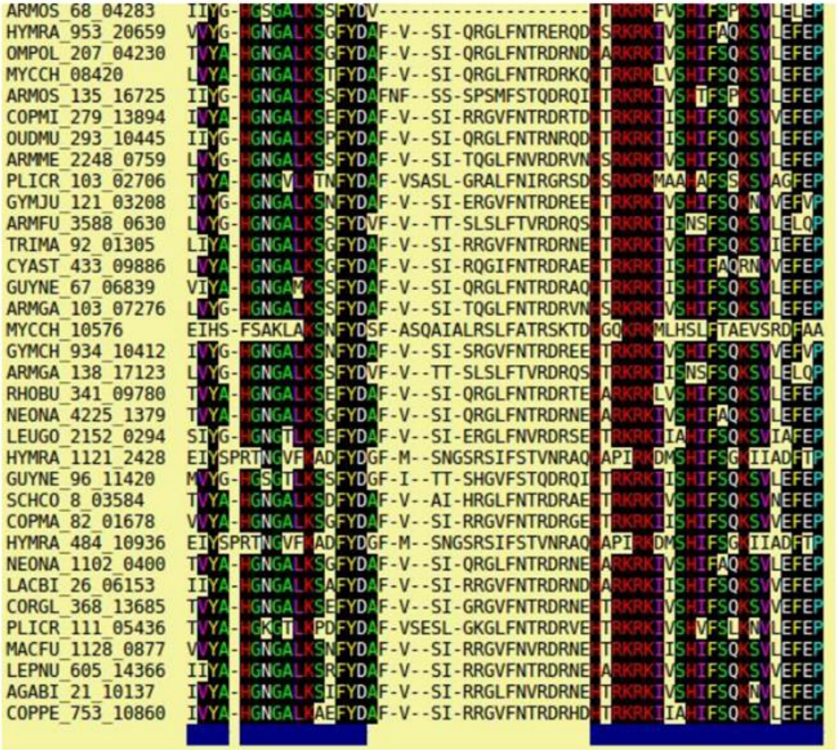
□ Allow gap positions within the final blocks

□ Allow less strict flanking positions

Options for a more stringent selection:

□ Do not allow many contiguous nonconserved positions

The is an on-line server that you can use if you only want to trim one alignment.



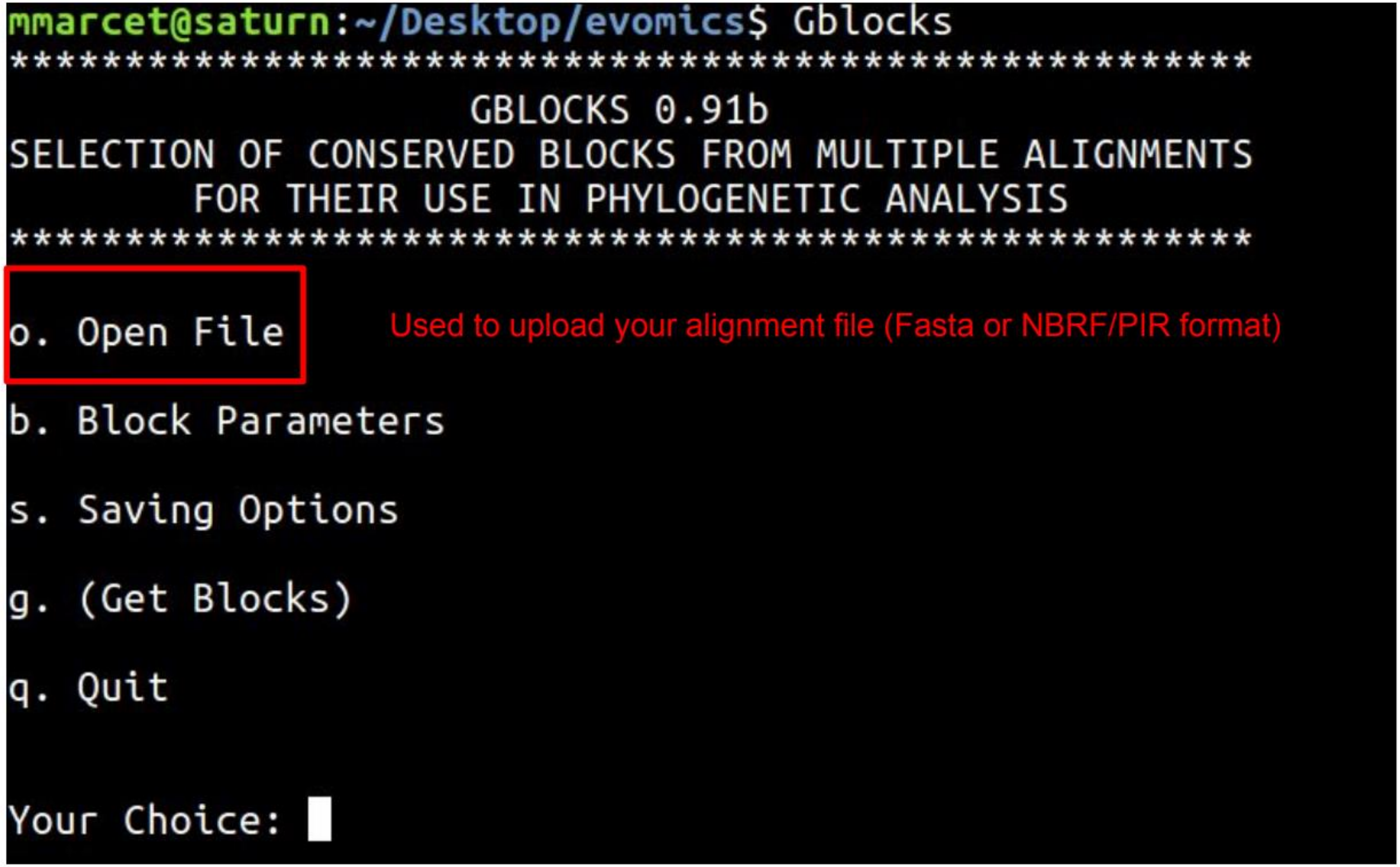
```
ARMOS_68_04283   IIYG-HGSGALKSSFYDV--------------------HTRKRKFVSHIFSPKSVLELEP
HYMRA_953_20659  VVYG-HGNGALKSGFYDAF-V--SI-QRGLFNTRERQDHSRKRKIVSHIFAQKSVLEFEP
OMPOL_207_04230  TVYA-HGNGALKSGFYDAF-V--SI-QRGLFNTRDRNDHARKRKIVSHIFSQKSVLEFEP
MYCCH_08420      LVYA-HGNGALKSTFYDAF-V--SI-QRGLFNTRDRKQHTRKRKLVSHIFSQKSVLEFEP
ARMOS_135_16725  IIYG-HGNGALKSSFYDAFNF--SS-SPSMFSTQDRQIHTRKRKIVSHTFSPKSVLEFEP
COPMI_279_13894  IVYA-HGNGALKSEFYDAF-V--SI-RRGVFNTRDRTDHTRKRKIISHIFSQKSVVEFEP
OUDMU_293_10445  IIYG-HGNGALKSPFYDAF-V--SI-QRGLFNTRNRQDHTRKRKIISHIFSQKSVLEFEP
ARMME_2248_0759  LVYG-HGNGALKSSFYDAF-V--SI-TQGLFNVRDRVNHSRKRKIVSHIFSQKSVLEFEP
PLICR_103_02706  TVYA-HGNGVLKTNFYDAF-VSASL-GRALFNIRGRSDHSRKRKMAAHAFSSKSVAGFEP
GYMJU_121_03208  IVYG-HGNGALKSNFYDAF-V--SI-ERGVFNTRDREEHTRKRKIVSHIFSQKNVVEFVP
ARMFU_3588_0630  LVYG-HGNGALKSSFYDVF-V--TT-SLSLFTVRDRQSHTRKRKIISNSFSQKSVLELQP
TRIMA_92_01305   LIYA-HGNGALKSGFYDAF-V--SI-RRGVFNTRDRNEHTRKRKIVSHIFSQKSVIEFEP
CYAST_433_09886  LVYA-HGNGALKSGFYDAF-V--SI-RQGIFNTRDRAEHTRKRKIISHIFAQRNVVEFEP
GUYNE_67_06839   VIYA-HGNGAMKSSFYDAF-V--SI-QRGLFNTRDRAQHTRKRKIISHIFSQKSVLEFEP
ARMGA_103_07276  LVYG-HGNGALKSSFYDAF-V--SI-TQGLFNTRDRVNHSRKRKIVSHIFSQKSVLEFEP
MYCCH_10576      EIHS-FSAKLAKSNFYDSF-ASQAIALRSLFATRSKTDHGQKRKMLHSLFTAEVSRDFAA
GYMCH_934_10412  IVYG-HGNGALKSNFYDAF-V--SI-SRGVFNTRDREEHTRKRKIVSHIFSQKSVVEFVP
ARMGA_138_17123  LVYG-HGNGALKSSFYDVF-V--TT-SLSLFTVRDRQSHTRKRKIISNSFSQKSVLELQP
RHOBU_341_09780  TVYA-HGNGALKSEFYDAF-V--SI-QRGLFNTRDRTEHARKRKLVSHIFSQKSVLEFEP
NEONA_4225_1379  TVYA-HGNGALKSGFYDAF-V--SI-QRGLFNTRDRNEHARKRKIVSHIFAQKSVLEFEP
LEUGO_2152_0294  SIYG-HGNGTLKSEFYDAF-V--SI-ERGLFNVRDRSEHTRKRKIIAHIFSQKSVIAFEP
HYMRA_1121_2428  EIYSPRTNGVFKADFYDGF-M--SNGSRSIFSTVNRAQHAPIRKDMSHIFSGKIIADFTP
GUYNE_96_11420   MVYG-HGSGTLKSSFYDGF-I--TT-SHGVFSTQDRQIHTRKRKIISHIFSQKSVLEFEP
SCHCO_8_03584    TVYA-HGNGALKSDFYDAF-V--AI-HRGLFNTRDRAEHTRKRKIVSHIFSQKSVNEFEP
COPMA_82_01678   VVYA-HGNGALKSGFYDAF-V--SI-RRGVFNTRDRGEHTRKRKIISHIFSQKSVVEFEP
HYMRA_484_10936  EIYSPRTNGVFKADFYDGF-M--SNGSRSIFSTVNRAQHAPIRKDMSHIFSGKIIADFTP
NEONA_1102_0400  TVYA-HGNGALKSGFYDAF-V--SI-QRGLFNTRDRNEHARKRKIVSHIFAQKSVLEFEP
LACBI_26_06153   IIYA-HGNGALKSAFYDAF-V--SI-RRGVFNTRDRNDHARKRKIISHIFSQKSVVEFEP
CORGL_368_13685  TVYA-HGNGALKSEFYDAF-V--SI-GRGVFNTRDRNEHTRKRKIVSHIFSQKSVVEFEP
PLICR_111_05436  TVYA-HGKGTLKPDFYDAF-VSESL-GKGLFNTRDRVEHTRKRKIVSHVFSLKNVLEFEP
MACFU_1128_0877  VVYA-HGNGALKSNFYDAF-V--SI-RRGVFNVRDRNEHTRKRKIISHIFSQKSVLEFEP
LEPNU_605_14366  IIYA-HGNGALKSRFYDAF-V--SI-RRGVFNTRDRNEHARKRKIVSHIFSQKSVVEFEP
AGABI_21_10137   IVYA-HGNGALKSIFYDAF-V--SI-RRGLFNVRDRNEHTRKRKIVSHIFSQKNVLEFEP
COPPE_753_10860  IVYA-HGNGALKAEFYDAF-V--SI-RRGVFNTRDRHDHTRKRKIIAHIFSQKSVLEFEP
```

At the end of the alignment representation there's a link to obtain the trimmed alignment.

http://molevol.cmima.csic.es/castresana/Gblocks.html

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

How to run Gblocks: command line



```
mmarcet@saturn:~/Desktop/evomics$ Gblocks
******************************************************
                    GBLOCKS 0.91b
SELECTION OF CONSERVED BLOCKS FROM MULTIPLE ALIGNMENTS
        FOR THEIR USE IN PHYLOGENETIC ANALYSIS
******************************************************

o. Open File

b. Block Parameters

s. Saving Options

g. (Get Blocks)

q. Quit


Your Choice: 
```

Used to upload your alignment file (Fasta or NBRF/PIR format)

http://molevol.cmima.csic.es/castresana/Gblocks.html

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

The probable truth: it depends on the dataset and the methodology used.

Program	Number of citations (Google scholar)
BMGE	381
trimAl	1737
Gblocks	5736

Trimming alignments tends to be part of a normal phylogenetic reconstruction pipeline.