H3ABioNet

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Multiple Sequence Alignment

Building Multiple Sequence Alignment

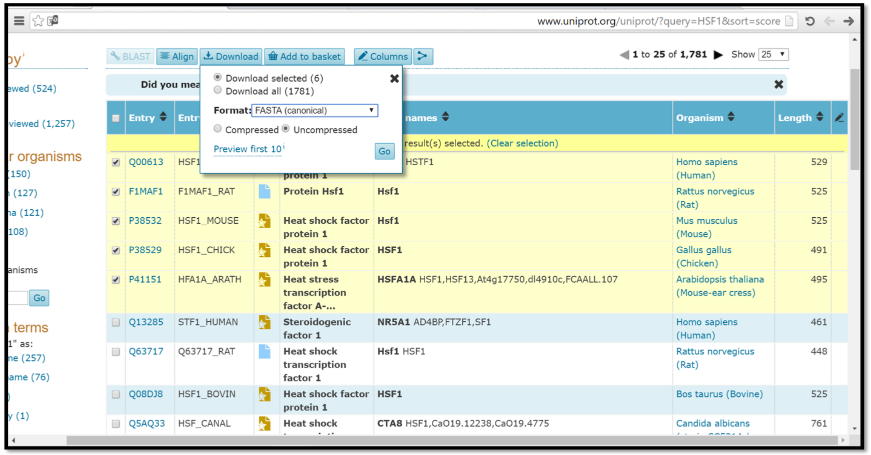
Lec2 Choosing the Right Sequences

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Choosing the Right Sequences

“Before you build your alignment, you must **carefully select the sequences** you want to align. These sequences are members of the **same protein family**, and they all **share a common ancestor**. The family is usually **too large** to be entirely included in your multiple alignment, and **picking the right sequences is an art**.”



◀ 1 to 25 of 1,781 ▶ Show 25 ▾

Did you mea

○ Download selected (6)
○ Download all (1781)

Format: FASTA (canonical) ▾

○ Compressed ● Uncompressed

Preview first 10

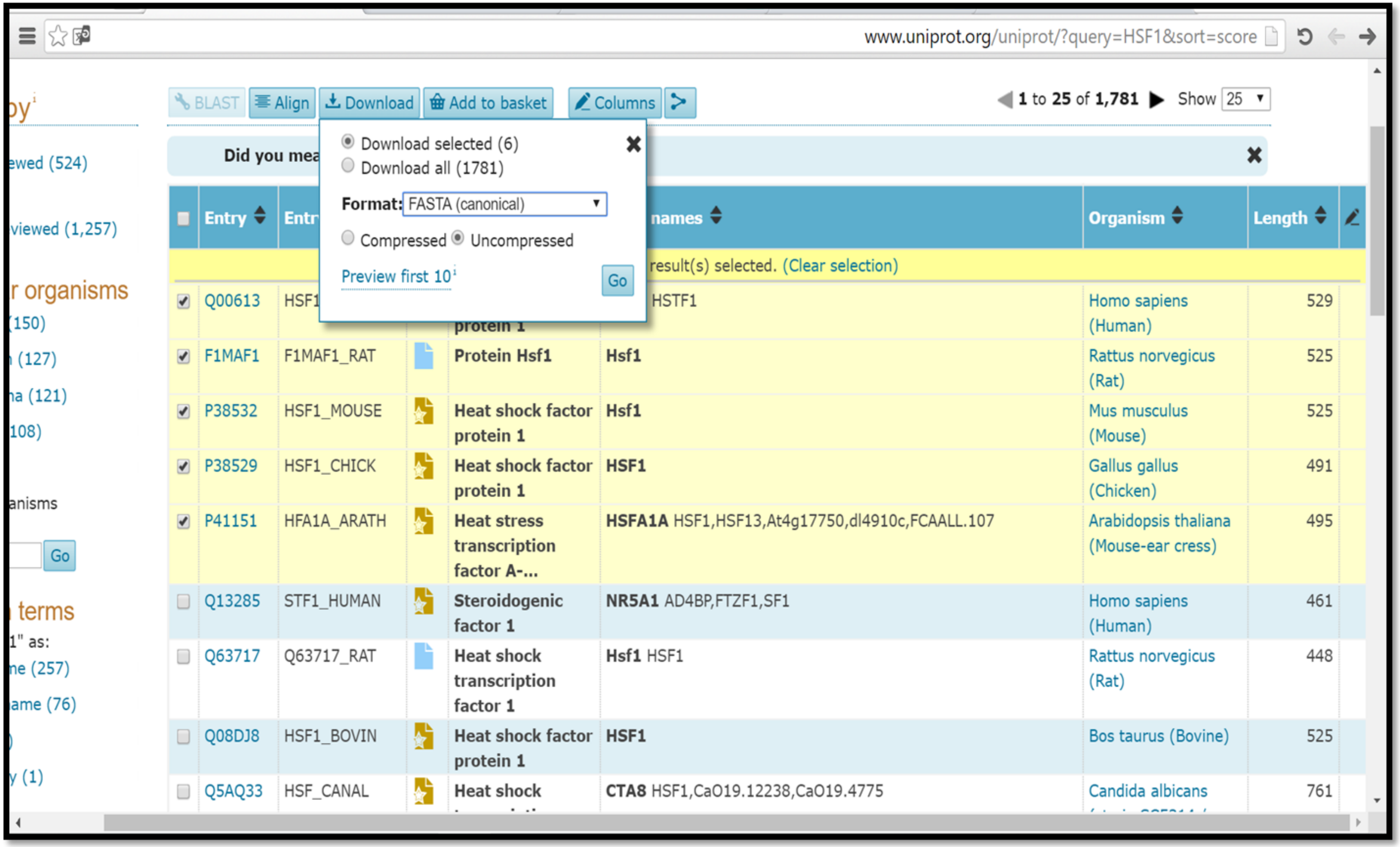
Go

result(s) selected. (Clear selection)

☐	Entry	Entr			names	Organism	Length
☑	Q00613	HSF1		protein 1	HSTF1	Homo sapiens (Human)	529
☑	F1MAF1	F1MAF1_RAT		Protein Hsf1	Hsf1	Rattus norvegicus (Rat)	525
☑	P38532	HSF1_MOUSE		Heat shock factor protein 1	Hsf1	Mus musculus (Mouse)	525
☑	P38529	HSF1_CHICK		Heat shock factor protein 1	HSF1	Gallus gallus (Chicken)	491
☑	P41151	HFA1A_ARATH		Heat stress transcription factor A-...	**HSFA1A** HSF1,HSF13,At4g17750,dl4910c,FCAALL.107	Arabidopsis thaliana (Mouse-ear cress)	495
☐	Q13285	STF1_HUMAN		Steroidogenic factor 1	**NR5A1** AD4BP,FTZF1,SF1	Homo sapiens (Human)	461
☐	Q63717	Q63717_RAT		Heat shock transcription factor 1	**Hsf1** HSF1	Rattus norvegicus (Rat)	448
☐	Q08DJ8	HSF1_BOVIN		Heat shock factor protein 1	HSF1	Bos taurus (Bovine)	525
☐	Q5AQ33	HSF_CANAL		Heat shock	**CTA8** HSF1,CaO19.12238,CaO19.4775	Candida albicans	761

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Retrieve Sequences from Uniprot (www.uniprot.org



○ Download selected (6)
○ Download all (1781)

Format: FASTA (canonical)

○ Compressed ● Uncompressed

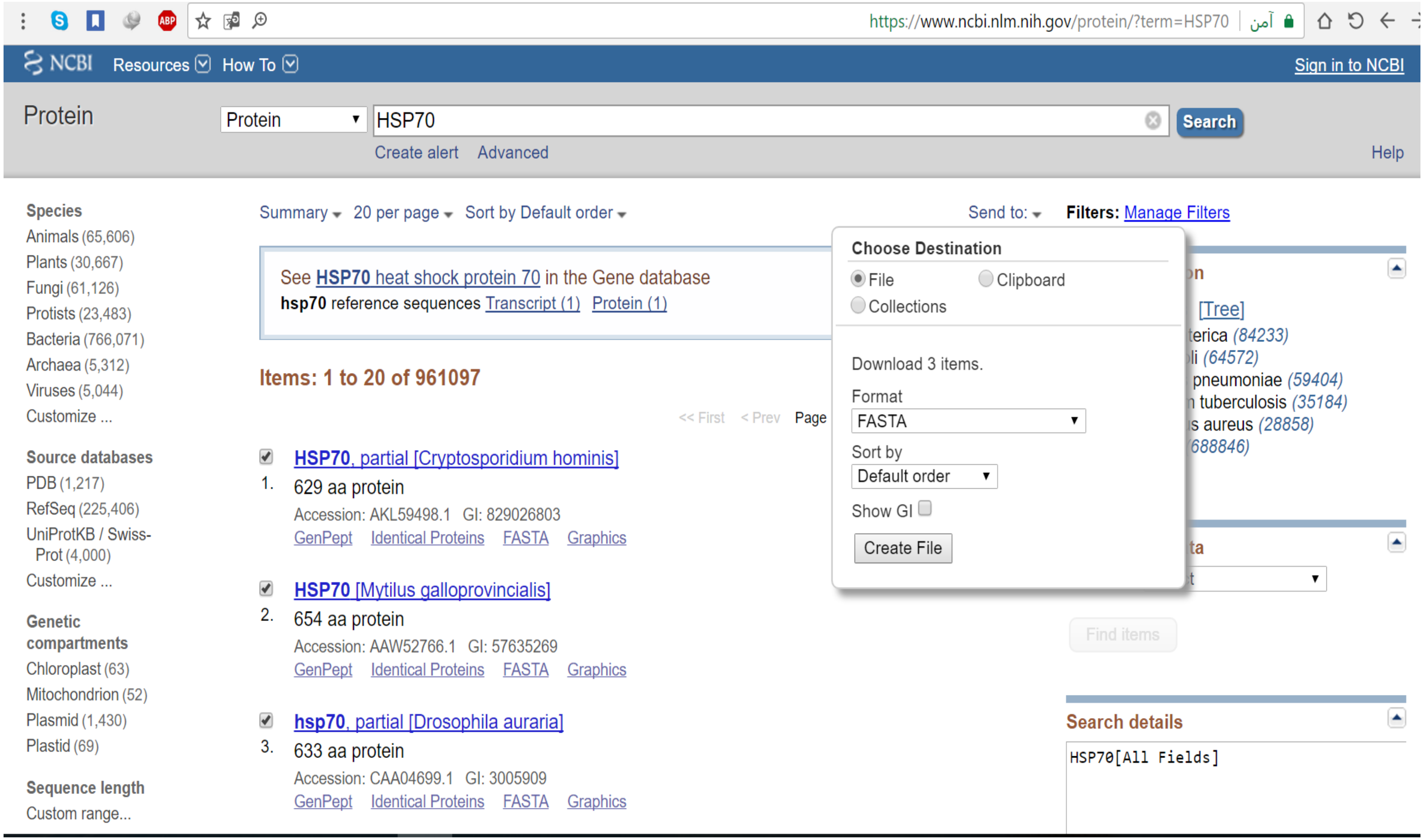
Preview first 10

Go

result(s) selected. (Clear selection)

☐	Entry	Entr			names	Organism	Length
☑	Q00613	HSF1		protein 1	HSTF1	Homo sapiens (Human)	529
☑	F1MAF1	F1MAF1_RAT	📄	Protein Hsf1	**Hsf1**	Rattus norvegicus (Rat)	525
☑	P38532	HSF1_MOUSE	★	Heat shock factor protein 1	**Hsf1**	Mus musculus (Mouse)	525
☑	P38529	HSF1_CHICK	★	Heat shock factor protein 1	**HSF1**	Gallus gallus (Chicken)	491
☑	P41151	HFA1A_ARATH	★	Heat stress transcription factor A-...	**HSFA1A** HSF1,HSF13,At4g17750,dl4910c,FCAALL.107	Arabidopsis thaliana (Mouse-ear cress)	495
☐	Q13285	STF1_HUMAN	★	Steroidogenic factor 1	**NR5A1** AD4BP,FTZF1,SF1	Homo sapiens (Human)	461
☐	Q63717	Q63717_RAT	📄	Heat shock transcription factor 1	**Hsf1** HSF1	Rattus norvegicus (Rat)	448
☐	Q08DJ8	HSF1_BOVIN	★	Heat shock factor protein 1	**HSF1**	Bos taurus (Bovine)	525
☐	Q5AQ33	HSF_CANAL	★	Heat shock	**CTA8** HSF1,CaO19.12238,CaO19.4775	Candida albicans	761

Retrieve Sequences from NCBI (https://www.ncbi.nlm.nih.gov/protein)



https://www.ncbi.nlm.nih.gov/protein/?term=HSP70 آمن

NCBI Resources How To Sign in to NCBI

Protein | Protein ▾ | HSP70 | Search

Create alert Advanced Help

Species
- Animals (65,606)
- Plants (30,667)
- Fungi (61,126)
- Protists (23,483)
- Bacteria (766,071)
- Archaea (5,312)
- Viruses (5,044)
- Customize ...

Source databases
- PDB (1,217)
- RefSeq (225,406)
- UniProtKB / Swiss-Prot (4,000)
- Customize ...

Genetic compartments
- Chloroplast (63)
- Mitochondrion (52)
- Plasmid (1,430)
- Plastid (69)

Sequence length
- Custom range...

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: Manage Filters

See HSP70 heat shock protein 70 in the Gene database
hsp70 reference sequences Transcript (1) Protein (1)

Items: 1 to 20 of 961097

<< First < Prev Page

☑ 1. HSP70, partial [Cryptosporidium hominis]
629 aa protein
Accession: AKL59498.1 GI: 829026803
GenPept Identical Proteins FASTA Graphics

☑ 2. HSP70 [Mytilus galloprovincialis]
654 aa protein
Accession: AAW52766.1 GI: 57635269
GenPept Identical Proteins FASTA Graphics

☑ 3. hsp70, partial [Drosophila auraria]
633 aa protein
Accession: CAA04699.1 GI: 3005909
GenPept Identical Proteins FASTA Graphics

Choose Destination
● File ○ Clipboard
○ Collections

Download 3 items.

Format: FASTA ▾

Sort by: Default order ▾

Show GI □

Create File

on
[Tree]
terica (84233)
li (64572)
pneumoniae (59404)
n tuberculosis (35184)
s aureus (28858)
688846)

ta
t ▾
Find items

Search details
HSP70[All Fields]

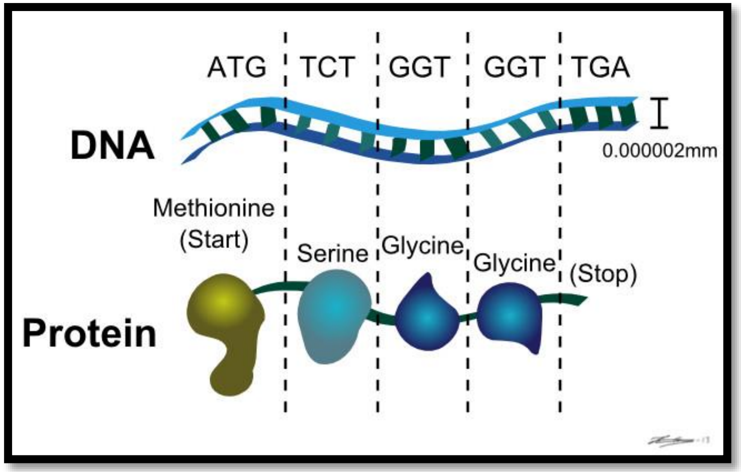
H3ABioNet
Pan African Bioinformatics Network for H3Africa

A Few Guidelines for Selecting Sequences

Proteins or DNA

Use **proteins whenever possible**. You can turn them back into DNA after doing the multiple alignment.

If the sequences are **non-coding sequences**, you must **use DNA**

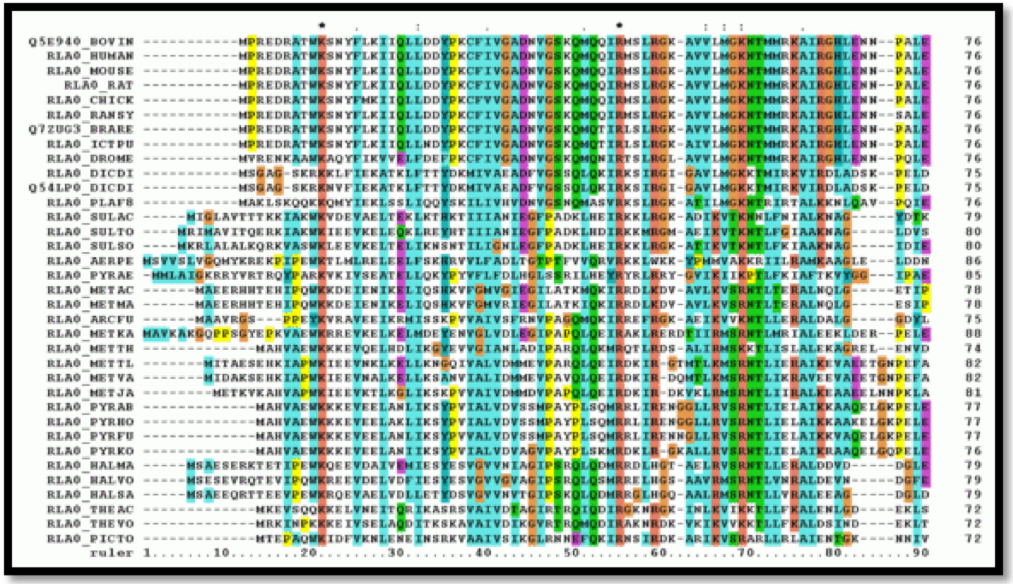


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

Many sequences

Start with 10–15 sequences; **avoid** aligning **more than 50** sequences.



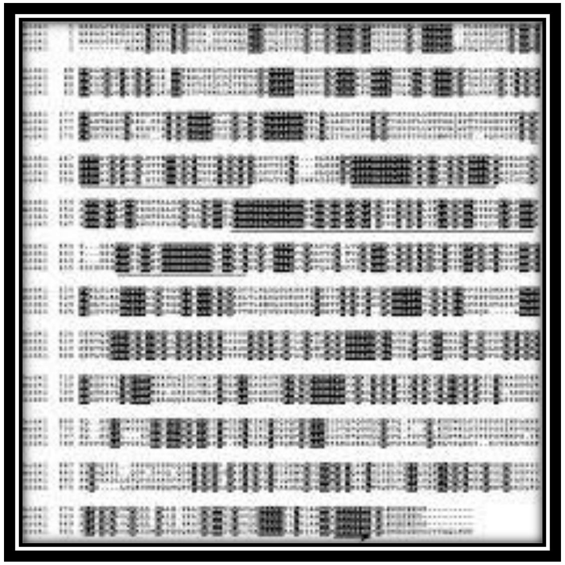
```
                                *   .       :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE  76
   RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE  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 -----MIGLAVTTTKKIAKMKYDEVAELTEKLKTRKTIIIAHIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK  79
   RLA0_SULTO ----MRIMAVITQERKIAKMKIEEVKELEGKLREYHTIIIAHIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS  80
   RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE  80
   RLA0_AERPE MSVVSLVGQHYKREKPIPEWKTLMLRELEELFSKNRVVLFADLTGTPTFVVQRVRKKLWKK-YPHMVAKKRIIILRAMKAAGLE---LDDN  86
   RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKYKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-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 -------MIDAKSEHKIAPWKIEEVNALKSLLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA  82
   RLA0_METJA --------METKYKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-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
```

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

Very different sequences

Sequences that are **less than 30 percent** identical to more than half the other sequences in the set often **cause troubles**.

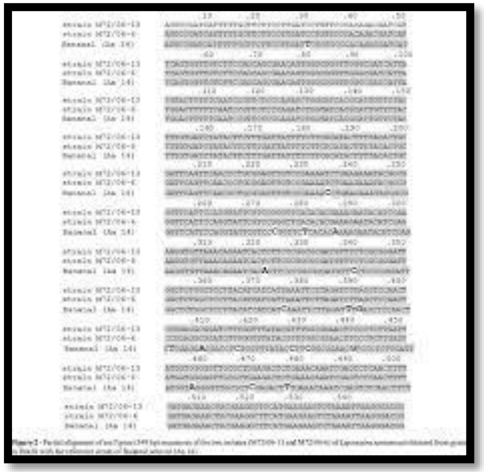


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing*, *Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

Identical sequences

They **never help**. Unless you have a very good reason to do so, **avoid incorporating** into your multiple alignment any sequence that’s more than **90 percent identical** to another sequence in the set.

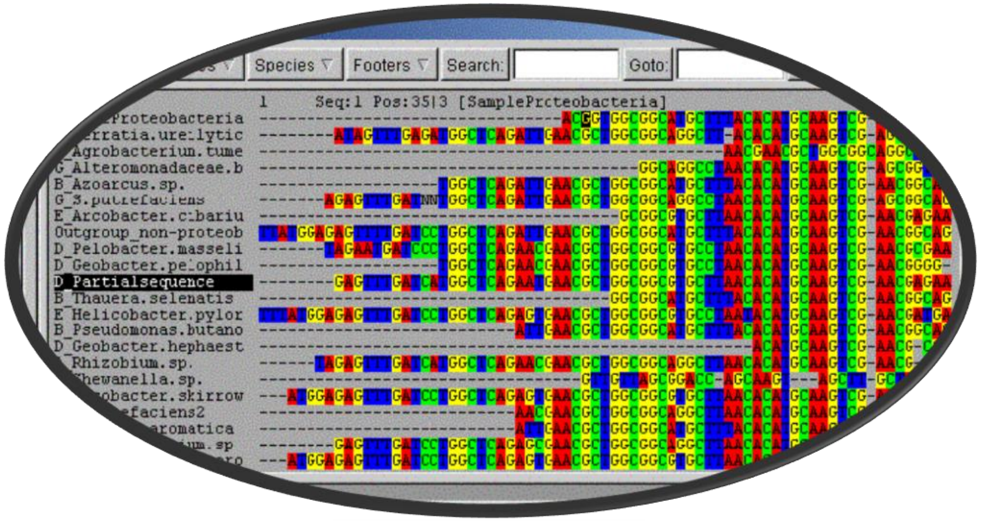


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

Partial sequences

Multiple-sequence-alignment programs prefer sequences that are roughly the same length. Programs often have **difficulties comparing** items in a **mixture of complete sequences** and **shorter fragments**.

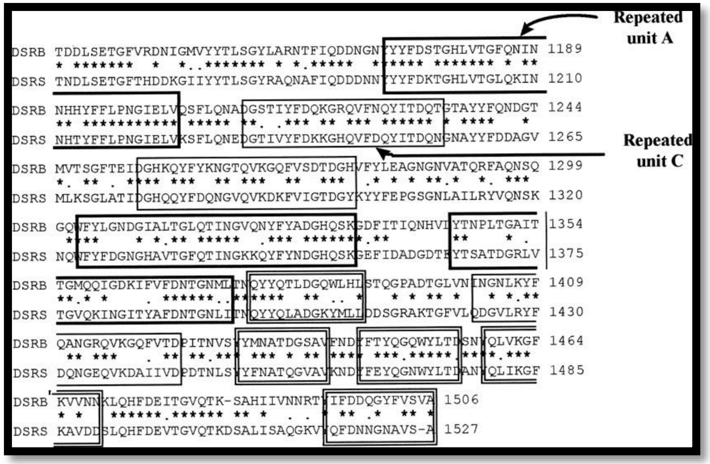


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing*, *Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

Repeated domains

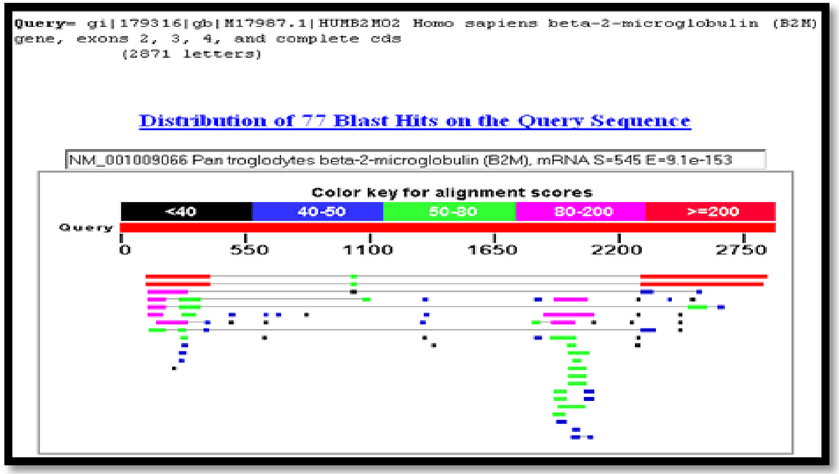
Sequences with repeated domains **cause trouble** for most multiple-alignment programs — especially if the **number of domains is different**. When this happens, you may be better off extracting the domains yourself with Dotlet or Lalign and making a multiple alignment of those segments.



```
DSRB TDDLSETGFVRDNIGMVYYTLSGYLARNTFIQDDNGNYYYFDSTGHLVTGFQNIN 1189
     * *******  *  *..******* * *.*****  ****** ******* * **
DSRS TNDLSETGFTHDDKGIIYYTLSGYRAQNAFIQDDDNNYYYFDKTGHLVTGLQKIN 1210

DSRB NHHYFFLPNGIELVQSFLQNADGSTIYFDQKGRQVFNQYITDQTGTAYYFQNDGT 1244
     ** *********** ***** **. .*** ** *** ****** * ****    *
DSRS NHTYFFLPNGIELVKSFLQNEDGTIVYFDKKGHQVFDQYITDQNGNAYYFDDAGV 1265

DSRB MVTSGFTEIDGHKQYFYKNGTQVKGQFVSDTDGHVFYLEAGNGNVATQRFAQNSQ 1299
     *. ** . **** ***  ** ***  **  ***  .* * * **.*  *. ***
DSRS MLKSGLATIDGHQQYFDQNGVQVKDKFVIGTDGYKYYFEPGSGNLAILRYVQNSK 1320

DSRB GQWFYLGNDGIALTGLQTINGVQNYFYADGHQSKGDFITIQNHVIYTNPLTGAIT 1354
      ****    * *.** *****  .*** ********.**      **    * .
DSRS NQWFYFDGNGHAVTGFQTINGKKQYFYNDGHQSKGEFIDADGDTEYTSATDGRLV 1375

DSRB TGMQQIGDKIFVFDNTGNMLTNQYYQTLDGQWLHLSTQGPADTGLVNINGNLKYF 1409
     **.* *      . ******..******  **  .. *    * * ** *   * *.**
DSRS TGVQKINGITYAFDNTGNLITNQYYQLADGKYMLLDDSGRAKTGFVLQDGVLRYF 1430

DSRB QANGRQVKGQFVTDPITNVSYYMNATDGSAVFNDYFTYQGQWYLTDSNYQLVKGF 1464
      ** ***    . ** **.*** *** * ** **** ***.*****.*****.***
DSRS DQNGEQVKDAIIVDPDTNLSYYFNATQGVAVKNDYFEYQGNWYLTDANYQLIKGF 1485

DSRB KVVNNKLQHFDEITGVQTK-SAHIIVNNRTYIFDDQGYFVSVA 1506
     * *   ******.****** ** *   . . * ** .*  ** *
DSRS KAVDDSLQHFDEVTGVQTKDSALISAQGKVYQFDNNGNAVS-A 1527
```

Repeated unit A

Repeated unit C



Query= gi|179316|gb|M17987.1|HUMB2M02 Homo sapiens beta-2-microglobulin (B2M) gene, exons 2, 3, 4, and complete cds
(2871 letters)

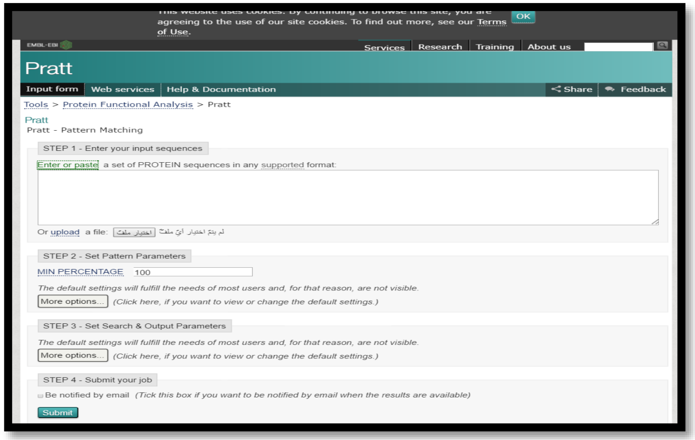
Distribution of 77 Blast Hits on the Query Sequence

NM_001009066 Pan troglodytes beta-2-microglobulin (B2M), mRNA S=545 E=9.1e-153

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing, Inc*. 436 pp.

A Few Guidelines for Selecting Sequences

If you still **cannot generate a proper alignment** from sequences that you know are related, you could **use a local** multiple alignment method, such as the **Gibbs sampler**, or a pattern extraction motif, such as **Pratt**.



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use. OK

Pratt

Tools > Protein Functional Analysis > Pratt

Pratt

Pratt - Pattern Matching

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

Or upload a file: اختيار ملف لم يتمّ اختيار أيّ ملفّ

STEP 2 - Set Pattern Parameters

MIN PERCENTAGE 100

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 3 - Set Search & Output Parameters

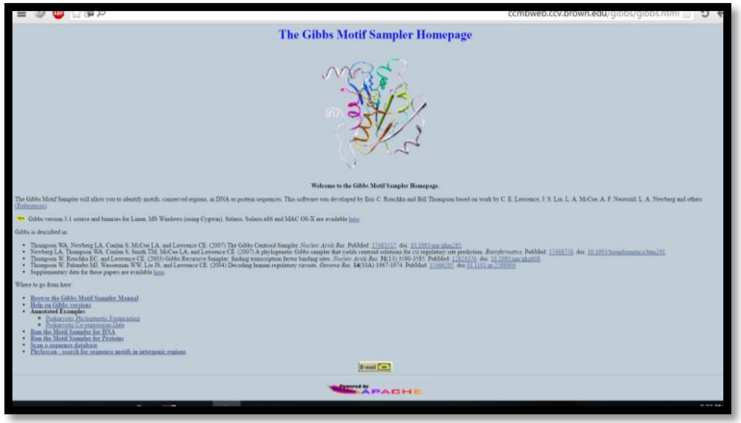
The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 4 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit



The Gibbs Motif Sampler Homepage

Welcome to the Gibbs Motif Sampler Homepage.

The Gibbs Motif Sampler will allow you to identify motifs, conserved regions, in DNA or protein sequences. This software was developed by Eric C. Rouchka and Bill Thompson based on work by C. E. Lawrence, J. S. Liu, L. A. McCue, A. F. Neuwald, L. A. Newberg and others (References)

Gibbs version 3.1 source and binaries for Linux, MS Windows (using Cygwin), Solaris, Solaris x86 and MAC OS-X are available here

Gibbs is described in:

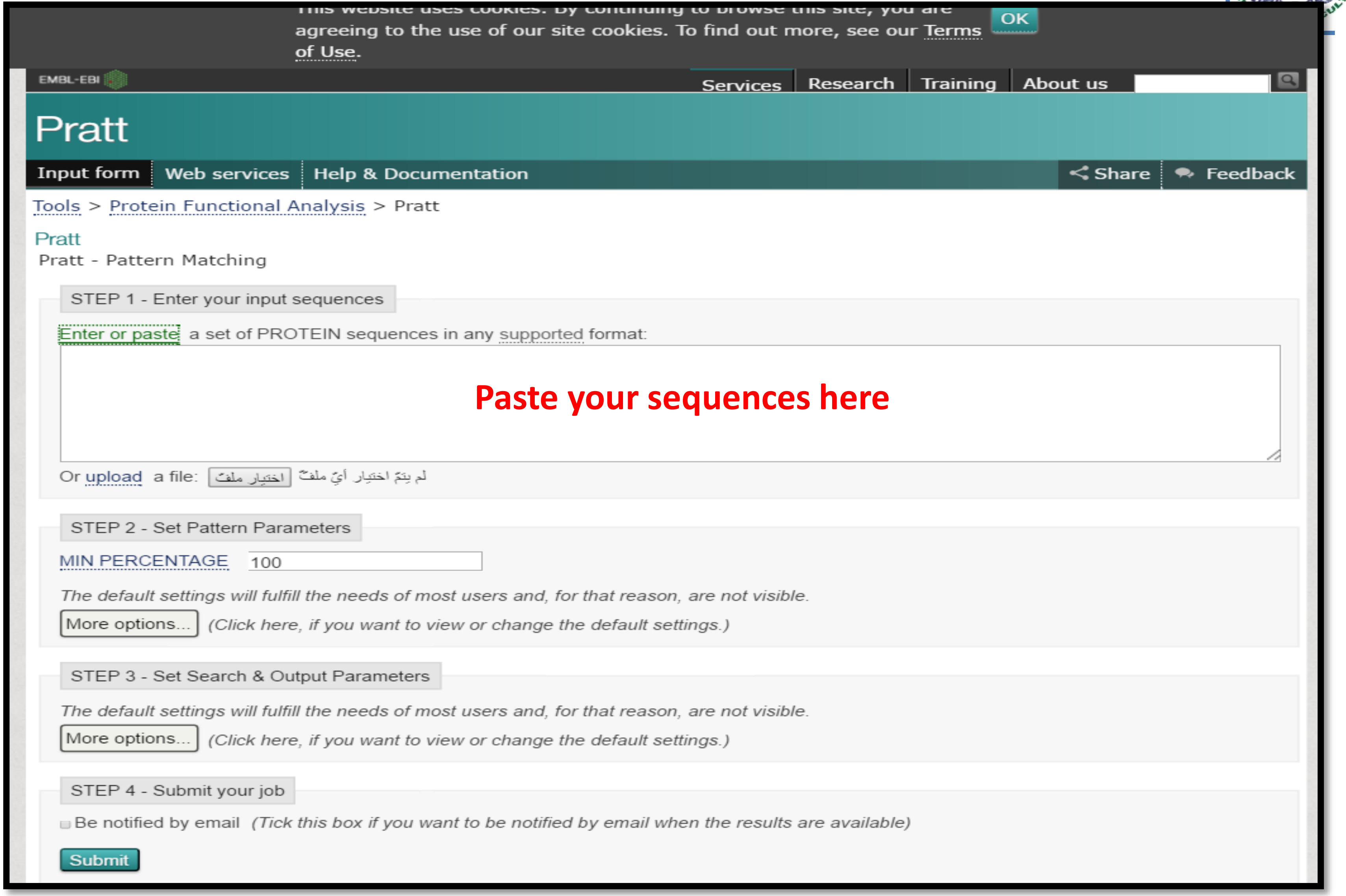
- Thompson WA, Newberg LA, Conlan S, McCue LA, and Lawrence CE. (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res.* PubMed: 17483517. doi: 10.1093/nar/gkm265
- Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, and Lawrence CE. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis* regulatory site prediction. *Bioinformatics.* PubMed: 17488758. doi: 10.1093/bioinformatics/btm241
- Thompson W, Rouchka EC, and Lawrence CE. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* **31**(13):3580-3585. PubMed: 12824370. doi: 10.1093/nar/gkg608
- Thompson W, Palumbo MJ, Wasserman WW, Liu JS, and Lawrence CE. (2004) Decoding human regulatory circuits. *Genome Res.* **14**(10A):1967-1974. PubMed: 15466295. doi: 10.1101/gr.2589004
- Supplementary data for these papers are available here

Where to go from here:

- Browse the Gibbs Motif Sampler Manual
- Help on Gibbs versions
- **Annotated Examples**
  - Prokaryotic Phylogenetic Footprinting
  - Prokaryotic Co-expression Data
- Run the Motif Sampler for DNA
- Run the Motif Sampler for Proteins
- Scan a sequence database
- Phyloscan - search for sequence motifs in intergenic regions

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

H3ABioNet
Pan African Bioinformatics Network for H3Africa



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

Pratt

Tools > Protein Functional Analysis > Pratt

Pratt

Pratt - Pattern Matching

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

Paste your sequences here

Or upload a file: اختيار ملف لم يتمّ اختيار أيّ ملف

STEP 2 - Set Pattern Parameters

MIN PERCENTAGE 100

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 3 - Set Search & Output Parameters

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 4 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

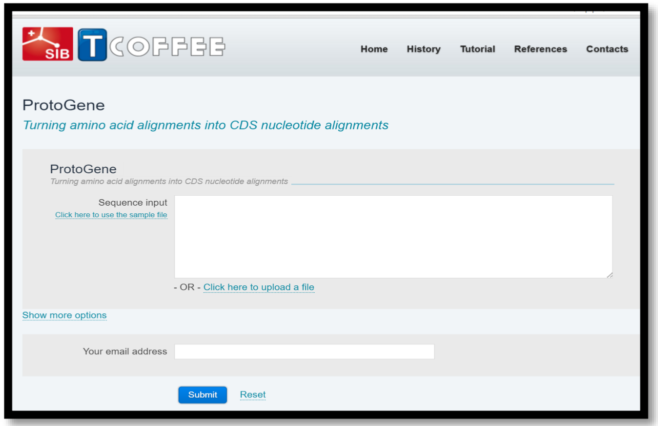
Bioinformatics Training Unit

For carrying out a **phylogenetic analysis** on a set of **coding DNA** sequences, do the following:

1. **Translate your DNA sequences into proteins.**

2. **Perform the multiple alignments on the proteins.**

3. **Thread the DNA back onto the protein multiple sequence alignment framework using pal2nal (coot.embl.de/pal2nal) or Protogene if you do not have the original DNA sequence (www.tcoffee.org).**



ProtoGene

Turning amino acid alignments into CDS nucleotide alignments

ProtoGene

Turning amino acid alignments into CDS nucleotide alignments

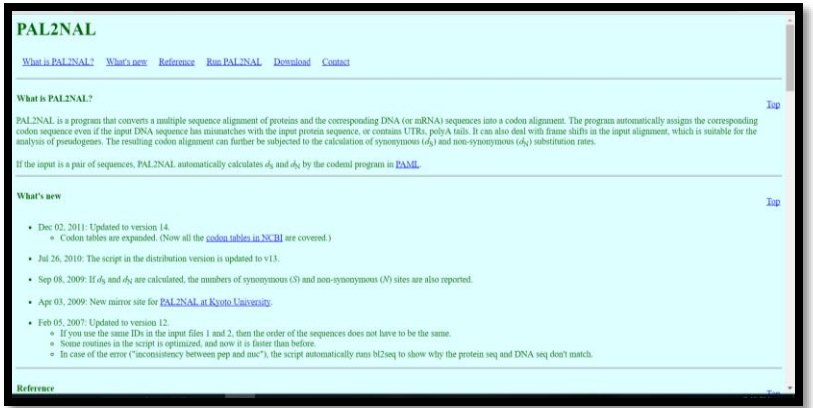
Sequence input

Click here to use the sample file

- OR - Click here to upload a file

Show more options

Your email address



PAL2NAL

What is PAL2NAL?

PAL2NAL is a program that converts a multiple sequence alignment of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment. The program automatically assigns the corresponding codon sequence even if the input DNA sequence has mismatches with the input protein sequence, or contains UTRs, polyA tails. It can also deal with frame shifts in the input alignment, which is suitable for the analysis of pseudogenes. The resulting codon alignment can further be subjected to the calculation of synonymous (d_S) and non-synonymous (d_N) substitution rates.

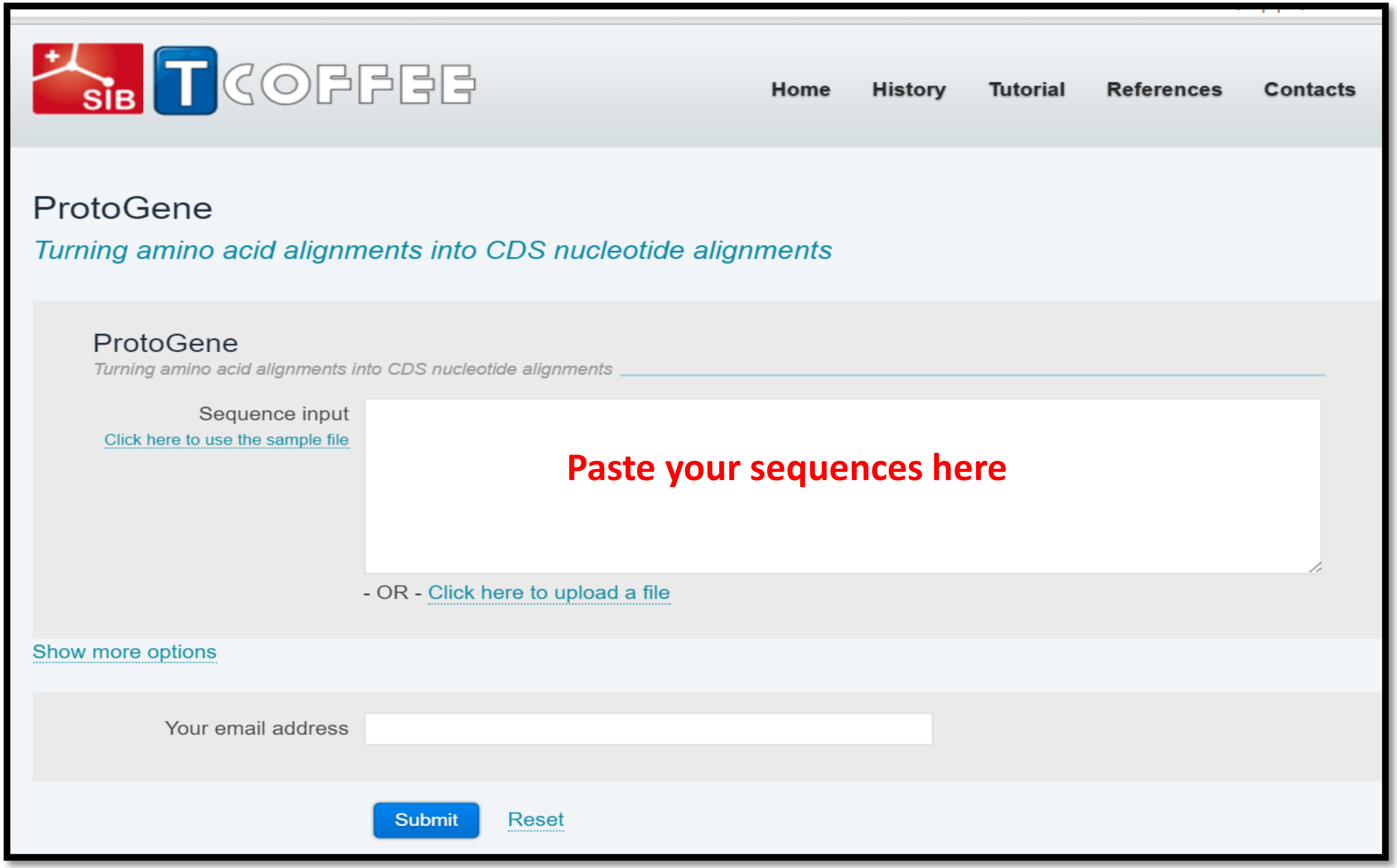
If the input is a pair of sequences, PAL2NAL automatically calculates d_S and d_N by the codeml program in PAML.

What's new

- Dec 02, 2011: Updated to version 14.
  - Codon tables are expanded. (Now all the codon tables in NCBI are covered.)
- Jul 26, 2010: The script in the distribution version is updated to v13.
- Sep 08, 2009: If d_S and d_N are calculated, the numbers of synonymous (*S*) and non-synonymous (*N*) sites are also reported.
- Apr 03, 2009: New mirror site for PAL2NAL at Kyoto University.
- Feb 05, 2007: Updated to version 12.
  - If you use the same IDs in the input files 1 and 2, then the order of the sequences does not have to be the same.
  - Some routines in the script is optimized, and now it is faster than before.
  - In case of the error ("inconsistency between pep and nuc"), the script automatically runs bl2seq to show why the protein seq and DNA seq don't match.

Reference

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing, Inc*. 436 pp.



ProtoGene

Turning amino acid alignments into CDS nucleotide alignments

ProtoGene

Turning amino acid alignments into CDS nucleotide alignments

Sequence input

Click here to use the sample file

Paste your sequences here

- OR - Click here to upload a file

Show more options

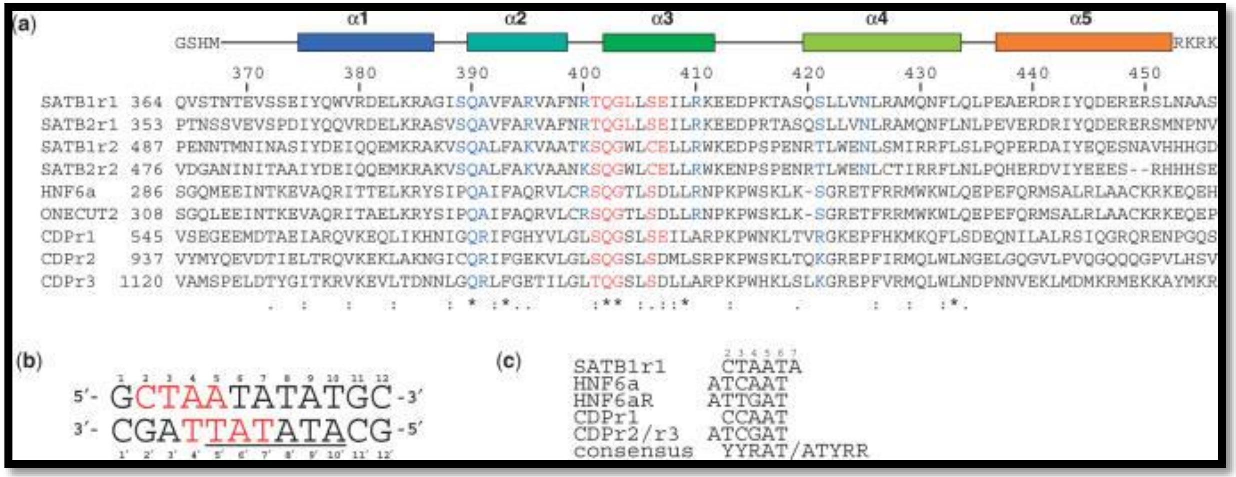
Your email address

Choosing the right number of sequences

1- you should start with a relatively small number of sequences — between 10 and 15 sequences would be suitable for most cases.

2- After you get something interesting happening with this small set, you can always increase its size.

3- it’s hard to see any reason for generating a multiple alignment with more than 50 sequences, unless you’re interested in building some extensive phylogenetic tree.



```
                  370       380       390       400       410       420       430       440       450
                    |         |         |         |         |         |         |         |         |
SATB1r1  364  QVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAAS
SATB2r1  353  PTNSSVEVSPDIYQQVRDELKRASVSQAVFARVAFNRTQGLLSEILRKEEDPRTASQSLLVNLRAMQNFLNLPEVERDRIYQDERERSMNPNV
SATB1r2  487  PENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGD
SATB2r2  476  VDGANINITAAIYDEIQQEMKRAKVSQALFAKVAANKSQGWLCELLRWKENPSPENRTLWENLCTIRRFLNLPQHERDVIYEEES--RHHHSE
HNF6a    286  SGQMEEINTKEVAQRITTELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEH
ONECUT2  308  SGQLEEINTKEVAQRITAELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEP
CDPr1    545  VSEGEEMDTAEIARQVKEQLIKHNIGQRIFGHYVLGLSQGSLSEILARPKPWNKLTVRGKEPFHKMKQFLSDEQNILALRSIQGRQRENPGQS
CDPr2    937  VYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSLSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
CDPr3   1120  VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSLSDLLARPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
                      .  :   :   :    : * :*..    :** :.::*   :     .      :  :  :*.
```

```
   1  2  3  4  5  6  7  8  9  10  11  12
5'- G  C  T  A  A  T  A  T  A  T   G   C -3'
3'- C  G  A  T  T  A  T  A  T  A   C   G -5'
   1' 2' 3' 4' 5' 6' 7' 8' 9' 10' 11' 12'
```

```
               234567
SATB1r1         CTAATA
HNF6a          ATCAAT
HNF6aR         ATTGAT
CDPr1           CCAAT
CDPr2/r3       ATCGAT
consensus       YYRAT/ATYRR
```

Why you should not use too much sequences to align?

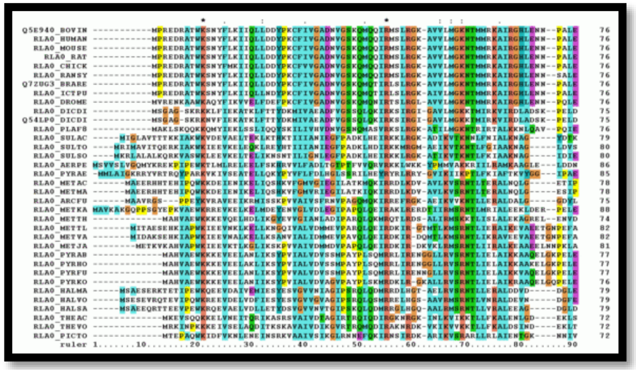
1- *Computing* **big alignments is difficult.**

2- *Building* big alignments is difficult

3- ***Displaying*** **big alignments is difficult.**

4-***Using*** **big alignments is difficult.**

5-**Making *accurate* big alignments is difficult**



```
                                 *   .      :       .          .   *         : : :      .
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-YPHMVAKKRIILRAMKAAGLE---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
```

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

Should you choose sequences that are very similar or very different?

- Make the right compromise between similarity and new information
- An alignment that only contains very similar sequences brings little information.
- You can use it to extrapolate annotations, but you can't do phylogeny, structure prediction, function perdiction, or any of the other useful applications that we mentioned before

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

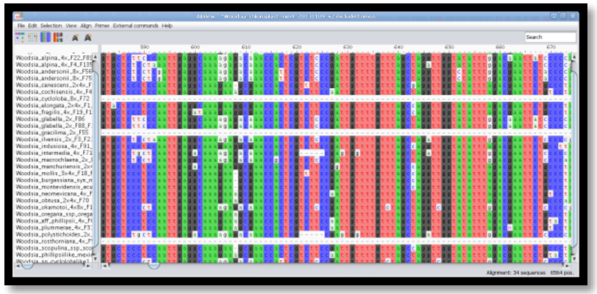
The general rule is that you want them to be as distantly related as possible - without requiring too many gaps in order to be properly aligned.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing*, *Inc*. 436 pp.

Two things multiple-sequence-alignment programs *really* don’t like are

1- Sequences that are very different from every other sequence in the group

2- Sequences that need long insertions/deletions to be properly aligned


File Edit Selection View Align Primer External commands Help
590
600
610
620
Woodsia_alpina_4x_F22_F8
Woodsia_alpina_4x_F4_F135
Woodsia_andersonii_8x_F56
Woodsia_andersonii_8x_F75
Woodsia_canescens_2x4x_F
Woodsia_cochisensis_4x_F4
Woodsia_cycloloba_8x_F72
Woodsia_elongata_2x4x_F1
Woodsia_fragilis_4x_F19_F1
Woodsia_glabella_2x_F86
Woodsia_glabella_2x_F88_F
Woodsia_gracilima_2x_F55
Woodsia_ilvensis_2x_F3_F2
Woodsia_indusiosa_4x_F91
Woodsia_intermedia_4x_F71
Woodsia_macrochlaena_2x
Woodsia_manchuriensis_2x
Woodsia_mollis_3x4x_F18
Woodsia_burgessiana_syn_n
Woodsia_montevidensis_ecu
Woodsia_neomexicana_4x_F
Woodsia_obtusa_2x4x_F70
Woodsia_okamotoi_4x8x_F1
Woodsia_oregana_ssp_orega
Woodsia_aff_phillipsii_4x_F
Woodsia_plummerae_4x_F3
Woodsia_polystichoides_2x
Woodsia_rosthorniana_4x_F
Woodsia_scopulina_ssp_sco
Woodsia_phillipsiilike_mexi
Woodsia_sp_cyclolobalike1
Search
Alignment: 34 sequences

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

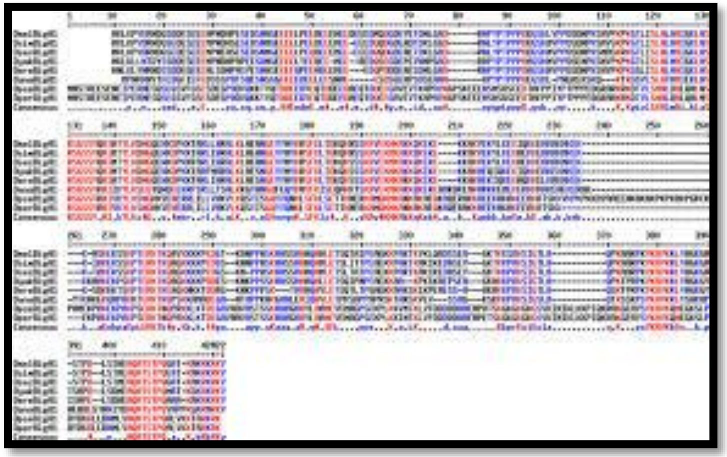
Gathering your sequences with online BLAST servers

Characterized sequences: Try to include sequences with **good annotations and experimental information** in your alignment because they bring biological information with them — and also **allow feature propagation**.

Uncharacterized sequences: including them in your multiple alignment is to **distinguish between the conserved positions** that cannot mutate and the other, less-important columns. They help in **getting some contrast** on your sequence of interest.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing, Inc*. 436 pp.

Sequences that are **so similar** to the query are **probably homologous**. We commonly refer to such sequences ***as hits*** or ***matches***.

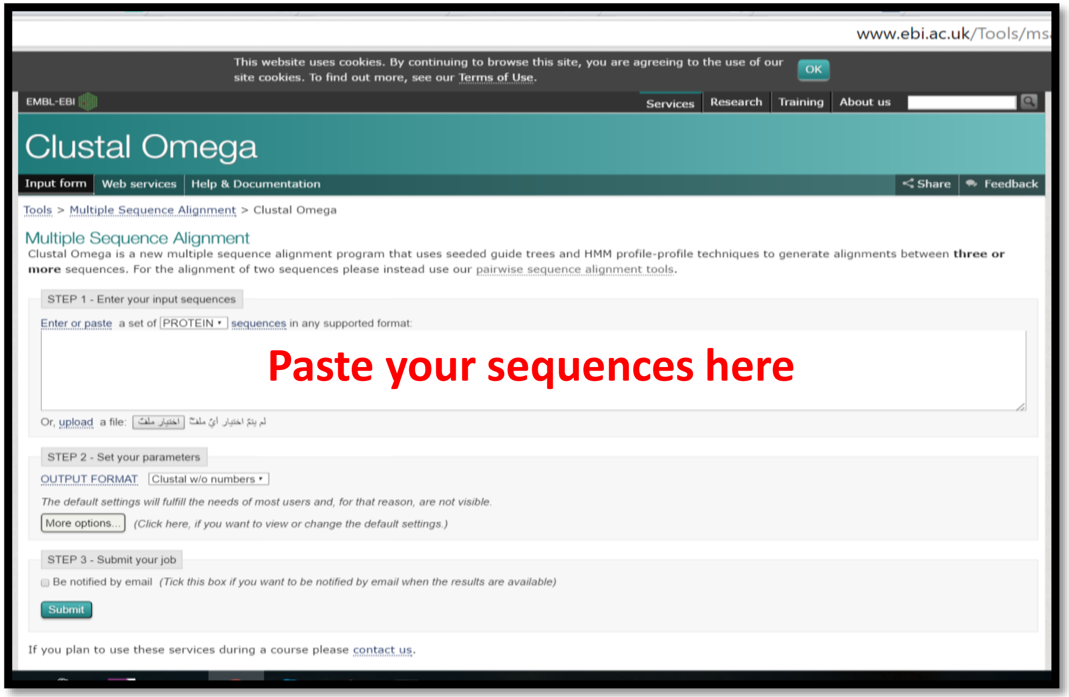


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

Choosing the Right Method of Multiple Sequence Alignment

ClustalOmega

The **most commonly** used multiple sequence alignment package. 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



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

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.

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

Paste your sequences here

Or, upload a file: اختيار ملف لم يتم اختيار أيّ ملف

STEP 2 - Set your parameters

OUTPUT FORMAT: Clustal w/o numbers

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

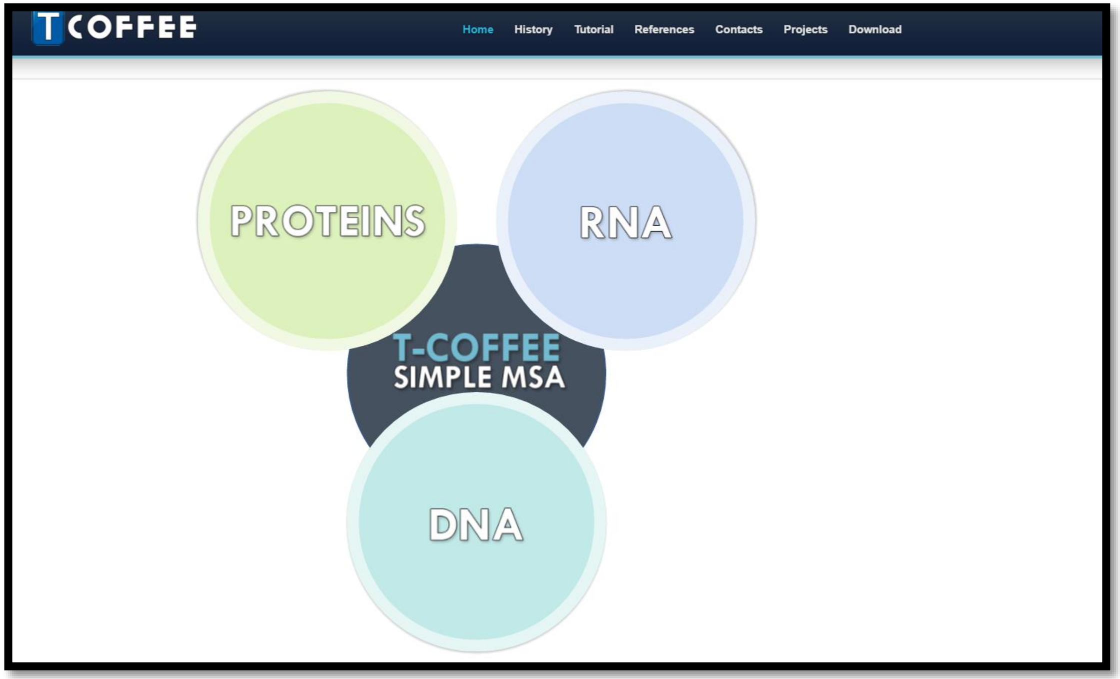
If you plan to use these services during a course please contact us.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2^{nd} Edn). *Wiley publishing, Inc*. 436 pp.

Choosing the Right Method of Multiple Sequence Alignment

2-Tcoffee

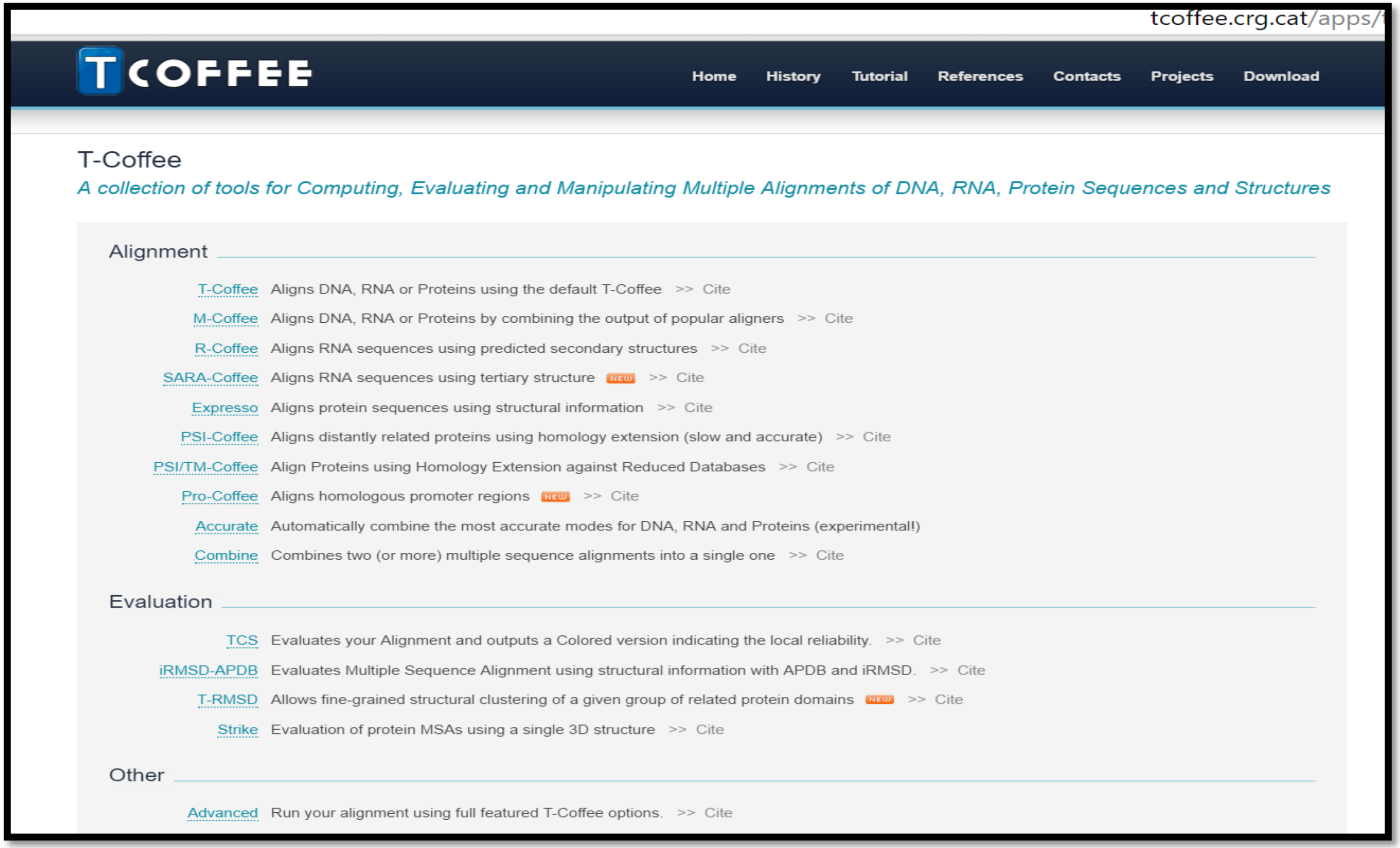
One of the latest multiple-sequence-alignment packages that you can use. With Tcoffee, you **can combine sequences** and **structures**, **evaluate an alignment**, or **merge** several alternative multiple alignments into a single unified result.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

Choosing the Right Method of Multiple Sequence Alignment

Continue……..



tcoffee.crg.cat/apps/t

T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite

Pro-Coffee Aligns homologous promoter regions NEW >> Cite

Accurate Automatically combine the most accurate modes for DNA, RNA and Proteins (experimental!)

Combine Combines two (or more) multiple sequence alignments into a single one >> Cite

Evaluation

TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Strike Evaluation of protein MSAs using a single 3D structure >> Cite

Other

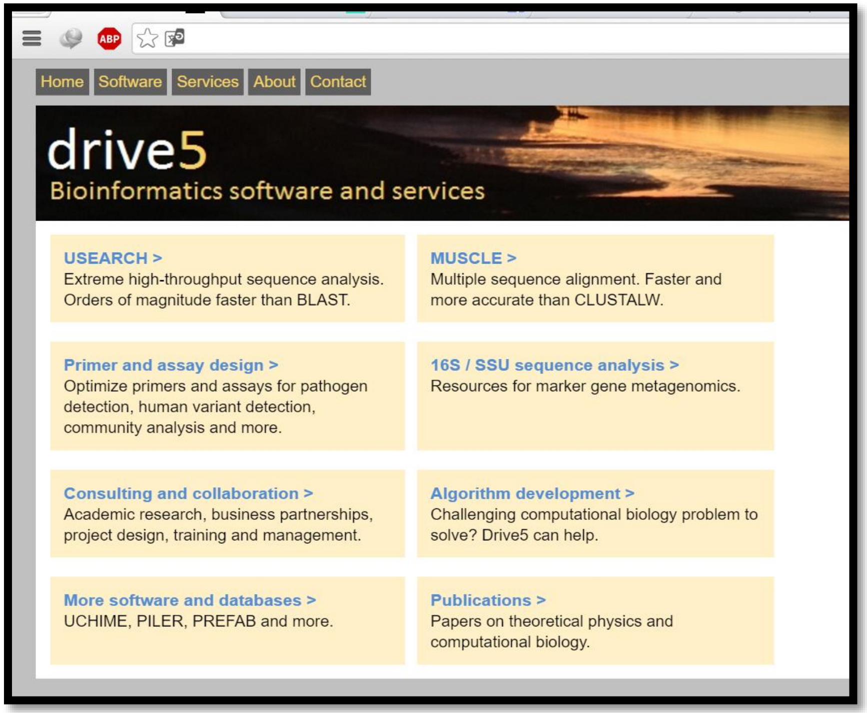
Advanced Run your alignment using full featured T-Coffee options. >> Cite

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing*, *Inc*. 436 pp.

Choosing the Right Method of Multiple Sequence Alignment

3- MUSCLE

One of **the fastest** alignment methods around for aligning large set of sequences



drive5
Bioinformatics software and services

USEARCH >
Extreme high-throughput sequence analysis. Orders of magnitude faster than BLAST.

MUSCLE >
Multiple sequence alignment. Faster and more accurate than CLUSTALW.

Primer and assay design >
Optimize primers and assays for pathogen detection, human variant detection, community analysis and more.

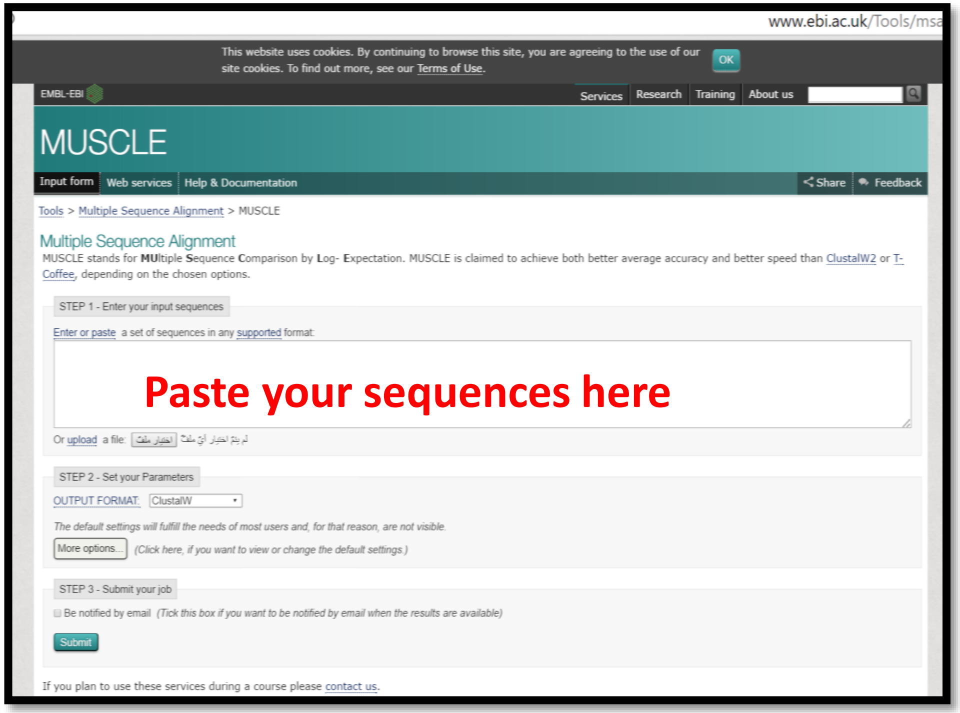
16S / SSU sequence analysis >
Resources for marker gene metagenomics.

Consulting and collaboration >
Academic research, business partnerships, project design, training and management.

Algorithm development >
Challenging computational biology problem to solve? Drive5 can help.

More software and databases >
UCHIME, PILER, PREFAB and more.

Publications >
Papers on theoretical physics and computational biology.



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

MUSCLE

Tools > Multiple Sequence Alignment > MUSCLE

Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Paste your sequences here

Or upload a file: اختيار ملف لم يتم اختيار أيّ ملف

STEP 2 - Set your Parameters

OUTPUT FORMAT: ClustalW

The default settings will fulfill the needs of most users and, for that reason, are not visible.

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

STEP 3 - Submit your job

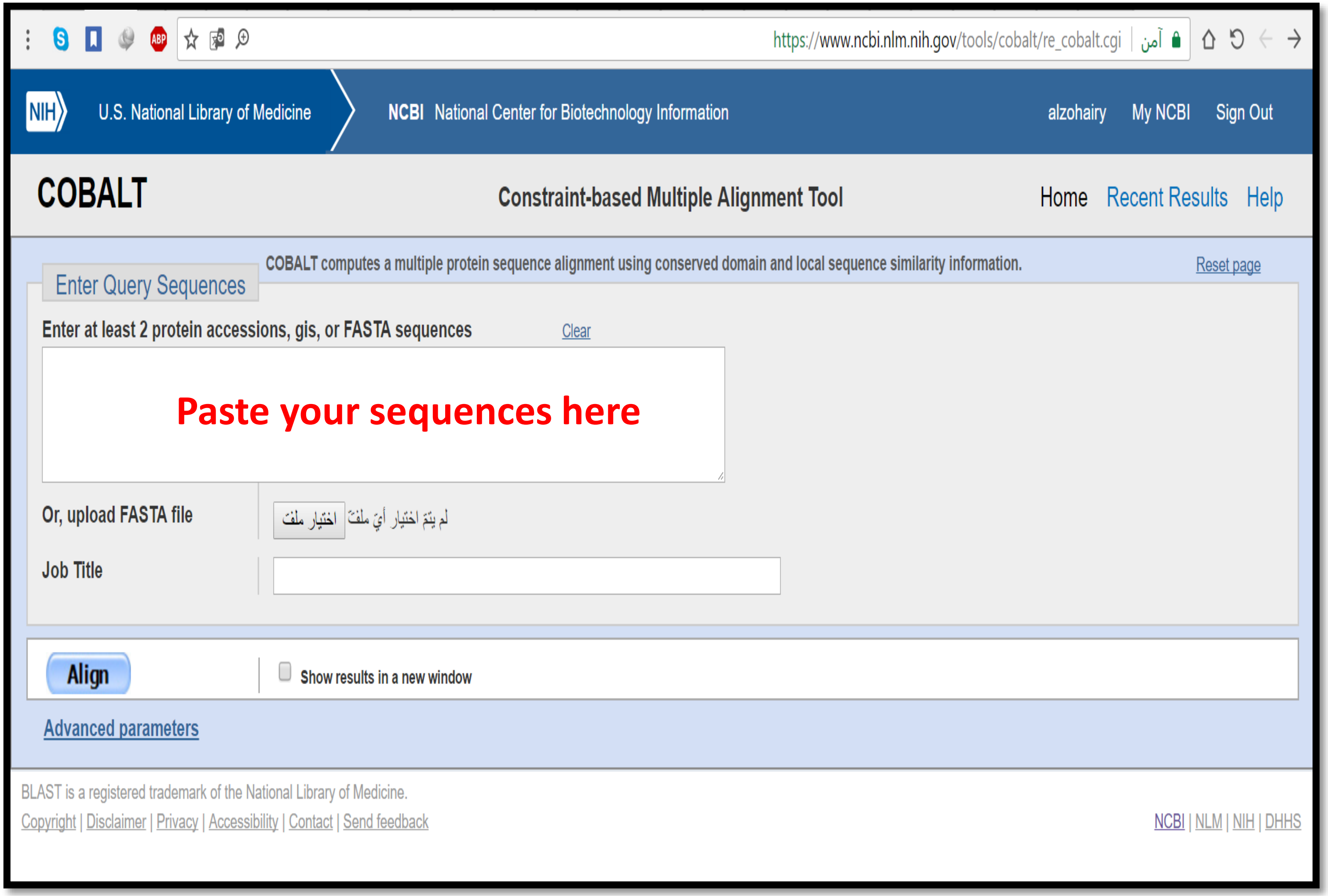
□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

If you plan to use these services during a course please contact us.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

4- COBALT

Constraint-based Multiple Alignment Tool



https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi | آمن

NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
alzohairy
My NCBI
Sign Out

COBALT

Constraint-based Multiple Alignment Tool

COBALT computes a multiple protein sequence alignment using conserved domain and local sequence similarity information.

Enter Query Sequences

Enter at least 2 protein accessions, gis, or FASTA sequences

Paste your sequences here

Or, upload FASTA file — اختيار ملف | لم يتم اختيار أي ملف

Job Title

□ Show results in a new window

BLAST is a registered trademark of the National Library of Medicine.

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

BTU BIOINFORMATICS
TRAINNING UNIT

GENOMICS

1st Module What Bioinformatics Can Do for You

2nd Module Manipulation of Biological Sequences

3rd Module Working with single DNA sequence

4th Module How to Build a Multiple Sequence Alignment?

5th Module Inferring Phylogenetics analysis using Jalview

6th Module Advanced Molecular Concepts

7th Module Inferring Protein Sequence (structure & Function)

8th Module RNAanalysis and Function

9th Module Editing and Publishing Alignments in your Manuscript

10th Module Building and Publishing Phylogenetic Trees

11th Module Working with Protein 3-D Structures

12th Module Advanced Bioinformatics Using R

Bioinformatics Workshops Series

EST by, Dr. Ahmed Mansour Alzohairy
Call us: 01026060324 / 01000727270
Facebook: https://www.facebook.com/Bioinformaticsunit

design by, Mohamed Fayed, 01026060324

By

Ahmed Mansour Alzohairy

Department of Genetics, Zagazig University, Zagazig, Egypt