Выравнивания

Форматы файлов, используемых в биоинформатике

FASTA

>roa1_drome Rea guano receptor type III >> 0.1
MVNSNQNQNGNSNGHDDDFPQDSITEPEHMRKLFIGGLDYRTTDENLKAHEKWGNIVDV
VVMKDPRTKRSRGFGFITYSHSSMIDEAQKSRPHKIDGRVEPKRAVPRQDIDSPNAGATV
KKLFVGALKDDHDEQSIRDYFQHFGNIVDNIVIDKETGKKRGFAFVEFDDYDPVDKVVLQ
KQHQLNGKMVDVKKALPKNDQQGGGGGRGGPGGRAGGNRGNMGGGNYGNQNGGGN
WNNGGNNWGNNRGNDNWGNNSFGGGGGGGGGYGGGNNSWGNNNPWDNGNGGGNF
GGGGNNWNGGNDFGGYQQNYGGGPQRGGGNFNNNRMQPYQGGGGFKAGGGNQGNY
GNNQGFNNGGNNRRY

>roa2_drome Rea guano ligand
MVNSNQNQNGNSNGHDDDFPQDSITEPEHMRKLFIGGLDYRTTDENLKAHEKWGNIVDV
VVMKDPTSTSTSTSTSTSTSTSTMIDEAQKSRPHKIDGRVEPKRAVPRQDIDSPNAGATVK
KLFVGALKDDHDEQSIRDYFQHLLLLLLLDLLLLDLLLLDLLLFVEFDDYDPVDKVVLQK
QHQLNGKMVDVKKALPKNDQQGGGGGRGGPGGRAGGNRGNMGGGNYGNQNGGGNW
NNGGNNWGNNRGNDNWGNNSFGGGGGGGGGYGGGNNSWGNNNPWDNGNGGGNFG
GGGNNWNGGNDFGGYQQNYGGGPQRGGGNFNNNRMQPYQGGGGFKAGGGNQGNYG
NNQGFNNGGNNRRY

ClustalW

- Очень известная и широко распространённая программа: UNIX, Internet, Windows.

- Выполняет MSA; может строить филогенетические деревья.

- Входной файл – формат multi-fasta.

ClustalW

- ToTo To fasta @ list

Making the file in unix

```
>IPNS_STRJU P18286
MPILMPSAEVPTIDISPLSGDDAKAKQRVAQEINKAARGSGFFYASNHGVDVQLLQDVVN
EFHRNMSDQEKHDLAINAYNKDNPHVRNGYYKAIKGKKAVESFCYLNPSFSDDHPMIKSE
TPMHEVNLWPDEEKHPRFRPFCEDYYRQLLRLSTVIMRGYALALGRREDFFDEALAEADT
LSSVSLIRYPYLEEYPPVKTGADGTKLSFEDHLDVSMITVLYQTEVQNLQVETVDGWQDI
PRSDEDFLVNCGTYMGHITHDYFPAPNHRVKFINAERLSLPFFLNAGHNSVIEPFVPEGA
AGTVKNPTTSYGEYLQHGLRALIVKNGQT
>IPNS_STRCL P10621
MPVLMPSAHVPTIDISPLFGTDAAAKKRVAEEIHGACRGSGFFYATNHGVDVQQLQDVVN
EFHGAMTDQEKHDLAIHAYNPDNPHVRNGYYKAVPGRKAVESFCYLNPDFGEDHPMIAAG
TPMHEVNLWPDEERHPRFRPFCEGYYRQMLKLSTVLMRGLALALGRPEHFFDAALAEQDS
LSSVSLIRYPYLEEYPPVKTGPDGQLLSFEDHLDVSMITVLFQTQVQNLQVETVDGWRDI
PTSENDFLVNCGTYMAHVTNDYFPAPNHRVKFVNAERLSLPFFLNGGHEAVIEPFVPEGA
SEEVRNEALSYGDYLQHGLRALIVKNGQT
```

input file:
Multi-fasta

ClustalW

```
CLUSTAL W (1.7) multiple sequence alignment


IPNS_STRJU     -MPILMPSAEVPTIDISPLSGDDAKAKQRVAQEINKAARGSGFFYASNHGVDVQLLQDVV
IPNS_STRGR     -MPIPMLPAHVPTIDISPLSGGDADDKKRVAQEINKACRESGFFYASHHGIDVQLLKDVV
IPNS_FLASS     ----MNRHADVPVIDISGLSGNDMDVKKDIAARIDRACRGSGFFYAANHGVDLAALQKFT
IPNS_PENCH     --MASTPKANVPKIDVSPLFGDNMEEKMKVARAIDAASRDTGFFYAVNHGVDVKRLSNKT
IPNS_CEPAC     MGSVPVPVANVPRIDVSPLFGDDKEKKLEVARAIDAASRDTGFFYAVNHGVDLPWLSRET
                   *.** **:* * *.: . *  :*  *: *.* :***** :**:*:  *.  .

IPNS_STRJU     NEFHRNMSDQEKHDLAINAYNKDNP-HVRNGYYKAIKGKKAVESFCYLNPSFSDDHPMIK
IPNS_STRGR     NEFHRTMTDEEKYDLAINAYNKNNP-RTRNGYYMAVKGKKAVESWCYLNPSFSEDHPQIR
IPNS_FLASS     TDWHMAMSAEEKWELAIRAYNPANP-RNRNGYYMAVEGKKANESFCYLNPSFDADHATIK
IPNS_PENCH     REFHFSITDEEKWDLAIRAYNKEHQDQIRAGYYLSIPEKKAVESFCYLNPNFKPDHPLIQ
IPNS_CEPAC     NKFHMSITDEEKWQLAIRAYNKEHESQIRAGYYLPIPGKKAVESFCYLNPSFSPDHPRIK
                 .:*  :: :** :***.***  :  : * *** .:  *** **:*****.*. **. *:
```

Выходной файл: aln format

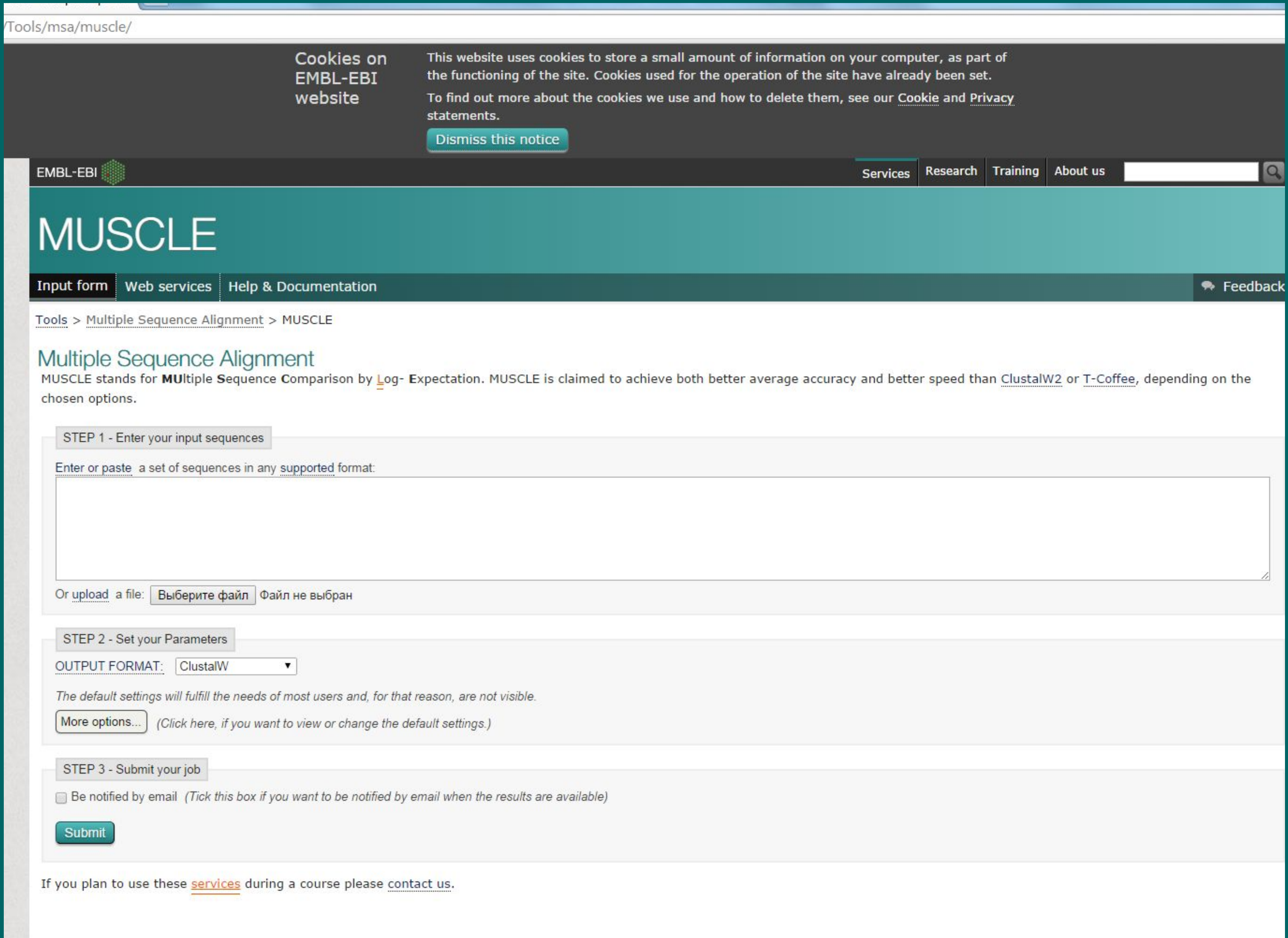
форматы

http://www.ebi.ac.uk/help/formats.html

ClustalW - результат

```
IPNS_STRJU      -----------------SYNTHASEMP-ILMPSAEVPTIDISPLSGDDAKAKQRVAQEIN 42
IPNS_STRCL      ------------------YNTHASEMP-VLMPSAHVPTIDISPLFGTDAAAKKRVAEEIH 41
IPNS_NOCLA      ----------------------ASEM---KMPSAEVPTIDVSPLFGDDAQEKVRVGQEIN 35
IPNS_CEPAC      IPNSISPENICILLINNSYNTHASEMGSVPVPVANVPRIDVSPLFGDDKEKKLEVARAID 60
                                      ****    :* *.** **:*** * *   * .*.. *.

IPNS_STRJU      KAARGSGFFYASNHGVDVQLLQDVVNEFHRNMSDQEKHDLAINAYNKDNP-HVRNGYYKA 101
IPNS_STRCL      GACRGSGFFYATNHGVDVQQLQDVVNEFHGAMTDQEKHDLAIHAYNPDNP-HVRNGYYKA 100
IPNS_NOCLA      KACRGSGFFYAANHGVDVQRLQDVVNEFHRTMSPQEKYDLAIHAYNKNNS-HVRNGYYMA 94
IPNS_CEPAC      AASRDTGFFYAVNHGVDLPWLSRETNKFHMSITDEEKWQLAIRAYNKEHESQIRAGYYLP 120
                 *.*.:***** *****:  *.  .*:**  :: :** :***.*** ::  ::* *** .

IPNS_STRJU      IKGKKAVESFCYLNPSFSDDHPMIKSETPMHEVNLWPDEEKHPRFRPFCEDYYRQLLRLS 161
IPNS_STRCL      VPGRKAVESFCYLNPDFGEDHPMIAAGTPMHEVNLWPDEERHPRFRPFCEGYYRQMLKLS 160
IPNS_NOCLA      IEGKKAVESFCYLNPSFSEDHPEIKAGTPMHEVNSWPDEEKHPSFRPFCEEYYWTMHRLS 154
IPNS_CEPAC      IPGKKAVESFCYLNPSFSPDHPRIKEPTPMHEVNVWPDEAKHPGFRAFAEKYYWDVFGLS 180
                : *:***********.*. *** *   ******* **** :** **.*.* **  :  **

IPNS_STRJU      TVIMRGYALALGRREDFFDEALAEADTLSSVSLIRYPYLEEYPP--VKTGADGTKLSFED 219
IPNS_STRCL      TVLMRGLALALGRPEHFFDAALAEQDSLSSVSLIRYPYLEEYPP--VKTGPDGQLLSFED 218
IPNS_NOCLA      KVLMRGFALALGKDERFFEPELKEADTLSSVSLIRYPYLEDYPP--VKTGPDGEKLSFED 212
IPNS_CEPAC      SAVLRGYALALGRDEDFFTRHSRRDTTLSSVVLIRYPYLDPYPEPAIKTADDGTKLSFEW 240
                ..::** *****: * **     .  :**** *******: **   :**. **  ****

IPNS_STRJU      HLDVSMITVLYQTEVQNLQVETVDGWQDIPRSDEDFLVNCGTYMGHITHDYFPAPNHRVK 279
IPNS_STRCL      HLDVSMITVLFQTQVQNLQVETVDGWRDIPTSENDFLVNCGTYMAHVTNDYFPAPNHRVK 278
IPNS_NOCLA      HFDVSMITVLYQTQVQNLQVETVDGWRDLPTSDTDFLVNAGTYLGHLTNDYFPSPLHRVK 272
IPNS_CEPAC      HEDVSLITVLYQSDVQNLQVKTPQGWQDIQADDTGFLINCGSYMAHITDDYYPAPIHRVK 300
                * ***:****:*::******:* :**:*:  .: .**:*.*:*:.*:*.**:*:* ****

IPNS_STRJU      FINAERLSLPFFLNAGHNSVIEPFVP----EGAAGTVK-NPTTSYGEYLQHGLRALIVKN 334
IPNS_STRCL      FVNAERLSLPFFLNGGHEAVIEPFVP----EGASEEVR-NEALSYGDYLQHGLRALIVKN 333
IPNS_NOCLA      FVNAERLSLPFFFHAGQHTLIEPFFP----DGAPEGKQGNEAVRYGDYLNHGLHSLIVKN 328
IPNS_CEPAC      WVNEERQSLPFFVNLGWEDTIQPWDPATAKDGAKDAAKDKPAISYGEYLQGGLRGLINKN 360
                ::* ** *****.: * .  *:*: *    :**    : : :  **:**: **:.** **

IPNS_STRJU      GQT 337
IPNS_STRCL      GQT 336
IPNS_NOCLA      GQT 331
IPNS_CEPAC      GQT 363
                ***
```

http://www.ebi.ac.uk/Tools/msa/muscle/



Cookies on EMBL-EBI website

This website uses cookies to store a small amount of information on your computer, as part of the functioning of the site. Cookies used for the operation of the site have already been set.

To find out more about the cookies we use and how to delete them, see our Cookie and Privacy statements.

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:

Or upload a file: Выберите файл Файл не выбран

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.

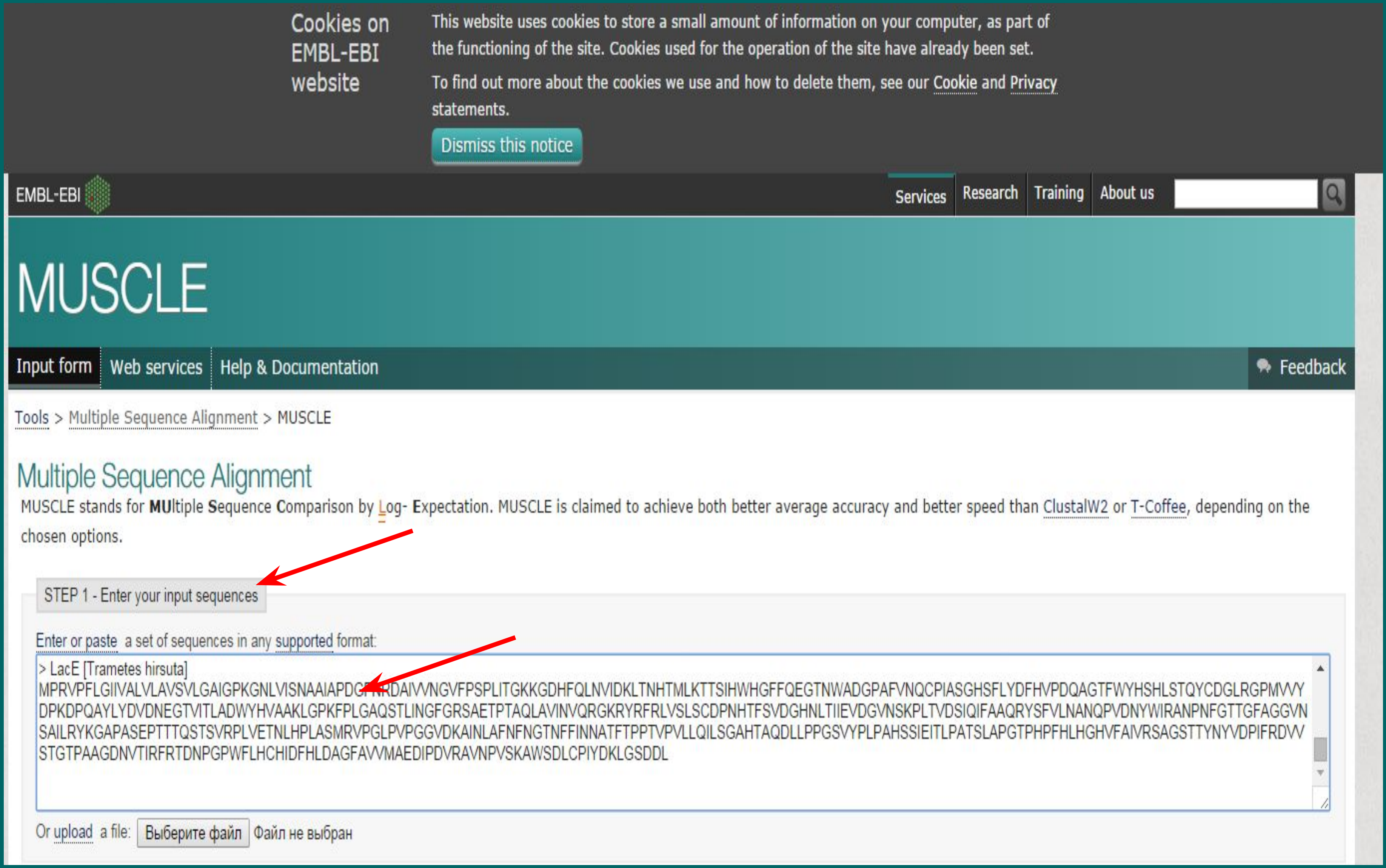
More options... *(Click here, if you want to view or change the default settings.)*

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.

Step 1



Cookies on EMBL-EBI website

This website uses cookies to store a small amount of information on your computer, as part of the functioning of the site. Cookies used for the operation of the site have already been set.

To find out more about the cookies we use and how to delete them, see our Cookie and Privacy statements.

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.

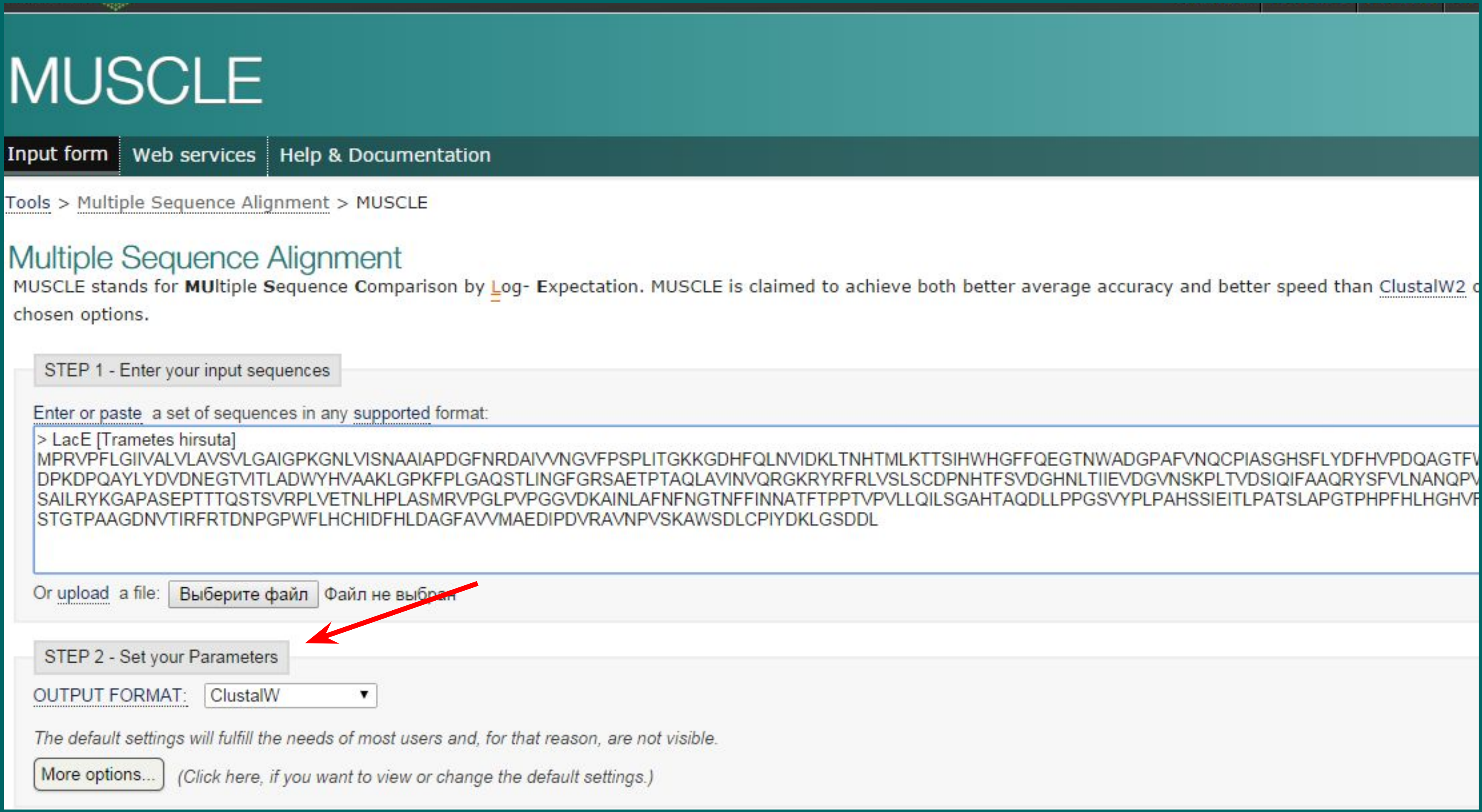
STEP 1 - Enter your input sequences

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

```
> LacE [Trametes hirsuta]
MPRVPFLGIIVALVLAVSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLITGKKGDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPMVVY
DPKDPQAYLYDVDNEGTVITLADWYHVAAKLGPKFPLGAQSTLINGFGRSAETPTAQLAVINVQRGKRYRFRLVSLSCDPNHTFSVDGHNLTIIEVDGVNSKPLTVDSIQIFAAQRYSFVLNANQPVDNYWIRANPNFGTTGFAGGVN
SAILRYKGAPASEPTTTQSTSVRPLVETNLHPLASMRVPGLPVPGGVDKAINLAFNFNGTNFFINNATFTPPTVPVLLQILSGAHTAQDLLPPGSVYPLPAHSSIEITLPATSLAPGTPHPFHLHGHVFAIVRSAGSTTYNYVDPIFRDVV
STGTPAAGDNVTIRFRTDNPGPWFLHCHIDFHLDAGFAVVMAEDIPDVRAVNPVSKAWSDLCPIYDKLGSDDL
```

Or upload a file: Выберите файл Файл не выбран

Step 2



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
chosen options.

STEP 1 - Enter your input sequences

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

```
> LacE [Trametes hirsuta]
MPRVPFLGIIVALVLAVSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLITGKKGDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTF
DPKDPQAYLYDVDNEGTVITLADWYHVAAKLGPKFPLGAQSTLINGFGRSAETPTAQLAVINVQRGKRYRFRLVSLSCDPNHTFSVDGHNLTIIEVDGVNSKPLTVDSIQIFAAQRYSFVLNANQP
SAILRYKGAPASEPTTTQSTSVRPLVETNLHPLASMRVPGLPVPGGVDKAINLAFNFNGTNFFINNATFTPPTVPVLLQILSGAHTAQDLLPPGSVYPLPAHSSIEITLPATSLAPGTPHPFHLHGHV
STGTPAAGDNVTIRFRTDNPGPWFLHCHIDFHLDAGFAVVMAEDIPDVRAVNPVSKAWSDLCPIYDKLGSDDL
```

Or upload a file: Выберите файл Файл не выбран

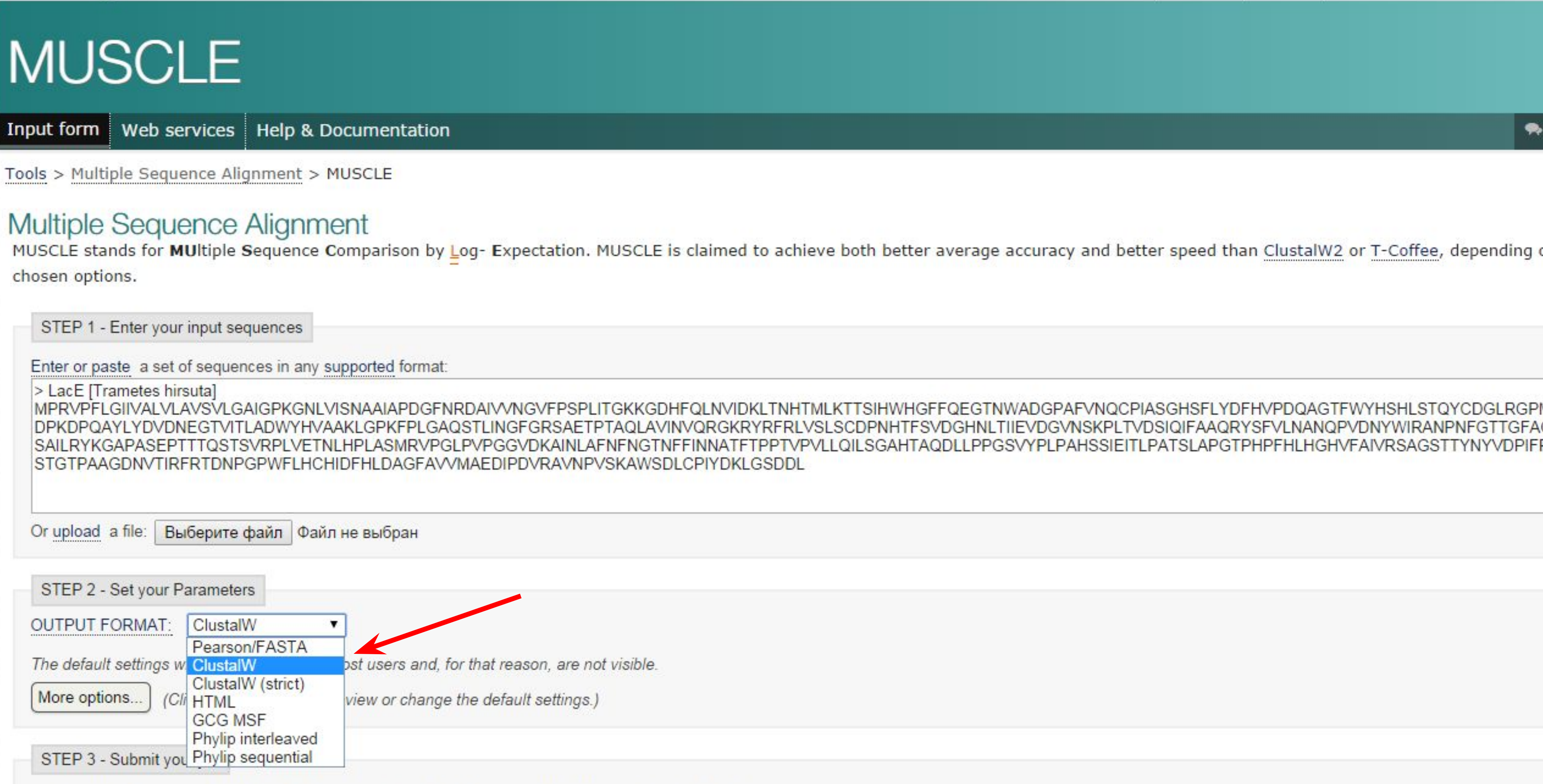
STEP 2 - Set your Parameters

OUTPUT FORMAT: ClustalW ▼

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

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

Step 2 cont…



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
chosen options.

STEP 1 - Enter your input sequences

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

```
> LacE [Trametes hirsuta]
MPRVPFLGIIVALVLAVSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLITGKKGDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGP
DPKDPQAYLYDVDNEGTVITLADWYHVAAKLGPKFPLGAQSTLINGFGRSAETPTAQLAVINVQRGKRYRFRLVSLSCDPNHTFSVDGHNLTIIEVDGVNSKPLTVDSIQIFAAQRYSFVLNANQPVDNYWIRANPNFGTTGFA
SAILRYKGAPASEPTTTQSTSVRPLVETNLHPLASMRVPGLPVPGGVDKAINLAFNFNGTNFFINNATFTPPTVPVLLQILSGAHTAQDLLPPGSVYPLPAHSSIEITLPATSLAPGTPHPFHLHGHVFAIVRSAGSTTYNYVDPIF
STGTPAAGDNVTIRFRTDNPGPWFLHCHIDFHLDAGFAVVMAEDIPDVRAVNPVSKAWSDLCPIYDKLGSDDL
```

Or upload a file: Выберите файл Файл не выбран

STEP 2 - Set your Parameters

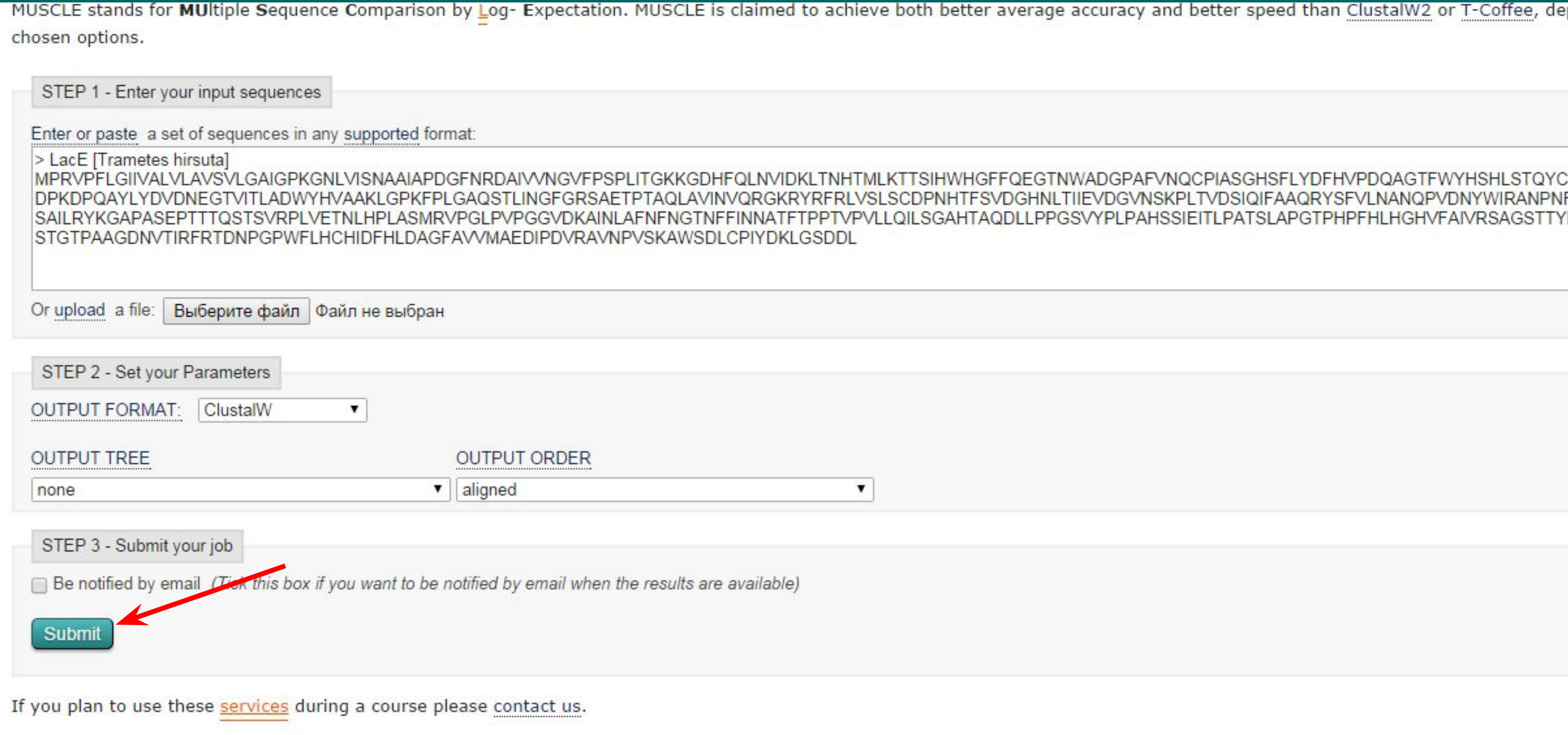
OUTPUT FORMAT: ClustalW

The default settings w st users and, for that reason, are not visible.

(Cl view or change the default settings.)

STEP 3 - Submit you

Step 3



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, de
chosen options.

STEP 1 - Enter your input sequences

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

```
> LacE [Trametes hirsuta]
MPRVPFLGIIVALVLAVSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLITGKKGDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYC
DPKDPQAYLYDVDNEGTVITLADWYHVAAKLGPKFPLGAQSTLINGFGRSAETPTAQLAVINVQRGKRYRFRLVSLSCDPNHTFSVDGHNLTIIEVDGVNSKPLTVDSIQIFAAQRYSFVLNANQPVDNYWIRANPN
SAILRYKGAPASEPTTTQSTSVRPLVETNLHPLASMRVPGLPVPGGVDKAINLAFNFNGTNFFINNATFTPPTVPVLLQILSGAHTAQDLLPPGSVYPLPAHSSIEITLPATSLAPGTPHPFHLHGHVFAIVRSAGSTTY
STGTPAAGDNVTIRFRTDNPGPWFLHCHIDFHLDAGFAVVMAEDIPDVRAVNPVSKAWSDLCPIYDKLGSDDL
```

Or upload a file: Файл не выбран

STEP 2 - Set your Parameters

OUTPUT FORMAT: ClustalW

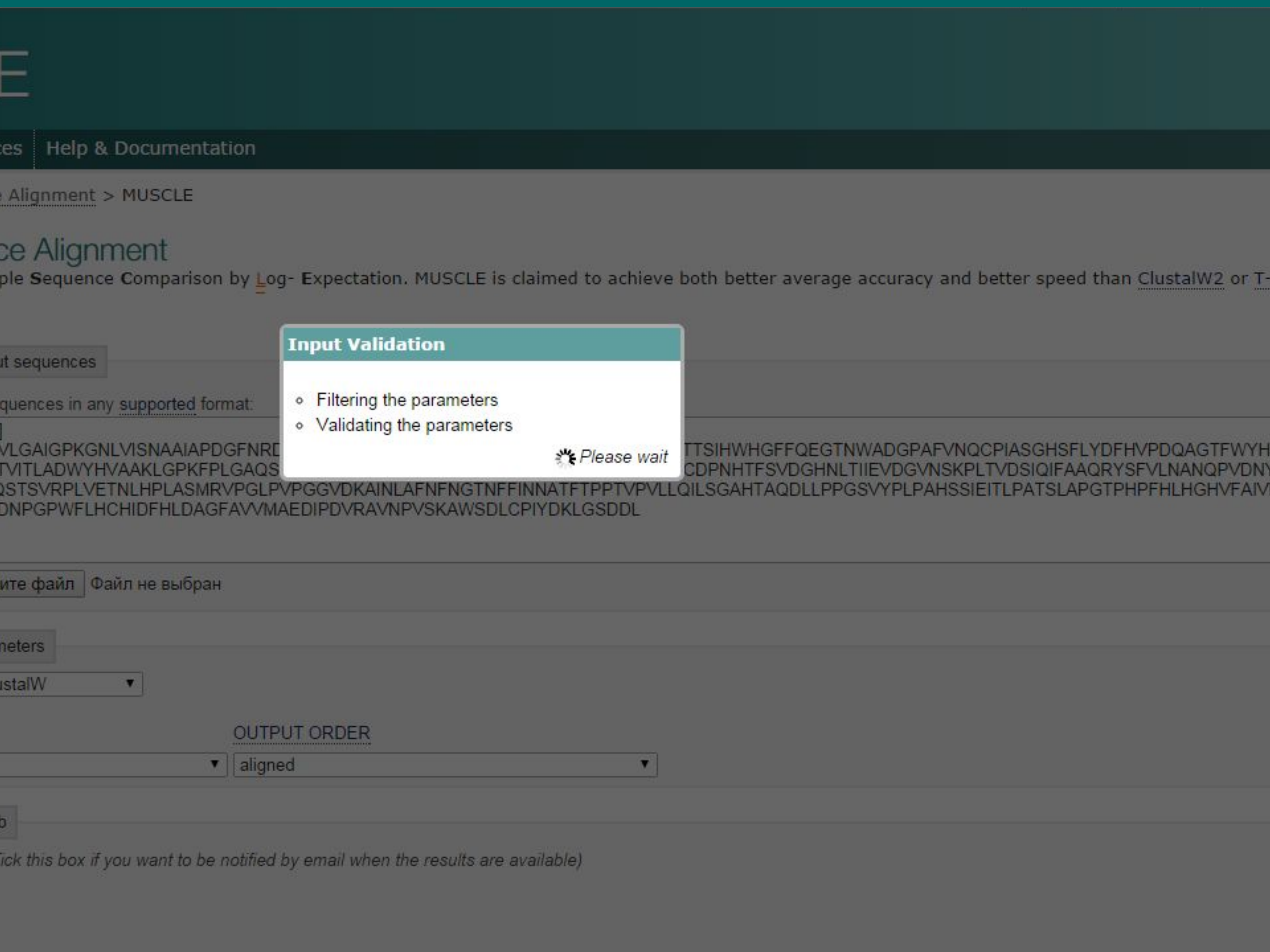
OUTPUT TREE: none

OUTPUT ORDER: aligned

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.



Alignment > MUSCLE

ce Alignment

ple **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-

Input Validation

- Filtering the parameters
- Validating the parameters

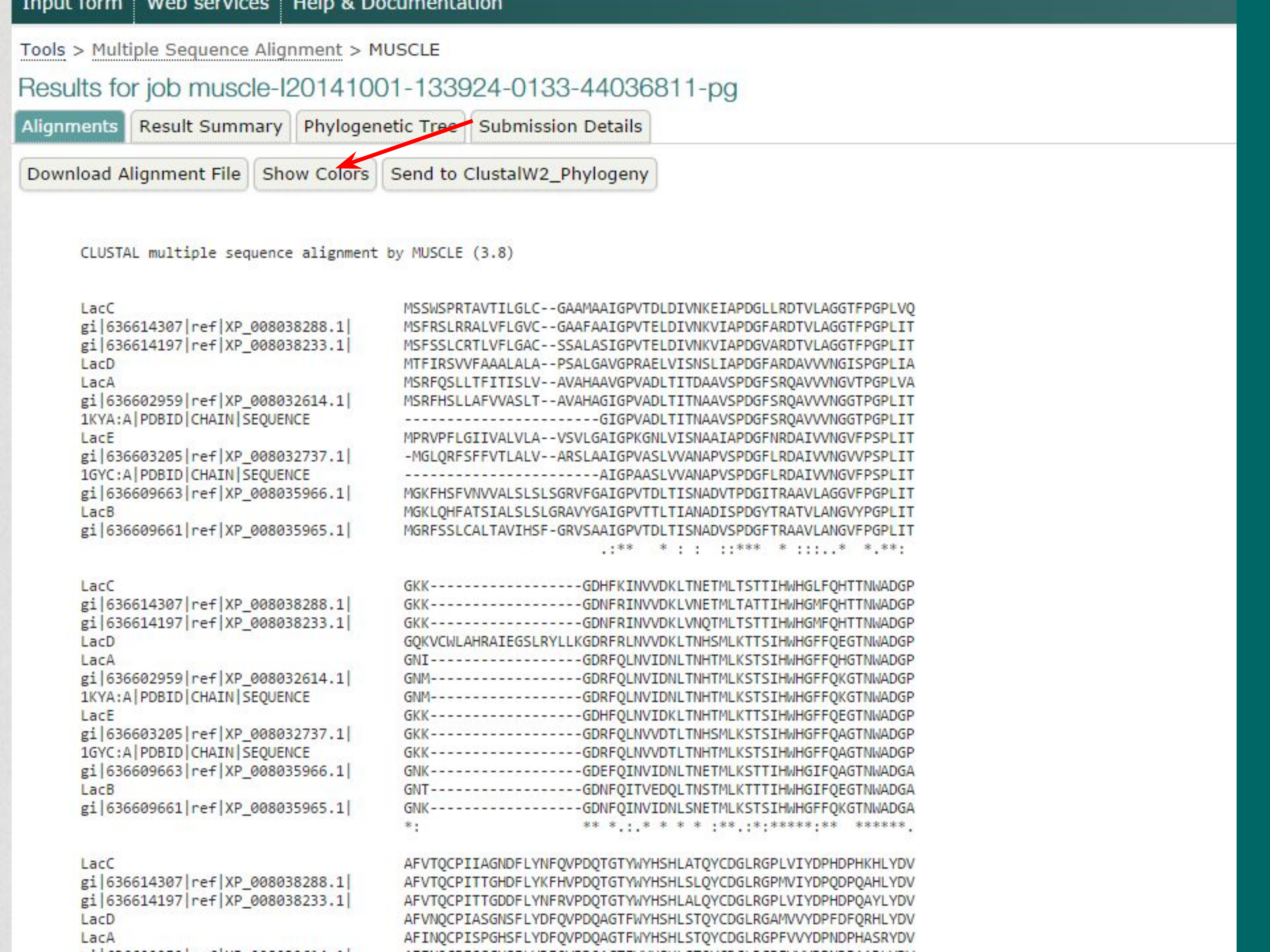
Please wait

quences in any supported format:

```
]
VLGAIGPKGNLVISNAAIAPDGFNRD
TVITLADWYHVAAKLGPKFPLGAQS
QSTSVRPLVETNLHPLASMRVPGLPVPGGVDKAINLAFNFNGTNFFINNATFTPPTVPVLLQILSGAHTAQDLLPPGSVYPLPAHSSIEITLPATSLAPGTPHPFHLHGHVFAIV
DNPGPWFLHCHIDFHLDAGFAVVMAEDIPDVRAVNPVSKAWSDLCPIYDKLGSDDL
```

```
TTSIHWHGFFQEGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYH
CDPNHTFSVDGHNLTIIEVDGVNSKPLTVDSIQIFAAQRYSFVLNANQPVDNY
```

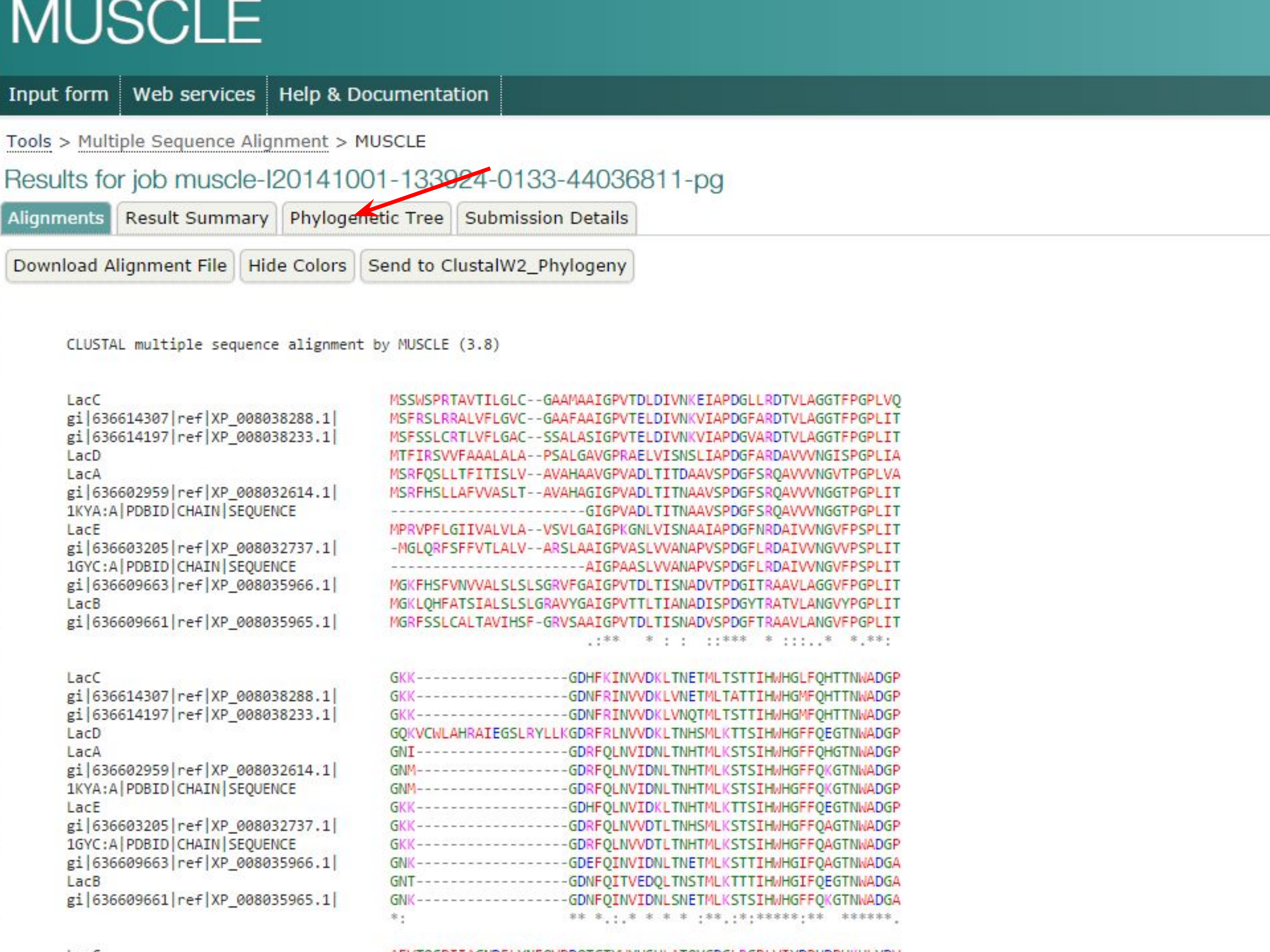
Tick this box if you want to be notified by email when the results are available)



Tools > Multiple Sequence Alignment > MUSCLE

Results for job muscle-I20141001-133924-0133-44036811-pg

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


LacC                                  MSSWSPRTAVTILGLC--GAAMAAIGPVTDLDIVNKEIAPDGLLRDTVLAGGTFPGPLVQ
gi|636614307|ref|XP_008038288.1|      MSFRSLRRALVFLGVC--GAAFAAIGPVTELDIVNKVIAPDGFARDTVLAGGTFPGPLIT
gi|636614197|ref|XP_008038233.1|      MSFSSLCRTLVFLGAC--SSALASIGPVTELDIVNKVIAPDGVARDTVLAGGTFPGPLIT
LacD                                  MTFIRSVVFAAALALA--PSALGAVGPRAELVISNSLIAPDGFARDAVVVNGISPGPLIA
LacA                                  MSRFQSLLTFITISLV--AVAHAAVGPVADLTITDAAVSPDGFSRQAVVVNGVTPGPLVA
gi|636602959|ref|XP_008032614.1|      MSRFHSLLAFVVASLT--AVAHAGIGPVADLTITNAAVSPDGFSRQAVVVNGGTPGPLIT
1KYA:A|PDBID|CHAIN|SEQUENCE           -----------------------GIGPVADLTITNAAVSPDGFSRQAVVVNGGTPGPLIT
LacE                                  MPRVPFLGIIVALVLA--VSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLIT
gi|636603205|ref|XP_008032737.1|      -MGLQRFSFFVTLALV--ARSLAAIGPVASLVVANAPVSPDGFLRDAIVVNGVVPSPLIT
1GYC:A|PDBID|CHAIN|SEQUENCE           -----------------------AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLIT
gi|636609663|ref|XP_008035966.1|      MGKFHSFVNVVALSLSLSGRVFGAIGPVTDLTISNADVTPDGITRAAVLAGGVFPGPLIT
LacB                                  MGKLQHFATSIALSLSLGRAVYGAIGPVTTLTIANADISPDGYTRATVLANGVYPGPLIT
gi|636609661|ref|XP_008035965.1|      MGRFSSLCALTAVIHSF-GRVSAAIGPVTDLTISNADVSPDGFTRAAVLANGVFPGPLIT
                                                             .:**   * : :  ::***  * :::..*  *.**:

LacC                                  GKK------------------GDHFKINVVDKLTNETMLTSTTIHWHGLFQHTTNWADGP
gi|636614307|ref|XP_008038288.1|      GKK------------------GDNFRINVVDKLVNETMLTATTIHWHGMFQHTTNWADGP
gi|636614197|ref|XP_008038233.1|      GKK------------------GDNFRINVVDKLVNQTMLTSTTIHWHGMFQHTTNWADGP
LacD                                  GQKVCWLAHRAIEGSLRYLLKGDRFRLNVVDKLTNHSMLKTTSIHWHGFFQEGTNWADGP
LacA                                  GNI------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQHGTNWADGP
gi|636602959|ref|XP_008032614.1|      GNM------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQKGTNWADGP
1KYA:A|PDBID|CHAIN|SEQUENCE           GNM------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQKGTNWADGP
LacE                                  GKK------------------GDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGP
gi|636603205|ref|XP_008032737.1|      GKK------------------GDRFQLNVVDTLTNHSMLKSTSIHWHGFFQAGTNWADGP
1GYC:A|PDBID|CHAIN|SEQUENCE           GKK------------------GDRFQLNVVDTLTNHTMLKSTSIHWHGFFQAGTNWADGP
gi|636609663|ref|XP_008035966.1|      GNK------------------GDEFQINVIDNLTNETMLKSTTIHWHGIFQAGTNWADGA
LacB                                  GNT------------------GDNFQITVEDQLTNSTMLKTTTIHWHGIFQEGTNWADGA
gi|636609661|ref|XP_008035965.1|      GNK------------------GDNFQINVIDNLSNETMLKSTSIHWHGFFQKGTNWADGA
                                      *:                   ** *.:.* * * * :**.:*:*****:**  ******.

LacC                                  AFVTQCPIIAGNDFLYNFQVPDQTGTYWYHSHLATQYCDGLRGPLVIYDPHDPHKHLYDV
gi|636614307|ref|XP_008038288.1|      AFVTQCPITTGHDFLYKFHVPDQTGTYWYHSHLSLQYCDGLRGPMVIYDPQDPQAHLYDV
gi|636614197|ref|XP_008038233.1|      AFVTQCPITTGDDFLYNFRVPDQTGTYWYHSHLALQYCDGLRGPLVIYDPHDPQAYLYDV
LacD                                  AFVNQCPIASGNSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGAMVVYDPFDFQRHLYDV
LacA                                  AFINQCPISPGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASRYDV
```



MUSCLE

Input form | Web services | Help & Documentation

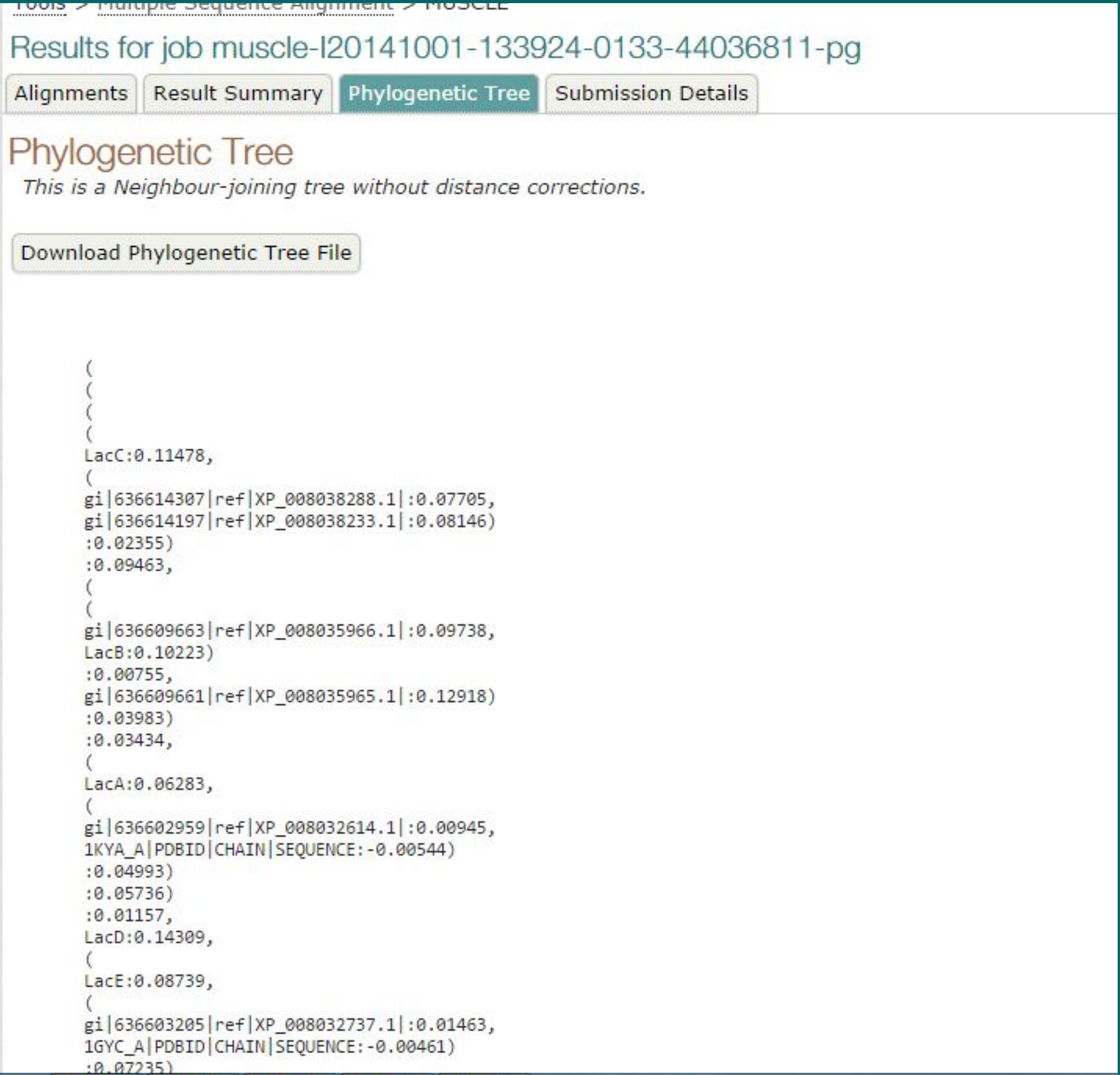
Tools > Multiple Sequence Alignment > MUSCLE

Results for job muscle-I20141001-133924-0133-44036811-pg

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Hide Colors | Send to ClustalW2_Phylogeny

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


LacC                                  MSSWSPRTAVTILGLC--GAAMAAIGPVTDLDIVNKEIAPDGLLRDTVLAGGTFPGPLVQ
gi|636614307|ref|XP_008038288.1|      MSFRSLRRALVFLGVC--GAAFAAIGPVTELDIVNKVIAPDGFARDTVLAGGTFPGPLIT
gi|636614197|ref|XP_008038233.1|      MSFSSLCRTLVFLGAC--SSALASIGPVTELDIVNKVIAPDGVARDTVLAGGTFPGPLIT
LacD                                  MTFIRSVVFAAALALA--PSALGAVGPRAELVISNSLIAPDGFARDAVVVNGISPGPLIA
LacA                                  MSRFQSLLTFITISLV--AVAHAAVGPVADLTITDAAVSPDGFSRQAVVVNGVTPGPLVA
gi|636602959|ref|XP_008032614.1|      MSRFHSLLAFVVASLT--AVAHAGIGPVADLTITNAAVSPDGFSRQAVVVNGGTPGPLIT
1KYA:A|PDBID|CHAIN|SEQUENCE           -----------------------GIGPVADLTITNAAVSPDGFSRQAVVVNGGTPGPLIT
LacE                                  MPRVPFLGIIVALVLA--VSVLGAIGPKGNLVISNAAIAPDGFNRDAIVVNGVFPSPLIT
gi|636603205|ref|XP_008032737.1|      -MGLQRFSFFVTLALV--ARSLAAIGPVASLVVANAPVSPDGFLRDAIVVNGVVPSPLIT
1GYC:A|PDBID|CHAIN|SEQUENCE           -----------------------AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLIT
gi|636609663|ref|XP_008035966.1|      MGKFHSFVNVVALSLSLSGRVFGAIGPVTDLTISNADVTPDGITRAAVLAGGVFPGPLIT
LacB                                  MGKLQHFATSIALSLSLGRAVYGAIGPVTTLTIANADISPDGYTRATVLANGVYPGPLIT
gi|636609661|ref|XP_008035965.1|      MGRFSSLCALTAVIHSF-GRVSAAIGPVTDLTISNADVSPDGFTRAAVLANGVFPGPLIT
                                                             .:**   * : :  ::***  * ::::.*  *.**: 

LacC                                  GKK------------------GDHFKINVVDKLTNETMLTSTTIHWHGLFQHTTNWADGP
gi|636614307|ref|XP_008038288.1|      GKK------------------GDNFRINVVDKLVNETMLTATTIHWHGMFQHTTNWADGP
gi|636614197|ref|XP_008038233.1|      GKK------------------GDNFRINVVDKLVNQTMLTSTTIHWHGMFQHTTNWADGP
LacD                                  GQKVCWLAHRAIEGSLRYLLKGDRFRLNVVDKLTNHSMLKTTSIHWHGFFQEGTNWADGP
LacA                                  GNI------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQHGTNWADGP
gi|636602959|ref|XP_008032614.1|      GNM------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQKGTNWADGP
1KYA:A|PDBID|CHAIN|SEQUENCE           GNM------------------GDRFQLNVIDNLTNHTMLKSTSIHWHGFFQKGTNWADGP
LacE                                  GKK------------------GDHFQLNVIDKLTNHTMLKTTSIHWHGFFQEGTNWADGP
gi|636603205|ref|XP_008032737.1|      GKK------------------GDRFQLNVVDTLTNHSMLKSTSIHWHGFFQAGTNWADGP
1GYC:A|PDBID|CHAIN|SEQUENCE           GKK------------------GDRFQLNVVDTLTNHTMLKSTSIHWHGFFQAGTNWADGP
gi|636609663|ref|XP_008035966.1|      GNK------------------GDEFQINVIDNLTNETMLKSTTIHWHGIFQAGTNWADGA
LacB                                  GNT------------------GDNFQITVEDQLTNSTMLKTTTIHWHGIFQEGTNWADGA
gi|636609661|ref|XP_008035965.1|      GNK------------------GDNFQINVIDNLSNETMLKSTSIHWHGFFQKGTNWADGA
                                      *:                   ** *.:.* * * * :**.:*:*****:**  ******.
```

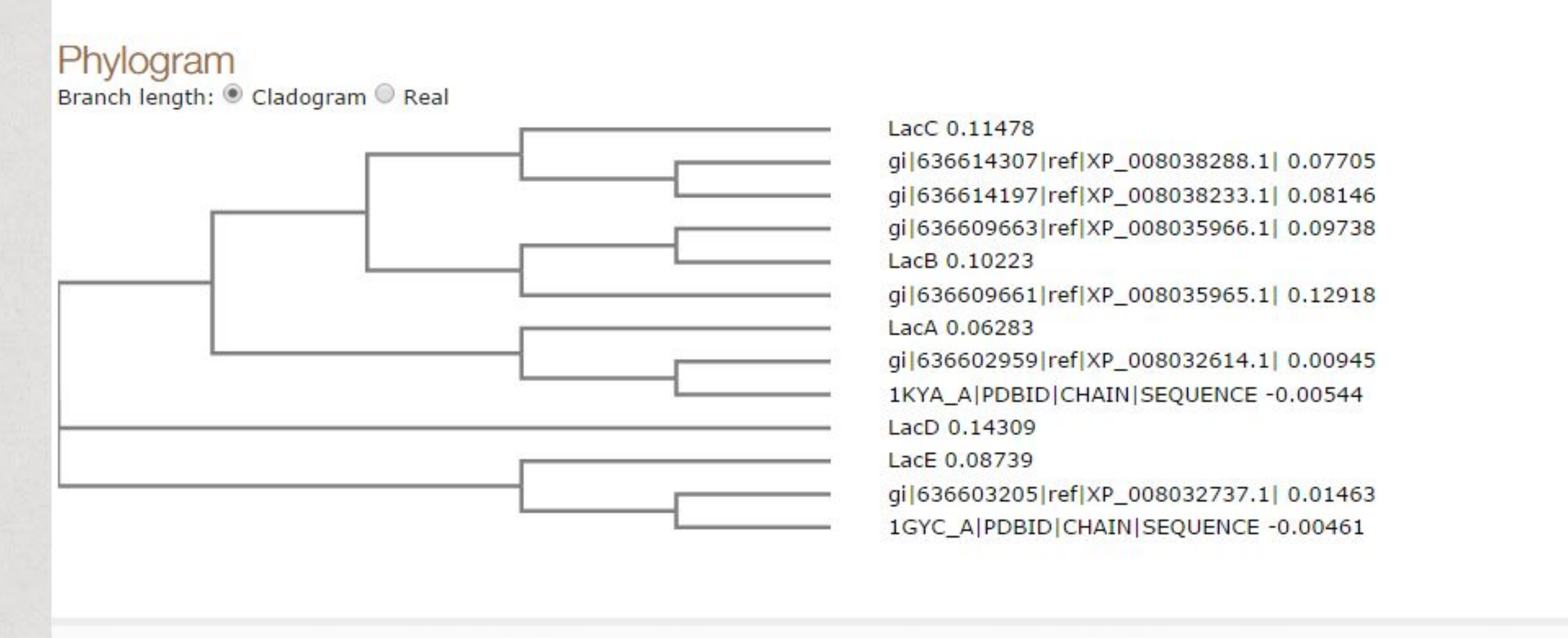


Results for job muscle-I20141001-133924-0133-44036811-pg

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

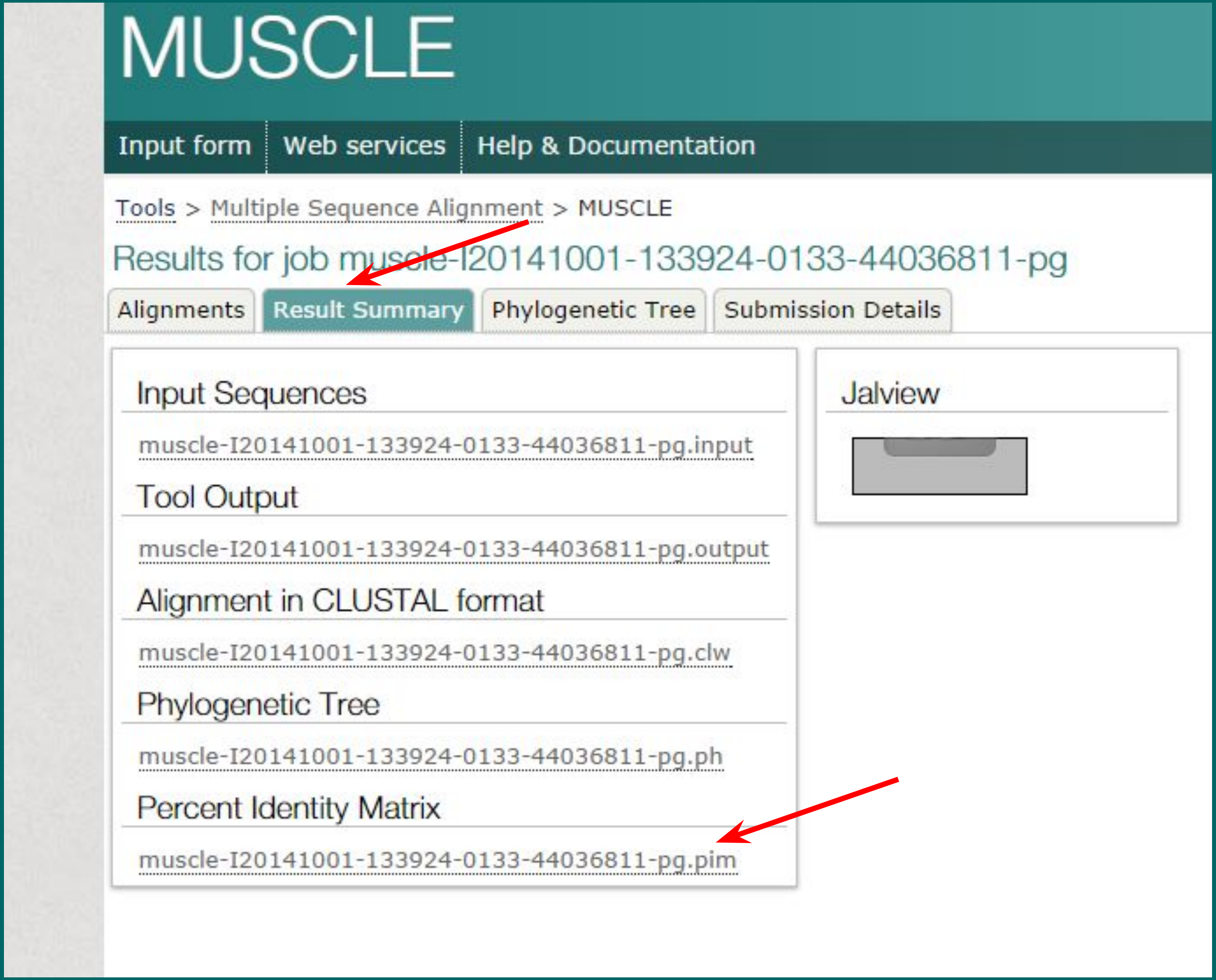
```
(
(
(
(
LacC:0.11478,
(
gi|636614307|ref|XP_008038288.1|:0.07705,
gi|636614197|ref|XP_008038233.1|:0.08146)
:0.02355)
:0.09463,
(
(
gi|636609663|ref|XP_008035966.1|:0.09738,
LacB:0.10223)
:0.00755,
gi|636609661|ref|XP_008035965.1|:0.12918)
:0.03983)
:0.03434,
(
LacA:0.06283,
(
gi|636602959|ref|XP_008032614.1|:0.00945,
1KYA_A|PDBID|CHAIN|SEQUENCE:-0.00544)
:0.04993)
:0.05736)
:0.01157,
LacD:0.14309,
(
LacE:0.08739,
(
gi|636603205|ref|XP_008032737.1|:0.01463,
1GYC_A|PDBID|CHAIN|SEQUENCE:-0.00461)
:0.07235)
```



Phylogram

Branch length: ● Cladogram ○ Real

Статистика выравнивания



MUSCLE

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > MUSCLE

Results for job muscle-I20141001-133924-0133-44036811-pg

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Input Sequences

muscle-I20141001-133924-0133-44036811-pg.input

Tool Output

muscle-I20141001-133924-0133-44036811-pg.output

Alignment in CLUSTAL format

muscle-I20141001-133924-0133-44036811-pg.clw

Phylogenetic Tree

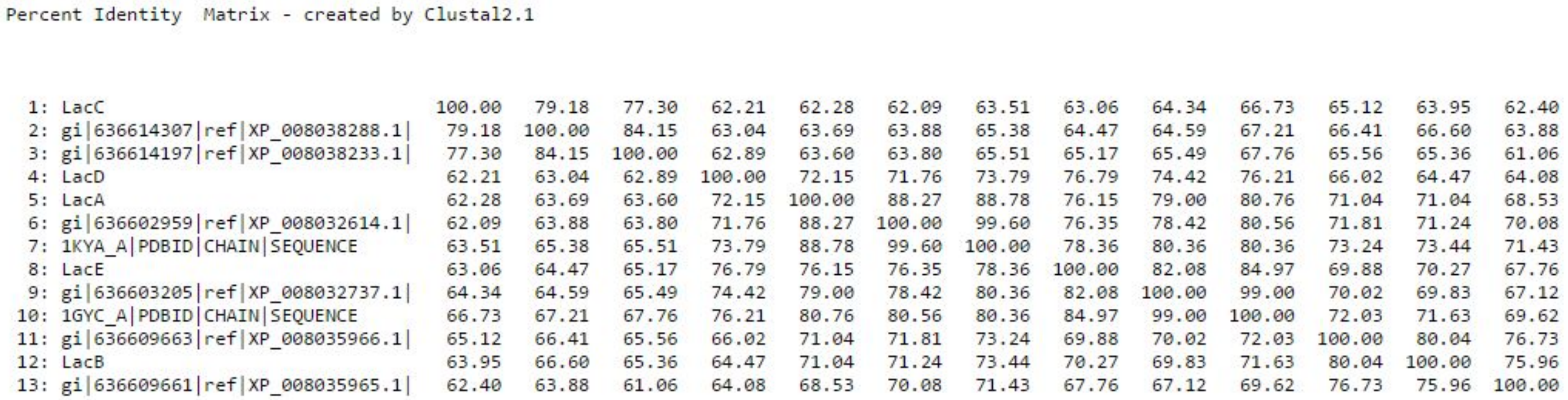
muscle-I20141001-133924-0133-44036811-pg.ph

Percent Identity Matrix

muscle-I20141001-133924-0133-44036811-pg.pim

Jalview

Таблица



```
Percent Identity  Matrix - created by Clustal2.1


  1: LacC                               100.00   79.18   77.30   62.21   62.28   62.09   63.51   63.06   64.34   66.73   65.12   63.95   62.40
  2: gi|636614307|ref|XP_008038288.1|    79.18  100.00   84.15   63.04   63.69   63.88   65.38   64.47   64.59   67.21   66.41   66.60   63.88
  3: gi|636614197|ref|XP_008038233.1|    77.30   84.15  100.00   62.89   63.60   63.80   65.51   65.17   65.49   67.76   65.56   65.36   61.06
  4: LacD                                62.21   63.04   62.89  100.00   72.15   71.76   73.79   76.79   74.42   76.21   66.02   64.47   64.08
  5: LacA                                62.28   63.69   63.60   72.15  100.00   88.27   88.78   76.15   79.00   80.76   71.04   71.04   68.53
  6: gi|636602959|ref|XP_008032614.1|    62.09   63.88   63.80   71.76   88.27  100.00   99.60   76.35   78.42   80.56   71.81   71.24   70.08
  7: 1KYA_A|PDBID|CHAIN|SEQUENCE         63.51   65.38   65.51   73.79   88.78   99.60  100.00   78.36   80.36   80.36   73.24   73.44   71.43
  8: LacE                                63.06   64.47   65.17   76.79   76.15   76.35   78.36  100.00   82.08   84.97   69.88   70.27   67.76
  9: gi|636603205|ref|XP_008032737.1|    64.34   64.59   65.49   74.42   79.00   78.42   80.36   82.08  100.00   99.00   70.02   69.83   67.12
 10: 1GYC_A|PDBID|CHAIN|SEQUENCE         66.73   67.21   67.76   76.21   80.76   80.56   80.36   84.97   99.00  100.00   72.03   71.63   69.62
 11: gi|636609663|ref|XP_008035966.1|    65.12   66.41   65.56   66.02   71.04   71.81   73.24   69.88   70.02   72.03  100.00   80.04   76.73
 12: LacB                                63.95   66.60   65.36   64.47   71.04   71.24   73.44   70.27   69.83   71.63   80.04  100.00   75.96
 13: gi|636609661|ref|XP_008035965.1|    62.40   63.88   61.06   64.08   68.53   70.08   71.43   67.76   67.12   69.62   76.73   75.96  100.00
```

Задание на дом

Провести выравнивание последовательностей из файла (Muscle)

Дать статистику выравнивания: описать, какие белки имеют большее сходство (по результатам таблицы).

Найти участки консервативные (глазами и руками☺)

Поиск консервативных участков глазами и руками ☺

```
CcL1A65         GPDGVTQCPIAQSGDSFTYSFDAGNEAGTFWYHSHYGTQYCDGLRGPLVIYDDNDPYKNL
Cc1HFU_A        GADGVNQCPIS-PGHAFLYKFTPAGHAGTFWYHSHFGTQYCDGLRGPMVIYDDNDPHAAL
LacC_Trame      GPAFVTQCPII-AGNDFLYNFQVPDQTGTYWYHSHLATQYCDGLRGPLVIYDPHDPHKHL
Tsp3KW7_A       GPAFVNQCPIA-SGNSFLYDFTVPDQAGTFWYHSHLSTQYCDGLRGPLVVYDPSDPYASM
LacB_Trame      GAAFVNQCPIA-SGNSFLYDFNVTDQAGTFWYHSHLSTQYCDGLRGPMVVYDPNDPHADL
LacD_Trame      GPAFVNQCPIA-SGNSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGAMVVYDPFDFQRHL
Tt2HRH_A        GPAFVNQCPIS-TGHAFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPIVVYDPQDPHKSL
TvL1KYA         GPAFINQCPIS-SGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPAADL
Tv1KYA_A        GPAFINQCPIS-SGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPAADL
Th3V9C          GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
Th3FPX          GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
LacA_Trame      GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
LacE_Trame      GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPMVVYDPKDPQAYL
TvL1GYC         GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASR
Tv1GYC_A_A      GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASR
                *.  :.****  .*  * *.*   . :**:***** .*********.:*:**  *
```

«*» - строго одна и та же ак у всех последовательностей;

« . » - замена ак из одной функциональной категории

« : » - замена ак из разных функциональных категорий

Система оценки - белки

Category	**Amino Acid**
Кислоты\амиды	Asp (D) Glu(E) Asn (N) Gln (Q)
Основания	His (H) Lys (K) Arg (R)
Ароматические	Phe (F) Tyr (Y) Trp (W)
Гидрофильные	Ala (A) Cys (C) Gly (G) Pro (P) Ser (S) Thr (T)
Гидрофобные	Ile (I) Leu (L) Met (M) Val (V)

Задание на дом

Провести выравнивание последовательностей из файла (Muscle)

Дать статистику выравнивания:

Найти участки консервативные (глазами и руками☺)

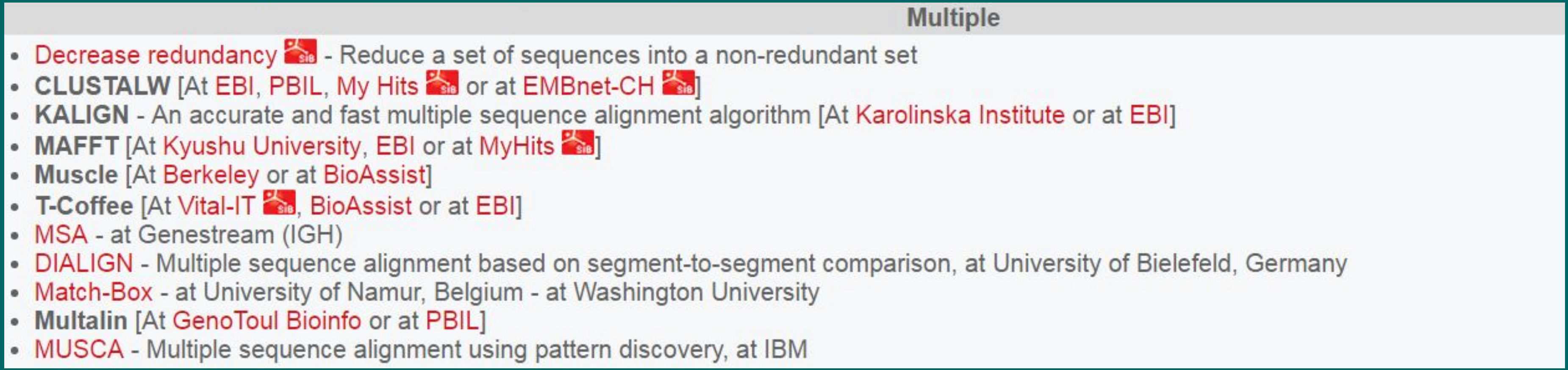
Построить LOGO консервативных участков (Weblogo) (не менее трех участков, а лучше все, какие найдете.

Задание на дом

Провести выравнивание последовательностей из файла
с использованием других программ:
T-coffee
ClustalW
ProbCons

Сравнить результаты выравнивания последовательностей разными программами (есть ли отличия – сделать
и описать скриншоты выравниваний)

Программы на http://www.expasy.org/tools/



Multiple

- Decrease redundancy - Reduce a set of sequences into a non-redundant set
- **CLUSTALW** [At EBI, PBIL, My Hits or at EMBnet-CH]
- **KALIGN** - An accurate and fast multiple sequence alignment algorithm [At Karolinska Institute or at EBI]
- **MAFFT** [At Kyushu University, EBI or at MyHits]
- **Muscle** [At Berkeley or at BioAssist]
- **T-Coffee** [At Vital-IT, BioAssist or at EBI]
- MSA - at Genestream (IGH)
- DIALIGN - Multiple sequence alignment based on segment-to-segment comparison, at University of Bielefeld, Germany
- Match-Box - at University of Namur, Belgium - at Washington University
- **Multalin** [At GenoToul Bioinfo or at PBIL]
- MUSCA - Multiple sequence alignment using pattern discovery, at IBM

Построение Logo
Weblogo –

http://weblogo.berkeley.edu/

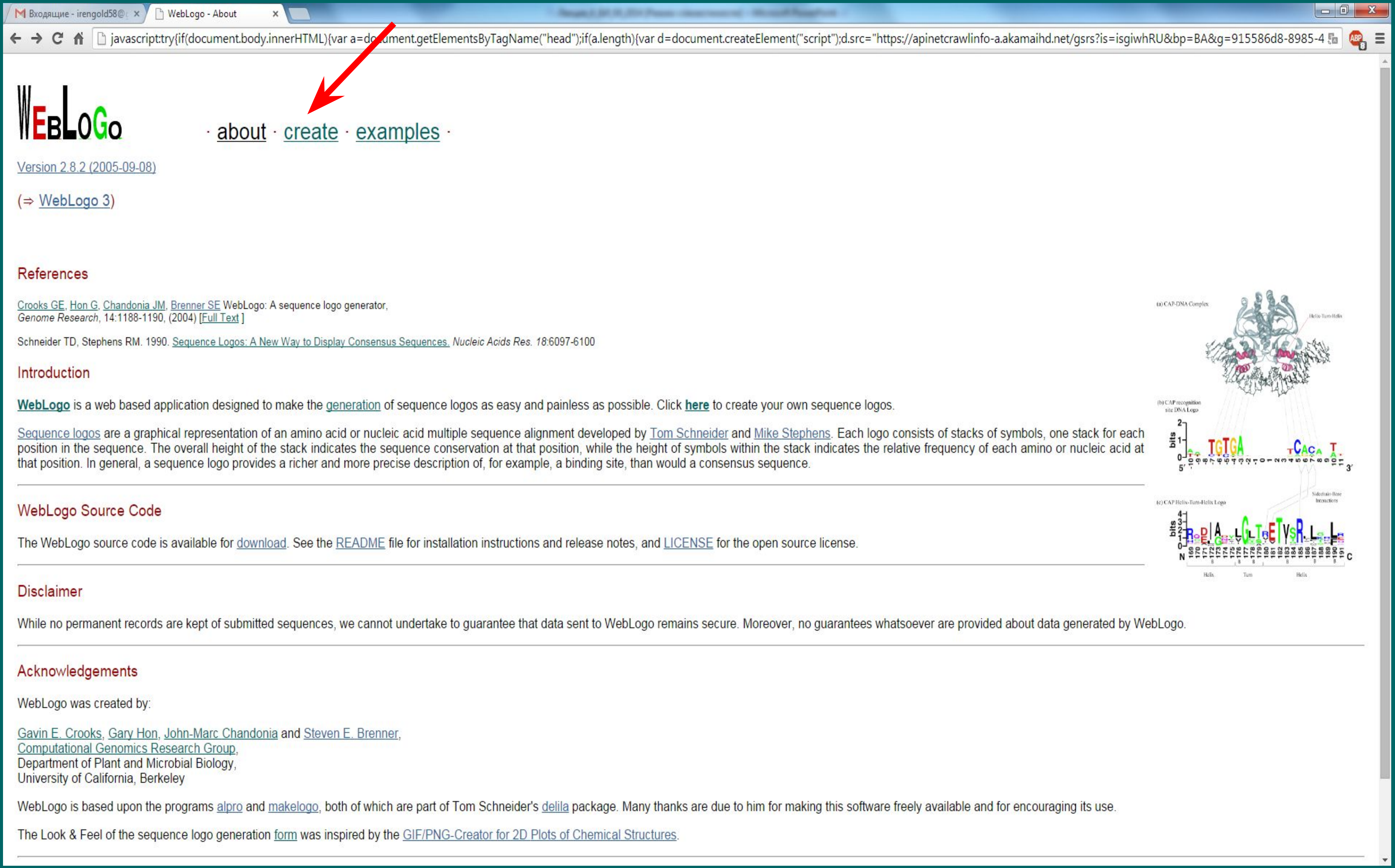
1 этап – определяем консервативный участок

```
CcL1A65         GPDGVTQCPIAQSGDSFTYSFDAGNEAGTFWYHSHYGTQYCDGLRGPLVIYDDNDPYKNL
Cc1HFU_A        GADGVNQCPIS-PGHAFLYKFTPAGHAGTFWYHSHFGTQYCDGLRGPMVIYDDNDPHAAL
LacC_Trame      GPAFVTQCPII-AGNDFLYNFQVPDQTGTYWYHSHLATQYCDGLRGPLVIYDPHDPHKHL
Tsp3KW7_A       GPAFVNQCPIA-SGNSFLYDFTVPDQAGTFWYHSHLSTQYCDGLRGPLVVYDPSDPYASM
LacB_Trame      GAAFVNQCPIA-SGNSFLYDFNVTDQAGTFWYHSHLSTQYCDGLRGPMVVYDPNDPHADL
LacD_Trame      GPAFVNQCPIA-SGNSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGAMVVYDPFDFQRHL
Tt2HRH_A        GPAFVNQCPIS-TGHAFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPIVVYDPQDPHKSL
TvL1KYA         GPAFINQCPIS-SGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPAADL
Tv1KYA_A        GPAFINQCPIS-SGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPAADL
Th3V9C          GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
Th3FPX          GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
LacA_Trame      GPAFINQCPIS-PGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPHASR
LacE_Trame      GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPMVVYDPKDPQAYL
TvL1GYC         GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASR
Tv1GYC_A_A      GPAFVNQCPIA-SGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASR
                *.  :.****  .*  * *.*   . :**:***** .*********.:*:**  *
```

2 этап – выбираем общее выравнивание этого участка

```
GTFWYHSHLSTQYCDGLRGPL
GTFWYHSHLSTQYCDGLRGPI
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHFGTQYCDGLRGPM
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPM
GTYWYHSHLATQYCDGLRGPL
GTFWYHSHLSTQYCDGLRGAM
GTFWYHSHLSTQYCDGLRGPM
```

3 этап – запускаем WebLogo



· about · create · examples ·

Version 2.8.2 (2005-09-08)

(⇒ WebLogo 3)

References

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator,
Genome Research, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res. 18*:6097-6100

Introduction

WebLogo is a web based application designed to make the generation of sequence logos as easy and painless as possible. Click **here** to create your own sequence logos.

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

WebLogo Source Code

The WebLogo source code is available for download. See the README file for installation instructions and release notes, and LICENSE for the open source license.

Disclaimer

While no permanent records are kept of submitted sequences, we cannot undertake to guarantee that data sent to WebLogo remains secure. Moreover, no guarantees whatsoever are provided about data generated by WebLogo.

Acknowledgements

WebLogo was created by:

Gavin E. Crooks, Gary Hon, John-Marc Chandonia and Steven E. Brenner,
Computational Genomics Research Group,
Department of Plant and Microbial Biology,
University of California, Berkeley

WebLogo is based upon the programs alpro and makelogo, both of which are part of Tom Schneider's delila package. Many thanks are due to him for making this software freely available and for encouraging its use.

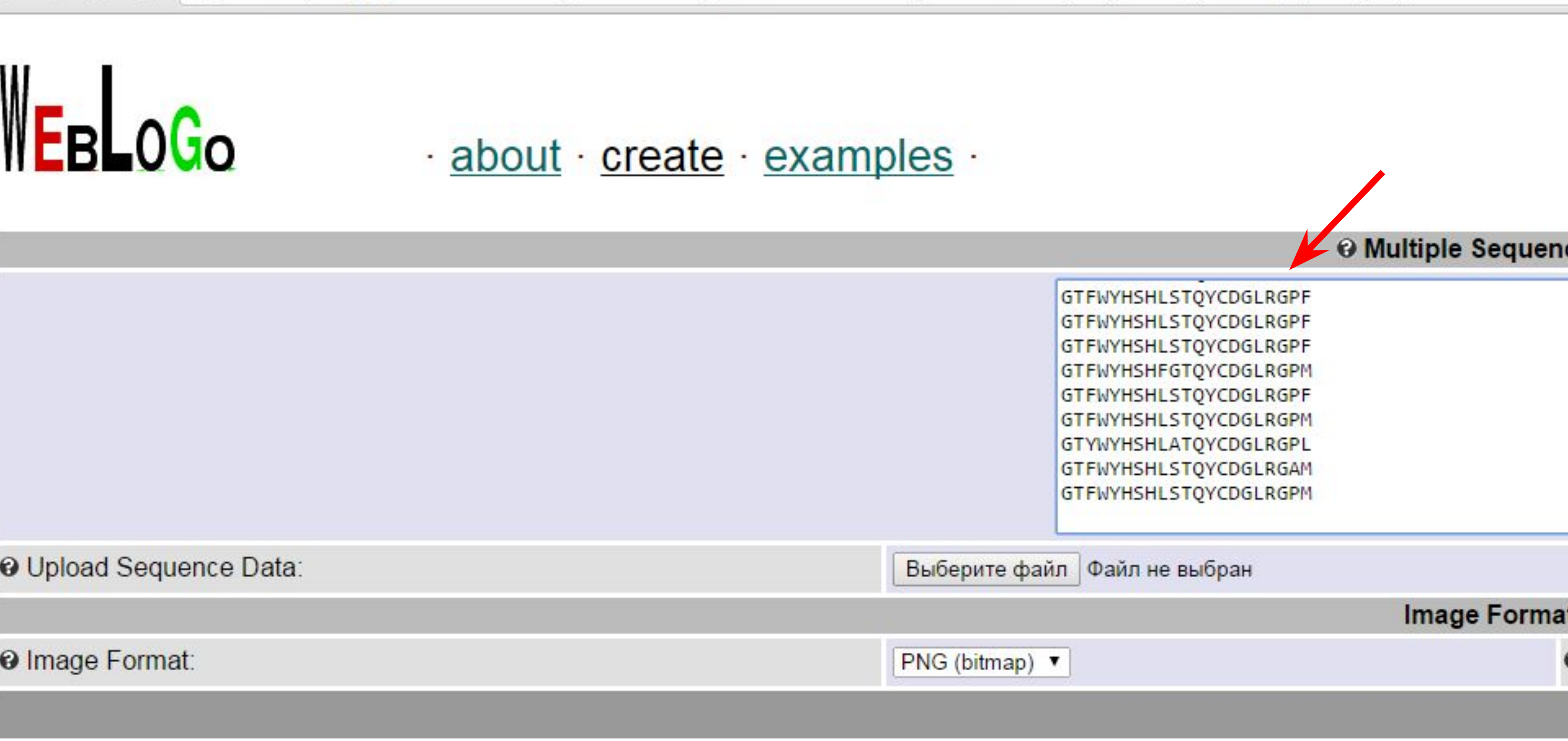
The Look & Feel of the sequence logo generation form was inspired by the GIF/PNG-Creator for 2D Plots of Chemical Structures.

4 этап – последовательности вставляем в окно

Upload Sequence Data: Выберите файл Файл не выбран

Image Format: PNG (bitmap)

4 этап – последовательности вставляем в окно

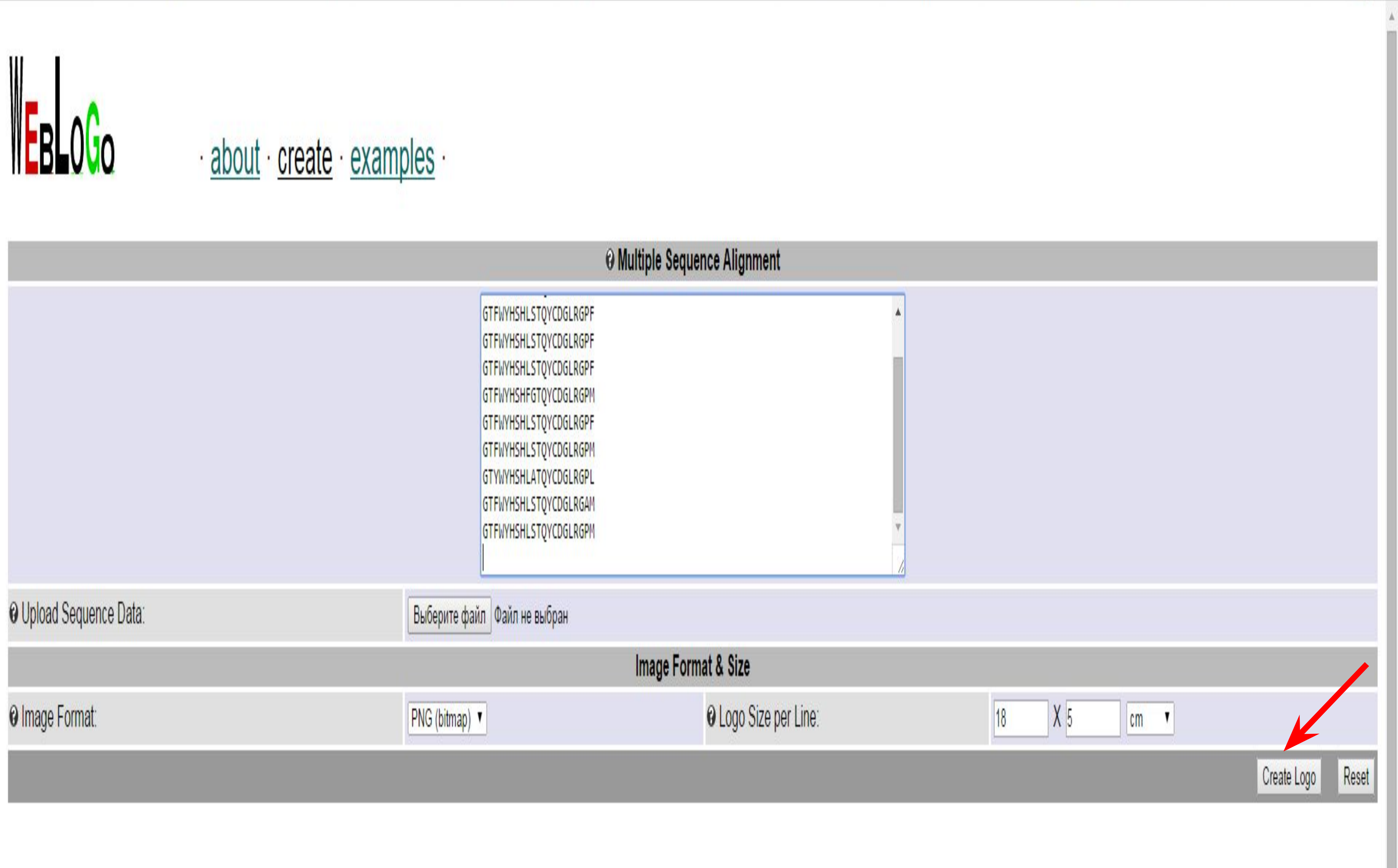


```
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHFGTQYCDGLRGPM
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPM
GTYWYHSHLATQYCDGLRGPL
GTFWYHSHLSTQYCDGLRGAM
GTFWYHSHLSTQYCDGLRGPM
```

Upload Sequence Data: Выберите файл Файл не выбран

Image Format: PNG (bitmap)

5 этап – создание Logo



Multiple Sequence Alignment

```
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHFGTQYCDGLRGPM
GTFWYHSHLSTQYCDGLRGPF
GTFWYHSHLSTQYCDGLRGPM
GTYWYHSHLATQYCDGLRGPL
GTFWYHSHLSTQYCDGLRGAM
GTFWYHSHLSTQYCDGLRGPM
```

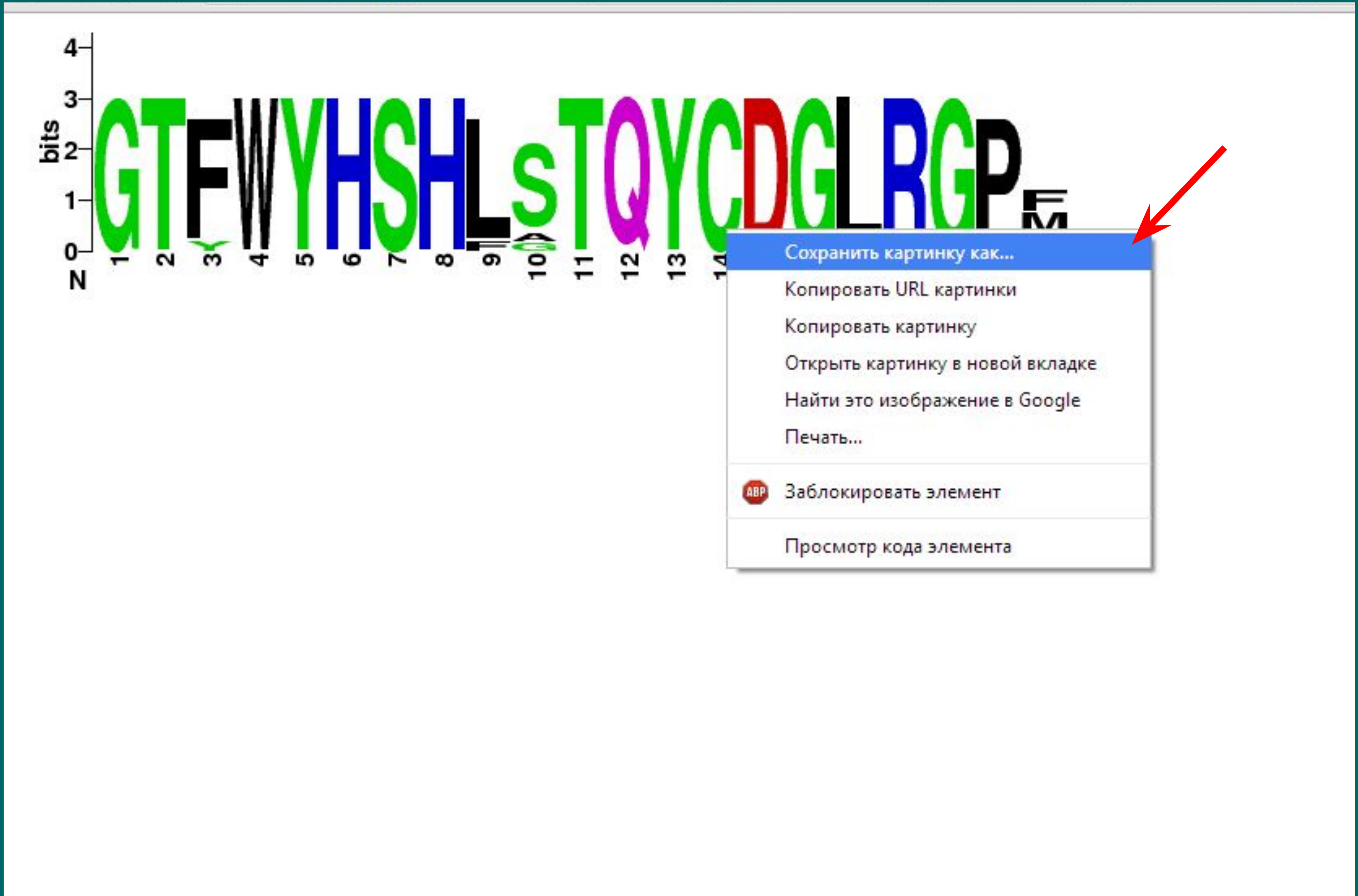
Upload Sequence Data: Выберите файл Файл не выбран

Image Format & Size

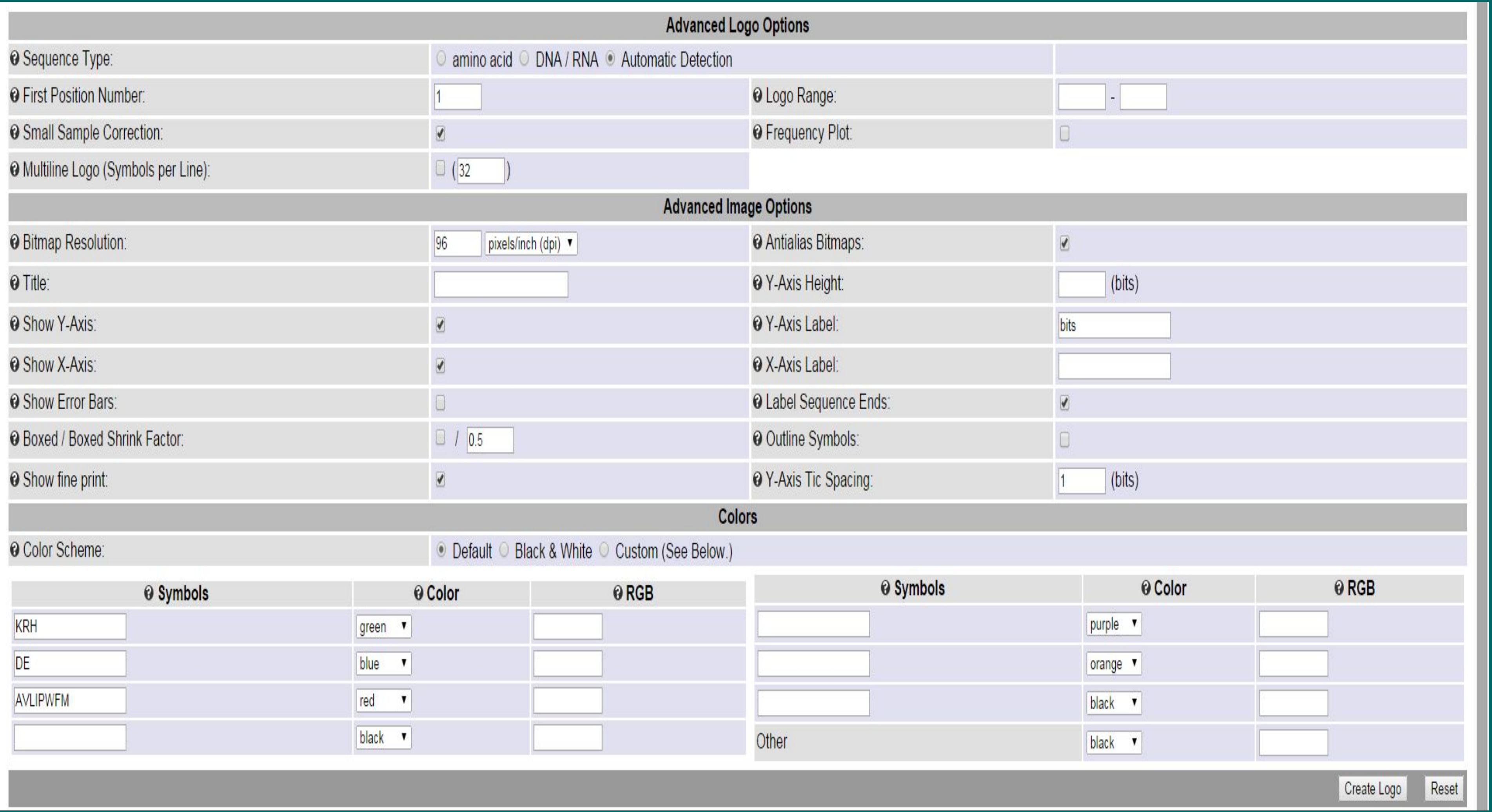
Image Format: PNG (bitmap)

Logo Size per Line: 18 X 5 cm

6 этап – готовый Logo



Можно поиграть с настройками (при желании ☺)



Advanced Logo Options

Sequence Type:	○ amino acid ○ DNA / RNA ● Automatic Detection		
First Position Number:	1	Logo Range:	-
Small Sample Correction:	☑	Frequency Plot:	☐
Multiline Logo (Symbols per Line):	☐ (32)		

Advanced Image Options

Bitmap Resolution:	96 pixels/inch (dpi)	Antialias Bitmaps:	☑
Title:		Y-Axis Height:	(bits)
Show Y-Axis:	☑	Y-Axis Label:	bits
Show X-Axis:	☑	X-Axis Label:	
Show Error Bars:	☐	Label Sequence Ends:	☑
Boxed / Boxed Shrink Factor:	☐ / 0.5	Outline Symbols:	☐
Show fine print:	☑	Y-Axis Tic Spacing:	1 (bits)

Colors

Color Scheme: ● Default ○ Black & White ○ Custom (See Below.)

Symbols	Color	RGB	Symbols	Color	RGB
KRH	green			purple	
DE	blue			orange	
AVLIPWFM	red			black	
	black		Other	black	

Задание на дом

Провести выравнивание последовательностей из файла (разные программы Muscle, T-coffee, ClustalW, ProbCons)

-Сравнить результаты выравнивания, полученные разными программами

-Сделать отдельный файл (Word) с результатами выравнивания Muscle

-Описать результат (выделить области, имеющие идентичность и сходство, где они локализованы)

Дать статистику выравнивания: описать, какие белки имеют большее сходство (по результатам таблицы).

Построить LOGO консервативных участков (Weblogo) (не менее трех участков, а лучше все, какие найдете.

Высылаете 2 файла – один в Word с результатами выравнивания Muscle; второй – в ppt – остальные результаты