Section II

The Structure of Macromolecules: *Human muscarinic M3 and bovine rhodopsin sequences alignment - tutorial*

Liliana Halip, Dan C. Fara and Tudor I. Oprea

Objective

To learn how to use T-Coffee for sequences alignment.

Brief info

A sequence alignment consists in the primary sequences arrangement of proteins, DNA or RNA, in order to identify regions of similarity that may be a result of functional, structural, or evolutionary relationships between those sequences. Aligned sequences are typically represented as rows within a matrix, and identical or similar characters must be aligned in successive columns. Gaps may be inserted between the residues if it is necessary. There are several web-based programs that can be used for sequences alignment. The main ones are BLAST[1,2], ClustalW, and T-Coffee[3].

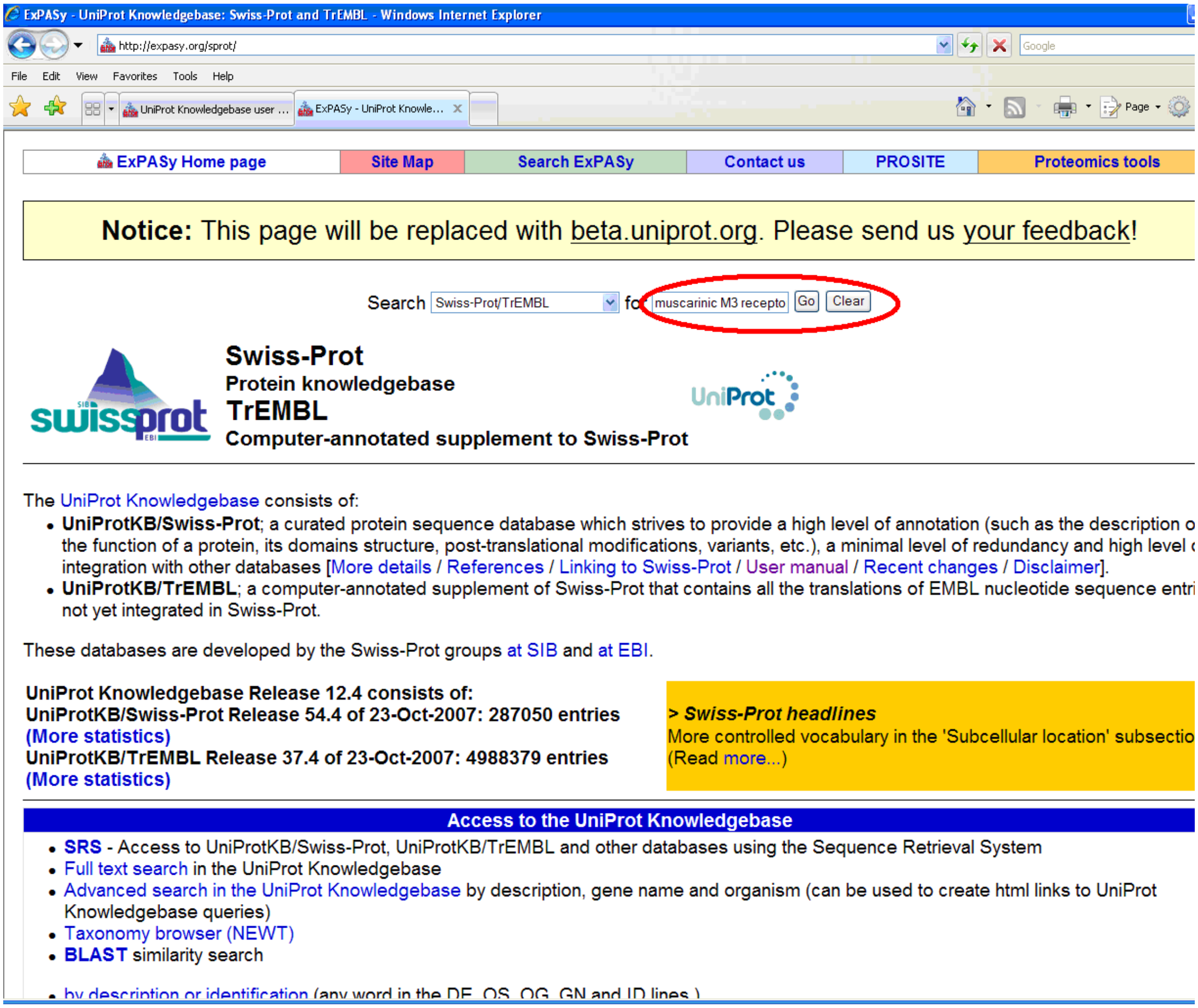
T-Coffee runs by assembling a library - a list of potential pairs of residues and then turns this library into an alignment. Each library line is a constraint and the purpose is to assemble the alignment that accommodates the more all the constraints.

For this tutorial, the human muscarinic M3 (hM3) and bovine rhodopsin (RHO) sequences were chosen to be aligned using T-Coffee. The bovine rhodopsin is the only protein belonging to the same family with hM3 receptor (G-protein coupled receptors, GPCR) having the 3D structure available from experimental studies (X-RAY or RMN studies).

How to

1. Obtain the proteins sequences

- go to Swiss-PROT[4,5] (the protein sequence database).
- enter the protein name (e.g. muscarinic M3 receptor) in the query field - see Figure 1; you will be directed to a new page where you can find the results of your query.



Notice: This page will be replaced with beta.uniprot.org. Please send us your feedback!

Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description o the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level (integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entri not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 12.4 consists of:
UniProtKB/Swiss-Prot Release 54.4 of 23-Oct-2007: 287050 entries
(More statistics)
UniProtKB/TrEMBL Release 37.4 of 23-Oct-2007: 4988379 entries
(More statistics)

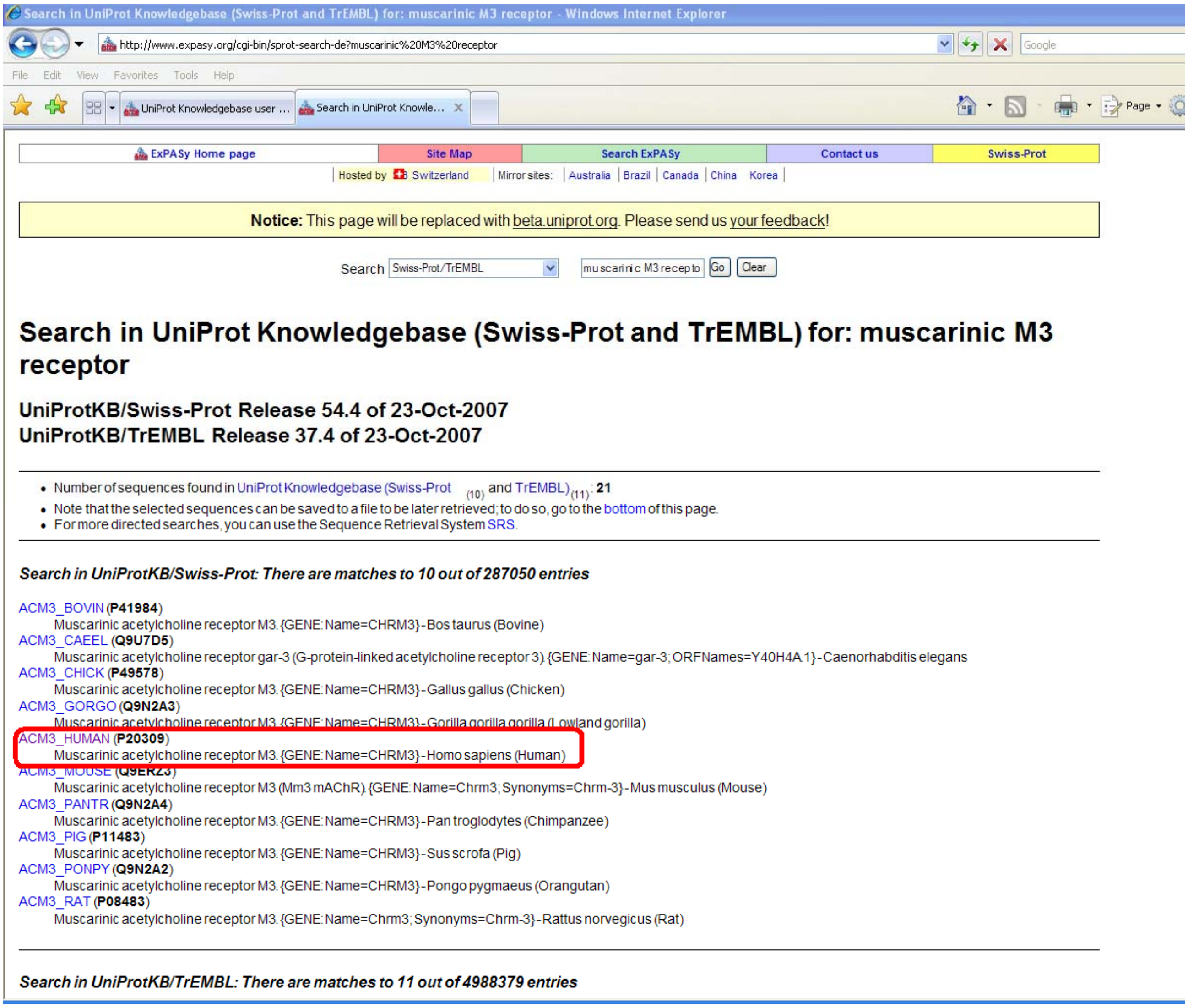
> Swiss-Prot headlines
More controlled vocabulary in the 'Subcellular location' subsectio
(Read more...)

Access to the UniProt Knowledgebase

- **SRS** - Access to UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and other databases using the Sequence Retrieval System
- Full text search in the UniProt Knowledgebase
- Advanced search in the UniProt Knowledgebase by description, gene name and organism (can be used to create html links to UniProt Knowledgebase queries)
- Taxonomy browser (NEWT)
- **BLAST** similarity search
- by description or identification (any word in the DE, OS, OG, GN and ID lines.)

Figure 1

- select/open the record for the human M3 sequence (SWISS PROT ID: P20309) – see Figure 2. At the end of this page, in the sequence section, you have the option to view the protein sequence as FASTA format – see Figure 3.


Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: muscarinic M3 receptor - Windows Internet Explorer
http://www.expasy.org/cgi-bin/sprot-search-de?muscarinic%20M3%20receptor
Google
File
Edit
View
Favorites
Tools
Help
UniProt Knowledgebase user ...
Search in UniProt Knowle...
Page
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Hosted by Switzerland
Mirror sites:
Australia
Brazil
Canada
China
Korea

Notice: This page will be replaced with beta.uniprot.org. Please send us your feedback!

Search: Swiss-Prot/TrEMBL | muscarinic M3 recepto | Go | Clear

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: muscarinic M3 receptor

UniProtKB/Swiss-Prot Release 54.4 of 23-Oct-2007
UniProtKB/TrEMBL Release 37.4 of 23-Oct-2007

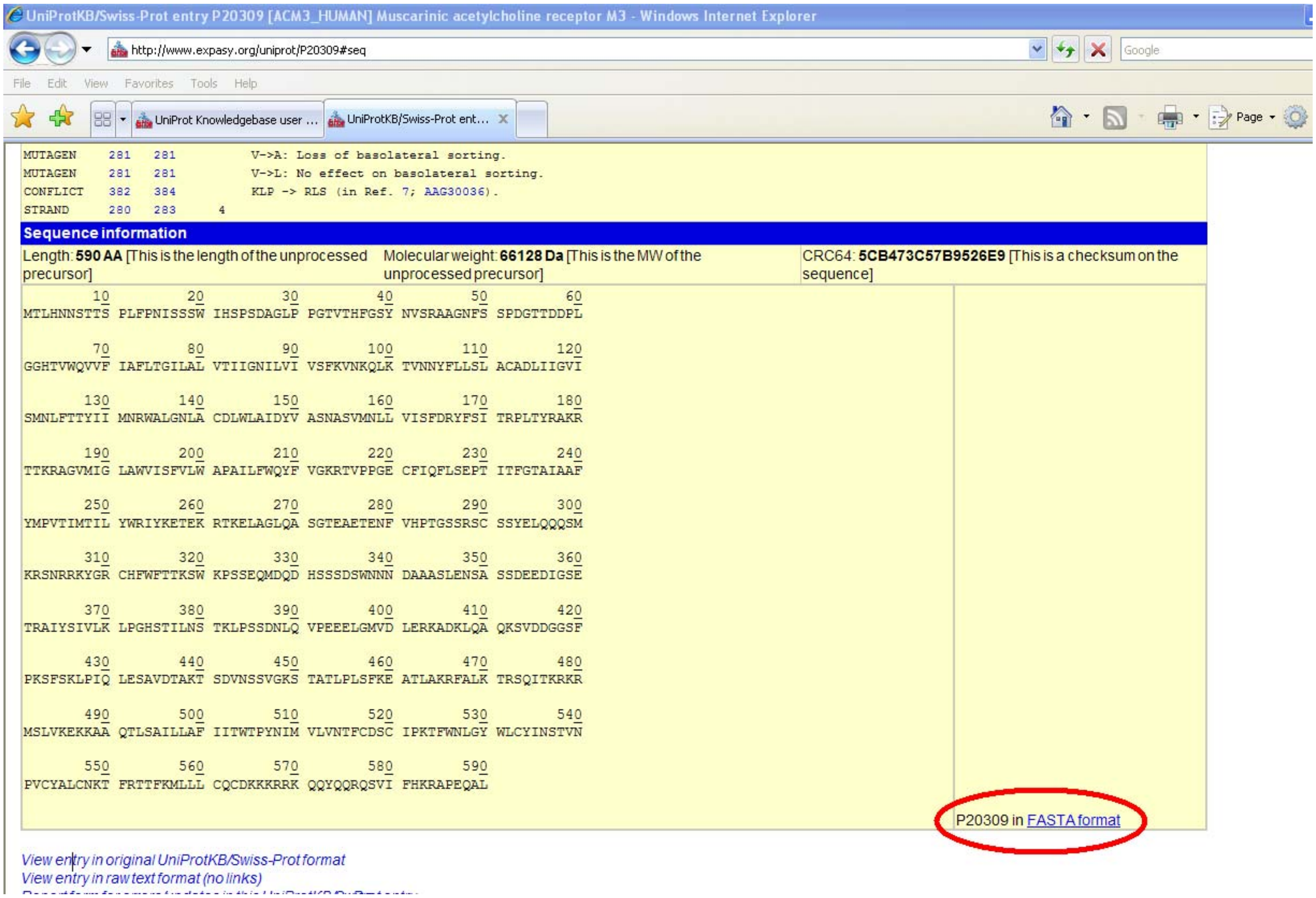
- Number of sequences found in UniProt Knowledgebase (Swiss-Prot $_{(10)}$ and TrEMBL) $_{(11)}$: **21**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 10 out of 287050 entries

ACM3_BOVIN (**P41984**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Bos taurus (Bovine)
ACM3_CAEEL (**Q9U7D5**)
Muscarinic acetylcholine receptor gar-3 (G-protein-linked acetylcholine receptor 3). {GENE: Name=gar-3; ORFNames=Y40H4A.1} - Caenorhabditis elegans
ACM3_CHICK (**P49578**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Gallus gallus (Chicken)
ACM3_GORGO (**Q9N2A3**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Gorilla gorilla gorilla (Lowland gorilla)
ACM3_HUMAN (**P20309**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Homo sapiens (Human)
ACM3_MOUSE (**Q9ERZ3**)
Muscarinic acetylcholine receptor M3 (Mm3 mAChR). {GENE: Name=Chrm3; Synonyms=Chrm-3} - Mus musculus (Mouse)
ACM3_PANTR (**Q9N2A4**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Pan troglodytes (Chimpanzee)
ACM3_PIG (**P11483**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Sus scrofa (Pig)
ACM3_PONPY (**Q9N2A2**)
Muscarinic acetylcholine receptor M3. {GENE: Name=CHRM3} - Pongo pygmaeus (Orangutan)
ACM3_RAT (**P08483**)
Muscarinic acetylcholine receptor M3. {GENE: Name=Chrm3; Synonyms=Chrm-3} - Rattus norvegicus (Rat)

Search in UniProtKB/TrEMBL: There are matches to 11 out of 4988379 entries

Figure 2


UniProtKB/Swiss-Prot entry P20309 [ACM3_HUMAN] Muscarinic acetylcholine receptor M3 - Windows Internet Explorer
http://www.expasy.org/uniprot/P20309#seq
Google
File Edit View Favorites Tools Help
UniProt Knowledgebase user ...
UniProtKB/Swiss-Prot ent...
Page

```
MUTAGEN     281   281          V->A: Loss of basolateral sorting.
MUTAGEN     281   281          V->L: No effect on basolateral sorting.
CONFLICT    382   384          KLP -> RLS (in Ref. 7; AAG30036).
STRAND      280   283      4
```

Sequence information

Length: **590 AA** [This is the length of the unprocessed precursor] Molecular weight: **66128 Da** [This is the MW of the unprocessed precursor] CRC64: **5CB473C57B9526E9** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MTLHNNSTTS PLFPNISSSW IHSPSDAGLP PGTVTHFGSY NVSRAAGNFS SPDGTTDDPL

        70         80         90        100        110        120
GGHTVWQVVF IAFLTGILAL VTIIGNILVI VSFKVNKQLK TVNNYFLLSL ACADLIIGVI

       130        140        150        160        170        180
SMNLFTTYII MNRWALGNLA CDLWLAIDYV ASNASVMNLL VISFDRYFSI TRPLTYRAKR

       190        200        210        220        230        240
TTKRAGVMIG LAWVISFVLW APAILFWQYF VGKRTVPPGE CFIQFLSEPT ITFGTAIAAF

       250        260        270        280        290        300
YMPVTIMTIL YWRIYKETEK RTKELAGLQA SGTEAETENF VHPTGSSRSC SSYELQQQSM

       310        320        330        340        350        360
KRSNRRKYGR CHFWFTTKSW KPSSEQMDQD HSSSDSWNNN DAAASLENSA SSDEEDIGSE

       370        380        390        400        410        420
TRAIYSIVLK LPGHSTILNS TKLPSSDNLQ VPEEELGMVD LERKADKLQA QKSVDDGGSF

       430        440        450        460        470        480
PKSFSKLPIQ LESAVDTAKT SDVNSSVGKS TATLPLSFKE ATLAKRFALK TRSQITKRKR

       490        500        510        520        530        540
MSLVKEKKAA QTLSAILLAF IITWTPYNIM VLVNTFCDSC IPKTFWNLGY WLCYINSTVN

       550        560        570        580        590
PVCYALCNKT FRTTFKMLLL CQCDKKKRRK QQYQQRQSVI FHKRAPEQAL
```

P20309 in FASTA format

View entry in original UniProtKB/Swiss-Prot format
View entry in raw text format (no links)

Figure 3

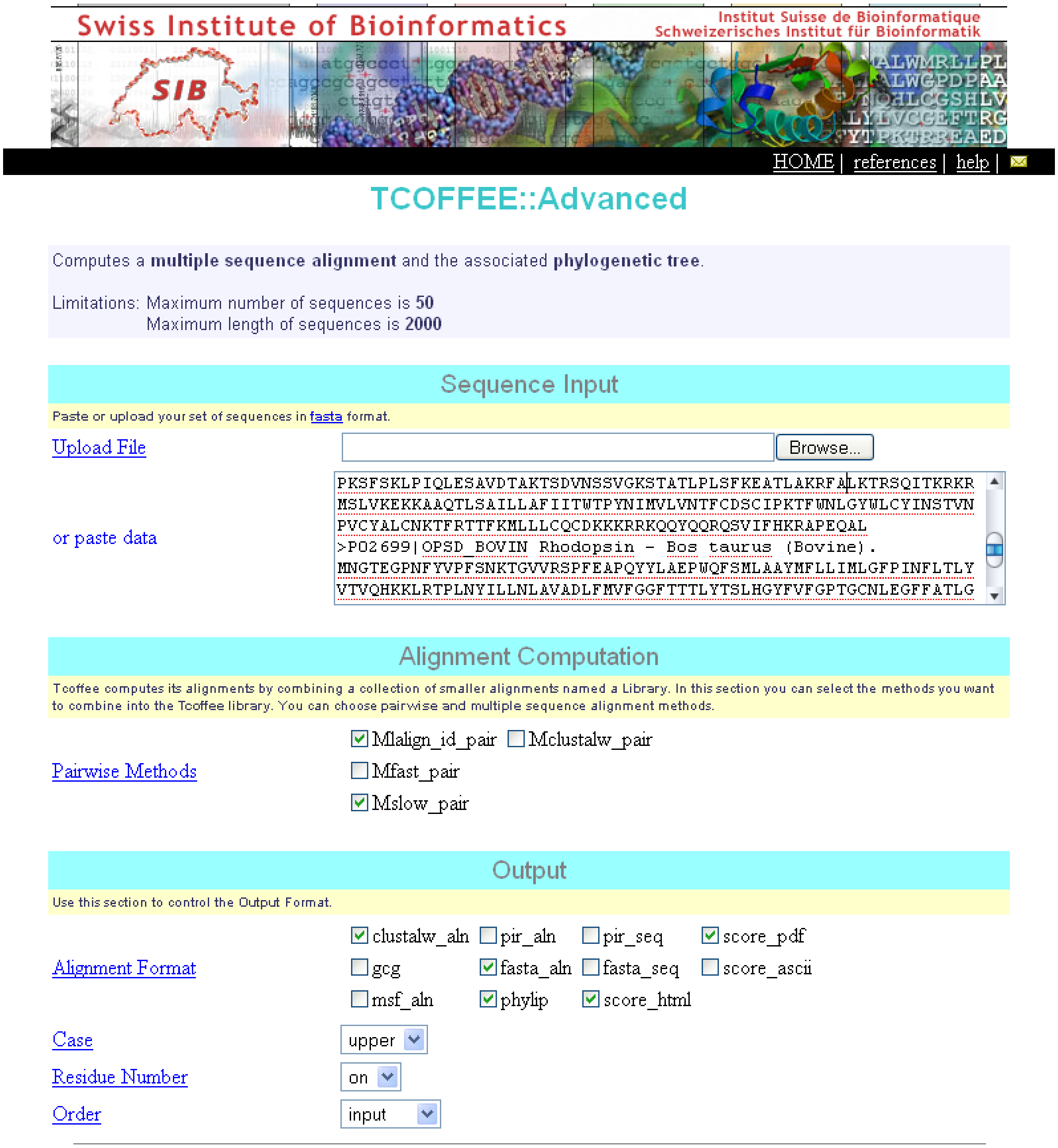
- save the sequence (in FASTA format) as a text file.
- repeat the above-given steps for bovine rhodopsin and save the sequence OPSD_BOVIN (SWISS PROT ID: P02699)
- merge the two text files into a single one and save it as “sequences.txt”

2. Align the proteins sequences

- access the T-Coffee server

Note that the following five modules are available for proteins sequences alignment: (i) T-Coffee, (ii) Expresso (it replaces 3DCoffee), (iii) M-Coffee, (iv) Rcoffee (beta version), and (v) Combine. *T-Coffee* computes a multiple sequence alignment and the associated phylogenetic tree. *Expresso* computes **structure** based multiple sequence alignments by running a BLAST alignment between every sequence in the query against the PDB database. If it finds one structure similar enough to a sequence in your dataset (>60% identity), it will use it as a template for your sequence. *M-Coffee* computes a multiple sequence alignment and the associated phylogenetic tree by combining the output of several multiple sequence alignment packages (PCMA, Poa, Mafft, Muscle, T-Coffee, ClustalW, ProbCons, DialignT). *Rcoffee* computes multiple sequence alignment of **non coding** RNA sequences using RNAplfold predicted secondary structures. *Combine* combines two (or more) multiple sequence alignments into a single one. Details about each of these modules can be found by following the link “cite” to the original papers.

- go to T-Coffee > Advanced and input the sequences in Fasta format. This can be done in two ways: (i) either upload the text file containing both sequences in Fasta format (sequences.txt), or (ii) paste both proteins sequences in Fasta format (no empty line between the sequences). Keep the default options for the "Alignment computation" and the "Output", as can be seen in Figure 4:


Swiss Institute of Bioinformatics
Institut Suisse de Bioinformatique
Schweizerisches Institut für Bioinformatik
SIB
HOME | references | help |

TCOFFEE::Advanced

Computes a **multiple sequence alignment** and the associated **phylogenetic tree**.

Limitations: Maximum number of sequences is **50**
Maximum length of sequences is **2000**

Sequence Input

Paste or upload your set of sequences in fasta format.

Upload File [ ] Browse...

or paste data

```
PKSFSKLPIQLESAVDTAKTSDVNSSVGKSTATLPLSFKEATLAKRFALKTRSQITKRKR
MSLVKEKKAAQTLSAILLAFIITWTPYNIMVLVNTFCDSCIPKTFWNLGYWLCYINSTVN
PVCYALCNKTFRTTFKMLLLCQCDKKKRRKQQYQQRQSVIFHKRAPEQAL
>P02699|OPSD_BOVIN Rhodopsin - Bos taurus (Bovine).
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
```

Alignment Computation

Tcoffee computes its alignments by combining a collection of smaller alignments named a Library. In this section you can select the methods you want to combine into the Tcoffee library. You can choose pairwise and multiple sequence alignment methods.

Pairwise Methods: ☑ Mlalign_id_pair ☐ Mclustalw_pair ☐ Mfast_pair ☑ Mslow_pair

Output

Use this section to control the Output Format.

Alignment Format: ☑ clustalw_aln ☐ pir_aln ☐ pir_seq ☑ score_pdf ☐ gcg ☑ fasta_aln ☐ fasta_seq ☐ score_ascii ☐ msf_aln ☑ phylip ☑ score_html

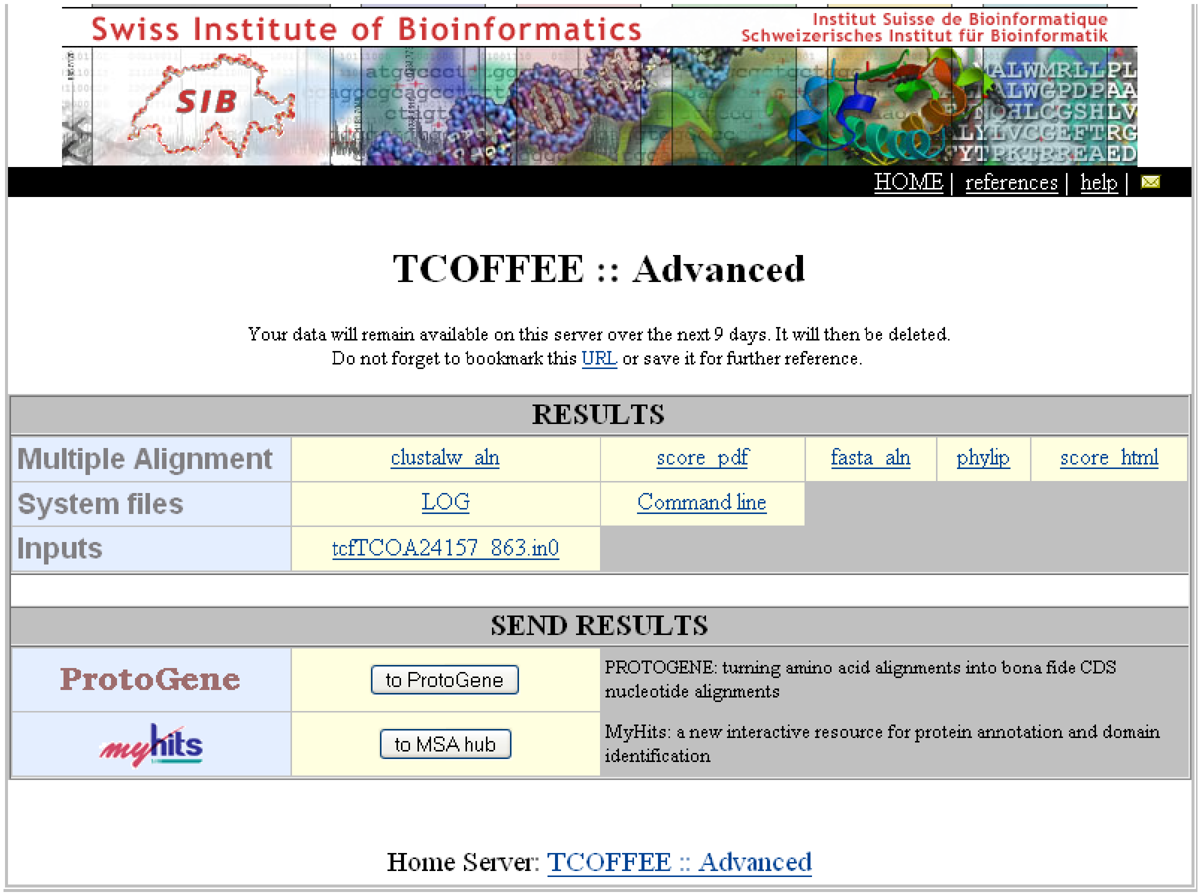
Case: upper

Residue Number: on

Order: input

Figure 4

- press the submit button and wait until the sequence alignment will be computed. The results are given in different formats – see Figure 5. The high-similarity regions in the alignment are marked yellow/red in the HTML or PDF format.


Swiss Institute of Bioinformatics
Institut Suisse de Bioinformatique
Schweizerisches Institut für Bioinformatik
SIB
HOME | references | help |

TCOFFEE :: Advanced

Your data will remain available on this server over the next 9 days. It will then be deleted.
Do not forget to bookmark this URL or save it for further reference.

RESULTS					
Multiple Alignment	clustalw_aln	score_pdf	fasta_aln	phylip	score_html
System files	LOG	Command line			
Inputs	tcfTCOA24157_863.in0				

SEND RESULTS		
ProtoGene	to ProtoGene	PROTOGENE: turning amino acid alignments into bona fide CDS nucleotide alignments
myhits	to MSA hub	MyHits: a new interactive resource for protein annotation and domain identification

Home Server: TCOFFEE :: Advanced

Figure 5

The sequence alignment is the most critical step in developing a homology model, for example a misalignment by just one residue will cause an error around 4Å in the model. For an accurate alignment the **structurally-conserved regions** (SCRs) have to be identified in both sequences. Structurally-conserved regions within a family proteins refer to the fragments for which an average structure or framework can be constructed for these regions of the proteins.

In rhodopsin-like family the highly conserved amino acids are[6]:

- Gly17, Asn 18 and Val21 on helix I,
- Asn or Ser9, Leu10, Ala11, Ala or Ser13, and Asp14 on helix II,
- Ser 14, Leu 18, Ile 21, Ser or Ala22, Asp or Glu24, Arg 25, Tyr 26, Ile or Val29, on helix III,
- Trp11, Ser or Ala14 and Pro20 on helix IV,
- Phe11, Pro14, Ile or Met18, Tyr22 and Ile or Val25 on helix V,
- Lys or Arg0, Phe12, Cys15, Trp16 and Pro18 on helix VI,
- Asn or Ser13, Ser or Cys14, Asn or Asp17, Pro18, Tyr21, Phe or Tyr28 and Arg or Lys29 on helix VII.

After highly conserved residues identification in the alignment (yellow background in figure 6) one can observe that T-coffee fails to find the common motif in the helices 5 (green background in figure 6). This is due to the fact that the third intracellular loops have obviously different lengths (218 residues in M3 receptor versus 25 residues in rhodopsin). To avoid this inconvenient, a new alignment is necessary, this time using a short sequence for hM3 receptor (the short sequence does not contain all the amino acid residues from the third loop). This deletion will not have a bad influence on following studies (e.g. docking) because the binding site is located in the transmembranar region. In Figure 6 bold letters indicate the transmembranar domains; the fifth transmembrane is pointed out with bold and italic letters.

```
CLUSTAL FORMAT for T-COFFEE Version_5.13 [http://www.tcoffee.org], CPU=0.09 sec, SCORE=63, Nseq=2, Len=595

P02699|OPSD_BOVIN ----MNGTEGPNF------YVPFSNKTGV---------------VRSPFEAPQYYLAEP- 34
P20309|ACM3_HUMAN MTLHNNSTTSPLFPNISSSWIHSPSDAGLPPGTVTHFGSYNVSRAAGNFSSPDGTTDDPL 60
                       *.* .* *      ::  ...:*:               . . *.:*:    :*
                                  TM I                             TM II
P02699|OPSD_BOVIN -----WQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFG 89
P20309|ACM3_HUMAN GGHTVWQVVFIAFLTGILALVTIIGNILVIVSFKVNKQLKTVNNYFLLSLACADL--IIG 118
                       **. ::*    :* :: :  *:*.:     :*:*:*  **:**.** ***  ::*
                                                      TM III
P02699|OPSD_BOVIN GFTTTLYTS--LHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMS-NF 146
P20309|ACM3_HUMAN VISMNLFTTYIIMNRWALGNLACDLWLAIDYVASNASVMNLLVISFDRYFSITRPLTYRA 178
                   :: .*:*:  : . :.:*  .*:*   :  :..: :: .*:*::::**. : :*:: .
                                TM IV
P02699|OPSD_BOVIN RFGENHAIMGVAFTWVMALACAAPPLVGWSRYI--------------------------- 179
P20309|ACM3_HUMAN KRTTKRAGVMIGLAWVISFVLWAPAILFWQYFVGKRTVPPGECFIQFLSEPTITFGTAIA 238
                  :   ::* : :.::**:::.  **.:: *. ::

P02699|OPSD_BOVIN --------------------------------------------PEGMQCSCGI------ 189
P20309|ACM3_HUMAN AFYMPVTIMTILYWRIYKETEKRTKELAGLQASGTEAETENFVHPTGSSRSCSSYELQQQ 298
                                                               * * . **.

P02699|OPSD_BOVIN --------------------------------------------------DYYTPHEETN 199
P20309|ACM3_HUMAN SMKRSNRRKYGRCHFWFTTKSWKPSSEQMDQDHSSSDSWNNNDAAASLENSASSDEEDIG 358
                                                                    .  : .*: .

P02699|OPSD_BOVIN NESFVIYMFVVHF----------------------------------------------- 212
P20309|ACM3_HUMAN SETRAIYSIVLKLPGHSTILNSTKLPSSDNLQVPEEELGMVDLERKADKLQAQKSVDDGG 418
                  .*: .** :*:::

P02699|OPSD_BOVIN -------IIPLIV---------------IFFCYGQLVFTVKEAA-----------AQQQE 239
P20309|ACM3_HUMAN SFPKSFSKLPIQLESAVDTAKTSDVNSSVGKSTATLPLSFKEATLAKRFALKTRSQITKR 478
                          :*: :               :  . . * ::.***:              :.
                                      TM VI
P02699|OPSD_BOVIN SATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSA 299
P20309|ACM3_HUMAN KRMSLVKEKKAAQTLSAILLAFIITWTPY-NIMVLVNTFCDSCIPKTFWNLGYWLCYINS 537
                  .  :   **:.:: :  :::**:* * **  .: . : *.  .* :   * .:  ::.   .:
               TM VII
P02699|OPSD_BOVIN VYNPVIYIMMNKQFRNCMVTTLCC------GKNPLGDDEASTTVSKTETSQVAPA 348
P20309|ACM3_HUMAN TVNPVCYALCNKTFRTTFKMLLLCQCDKKKRRKQQYQQRQSVIFHKRAPEQA--L 590
                  . *** * : ** **. :   * *       ::   ::. *. . *  ..*.
```

Figure 6

Therefore the improved alignment must be accomplished so that the SCRs from the two sequences are perfectly aligned. In the same time, each transmembrane alignment must not contain insertion or deletion at their level. If that happens, they have to be moved in loop regions for the sake of the accuracy of the alignment. As one can see from Figure 6 such operations have to be performed at helix 2 and 6 levels. In order to avoid the induction of perturbations it is better to eliminate the insertions before helix 1 and after helix 8. Also helices 5 must be manually realigned. Based on these modifications and by removing a part of CIII loop a new alignment is obtained – see Figure 7.

```
CLUSTAL FORMAT for T-COFFEE Version_5.13 [http://www.tcoffee.org], CPU=0.07 sec, SCORE=64, Nseq=2, Len=390
                                                                             TM I
        P02699|OPSD_BOVIN  MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTV
        P20309|ACM3_HUMAN  NNSTTSPLFWIHSPSDAGLAAGNFSSPDGTTDDPWQVVFIAFLTGILALVTIIGNILVIVSFK
```

```
                           TM II                                   TM III
P02699|OPSD_BOVIN QHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWS
P20309|ACM3_HUMAN VNKQLKTVNNYFLLSLACADLIIGVISMNLFTTYIIMNRWALGNLACDLWLAIDYVASNASVMN

                                                  TM IV
P02699|OPSD_BOVIN LVVLAIERYVVVCKPMS-NFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEGMQCSCGID
P20309|ACM3_HUMAN LLVISFDRYFSITRPLTYRAKRTTKRAGVMIGLAWVISFVLWAPAILFWQYFVGKRTVPPG---

                                    TM V
P02699|OPSD_BOVIN YYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVF-TVKEAAAQQQESATTQKAEKEVTRM
P20309|ACM3_HUMAN -ECFIQFLSEPTITFGTAIAAFYMPVTIMTILYWRIYKETEKRTKLQQQ-KRMSLVKEKKAAQT

                      TM VI                                     TM VII
P02699|OPSD_BOVIN VIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAVYNPVIYIMMNKQFRNCM
P20309|ACM3_HUMAN LSAILLAFIITWTPYNIMVLVNTFCDSCIP-KTFWNLGYWLCYINSTVNPVCYALCNKTFRTTF


P02699|OPSD_BOVIN VTTLCCGKNPLGDDEASTTVSKTETSQV
P20309|ACM3_HUMAN KMLLLCRRKQQYQQRQSVIFHKRAPEQA
```

Figure 7

- compare your newly obtained short sequence of amino acids for hM3 receptor with this one [insert link]
- correct the errors
- save the corrected sequence as text file, i.e. h3M_short_seq.txt (!keep the first line from the initial text file!), because you will need it in the next section of the course.

[1] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., ***Nucleic Acids Res*** *25* (**1997*)*** 3389-3402

[2] Jaroszewski L., Rychlewski L., Zhang B., Godzik A., ***Protein Sci*** 7 (**1998**) 1431-1440

[3] C. Notredame, D. Higgins, J. HeringaJournal of Molecular Biology, 302, 205-217, (2000)

[4] Bairoch A., Apweiler R., ***Nucleic Acids Res***., 28 (**2000**) 45-48

[5] http://www.expasy.org/sprot/sprot-top.html

[6] Baldwin, J.M.; Schertler, G.F.X.; Unger, V.M. J Mol Biol 1997**,** 272, 144–164