Section II

The Structure of Macromolecules: *Sequence alignment - tutorial*

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[3], and T-Coffee[4].

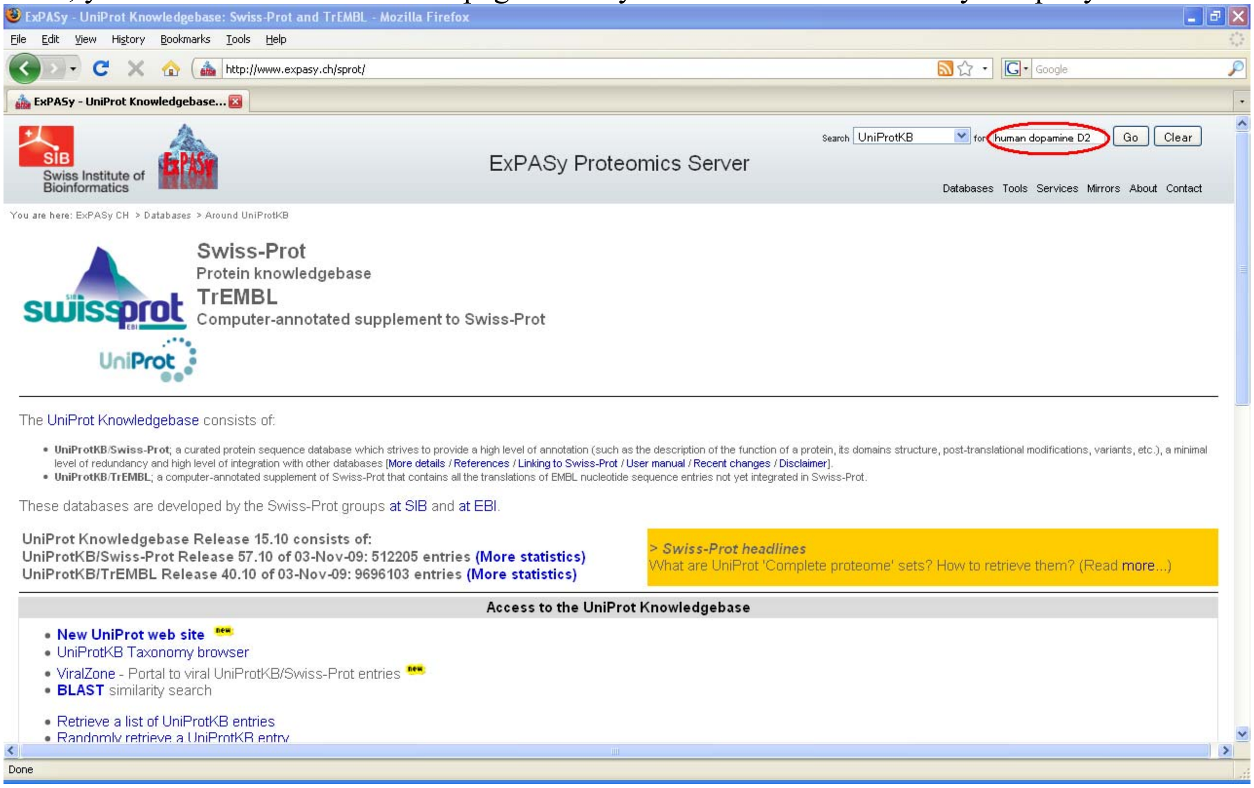
During this tutorial, you will learn the basic operation for building reliable sequence alignments using T-Coffee software and two G-Protein Coupled Receptors: human dopamine D2 receptor (hD2) and human beta2-adrenergic receptor (β2-AR).

How to

1. Obtain the protein sequences

✓ Go to Swiss-PROT[5, 6] (the protein sequence database).

✓ Enter the protein name (e.g. human dopamine D2 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.


ExPASy - UniProt Knowledgebase: Swiss-Prot and TrEMBL - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.expasy.ch/sprot/
Google
ExPASy - UniProt Knowledgebase...
SIB
Swiss Institute of Bioinformatics
ExPASy
Search UniProtKB for human dopamine D2
Go
Clear
ExPASy Proteomics Server
Databases Tools Services Mirrors About Contact
swissprot
UniProt
new
new
Done

You are here: ExPASy CH > Databases > Around UniProtKB

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 of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of 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 entries not yet integrated in Swiss-Prot.

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

UniProt Knowledgebase Release 15.10 consists of:
UniProtKB/Swiss-Prot Release 57.10 of 03-Nov-09: 512205 entries (More statistics)
UniProtKB/TrEMBL Release 40.10 of 03-Nov-09: 9696103 entries (More statistics)

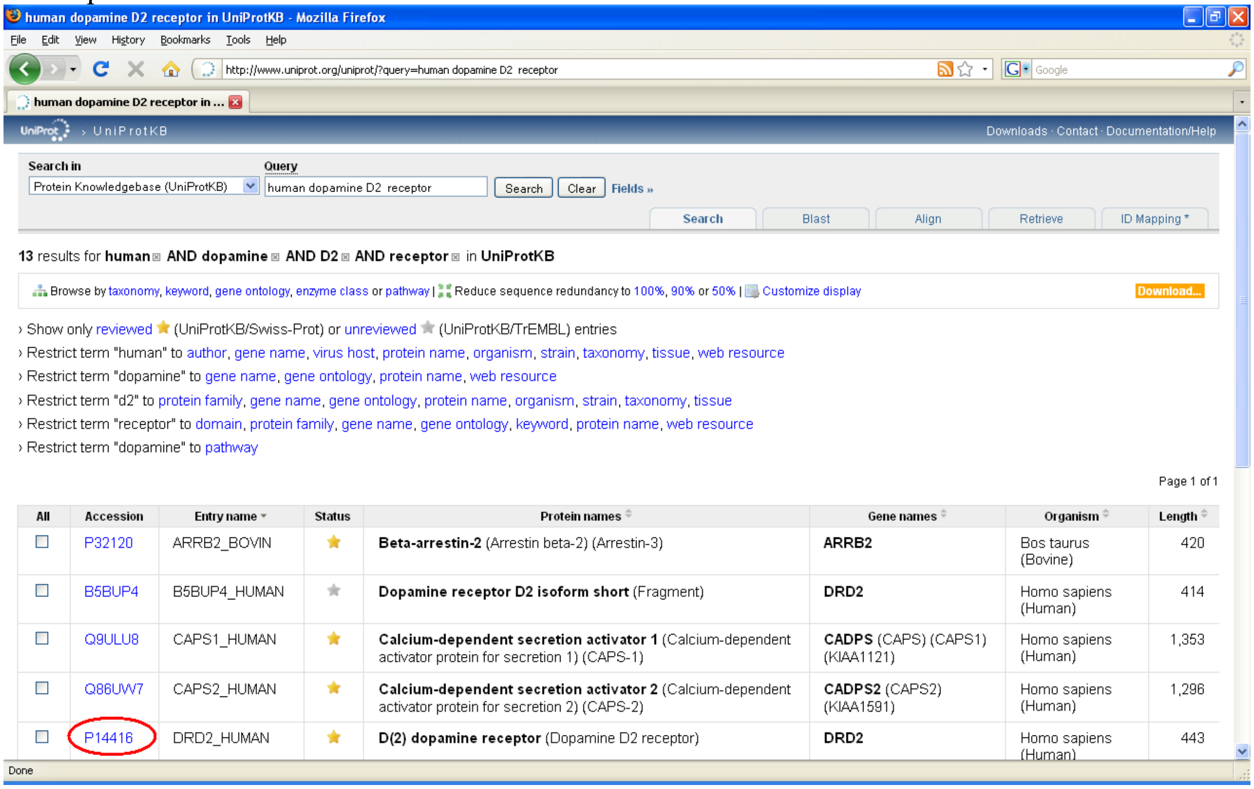
> *Swiss-Prot headlines*
What are UniProt 'Complete proteome' sets? How to retrieve them? (Read more...)

Access to the UniProt Knowledgebase

- New UniProt web site
- UniProtKB Taxonomy browser
- ViralZone - Portal to viral UniProtKB/Swiss-Prot entries
- BLAST similarity search
- Retrieve a list of UniProtKB entries
- Randomly retrieve a UniProtKB entry

Figure 1

✓ Select/open the record for the human D2 sequence (SWISS PROT ID: P14416) – see Figure 2. In the *Sequence* section of this page, you have the option to view the protein sequence as FASTA format.


human dopamine D2 receptor in UniProtKB - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.uniprot.org/uniprot/?query=human dopamine D2 receptor
Google
human dopamine D2 receptor in ...
UniProt › UniProtKB
Downloads · Contact · Documentation/Help
Search Clear Fields »
Search Blast Align Retrieve ID Mapping *
Download...
Done

Search in: Protein Knowledgebase (UniProtKB)
Query: human dopamine D2 receptor

13 results for **human** ☒ **AND dopamine** ☒ **AND D2** ☒ **AND receptor** ☒ in **UniProtKB**

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display

› Show only reviewed ★ (UniProtKB/Swiss-Prot) or unreviewed ★ (UniProtKB/TrEMBL) entries
› Restrict term "human" to author, gene name, virus host, protein name, organism, strain, taxonomy, tissue, web resource
› Restrict term "dopamine" to gene name, gene ontology, protein name, web resource
› Restrict term "d2" to protein family, gene name, gene ontology, protein name, organism, strain, taxonomy, tissue
› Restrict term "receptor" to domain, protein family, gene name, gene ontology, keyword, protein name, web resource
› Restrict term "dopamine" to pathway

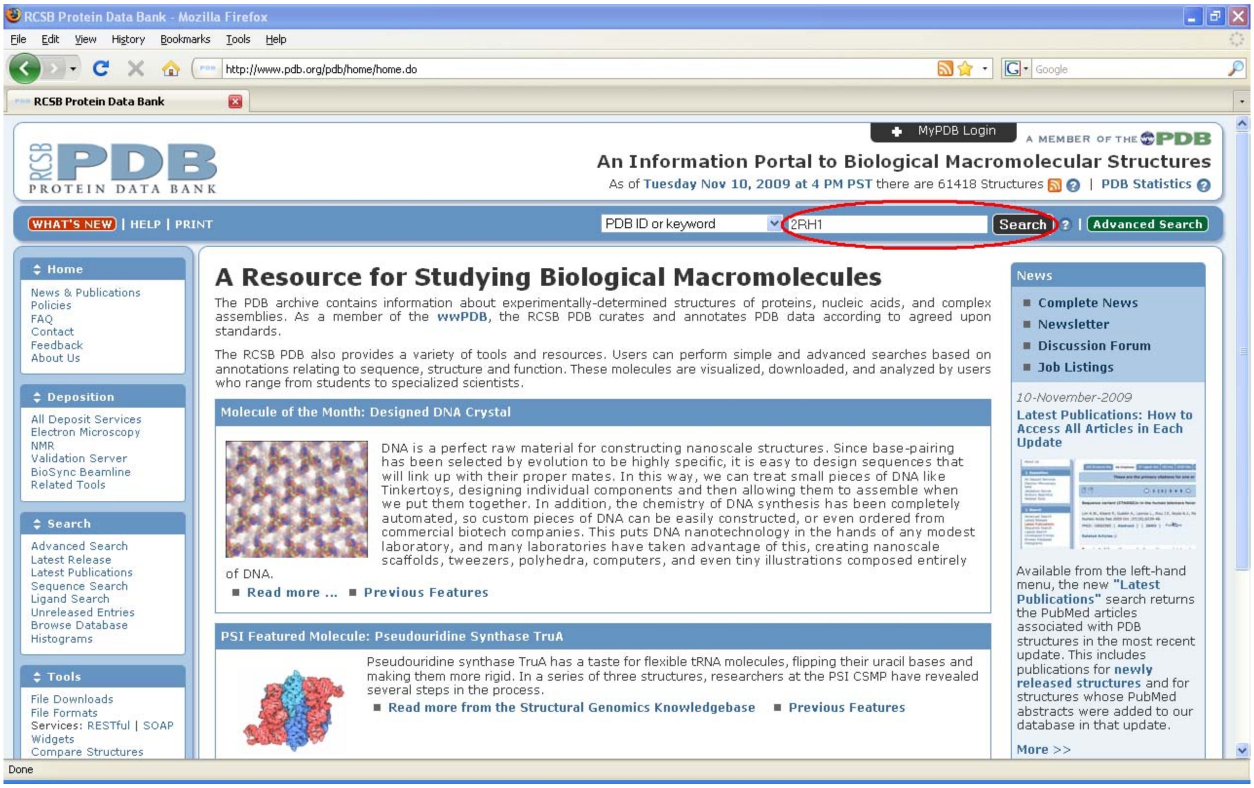
Page 1 of 1

All	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
□	P32120	ARRB2_BOVIN	★	**Beta-arrestin-2** (Arrestin beta-2) (Arrestin-3)	**ARRB2**	Bos taurus (Bovine)	420
□	B5BUP4	B5BUP4_HUMAN	★	**Dopamine receptor D2 isoform short** (Fragment)	**DRD2**	Homo sapiens (Human)	414
□	Q9ULU8	CAPS1_HUMAN	★	**Calcium-dependent secretion activator 1** (Calcium-dependent activator protein for secretion 1) (CAPS-1)	**CADPS** (CAPS) (CAPS1) (KIAA1121)	Homo sapiens (Human)	1,353
□	Q86UW7	CAPS2_HUMAN	★	**Calcium-dependent secretion activator 2** (Calcium-dependent activator protein for secretion 2) (CAPS-2)	**CADPS2** (CAPS2) (KIAA1591)	Homo sapiens (Human)	1,296
□	P14416	DRD2_HUMAN	★	**D(2) dopamine receptor** (Dopamine D2 receptor)	**DRD2**	Homo sapiens (Human)	443

Figure 2

✓ Save the sequence (in FASTA format) as a text file.

✓ you may repeat the above-given steps to obtained the text file for the β2-AR (SWISS-PROT ID: P08913) but because we plan to use the resulted alignment in a further homology modeling experiment, the Fasta sequence of the crystal structure of β2-AR deposited in Protein data Bank is going to be used. In this purpose go to PDB and enter 2RH1 code on the top bar of the page, and click Search – see Figure 3. The result is the Structure Summary page for 2RH1 structure, which is the structure ID for human β2-AR. Click on the Download File button on the right-upper corner of the page, choose Fasta Sequence Format and save the txt file. Repeat this operation to save the 2RH1 structure in PDB format.



http://www.pdb.org/pdb/home/home.do

An Information Portal to Biological Macromolecular Structures

As of Tuesday Nov 10, 2009 at 4 PM PST there are 61418 Structures

PDB ID or keyword: 2RH1

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month: Designed DNA Crystal

DNA is a perfect raw material for constructing nanoscale structures. Since base-pairing has been selected by evolution to be highly specific, it is easy to design sequences that will link up with their proper mates. In this way, we can treat small pieces of DNA like Tinkertoys, designing individual components and then allowing them to assemble when we put them together. In addition, the chemistry of DNA synthesis has been completely automated, so custom pieces of DNA can be easily constructed, or even ordered from commercial biotech companies. This puts DNA nanotechnology in the hands of any modest laboratory, and many laboratories have taken advantage of this, creating nanoscale scaffolds, tweezers, polyhedra, computers, and even tiny illustrations composed entirely of DNA.

PSI Featured Molecule: Pseudouridine Synthase TruA

Pseudouridine synthase TruA has a taste for flexible tRNA molecules, flipping their uracil bases and making them more rigid. In a series of three structures, researchers at the PSI CSMP have revealed several steps in the process.

10-November-2009

Latest Publications: How to Access All Articles in Each Update

Available from the left-hand menu, the new "Latest Publications" search returns the PubMed articles associated with PDB structures in the most recent update. This includes publications for newly released structures and for structures whose PubMed abstracts were added to our database in that update.

Figure 3

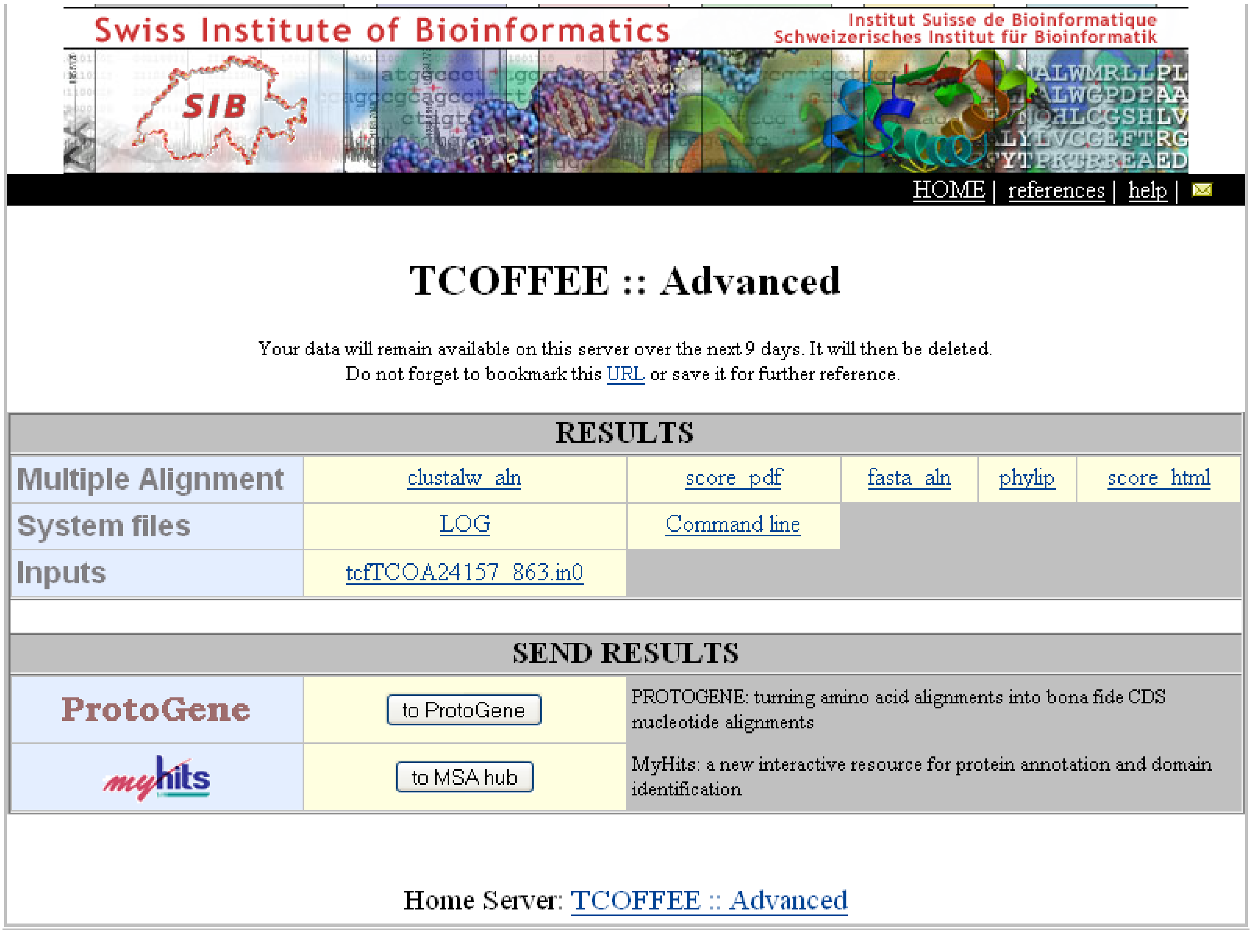
- ✓ merge the two text files into a single one and save it as “sequences.txt”

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, TCoffee, 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”;
- ✓ Press the submit button and wait until the sequence alignment will be computed. The results are given in different formats – see Figure 4. The high-similarity regions in the


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.

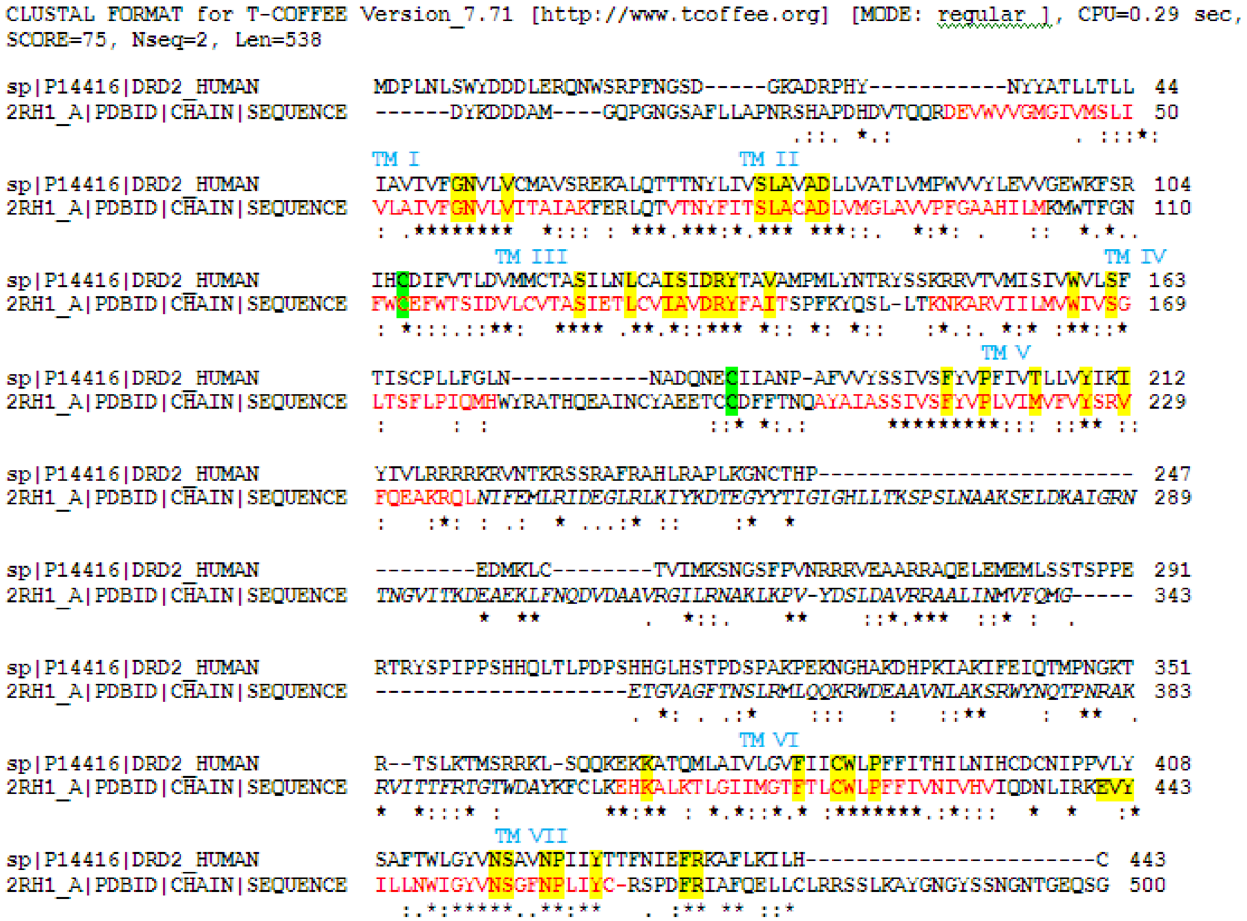
RESULTS					
Multiple Alignment	clustalw_aln	score_pdf	fasta_aln	phylip	score_html
System files	LOG	Command line			
Inputs	tcfTCOA24157_863.in0				

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 4

- ✓ Open PDB file saved in step 1 using any software for protein visualization. For the simpleness of handle, hide all the atoms and display the protein as a solid ribbon. In this way you can easily identify the seven transmembranes (depicted with red in figure 5) and the T4 lysozome (depicted with italic letters) on your alignment.
- ✓ 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[7]: 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.

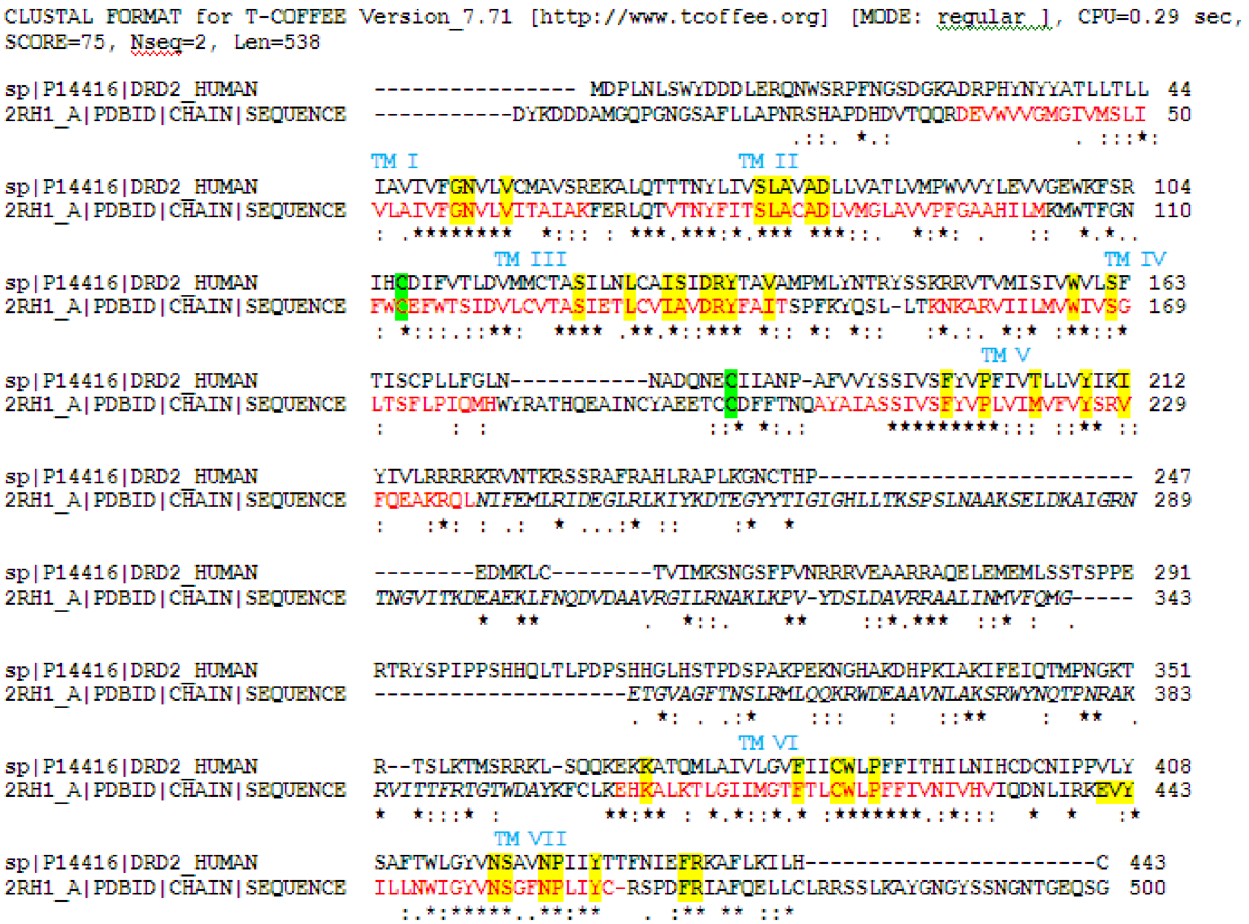


```
CLUSTAL FORMAT for T-COFFEE Version_7.71 [http://www.tcoffee.org] [MODE: regular ], CPU=0.29 sec,
SCORE=75, Nseq=2, Len=538

sp|P14416|DRD2_HUMAN         MDPLNLSWYDDDLERQNWSRPFNGSD-----GKADRPHY-----------NYYATLLTLL  44
2RH1_A|PDBID|CHAIN|SEQUENCE  ------DYKDDDAM----GQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVMSLI  50
                                                              .::. *.:              . :::*:
                             TM I                         TM II
sp|P14416|DRD2_HUMAN         IAVTVFGNVLVCMAVSREKALQTTTNYLIVSLAVADLLVATLVMPWVVYLEVVGEWKFSR  104
2RH1_A|PDBID|CHAIN|SEQUENCE  VLAIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGN  110
                             : .********* *::: : ***.***:*.*** ***::.  *:*:  .   :: *.*..
                                      TM III                                        TM IV
sp|P14416|DRD2_HUMAN         IHCDIFVTLDVMMCTASILNLCAISIDRYTAVAMPMLYNTRYSSKRRVTVMISIVWVLSF  163
2RH1_A|PDBID|CHAIN|SEQUENCE  FWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFKYQSL-LTKNKARVIILMVWIVSG  169
                             : *:::.::**:  **** .**.*::***  *:: *: *::   :*.:. *:* :**::*
                                                                             TM V
sp|P14416|DRD2_HUMAN         TISCPLLFGLN-----------NADQNECIIANP-AFVVYSSIVSFYVPFIVTLLVYIKI  212
2RH1_A|PDBID|CHAIN|SEQUENCE  LTSFLPIQMHWYRATHQEAINCYAEETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRV  229
                             :     :  :                 ::* *:.:     *********:::  ::** ::

sp|P14416|DRD2_HUMAN         YIVLRRRRKRVNTKRSSRAFRAHLRAPLKGNCTHP-------------------------  247
2RH1_A|PDBID|CHAIN|SEQUENCE  FQEAKRQLNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRN  289
                             :   :*:  : .:  * ...:* ::    :*   *

sp|P14416|DRD2_HUMAN         --------EDMKLC--------TVIMKSNGSFPVNRRRVEAARRAQELEMEMLSSTSPPE  291
2RH1_A|PDBID|CHAIN|SEQUENCE  TNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPV-YDSLDAVRRAALINMVFQMG-----  343
                                      *  **        .  *::.    **    ::*.***  ::* :  .

sp|P14416|DRD2_HUMAN         RTRYSPIPPSHHQLTLPDPSHHGLHSTPDSPAKPEKNGHAKDHPKIAKIFEIQTMPNGKT  351
2RH1_A|PDBID|CHAIN|SEQUENCE  --------------------ETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAK  383
                                                  . *: . .:*    :::   :   ::**    :  **  .
                                                          TM VI
sp|P14416|DRD2_HUMAN         R--TSLKTMSRRKL-SQQKEKKATQMLAIVLGVFIICWLPFFITHILNIHCDCNIPPVLY  408
2RH1_A|PDBID|CHAIN|SEQUENCE  RVITTFRTGTWDAYKFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQDNLIRKEVY  443
                             *  *:::*  :       **:** : *.*::*.* :*******.:*:::  *  *   :*
                                      TM VII
sp|P14416|DRD2_HUMAN         SAFTWLGYVNSAVNPIIYTTFNIEFRKAFLKILH----------------------C  443
2RH1_A|PDBID|CHAIN|SEQUENCE  ILLNWIGYVNSGFNPLIYC-RSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNTGEQSG  500
                               :.*:*****..**:**    . :** ** ::*
```

Figure 5

One should avoid insertion or deletion in the alpha-helices or beta-sheets arrangements so a further manual refinement of the alignment is recommended since we have two deletions in helices 1 and 4. In order to avoid the induction of perturbations it is better to move the deletions before helix 1 and after helix 4. Also, during manual refinement of the alignment we should reassess all the insertions and deletions according to the three dimensional structure of the template. When necessary, the deletions and the insertions can be pasted into a single piece per loop and placed in the most adequate point in accordance with the template structure. Any other structural characteristic of the studied proteins should not be neglected, like the sulfur bridge between the second extracellular loop (EL2) and the third transmembrane in GPCR family (green in figure 5).

Based on these observations a new alignment is obtained – see Figure 6.



```
CLUSTAL FORMAT for T-COFFEE Version_7.71 [http://www.tcoffee.org] [MODE: regular ], CPU=0.29 sec,
SCORE=75, Nseq=2, Len=538

sp|P14416|DRD2_HUMAN         ---------------- MDPLNLSWYDDDLERQNWSRPFNGSDGKADRPHYNYYATLLTLL  44
2RH1_A|PDBID|CHAIN|SEQUENCE  -----------DYKDDDAMGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVMSLI  50
                                                         .::. *.:             . :::*:
                             TM I                                         TM II
sp|P14416|DRD2_HUMAN         IAVTVFGNVLVCMAVSREKALQTTTNYLIVSLAVADLLVATLVMPWVVYLEVVGEWKFSR  104
2RH1_A|PDBID|CHAIN|SEQUENCE  VLAIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGN  110
                             : .********* *::: : ***.***:*.*** ***::.  *:*: .    ::  *.*..
                                            TM III                                   TM IV
sp|P14416|DRD2_HUMAN         IHCDIFVTLDVMMCTASILNLCAISIDRYTAVAMPMLYNTRYSSKRRVTVMISIVWVLSF  163
2RH1_A|PDBID|CHAIN|SEQUENCE  FWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFKYQSL-LTKNKARVIILMVWIVSG  169
                             : *::::.::**:   **** .**.*::*** *:: *: *::    :*.:. *:* :**::*
                                                                                TM V
sp|P14416|DRD2_HUMAN         TISCPLLFGLN-----------NADQNECIIANP-AFVVYSSIVSFYVPFIVTLLVYIKI  212
2RH1_A|PDBID|CHAIN|SEQUENCE  LTSFLPIQMHWYRATHQEAINCYAEETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRV  229
                             :     : :                   ::* *:.:       *********::: ::** ::

sp|P14416|DRD2_HUMAN         YIVLRRRRKRVNTKRSSRAFRAHLRAPLKGNCTHP-------------------------  247
2RH1_A|PDBID|CHAIN|SEQUENCE  FQEAKRQLNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRN  289
                             :   :*: : .:  * ...:* ::     :*  *

sp|P14416|DRD2_HUMAN         --------EDMKLC--------TVIMKSNGSFPVNRRRVEAARRAQELEMEMLSSTSPPE  291
2RH1_A|PDBID|CHAIN|SEQUENCE  TNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPV-YDSLDAVRRAALINMVFQMG-----  343
                                     *  **         .  *::.    **    ::*.***  ::* :   .

sp|P14416|DRD2_HUMAN         RTRYSPIPPSHHQLTLPDPSHHGLHSTPDSPAKPEKNGHAKDHPKIAKIFEIQTMPNGKT  351
2RH1_A|PDBID|CHAIN|SEQUENCE  --------------------ETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAK  383
                                                 . *: . .:*    :::    :    ::**    :  **  .
                                                                          TM VI
sp|P14416|DRD2_HUMAN         R--TSLKTMSRRKL-SQQKEKKATQMLAIVLGVFIICWLPFFITHILNIHCDCNIPPVLY  408
2RH1_A|PDBID|CHAIN|SEQUENCE  RVITTFRTGTWDAYKFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQDNLIRKEVY  443
                             *  *:::* :          **:** : *.*::*.* :*******.:*:::  *  *   :*
                                            TM VII
sp|P14416|DRD2_HUMAN         SAFTWLGYVNSAVNPIIYTTFNIEFRKAFLKILH----------------------C  443
2RH1_A|PDBID|CHAIN|SEQUENCE  ILLNWIGYVNSGFNPLIYC-RSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNTGEQSG  500
                               :.*:*****..**:**    . :** ** ::*
```

Figure 6

In Figure 5 and 6 red letters indicate the transmembrane domains; the T4 lysozome is pointed out with italic letters; the highly conserved residues in GPCR family are highlighted in yellow and the sulfur bridge with green.

- ✓ Compare your alignment with the one in figure 6. Using the pdb file of β2-AR observe the the new positions of the insertions and deletions.
- ✓ Correct all the inconsistencies and save the alignment as text file, i.e. alignment.txt (!keep the first line from the initial textfile!), 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] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG., **Bioinformatics** 23, (2007) 2947-294

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

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

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

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