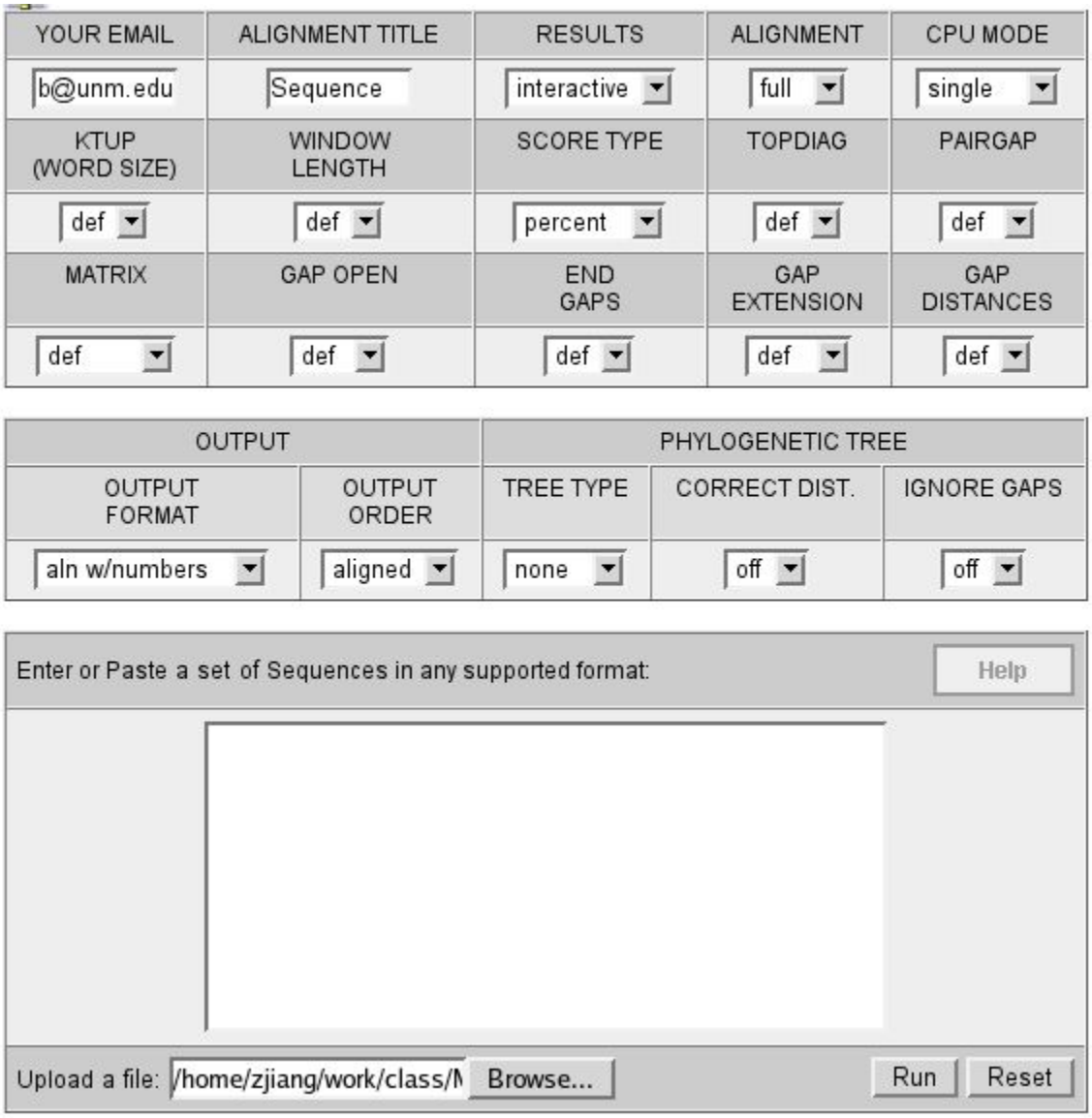
Multiple Sequence Alignment

Multiple sequence alignments can be used to study the relationship(s) between sequences in sets of more than two sequences. This application is particularly useful when studying the relationships between a similar type of gene product that is expressed by different organisms, like analyzing GPCR sequences from several different species, or when studying similar, yet divergent, sequences within the same organism, as the variance in neuropeptide Y receptor isoforms in *Homo sapiens*.

Often, a primary focus of a multiple sequence alignment is to identify, within several related sequences, regions that are highly conserved in identity or similarity, and therefore probably have functional and/or structural significance. Many factors affect the analysis of conserved regions within related sequences, such as the number of sequences included in the analysis, and the ratio of the number of very similar (almost identical) sequences to the number of more distantly related sequences. Divergent sequences can cause problems in a multiple sequence alignment. It is more difficult to identify the correct alignment when two sequences that are related throughout part of the sequence also contain large sections that diverge. Therefore, the error rates in the alignment increase as divergence increases. These errors in the alignment can cause the related part of the sequences to show lower similarity than they actually have, and this sort of error is often amplified in subsequent steps. ClustalW is a commonly used multiple sequence alignment program that addresses the problems associated with alignment of divergent sequences in several ways. Individual weights are assigned to each sequence in a partial alignment such that near-duplicate sequences are down-weighted and divergent sequences are up-weighted. Also, the amino acid substitution matrices at different alignment stages are chosen according to the divergence of the sequences to be aligned. Residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions increase the penalty for opening new gaps in regions of regular secondary structure. Therefore, it increases the likelihood that gaps will occur in loop regions than in highly structured regions such as alpha helices and beta sheets. Highly structured regions are commonly very important to the fold and function of a protein, and so divergence is often biologically less likely in these areas. For similar reasons, existing gaps receive locally reduced gap penalties to encourage the opening up of new gaps near the existing ones. These features have been designed into ClustalW to produce multiple sequence alignments that are biologically meaningful.

ClustalW is in fact pairwise alignments with some optimizations. It first performs all possible pairwise alignments between each pair of sequences; then calculates the "distance" [based on scoring methods we discussed earlier] between each pair of sequences based on these isolated pairwise alignments; and generates a distance matrix of all input sequences. From the distance matrix, a neighbor-joining "guide tree" is generated, and this "guide tree" gives the order in which the progressive alignment [a heuristic algorithm devised by Feng and Doolittle in 1987] will be carried out.

We will use the ClustalW web service at EBI to do an exercise. We'll use a set of neuropeptide Y Y1 receptors, which belong to the GPCR family of proteins. The sequences are in the fasta format. You can copy and paste into the the input form or save the sequences in a file and upload the file by clicking the "Browse..." button, then click



the "Run" button; Just wait a while and a new page will open which has the url to your results. Alternatively, you can fill up the "Your EMAIL" field with your email address in the submission page, and results will be delivered to you via email.

```
>Dog_NY1R D02813
MNSTSFSQVE NHSIFCNFSE NSQFLAFESD DCHLPLAMIF TLALAYGAVI ILGVTGNLAL
IMIILKQKEM RNVTNILIVN LSFSDLLVAI MCLPFTFVYT LMDHWVFGEA MCKLNPFVQC
VSITVSIFSL VLIAVERHQL IINPRGWRPN NRHAYVGIAV IWVLAVVSSL PFLIYQVLTD
EPFQNVTLDA FKDKYVCFDK FPSDSHRLSY TTLLLMLQYF GPLCFIFICY FKIYIRLKRR
NNMMDKMRDN KYRSSETKRI NIMLLSIVVA FAVCWLPLTI FNTVFDWNHQ IIATCNHNLL
FLLCHLTAMI STCVNPIFYG FLNKNFQRDL QFFFNFCDFR SRDDDYETIA MSTMHTDVSK
TSLKQASPVA FKKINNDDNE KI

>GuineaPig_NY1R Q9WVD0
MNSTSFSQLE NHSVHYNLSE EKPSFFAFEN DDCHLPLAVI FTLALAYGAV IILGVSGNLA
LILIILKQKE MRNVTNILIV NLSFSDLLVA IMCLPFTFVY TLMDHWIFGE IMCKLNPFVQ
CVSITVSIFS LVLIAVERHQ LIINPRGWRP NNRHAYIGIA VIWVLAVASS LPFMIYQVLT
DEPFQNVTLD AFKDKLVCFD QFPSDSHRLS YTTLLLVLQY FGPLCFIFIC YFKIYIRLKR
```

RNNMMDKMRD SKYRSSESKR INIMLLSIVV AFAVCWLPLT IFNTVFDWNH QIIATCNHNL
LFLLCHLTAM ISTCVNPIFY GFLNKNFQRD LQFFFNFCDF RSRDDDYETI AMSTMHTDVS
KTSLKQASPL AFKKISCVEN EKI

>Human_NY1R P25929
MNSTLFSQVE NHSVHSNFSE KNAQLLAFEN DDCHLPLAMI FTLALAYGAV IILGVSGNLA
LIIIILKQKE MRNVTNILIV NLSFSDLLVA IMCLPFTFVY TLMDHWVFGE AMCKLNPFVQ
CVSITVSIFS LVLIAVERHQ LIINPRGWRP NNRHAYVGIA VIWVLAVASS LPFLIYQVMT
DEPFQNVTLD AYKDKYVCFD QFPSDSHRLS YTTLLLVLQY FGPLCFIFIC YFKIYIRLKR
RNNMMDKMRD NKYRSSETKR INIMLLSIVV AFAVCWLPLT IFNTVFDWNH QIIATCNHNL
LFLLCHLTAM ISTCVNPIFY GFLNKNFQRD LQFFFNFCDF RSRDDDYETI AMSTMHTDVS
KTSLKQASPV AFKKINNNDD NEKI

>Mouse_NY1R Q04573
MNSTLFSKVE NHSIHYNASE NSPLLAFEND DCHLPLAVIF TLALAYGAVI ILGVSGNLAL
IIIILKQKEM RNVTNILIVN LSFSDLLVAV MCLPFTFVYT LMDHWVFGET MCKLNPFVQC
VSITVSIFSL VLIAVERHQL IINPRGWRPN NRHAYIGITV IWVLAVASSL PFVIYQILTD
EPFQNVSLAA FKDKYVCFDK FPSDSHRLSY TTLLLVLQYF GPLCFIFICY FKIYIRLKRR
NNMMDKIRDS KYRSSETKRI NIMLLSIVVA FAVCWLPLTI FNTVFDWNHQ IIATCNHNLL
FLLCHLTAMI STCVNPIFYG FLNKNFQRDL QFFFNFCDFR SRDDDYETIA MSTMHTDVSK
TSLKQASPVA FKKISMNDNE KV

>Pig_NY1R O02835
MNSTLSSQVE NHSIYYNFSE KNSQFLAFEN DDCHLPLAMI FTLALAYGAV IILGVSGNLA
LIIIILKQKE MRNVTNILIV NLSFSDLLVA IMCLPFTFVY TLMDHWVFGE VMCKLNPFVQ
CVSITVSIFS LVLIAVERHQ LIINPRGWRP SNRHAYVGIA VIWVLAVASS LPFLIYQVLT
DEPFQNVTLD AFKDKYVCFD KFLSDSHRLS YTTLLLVLQY FGPLCFIFIC YFKIYIRLKR
RNNMMDKMRD NKYRSSETKR INVMLLSIVV AFAVCWLPLT IFNTVFDWNH QIIATCNHNL
LFLLCHLTAM ISTCINPIFY GFLNKNFQRD LQFFFNFCDF RSRDDDYEVI AMSTMHTDVS
KTSLKQASPV ALKKIHSDDN EKI

>Rat_NY1R P21555
MNSTLFSRVE NYSVHYNVSE NSPFLAFEND DCHLPLAVIF TLALAYGAVI ILGVSGNLAL
IIIILKQKEM RNVTNILIVN LSFSDLLVAV MCLPFTFVYT LMDHWVFGET MCKLNPFVQC
VSITVSIFSL VLIAVERHQL IINPRGWRPN NRHAYIGITV IWVLAVASSL PFVIYQILTD
EPFQNVSLAA FKDKYVCFDK FPSDSHRLSY TTLLLVLQYF GPLCFIFICY FKIYIRLKRR
NNMMDKIRDS KYRSSETKRI NVMLLSIVVA FAVCWLPLTI FNTVFDWNHQ IIATCNHNLL
FLLCHLTAMI STCVNPIFYG FLNKNFQRDL QFFFNFCDFR SRDDDYETIA MSTMHTDVSK
TSLKQASPVA FKKISMNDNE KI

ClustalW Results

Results of search	
Number of sequences	6
Alignment score	34078
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.82
JalView	
Output file	clustalw-20040831-17173916.output
Alignment file	clustalw-20040831-17173916.aln
Guide tree file	clustalw-20040831-17173916.dnd
Your input file	clustalw-20040831-17173916.input

SUBMIT ANOTHER JOB

As you can see, four output files were generated, the *.input file is our input sequences, and the *.output file shows your the pairwise distance calculated between all the six input sequences, the *.dnd is the "guide-tree" file for progressive alignment, and finally the *.aln file is the alignment of the six input NY1R sequences as shown below.

```
CLUSTAL W (1.82) multiple sequence alignment


Dog_NY1R            MNSTSFSQVENHSIFCNFSE-NSQFLAFESDDCHLPLAMIFTLALAYGAVIILGVTGNLA 59
Pig_NY1R            MNSTLSSQVENHSIYYNFSEKNSQFLAFENDDCHLPLAMIFTLALAYGAVIILGVSGNLA 60
Human_NY1R          MNSTLFSQVENHSVHSNFSEKNAQLLAFENDDCHLPLAMIFTLALAYGAVIILGVSGNLA 60
Mouse_NY1R          MNSTLFSKVENHSIHYNASE-NSPLLAFENDDCHLPLAVIFTLALAYGAVIILGVSGNLA 59
Rat_NY1R            MNSTLFSRVENYSVHYNVSE-NSPFLAFENDDCHLPLAVIFTLALAYGAVIILGVSGNLA 59
GuineaPig_NY1R      MNSTSFSQLENHSVHYNLSEEKPSFFAFENDDCHLPLAVIFTLALAYGAVIILGVSGNLA 60
                    ****  *::**:*:. * ** :. ::***.********:****************:****

Dog_NY1R            LIMIILKQKEMRNVTNILIVNLSFSDLLVAIMCLPFTFVYTLMDHWVFGEAMCKLNPFVQ 119
Pig_NY1R            LIIIILKQKEMRNVTNILIVNLSFSDLLVAIMCLPFTFVYTLMDHWVFGEVMCKLNPFVQ 120
Human_NY1R          LIIIILKQKEMRNVTNILIVNLSFSDLLVAIMCLPFTFVYTLMDHWVFGEAMCKLNPFVQ 120
Mouse_NY1R          LIIIILKQKEMRNVTNILIVNLSFSDLLVAVMCLPFTFVYTLMDHWVFGETMCKLNPFVQ 119
Rat_NY1R            LIIIILKQKEMRNVTNILIVNLSFSDLLVAVMCLPFTFVYTLMDHWVFGETMCKLNPFVQ 119
GuineaPig_NY1R      LILIILKQKEMRNVTNILIVNLSFSDLLVAIMCLPFTFVYTLMDHWIFGEIMCKLNPFVQ 120
                    **:***************************:***************:*** *********

Dog_NY1R            CVSITVSIFSLVLIAVERHQLIINPRGWRPNNRHAYVGIAVIWVLAVVSSLPFLIYQVLT 179
Pig_NY1R            CVSITVSIFSLVLIAVERHQLIINPRGWRPSNRHAYVGIAVIWVLAVASSLPFLIYQVLT 180
Human_NY1R          CVSITVSIFSLVLIAVERHQLIINPRGWRPNNRHAYVGIAVIWVLAVASSLPFLIYQVMT 180
Mouse_NY1R          CVSITVSIFSLVLIAVERHQLIINPRGWRPNNRHAYIGITVIWVLAVASSLPFVIYQILT 179
Rat_NY1R            CVSITVSIFSLVLIAVERHQLIINPRGWRPNNRHAYIGITVIWVLAVASSLPFVIYQILT 179
GuineaPig_NY1R      CVSITVSIFSLVLIAVERHQLIINPRGWRPNNRHAYIGIAVIWVLAVASSLPFMIYQVLT 180
                    *****************************.*****:**:*******.*****:***::*
```

```
Dog_NY1R            DEPFQNVTLDAFKDKYVCFDKFPSDSHRLSYTTLLLMLQYFGPLCFIFICYFKIYIRLKR 239
Pig_NY1R            DEPFQNVTLDAFKDKYVCFDKFLSDSHRLSYTTLLLVLQYFGPLCFIFICYFKIYIRLKR 240
Human_NY1R          DEPFQNVTLDAYKDKYVCFDQFPSDSHRLSYTTLLLVLQYFGPLCFIFICYFKIYIRLKR 240
Mouse_NY1R          DEPFQNVSLAAFKDKYVCFDKFPSDSHRLSYTTLLLVLQYFGPLCFIFICYFKIYIRLKR 239
Rat_NY1R            DEPFQNVSLAAFKDKYVCFDKFPSDSHRLSYTTLLLVLQYFGPLCFIFICYFKIYIRLKR 239
GuineaPig_NY1R      DEPFQNVTLDAFKDKLVCFDQFPSDSHRLSYTTLLLVLQYFGPLCFIFICYFKIYIRLKR 240
                    *******:* *:*** ****:* *************:********************** 

Dog_NY1R            RNNMMDKMRDNKYRSSETKRINIMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 299
Pig_NY1R            RNNMMDKMRDNKYRSSETKRINVMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 300
Human_NY1R          RNNMMDKMRDNKYRSSETKRINIMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 300
Mouse_NY1R          RNNMMDKIRDSKYRSSETKRINIMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 299
Rat_NY1R            RNNMMDKIRDSKYRSSETKRINVMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 299
GuineaPig_NY1R      RNNMMDKMRDSKYRSSESKRINIMLLSIVVAFAVCWLPLTIFNTVFDWNHQIIATCNHNL 300
                    *******:**.******:****:************************************ 

Dog_NY1R            LFLLCHLTAMISTCVNPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYETIAMSTMHTDVS 359
Pig_NY1R            LFLLCHLTAMISTCINPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYEVIAMSTMHTDVS 360
Human_NY1R          LFLLCHLTAMISTCVNPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYETIAMSTMHTDVS 360
Mouse_NY1R          LFLLCHLTAMISTCVNPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYETIAMSTMHTDVS 359
Rat_NY1R            LFLLCHLTAMISTCVNPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYETIAMSTMHTDVS 359
GuineaPig_NY1R      LFLLCHLTAMISTCVNPIFYGFLNKNFQRDLQFFFNFCDFRSRDDDYETIAMSTMHTDVS 360
                    **************:**********************************.*********** 

Dog_NY1R            KTSLKQASPVAFKKINN-DDNEKI 382
Pig_NY1R            KTSLKQASPVALKKIHS-DDNEKI 383
Human_NY1R          KTSLKQASPVAFKKINNNDDNEKI 384
Mouse_NY1R          KTSLKQASPVAFKKISM-NDNEKV 382
Rat_NY1R            KTSLKQASPVAFKKISM-NDNEKI 382
GuineaPig_NY1R      KTSLKQASPLAFKKISC-VENEKI 383
                    *********:*:***    :***:
```

For this highly conserved sequences like NY1R you'll see very large portion of the sequences perfectly aligned, highly conserved functional domains can be revealed when using homologous sequences with lower percentage of identity.