Algorithms in Structural Bioinformatics
Fold Recognition Practical

TME-5

In this practical you will learn how to select and use appropriate tools for predicting the structure of a protein. You will work with the Influenza A virus (A/chicken/Jilin/9/2004(H5N1)). Each group will choose one protein sequence to work with, among those reported at the end of this document. No complete structure exists today for these proteins.

How many residues are in the protein sequence you have chosen (query sequence)?

Exercice - 1 **Search the pdb database using NCBI's *Blast* server**

Go to `http://www.ncbi.nlm.nih.gov/blast/` and perform a *blastp* search of the query sequence against the PDB database.

1- What is the E-value of the top-scoring hit?

2- Is the top-scoring hit statistically significant (do you believe the result)?

3- Which region of your sequence is aligned to the top-scoring template?

a- How many residues are in that region?
b- How many hits do you get with an E-value of less than 0.05?

4- Which region(s) of the query sequence is aligned to a template with an E-value of less than 0.05?

a- How many residues are in those regions?

5- For each of the regions where you did not find a significant hit, try to go through the different steps of this exercise again. If you got no hits at all go to the next exercise.

Exercice - 2 **Use *PSI-Blast* to find a hit**

If you did not find a statistically significant hit try *PSI-Blast*. Once again, go to `http://www.ncbi.nlm.nih.gov/blast/` and click on "protein BLAST (blastp)". Under "Program Selection" choose the "PSI-BLAST (Position-Specific Iterated BLAST)" option. Paste in your sequence, leave all other options as they are, and press "BLAST" as before.

1- How many hits with a E-value of less than 0.05 do you get?

2- You can find out the total number of hits by looking at the top of the results page. The web-based *Blast* has a limit on 500 hits that it shows. Is this limit reached?

The hits with names like	Are from	(look for keyword)
`gi|50365715|gb|AAT76158.1|`	GenBank	gb
`gi|438071|emb|CAA81463.1|`	EMBL	emb
`gi|73852958|ref|YP_308670.1|`	RefSeq	ref
`gi|138833|sp|P06821|M2_IAPUE`	SwissProt	sp
`gi|14278293|pdb|1EA3|B`	PDB	pdb

The SwissProt contains a nice annotation of many proteins. Write down the name of the first SwissProt hit. You will use this later.

3- Are there any hits from PDB? If the limit of 500 hits is reached, a hit may be found with a lower rank but you will never know unless you learn to run *Blast* in command line mode on an UNIX computer (but this is beyond the scope of this course)

4- If no hit from PDB is found, set HIT list size to 1000 (or more) press "Run PSI-Blast Iteration 2". Repeat the steps of the exercise until you either:
- find a hit in PDB (you can use Find from the Edit menu to search the page for "pdb" if there are many hits)
- reach iteration 5
- get a *Blast* report stating that the iterative procedure has converged (meaning that no new sequences were found and that another iteration therefore will not produce a result that is different from what you got in the current iteration)

5- Did you find a hit in PDB? If so, what is the name of the entry and the chain name?

The problem of a *PSI-BLAST* hit from PDB being hidden among many sequences without information has been addressed by the *PDB_BLAST* algorithm. It runs a *PSI-BLAST* search against the *nr* database. The server has unfortunately gone offline (`http://bioinformatics.ljcrf.edu/pages/`)

Exercice - 3 Try to use *CPHmodels* to make a model

Go to (`http://www.cbs.dtu.dk/services/CPHmodels/`) and paste in your sequence in fasta format.

1- Do you get any significant hits? If you did, save the *CPHmodels* ouput page and the model.

2- Repeat it for the regions of the sequence where *CPHmodels* did not produce a model

3- If you are unhappy with *CPHmodels* you can try to use one of the other homology modeling servers on the web such as *SWISSMODEL* (`http://swissmodel.expasy.org/`), *ESyPred3D* (`http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/`), or any other one you can find on the web.

Exercice - 4 Use fold recognition servers

If you did not find a hit for your protein or regions of it, then try to use fold recognition servers to predict the structure.

1- Start out with *HHpred* (`http://toolkit.tuebingen.mpg.de/hhpred`). What is the significance of the top-scoring hit?

2- If you still have time, and have not found a hit use the *FFAS* (`http://bioinformatics.ljcrf.edu/pages/`) server. Paste in your protein in fasta format: Note: Queries made from ALL users are shown on the results page, not just yours! The web page states that FFAS03 scores below -9.5 usually mean significant similarity (less than 3% of false positives).

a- What is the score of your query?
b- Is the score significant?
c- What is the template (see the "Covered by template(s)" column)?
d- Click on the template, then on "Biology&Chemistry". Does the description of the biological function increase or decrease your confidence in the hit?
e- Do you believe it even if the statistical score was not significant?

If the score is significant (or you want to make a model anyway) you can from the *FFAS* server output page
- Click on "PDB1105" in the row corresponding to your query
- Click on "ali" in the first row to get the alignment, and save the file to your computer
- Click on "scwrl", then "Submit query", then "Get all atom Model" to make a model using the *Scwrl* server.

3- If you still have time at the end of the exercise you can try *Phyre* (the successor of *3D-PSSM*; `http://www.sbg.bio.ic.ac.uk/phyre2/`). Click on the results link (you will also get the result back as a email).

a- What is the E-value and the estimated precision for the top scoring hit?
b- What is the percentage identity?
c- Does the description of the hit correspond to your protein?
d- Do you believe the hit?

You can download the model by clicking on the large picture in the "View Model" column.

Exercice - 5 Try new folds prediction

For the regions where the above methods failed there is probably no template-based method that will give you the structure of your protein. This may be because no proteins look like your protein (have the same fold) or because the current methods are not good enough to detect it. In any case you will have to settle for less ambitious goals such as predicting the secondary structure. The goal of this step is to collect as much structural and functional information about the protein as you can.

1- Predict the secondary structure of the protein using *PSIPRED* (`http://bioinf.cs.ucl.ac.uk/psipred/psiform.html`). If you were able to make a 3D model of the protein in the above steps you can see how many % of the secondary structure prediction correspond to the model above. On average, the secondary structure of homology models is correctly modelled (predicted) for 88% of the residues.

2- Try also to predict:

a- Transmembrane regions using *TMHMM* (`http://www.cbs.dtu.dk/services/TMHMM/`)
b- Disordered regions using *DisEMBL* (`http://dis.embl.de/`)
c- Coiled coil regions using *COILS* (`http://www.ch.embnet.org/software/COILS_form.html`)
d- Function using *ProtFun* (`http://www.cbs.dtu.dk/services/ProtFun/`)
e- Signal peptide using *SignalP* (`http://www.cbs.dtu.dk/services/SignalP/`)
f- Surface exposure using *NetSurP* (`http://www.cbs.dtu.dk/services/NetSurfP/`)

What does each of these servers predict and what is the output?

Sequences

```
>AAT76159.1 entry:AY653194 /gene="M" /product="matrix protein M1" [Influenza A
virus (A/chicken/Jilin/9/2004(H5N1))]
MSLLTEVETYVLSIIPSGPLKAEIAQRLEDVFAGKNTDLEALMEWLKTRPILSPLTKGIL
GFVFTLTVPSERGLQRRRFVQNALNGNGDPNNMDRAVKLYRKLKREITFHGAKEVALSYS
TGALASCMGLIYNRMGTVTTEVAFGLVCATCEQIADSQHRSHRQMVTITNPLIRHENRMV
LASTTAKAMEQMAGSSEQAAEAMEVANQARQMVQAMRTIGTHPSSSAGLRDDLLENLQAY
QKRMGVQMQRFK
>AAT76158.1 entry:AY653194 /gene="M" /product="matrix protein M2" [Influenza A
virus (A/chicken/Jilin/9/2004(H5N1))]
MSLLTEVETPTRNGWECRCSDSSDPLVVAASIIGILHLILWILDRLFFKCIYRRLKYGLK
RGPSTEGVPESMREEYRQEQQSAVDVDDGHFVNIELE
>AAT76162.1 entry:AY653197 /gene="NS" /product="nonstructural protein 2"
[Influenza A virus (A/chicken/Jilin/9/2004(H5N1))]
MDPNTVSSFQDILMRMSKMQLGSSSEDLNGMITRFESLKLYRDSLGEAVMRTGDLHSLQI
RNGKWREQLSQKFEEIRWLIEEVRHRLKITENSFEQITFMQALQLLLEVEQEIRTFSFQL
>AAT76164.1 entry:AY653198 /gene="PA" /product="polymerase" [Influenza A virus
(A/chicken/Jilin/9/2004(H5N1))]
MEDFVRQCFNPMTVELAEKAMKEYGEDPKIETNKFAAICTHLEVCFMYSDFHFIDERSES
IIVESGDQNALLKHRFEIIEGRDRTMAWTVVNSICNTTGVEKPKFLPDLYDYKENRFIEI
GVTRREVHTYYLEKANKIKSEKTHIHIFSFTGEEMATKADYTLDEESRARIKTRLFTIRQ
EMASRGLWDSFRQSERGEETIEEKFEITGTMRRLADQSLPPNFSSLENFRAYVDGFEPNG
CIEGKLSQMSKEVNARIEPFQKTTPRPLRLPDGPPCSQRSKFLLMDALKLSIEDPSHEGE
GIPLYDAIKCMKTFFGWKEPNVVKPHEKGINPNYLLAWKQVLAELQDIENEEKIPKTKNM
KKTSQLKWALGENMAPEKVDFEDCKDVSDLRQYDSDEPESRSLASWIQSEFNKACELTDS
SWIELDEIGEDVAPIEHIASMRRNYFTAEVSHCRATEYIMKGVYINTALLNASCAAMDDF
QLIPMISKCRTKEGRRKTNLYGFIIKGRSHLRNDTDVVNFVSMEFSLTDPRLEPHKWEKY
CVLEIGDMLLRTAVGQVSRPMFLYVRTNGTSKIKMKWGMEMRRCLLQSLQQIESMIEAES
SVKEKDMTKEFFENKSETWPIGESPKGVEEGSIGKVCRTLLAKSVFNSLYASPQLEGFSA
ESRKLLLIAQALRDNLEPGTFDLGGLYEAIEECLINDPWVLLNASWFNSFLTHALK
>AAT76165.1 entry:AY653199 /gene="PB1" /product="polymerase basic protein 1"
[Influenza A virus (A/chicken/Jilin/9/2004(H5N1))]
MDVNPTLLFLKVPVQNAISTTFPYTGDPPYSHGTGTGYTMDTVNRTHQYSEKGKWTTNTE
TGAPQLNPIDGPLPEDNEPSGYAQTDCVLEAMAFLEESHPGIFENSCLETMEIVQQTRVD
KLTQGRQTYDWTLNRNQPAATALANTIEIFRSNGLTANESGRLIDFLKDVMESMDKEEME
ITTHFQRKRRVRDNMTKKMVTQRTIGKEKQRLNKKSYLIRALTLNTMTKDAERGKLKRRA
IATPGMQIRGFVYFVETLARSICEKLEQSGLPVGGNEKKAKLANVVRKMMTNSQDTELSF
TITGDNTKWNENQNPRMFLAMITYITRNQPEWFRNVLSIAPIMFSNKMARLGKGYMFESK
SMKLRTQIPAEMLANIDLKYFNELTKKKIEKIRPLLIDGTASLSPGMMMGMFNMLSTVLG
VSILNLGQKKYTKTTYWWDGLQSSDDFALIVNAPNHEGIQAGVDRFYRTCKLVGINMSKK
KSYINRTGTFEFTSFFYRYGFVANFSMELPSFGVSGINESADMSIGVTVIKNNMINNDLG
PATAQMALQLFIKDYRYTYRCHRGDTQIQTRRAFELKKLWEQTHSKAGLLVSDGGPNLYN
IRNLHIPEVCLKWELMDEDYQGRLCNPLNPFVSHKEIESVNNAVVMPAHGPAKSMEYDAV
ATTHSWIPKRNRSILNTSQRGILEDEQMYQKCCNLFEKFFPSSSYRRPVGISSMVEAMMS
RARIDARIDFESGRIKKEEFAEIMKICSTIEELRRQK
>AAT76157.1 entry:AY653193 /gene="PB2" /product="polymerase basic protein 2"
[Influenza A virus (A/chicken/Jilin/9/2004(H5N1))]
MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALRMKWMMAMKYPITAD
KRIMEMVPERNEQGQTLWSKTNDAGSDRVMVSPLAVTWWNRNGPTTSAVHYPKVYKTYFE
KVERLKHGTFGPVHFRNQVKIRRRVDINPGHADLSAKEAQDVIMEVVFPNEVGAKILTSE
SQLAITKEKKEELQDCKIAPLMVAYMLERELVRKTRFLPVAGGTSSVYIEVLHLTQGTCW
EQMYTPGGEVRNDDVDQSLIIAARNIVRRATVSADPLASLLEMCHSTQIGGIRMVDILRQ
NPTEEQAVDICKAAMGLRISSSFSFGGFTFKRTSGSSVKKEEEVLTGNLQTLKIRVHEGY
EEFTMVGRRATAILRKATRRLIQLIVSGRDEQSIAEAIIVAMVFSQEDCMIKAVRGDLNF
VNRANQRLNPMHQLLRHFQKDAKVLFQNWGIEPIDNVMGMIGILPDMTPSTEMSLRGVRV
SKMGVDEYSSTERVVVSIDRFLRVRDQRGNVLLSPEEVSETQGTEKLTITYSSSMMWEIN
GPESVLVNTYQWIIRNWETVKIQWSQDPTMLYNKMEFEPFQSLVPKAARGQYSGFVRTLF
QQMRDVLGTFDTVQIIKLLPFAAAPPEQSRMQFSSLTVNVRGSGMRILVRGNSPVFNYNK
ATKRLTVLGKDAGALTEDPDEGTAGVESAVLRGFLILGKEDKRYGPALSINELSNLAKGE
KANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN
```