Name..... ID .....

Practicals on Protein Function Prediction (454)

In this practical, you will learn how use to bioinformatics tools for functional annotation of unknown/ uncharacterized/ hypothetical genes and proteins. First, we will use a conserved hypothetical nucleotide sequence obtained from *Neosartorya fischeri* NRRL 181.

>gi|119474158|ref|XM_001258954.1| Neosartorya fischeri NRRL 181 conserved hypothetical protein (NFIA_004140) partial

ATGGAGAGATCGAAATCGAAACCAGCCCCCGGCCAATCCTGCACCTATGGCCAAGCTTGC
ACGCATTGCTACAAGGCCAAGTGTCGGTGTGTGCGTGCCCCAAATGGTGATACTTGTGAA
AGGTGCCTTCGTCTCAAGAAAAGGTGTGAGCCATCAGAGTCGGTTCGCCGACGGAATGCT
CAAAATGCCCAGACAGCCAAAGTATCCGATAGGCGGATTGCCCGGCTGGAGGACAAAATG
GAGAGCTTGTTATCTGCCATGAACTCTTTCCTTGGTTCCACGGCGAGTTCGGGGTCTGCTG
TCAACGTCCTTCAATCTCTGCACGGAGACGGCATTTCATTATCCACGTCTTCTTCGAATACC
ACTCTAGTAACCCCTGCGAGCACTACTCCGCGCTTCACTGAGGGACTCAACTTCGCGACA
GACGCGCTTCATTATTCACTATCTACCCCCGCTCCATCGCCAAACCAAGCAGACGAAAGGC
TCAATTTCTTCCGGTCCCGAATGCTGCCTTCCTTTCCTTTCATCGATCTGACTCCAGACATT
ACAAGCTGCTATCTGCGTCAAAACAGGCCCTTCTTGCTCCAAGCCATCCACACTGTGACCA
CATTCTCGACCCAGGAAAGACTGACTCAGGTGGAGGAGCTGAAGCATCTGCTATTCACATC
GGCTTTGCTACAAGTCCAGTCAAATATCGACCTGCTGCTTGGATTGCTAACATACCTAGCA
TGGAGTACTGACCCCTTTCTTGGCCGAGCTGACCTCGTCTCTCGCCTCATGATGCTGGCAA
TTTCACTCGTCTATGATCTGCGATTGTTCAAACCATCCTCGCCGGACGTGGAACTCATGAT
GACTATTACCCAGGGGCGGGCGGATGACAATAATCAAAGCCACAACAATGAAACGCACCA
CGACTTATTGGAAAGACAGCGGGCAGTACTGGCGTGTTTCATTTTGAGCTCTAATATTGCG
TCCCACCTTGGGCGTCAAGACGCTCTACGATGGACACCTCAGATGGAAGAGGCGCTTCGA
GTCCTCACAATAAGCGAGGCATGTCCTGCAGATCGGCTATTCGTCTCTCAGGTCCGCTTGC
AGTTGCTAAAGCAAAGAGCAGATGATGTCCGACAACAGGACGAGGCTCACACAGGAACAG
CTCCTGCGGCGGTTTCAGCTCCTCGTCTATTGTATCTAAAGTCTTTACGAAGGGAGCTACA
CGAGCTAAGATCTTTGTTTCCTCCGGATCTCCCCCAGCTAAACATCCTCAATGCACACGCC
CAATACGTCGAATTATACATAAACCAGCTCGCCTATTCCGTCAGCCAAAACTCGCTTCCTCT
CAGTCTGACCGGACAACTGGGATTCGAACATCTGAAGTGTCTGTGGCAGTCGGTTGAGAA
CATCAAGTCGTGGCTGGACCATTTCTACCAGATCCCTTGCTCGGACCTTGTCGGCCAGCC
CTTTCATTTTTGGTCCCAGATGATTCTGACAGTTACACTGTTGAAATACCTCTCAACACTTCA
AGACCCTGAATGGGATTGCCAGGCGGTGCGGGGAACAGTCCACCTGATCTCGACGATGG
ACTGTATGATTCAGAAGCTCGATCTGAGCAGCAAAGAGCCGGAGCTTCAGTGCGACGACC
ATTTACTCAAGTTTTTATCCAAGCTTTTAACCAGATGTCGTCTGTGGGCCGAAGCTCGATGG
CATGATGAGGAGACCGGGCCGGGCCGGAGCGCCAGCTGTGACACCACTGGTCACAATCA
TCATATCCCGGAGCTGGATCAGATGGTCTGGATGCAGTCGATGGATTTGGGGGATGATCA
GTGGTTTGAAAATGTACTGGGTATGCCCACCACATTCTACTAG

Name..... ID .....

From the given nucleotide sequence, can you calculate %GC of this sequence? Also, tell the tool or program that you use.

Of course, we want to analyse the protein, but the lecturer has given you the nucleotide sequence. Can you convert or find the protein sequence corresponded to this gene? Also tell me how?

Once you obtained the protein sequence of this hypothetical protein. Please identify that which amino acid residues are mostly present in this protein and its chemical properties.

If we want to check whether this protein has homologs/ orthologs in the protein sequence databases, which database will you choose and which analysis will you perform? As you guess, we should do Blasting. But which BLAST option is suitable and why?

What your Blast result looks like? Do you understand it? Can you explain this to your friends? Surely, the first hit will be from *Neosartorya fischeri* NRRL 181. Can you conclude anything from this stage about this hypothetical protein? At what level of significance? Are there parameters that you should know, % identity, % coverage, e-value?

If you want to confirm this Blast result, you might try detailed characterization of this protein by a variety of tools that I have introduced to you. Easily, we can scan for the conserved domains/ motifs in the protein sequence by using PfamScan, InterProScan and SMART programs. Simply input your protein sequence into the text box and click submit. Can you explain the results? Did they return different ones? Please explain.

Until this step, you might be able to confirm some domains within this protein. We could try annotating the function of this protein by looking back to the BLAST result again. By using sequence-based function prediction, this protein hit with the C6 transcription factor several times. Can this protein be annotated as this transcription factor? To confirm it, you have to check identical level between the two proteins. Can you align your hypothetical protein sequence with the second hit from *Aspergilus*? Explain your method.

How do the aligned sequences look? Are they identical enough to be annotated as the same protein? If yes, please give reasons for your judgement.

Name..... ID .....

NOTE: To decide which method will be used for functional prediction. We would recommend to begin with the simple method first. If not enough information for functional annotation with confidence, more sophisticated method will be used. Sometimes multiple programs won't give much meaningful results back. Don't step back... You might have to try another programs or this protein is really the case of hypothetical one.

For the second approach of protein function prediction, we could predict the function of this hypothetical protein by inferring from the biochemical network/pathway that it is part of. Now let's predict the possible pathway for this protein using STRING program. Are there any interactions with this protein? By which evidence it is supported? Can you interpret this result?

Next, we will try to understand this protein by using the third approach of non-homology-based prediction. Try submitting your protein sequence to the WoLF PSORT program (http://www.genscript.com/wolf-psort.html) and do not forget to select the correct group of organism. SO from the result, which cellular location this protein should be? Does it support the previous predictions by other approaches? Any other tools that you can use to gain further information about the properties of this protein? Such as, lipid modification, transmembrane protein and signal sequence.

The last approach is the structure-based function prediction. Obviously, to do this you have to predict either 2D or 3D structure of this protein. Can you do this? How will you do, please explain?

While waiting for your structural prediction, you can try structural blast, HHPRED program which will try finding the protein whose structure matched with your hypothetical protein. Have you found any similar structures? What are they? Does this information support the results from the previous approaches? Please explain.

NOTE: The HHPRED program can also predict the structure of your hypothetical protein based on the closet structural matches by using MODELLER program.

Extra: Once you obtain the predicted structure, can you find the important domain within this structure? And also, can you explain the structure of this protein?

Name..... ID .....

At the end of this tutorial, please write a short summary of all analyses you have done. Explain if you have enough support for the annotation of this hypothetical protein.