Exercises

December 3, 2009

1 Introductory remarks

General comments We have attempted to compile a set of exercises and questions that represent a selection of typical steps you will perform, and problems you might encounter, when analyzing biological data. The data you will work with is, whenever possible, the same you will start the exercises with. It is thus as close to the reality as possible. However, in a few cases the data do not have all the features we would like them to have. We might then have to switch to a different data set, to allow for the planned analyses.

Web pages and data bases In the course of the exercise, you will need to go to several web-sites and databases to collect data for your analyses. Unfortunately we lack the time to mention every single of these web sites in detail. If you are not already familiar with these pages, please take some time and have a look what kind of information they provide. However, before you get completely lost, please ask!

Functions and programs Similar to the web pages and databases, a number of unix-functions, such as *less, chmod, sed, grep, head, tail, tr, sort, comm, uniq* might come in handy for data manipulation and quick data analysis. Again, we will not be able to give a thorough introduction into every single one of these functions. It is up to you what way you choose to complete the exercises. However, make sure that your approach is scalable to larger amounts of data than the one we will work with. If you are not sure how to complete a certain task, or if you are interested in a different way of doing things, please ask. Please keep in mind that now and then it's worthwhile to read and think for an hour finding out how a program can do something for you in less than a second, even

though you would only need five minutes to do it manually. Again, we can only suggest how to accomplish certain tasks. If you find other ways to be more efficient, go for it.

Documentation and solutions One part of the exercise is to document the individual steps of your analyses. Please enter the answers to the individual questions you'll find below at the appropriate place in your documentation. Furthermore, please add a remark to the individual questions, whether you find them

- trivial
- appropriate
- complex
- too difficult

We will collect your documentation at the end of the course(!), so please make sure that it is structured, complete and that you either do have it in electronic format or keep a hard copy for your own record.

2 Sequence retrieval and ortholog prediction

The first set of exercises and questions is concerned with putting together a data set to address the question of the phylogenetic position of our taxon of interest.

2.1 Collecting a dataset

1. ! Download the file *ENCCU_prot.fa.gz* from the following URL:

   http://www.cibiv.at/~ingo/Brno/data

   and unzip it.

   Note, the sequences are in *fasta-format*. In this format, each sequence has to be preceded by a sequence header starting with a >. Any information can be stored in this sequence header. The actual sequence starts in the next line following the header. Both, protein and DNA sequences can be stored in fasta format. Individual sequences, multiple unaligned sequences and alignments can be stored in fasta format.

2. How many sequences does this file contain?

3. ! Extract the sequences with the sequence identifiers *Q8SSC4* and *Q8SQP5* from the file *ENCCU_prot.fa* and address the following questions:

   (a) From what species are these sequences most likely derived?

   (b) What are the likely functions of these proteins?

   (c) Compare the evalues and the scores between the best BLAST hit and the next best Hit. Which one is informative in ranking the hits?

   Please use the BLAST tool at

   http://blast.ncbi.nlm.nih.gov/Blast.cgi

   for this purpose.

4. ! Find the complete protein sets for the following taxa:

   - Saccharomyces cerevisiae (SACCE)
   - Homo sapiens (HOMSA)
   - Arabidopsis thaliana (ARATH)

   Note: Use the diArk-web pages at

   http://www.diark.org

   as a start for your search.

   (a) How many choices for each taxon does diARK give you?

   (b) ! Download the protein sequences from an appropriate data repository in fasta format. *If you have the choice, choose ENSEMBL BioMart and use protein_id and gene_id as sequence identifier. Filter for genes with a protein_id assigned and a gene status known*

   (c) ! Rename the files to adhere to the following standard: SPECIES_prot.fa. (Please use for *SPECIES* the short versions given above, e.g. SACCE).

   (d) How many protein sequences are in each data set. Compare the values to the numbers of protein coding genes present in each of the genomes and explain your observation.

(e) ! Process the ARATH, HOMSA, and SACCE protein sets using the shell script *unique.sh* and *nentferner.pl* that you can download from

`http://www.cibiv.at/~ingo/Brno/scripts/`

Note, *unique.sh* will call the script *nentferner.pl* and expects it to be in your executable path.

(f) Describe the individual steps and functions in the script unique.sh. Note, *nentferner.pl* just removes newlines in a fasta file. Issue *nentferner.pl -h* for more info.

2.2 Orthology prediction

In this part of the exercises you will perform ortholog predictions using the sequences you have retrieved in the previous steps. The aim is to compile groups of evolutionarily related sequences that can be, in the next step, used to infer the phylogenetic relationships of the species under study.

5. ! Use the two sequences *Q8SSC4* and *Q8SQP* and perform a local reciprocal Blast search between

   (a) ENCCU and HOMSA

   (b) ENCCU and ARATH

   (c) ENCCU and SACCE

   Check whether you can identify orthologous sequences to the two query proteins in the three other species. If so, combine the orthologs for each protein in a file and name this file *Q8SSC4_ortho.fa* and *Q8SQP5_ortho.fa*, respectively.

   To complete this task, you need to use the local blast version *blastall* installed on the system. Perform the follwowing steps:

   (a) Reformat the fasta files for the four proteomes separately into blast databases. Use the program *formatdb* for this purpose.

   (b) Put the blast databases into a directory blast_db

(c) Use *blastall -p blastp* to search for hits in *SACCE* with the first query *Q8SSC4*. Issue the command *blastall* to obtain all the options for this program.

(d) Repeat the reciprocal blast search for *ARATH* and *HOMSA*.

(e) Repeat the entire procedure for the second protein *Q8SQP5*.

(f) Explain what you would need to do to confirm orthology for all proteins in Q8SSC4_ortho.fa.

6. ! Start an InParanoid ortholog search between all *ENCCU* proteins and all proteins from one(!) of the other three species. Talk to your neighbor and make sure that within the course all 3 searches are performed. Once the InParanoid searches are completed make sure that you obtain the results from the two searches you have not obtained.

7. ! Go to the OMA project at

   http://omabrowser.org

   and identify the OMA ortholog groups that correspond to the two genes you are analyzing. Use the protein search function in the OMA browser.

8. ! Download the two ortholog groups in fasta format and complement the files *Q8SSC4_ortho.fa* and *Q8SQP5_ortho.fa* with the taxa listed in the file *taxon_list.txt* you can download from

   http://www.cibiv.at/~ingo/Brno/taxon_list.txt

9. Use *mafft* to align the sequences in the two files. Use the most sensitive alignment settings *−−maxiterate 1000 −−localpair*.

10. Use *clustalw -convert* to convert the fasta format from the mafft output into phylip format. Issue the command *clustalw -help* for more info. What does clustalw usually do? What are the problems you may encounter with the sequence ids?

2.3 Maximum Parsimony tree reconstruction

Use the web page at

http://mobyle.pasteur.fr/

for computing the maximum parsimony trees. Perform 100 Bootstrap replicates to assess how good the data supports the tree. Please note, the 100 bootstrap replicates can take a while. As a suggestion, first compute the MP tree that you have something to work on and then afterwards compute the 100 boostrap maybe overnight.

1. ! Compute the maximum parsimony tree for the dataset *Q8SSC4_ortho.phy*
2. ! Compute the maximum parsimony tree for the dataset *Q8SQP5_ortho.phy*
3. Visualize the two trees with *figtree*, compare and discuss the differences.

2.4 Maximum likelihood tree reconstruction

Use *raxmlHPC* and the model PROTGAMMAIWAGF for tree reconstruction. Alternatively, you can also use the online tools at

http://mobyle.pasteur.fr/

for this purpose. If you choose the online tools then use the options

- substitution model: WAG
- estimated proportion of invariable sites *-v*

Please note, performing 100 bootstrap replicates can take a while. As a suggestion, first compute the ML tree that you have something to work on and then afterwards compute the 100 boostrap maybe overnight.

1. !Compute the maximum likelihood tree for the dataset *Q8SSC4_ortho.phy*
2. !Compute the maximum likelihood tree for the dataset *Q8SQP5_ortho.phy*
3. Visualize the two trees with *figtree*, compare and discuss the differences.
4. Discuss the results from the MP and the ML analysis.

2.5 Distance tree reconstruction

This part is optional! Use the web tools at

http://mobyle.pasteur.fr/

for this purpose. Use the programs *protdist* to construct a distance matrix and *bionj* for tree reconstruction.

1. Compute the distance tree for the dataset *Q8SSC4_ortho.phy*
2. Compute the distance tree for the dataset *Q8SQP5_ortho.phy*
3. Visualize the two trees with *figtree*, compare and discuss the differences.
4. Compare to the results from the MP and the ML analysis.