Exercises

February 4, 2008

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 throughout the course. However, in a few cases the data does not have all the features we would like them to have. We might then have to switch to a different data set then, 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* might be handy to use 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 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 the documentation of the individual steps you have done during 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 Data retrieval

The first set of exercises and questions is concerned with putting together an initial data set for the analysis of EST data from Xenoturbella bocki

2.1 Collecting a dataset

1. Visit the web site at

   http://www.ncbi.nih.gov/dbEST

   How many ESTs from Xenoturbella are available from this data source? In what formats is the data available? Can you get access to base quality values or trace data?

2. Find the trace archive at the NCBI home page.

   - How many traces are available from Xenoturbella?

- Is the data set the same as in dbEST?
- In what formats can you access the data?
- In what aspects does the data in the trace archive differ from the one in dbEST?
- Download the fasta files and the associated quality values for the Xenoturbella ESTs.

3. Unpack the downloaded information. Create a directory *Xeno_fasta_set*, change into this directory and create a soft link to the first 1000 Xenoturbella fasta files. We will use this subset for further analysis.

2.2 Cleaning of EST sequences

1. Get information about the vector sequence flanking the actual cloning site the Xenoturbella cDNA was inserted into from the information you have downloaded from the trace archive.

2. Generate a file *cloning-site.fasta* and put the sequence information retrieved in the previous step into this file.

3. The vector *pGEM-T was used for construction of the Xenoturbella EST library. Retrieve the sequence information for this cloning vector from the www. (Hint:* VecBase is a good data source).

4. Check for the presence of the programs *lucy and* cap3 on your computer. If it is not available on your computer, download the source from

   http://www.cibiv.at/ ingo/applied_bioinf/

   and perform a local installation of these programs.

5. Change to the directory *Xeno_fasta_set and run* lucy on the 1000 files. Use the following parameter settings:

   - -m 100
   - -cdna 15 4 450
   - -r 50 100 350
   - -b 10 0.02

- -w 50 0.03 10 0.3
- -e 1 1
- -v pGEM-T.fa cloning-site.fasta
- -i -d $i.info
- -output $i.out.fa $i.out.fa.qual

6. What is the meaning of the values chosen? What improvements would you suggest?

7. Screen for empty output files. Why do they occur? How many files remain after you have removed them?

8. Run the script remover.pl on each pair of lucy output (fasta and quality file). You can obtain this script from:

   http://www.cibiv.at/ ingo/applied_bioinf

   Alternatively, write your own script to remove the regions *lucy suggests for clipping.*

2.3 Clustering of ESTs

1. Put the sequences cleaned with *lucy into a multi-fasta file. Name this file* Xeno-clipped.fa. Use the actual name of the sequences as the fasta header thereby omitting the file ending. Do the same with the quality information. Name this file *Xeno-clipped.fa.qual. Make sure that the fasta headers of each sequence-quality pair are identical.*

2. Create a directory *cap3 and link the two files* Xeno-clipped.fa und *Xeno-clipped.fa.qual into it.*

3. Run the program *cap3 on Xeno-clipped.fa using the default values.*

4. How many contigs have been generated and how many sequence reads did make it into a contig?

5. What kind of information do you find in the file **.cap.info and* *.cap.ace? What is a *chimera in this context?*

6. Put all the names of the sequences that have been assembled into a contig into a file named *Xeno_seqs_in_contig.txt.*

2.4 CLC-workbench

You will now repeat the analysis you have done with open source software with a gui-based commercial software. Make sure to get an idea about possible advantages and disadvantages of this workbench.

1. Create a directory *CLC and a subdirectory* CLC/traces. Copy the traces for the first 99 filenames in your list *Xeno_seqs_in_contig.txt into the directory* CLC/traces. Copy also the file Xb_MM1_02B09.scf

2. Start the CLC-workbench and import the 100 traces. Have a look at the possibilities to visualize the sequences. What are the options in the *Toolbox menu.*

3. Trim the 100 sequences using the default options (screen agains VecBase). While this is processing, continue with the next exercises.

   - What are the differences to the clipping results obtained from *lucy. Focus on the sequence* Xb_MM1_02B09.
   - Repeat the clipping only for this sequence, this time providing the sequence information about the cloning site. Compare the results.

4. Load the sequence of the cloning vector *pGEM-T into the workbench. Display the vector in circular form.*

5. Generate a restriction map for this vector using the default enzymes. Where are the restriction sites located in this vector.

6. Perform an ORF prediction on the vector sequence. GIven the result, comment on the position of the restriction sites. Why are there no restriction sites on other positions of the vector?

7. Perform a BlastP search with the ORF, the restriction sites are located in. What protein obtains the best hit?