

# Exercises

February 7, 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'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/](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](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 against 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?

### 3 Characterization of the EST contigs

In this section, we will try to get an idea what the individual EST contigs, we have generated in the previous section might code for. This involves a number of steps starting from similarity searches and ORF-prediction and extending to characterization of the encoded protein and assignment of functional annotation. For performance reasons, it might be helpful to limit the blast search to 10 sequences at a time when using public services. However, if you find efficient ways to analyze all your sequences, feel free to go ahead. Please note, for many steps I cannot provide a clear-cut and well defined guide line how to perform the analyses. Rather, I try to point out what information exists, and where to look for it. It is up to you to gather as much as possible of information for the individual sequences in your data set. And... always check for alternatives for the suggested analyses!

#### 3.1 Blast and CDS identification

1. Blast the EST contigs you have generated in your *CAP3 analysis against a public database*. Which of the various variants of Blast, and which of the publicly available databases might be useful? Try the Blast using a Public Server e.g.

<http://www.ncbi.nlm.nih.gov>

2. Repeat the Blast with the CLC workbench. Compare the results to those obtained from the public data base.
3. You can also run the blast locally using the program *blastall*. *Play around with it and ask questions*.
4. *Extract the Identifier for the best hit for each EST contig and get the the corresponding sequence. Rename the Xenoturbella EST contig that triggered the hit according to the name or id of the best blast it. Watch out, change both the sequence name AND the fasta header and make sure to keep track of the original sequence name.*
5. *Login via SSH to the embnet server at*

*<http://emb1.bcc.univie.ac.at>*

*activate the module* Genewise. Find out what genewise does and why it might be useful in our analysis of ESTs.

6. Extract the protein coding part from your subset of ESTs. Check what approaches the workbench offer.

### **3.2 Annotation of the protein sequence**

7. Translate the CDS obtained in the previous steps. Find a tool that does it for you.
8. Somebody has told you that *Annotator is a good software to to protein annotation and also mentioned the name* Frank Eisenhaber. Find the web page for this program. What options does it give you?
9. Perform a standard annotation of your proteins and think about the results and discuss with the tutors. What do you learn about the proteins?
10. Move on with your annotation and proceed to the web page of the gene ontology project:

<http://www.geneontology.org>

What information can you get from this web page and how can you search for it?

11. Perform GO annotation with one or two of your favorite translated EST contigs. Monitor the results.
12. Find out about the meaning of Evidence codes.
13. Move on to the web page of the *Kyoto Encyclopedia of Genes and Genomes (KEGG)* at

*<http://www.genome.jp/kegg>*

*and move on to the* KEGG PATHWAY page.

14. What do you see? Do you find a pathway that might be interesting to look at given your results from the GO annotation of your proteins?
15. In what aspects might the KEGG database help you?

## 4 Orthology prediction

In this section we will continue with our analysis of the *Xenoturbella* EST contigs. After you have taken some efforts to learn something about the identity and the putative function of some of you contigs, we are now interested in the evolutionary relationships of *Xenoturbella bocki*. For this purpose we need to put together a set of orthologous proteins we can use for phylogenetic analysis.

### 4.1 Ortholog identification in the *Xenoturbella* ESTs

Rather than using the Blast-based approach applied in the previous exercises for annotation of the *Xenoturbella* ESTs, we will now use the *Hamster-approach* that was explained during the lecture.

1. Go to the Hamstr webpage at

*<http://www.deep-phylogeny.org/hamstr>*

2. Use the Hamstr pages to annotate the EST contigs you have built with CAP3. Choose modelorganisms for the data set, and *Homo sapiens* for the reference species/proteom. Since your ESTs are already clustered, make sure to uncheck the box 'trim and cluster EST sequences'. It will take a while until the analysis is done.
3. What does the Hamstr output tell you? What ist the difference to the blast analysis you have done on tuesday? Do you find cluster that have been annotated in both approaches? Do the results match?

### 4.2 Extension of the dataset

1. Follow the link to the Inparanoid-Homepage for five ortholog groups. Download the corresponding orthologs from

(a) *Apis mellifera*

(b) *Takifugu rubripes*



(c) *C. remani*

(d) *Anopheles gambiae*

*if they are present. Add the sequences to the corresponding ortholog groups you have retrieved from Hamstr.*

2. *Generate a multi fasta file each for each ortholog group. You will need these files for the phylogeny reconstruction section on friday.*

### **4.3 Data management**

*You have done a lot of analyses in the past days. Considering only the analyses performed on the EST data, please set up a relational database scheme that would be suitable to store the raw- and meta-data you have generated so far. Try to think of reasonable keys and constraints you want to put on individual fields in order to guarantee the integrity of your data base.*