BioNumerics Tutorial:

Annotating sequences in batch

1 Aim

The annotation application in BioNumerics has been designed for the annotation of coding regions on sequences. In this tutorial you will learn how to annotate multiple sequences with a single action based on an annotated sequence available in the BioNumerics database.

2 Example data

The features of the batch annotation functionality will be illustrated using query sequences which can be found on the download page of the Applied Maths website (Go to http://www.applied-maths.com/download/sample-data and select "Genome annotation in batch"). The query sequences are derived from publicly available and annotated bacterial chromosomes: annotated coding regions have been removed from these sequences in order to use them in this tutorial. One publicly available annotated bacterial chromosome sequence will be used to annotate these query sequences.

3 Preparing the database

1. Create a new database (see tutorial "Creating a new database") or open an existing database.

2. Import the two sequences stored in the `Genome_seq.txt` file using the instructions described in the tutorial "Importing sequences from FASTA files". Store the accession number, this is the first and only header tag, in the ***Key*** field and save the sequences in a new sequence type called **Complete genome**.

3. Click on the green colored dot of one of the entries in the *Experiment presence* panel to open the *Sequence editor* window.

The upper panel shows the numbered sequence, with bases grouped in blocks of 10. The middle panel shows a graphical representation of the sequence. The zoom slider allows one to continuously zoom in or out on the graphical sequence view. Zooming can be done up to base level. No annotation information is stored in the *Annotation* panel (see Figure 1).

4. Close the *Sequence editor* window.

5. Download the publicly available annotated bacterial chromosome sequence **AE004439** from EBI. Use the instructions described in the tutorial "Importing sequences from online repositories". Store the accession numbers in the ***Key*** field and save the sequences in the sequence type **Complete genome**.

6. Click on the green colored dot of entry **AE004439** in the *Experiment presence* panel to open the *Sequence editor* window.

The imported annotations are stored in the *Annotation* panel (see Figure 2).

7. Click on a particular feature, for example a **CDS**.

The selected feature is highlighted with an orange background in the upper and middle panels and is highlighted in the feature list in the lower panel. The *qualifiers* associated with the selected feature are given in

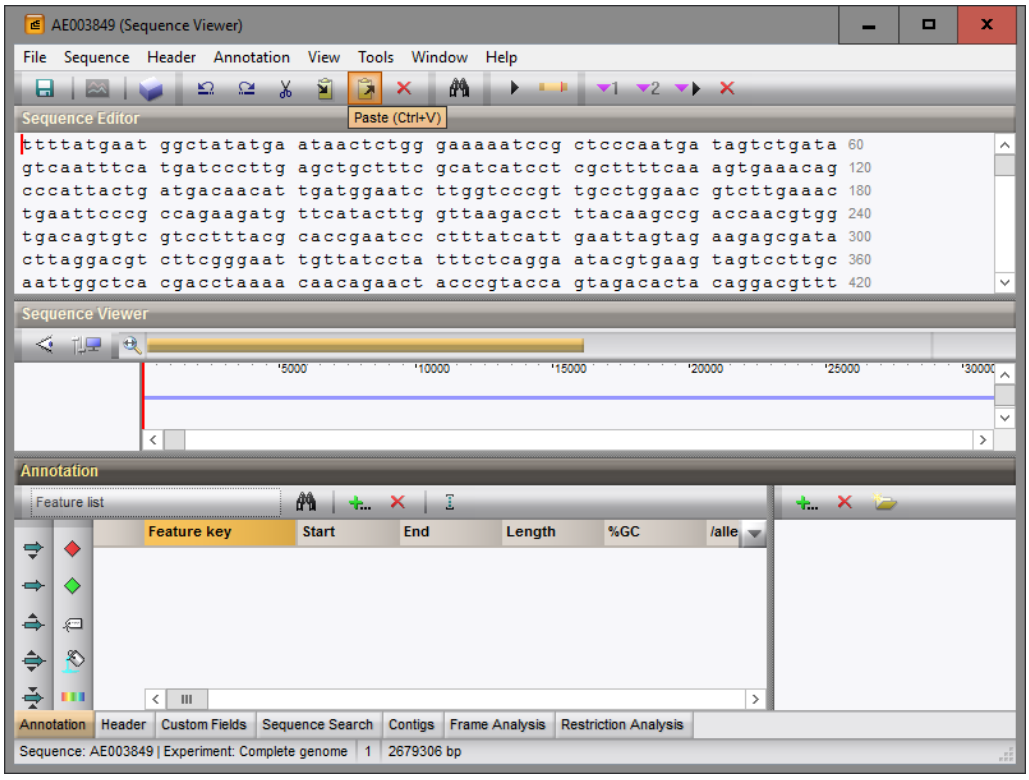

Figure 1: No annotations are present for entry **AE003849** and entry **AE003852**.

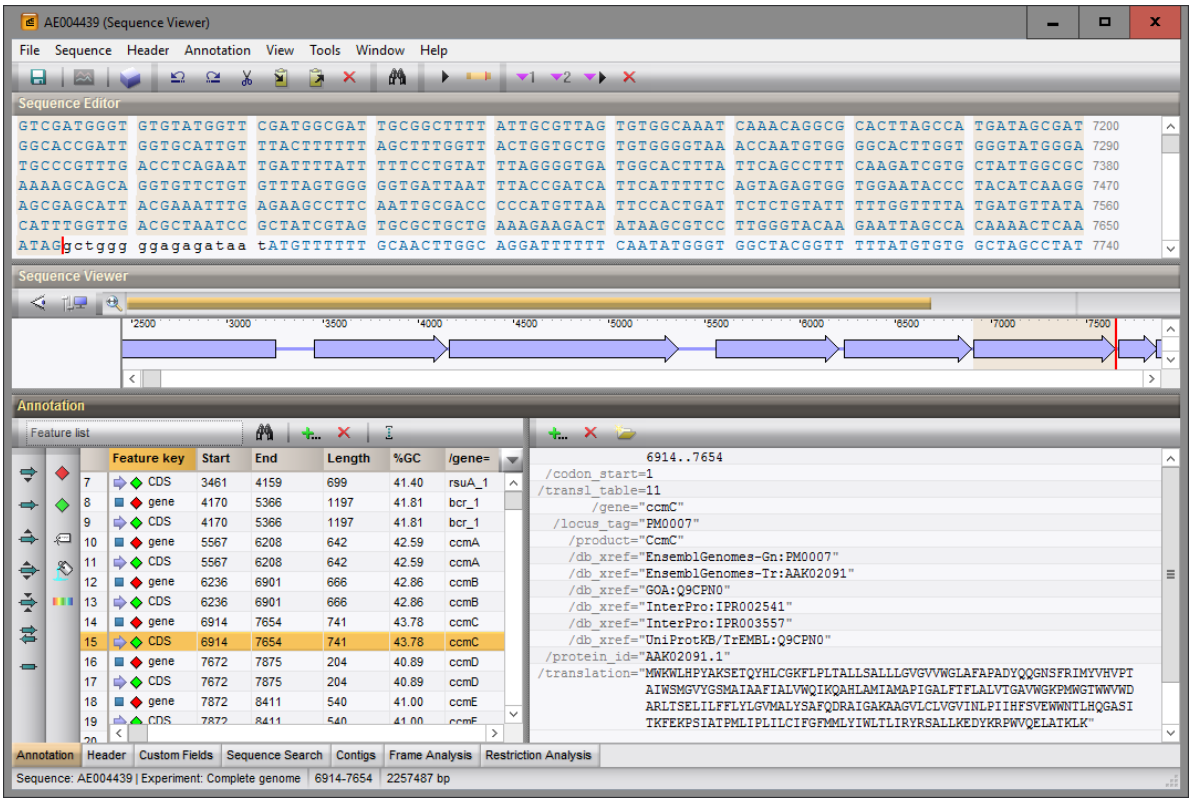

Figure 2: Annotations imported for entry **AE004439**.

the right panel.

8. Close the *Sequence editor* window.

The *Main* window should look like Figure 3 after import of the sequences.

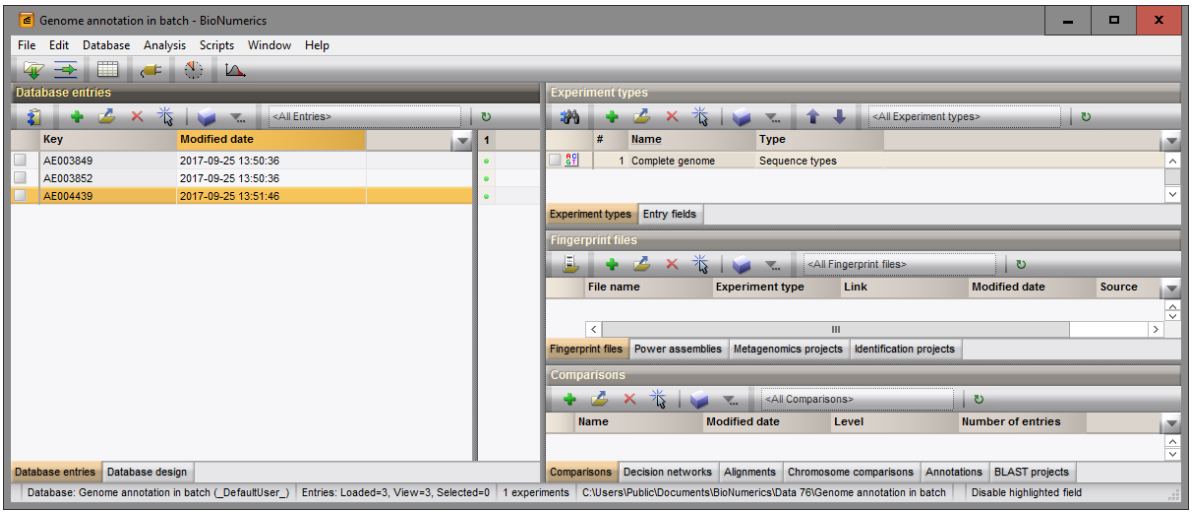

Figure 3: The *Main* window after import of the sequences.

4 Annotating in batch

1. Select entry **AE003849** and entry **AE003852** in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

2. Select ***File*** > ***Process...*** () to call the *Process data* dialog box (see Figure 4).

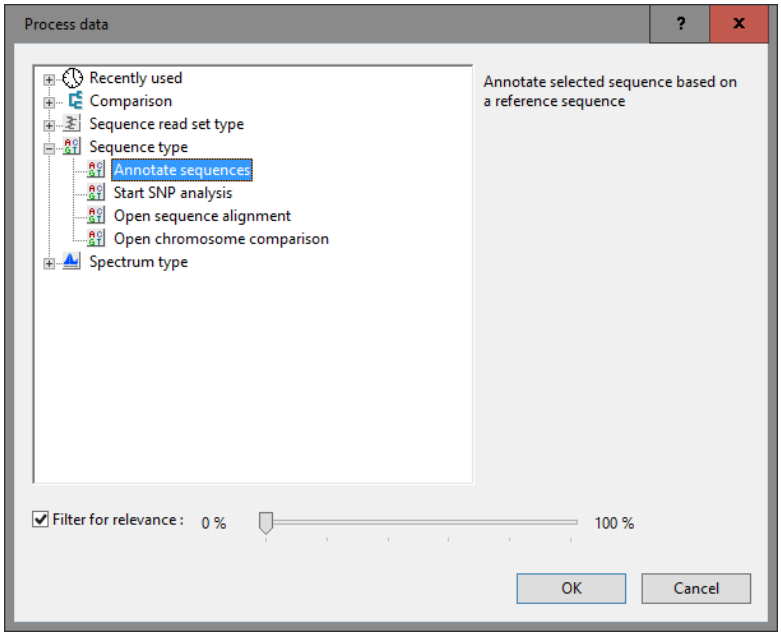

Figure 4: The *Process data* dialog box.

3. Select ***Annotate sequences*** under ***Sequence type*** and press <***OK***> to call the *Annotate sequences* dialog box (see Figure 5).

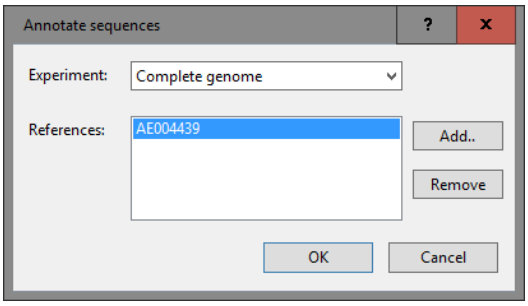

Figure 5: The *Annotate sequences* dialog box.

4. Select the sequence ***Experiment*** from the drop-down list and add the reference sequence containing the annotations (here: **AE004439**) using the <***Add***> button.

5. Press <***OK***> to start the annotation of the selected entries based on the annotations present in the selected reference.

The annotations are stored with the sequences in the *Sequence editor* window.

6. Click on the green colored dot of entry **AE003849** or entry **AE003852** in the *Experiment presence* panel to open the *Sequence editor* window.

The annotation hits that passed the annotation criteria are stored in the *Annotation* panel (see Figure 6).

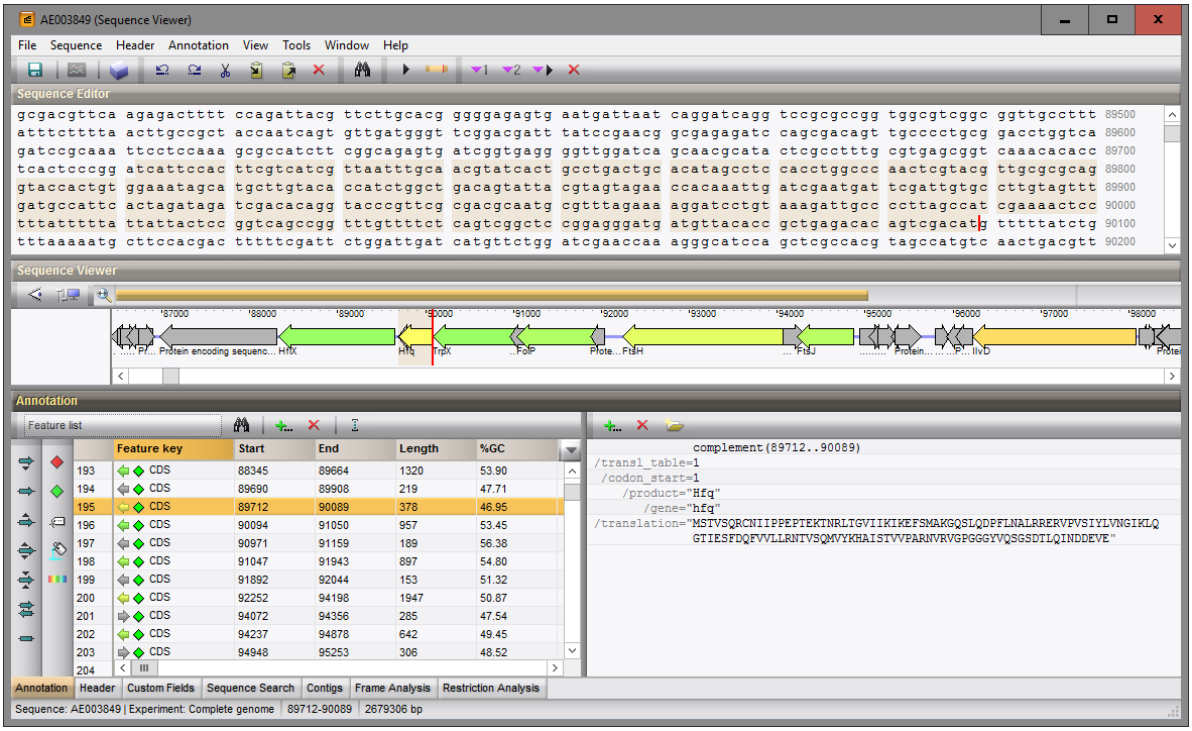

Figure 6: Annotations.

A color code indicates the quality of identification: the color range starts at red (100% identification score), goes over green (50%) and ends by blue (0%). The product description of the hit that is used for annotating the query sequence is displayed in the ***/product=*** column in the *Annotation* panel and below the plotted features in the *Sequence Viewer* panel. The plotted features are colored using the color obtained from the best scoring hit that is used to annotate the query feature. Open reading frames which did not show any hit

with any of the features from the template sequence are plotted in gray on the forward and reverse sequence.

7. Close the *Sequence editor* window.