BioNumerics Tutorial:

Annotating a single sequence

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 a single sequence based on one or more annotated sequences and how to edit the annotation.

2 Example data

The features of the annotation functionality will be illustrated using a query sequence 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"). The query sequence, stored in the EMBL formatted text file `annotation.txt`, is derived from a publicly available and annotated bacterial chromosome (accession number **U00096**). Annotated coding regions have been removed from this sequence in order to generate a bacterial template on which coding regions can be annotated. Three publicly available annotated bacterial chromosome sequences will be used to annotate this query sequence.

3 Preparing the database

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

2. Import the sequence stored in the `annotation.txt` file using the instructions described in the tutorial "Importing sequences from GenBank files". Store the accession number in the ***Key*** field and save the sequence in a new sequence type called **Complete genome**.

3. Download following three publicly available annotated bacterial chromosome sequences from EBI: **AE003849**, **AE003852**, and **AE004439**. 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**.

The *Main* window should look like Figure 1. The three sequences downloaded from EBI will be used as template sequences for the annotation of the query sequence.

4 Creating a new annotation project

In the *Main* window, the *Annotations* panel is displayed in default configuration as a tab in the lower right corner.

1. Make sure the four entries are selected in the *Main* window.

2. To create a new annotation project, select the *Annotations tab* in the *Main* window and select ***Edit*** > ***Create new object...*** ().

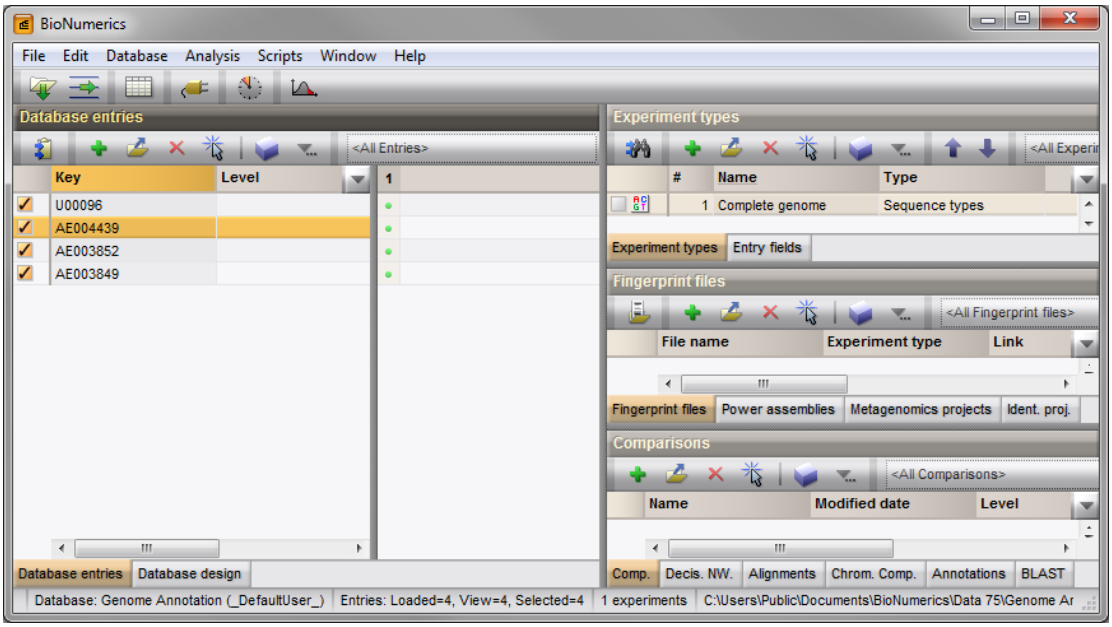

Figure 1: The *Main* window.

A name for the new annotation project is prompted for.

3. Specify a name and press <***OK***>.

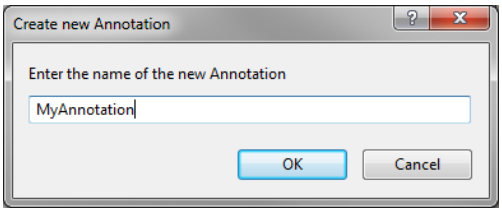

Figure 2: Specify a name for the annotation project.

The new annotation project is added to the *Annotations* panel in the *Main* window and the *Experiment types* dialog box opens. The *Experiment types* dialog box displays a list of available sequence types and the number of associated entries. From this list, the user can select the experiment type that should be included in the annotation project.

4. Leave the **Complete genome** type selected in the list and press <***OK***>.

5 Specifying a query sequence

Before the calculation of an annotation project can be started, one of the sequences in the *Project sequences panel* needs to be defined as the query sequence.

1. Select the row with key **U00096** in our example project and select ***File*** > ***Set query sequence*** ( ).

The appointed sequence is considered as the query sequence in the annotation project and is preceded with a blue dot in the *Project sequences panel* (see Figure 3). All other sequences in the *Project sequences panel* will serve as templates for the annotation of the query sequence in the annotation project.

The *Sequence display panel* shows the query sequence together with a frame analysis overview in the upper

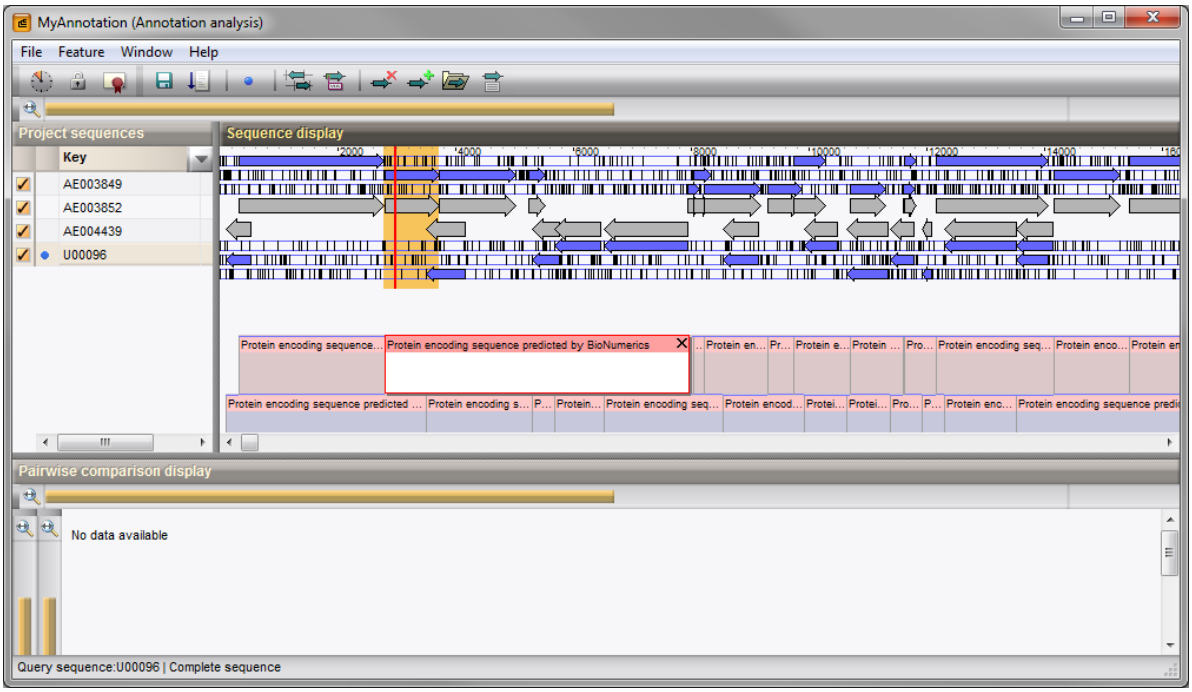

Figure 3: The *Annotation* window after specifying the query sequence.

part of the panel. The ***Standard Code*** translation table is default used to analyze the query sequence in function of its six translation frames, but this can be changed. The three reading frames of the forward strand are mapped above the forward query sequence, the three reading frames of the reverse strand are mapped below the reverse query sequence.

Open reading frames that fulfill the open reading frame settings are plotted in gray on the query sequence, and in blue on the six reading frames.

Within the lower part of the *Sequence display panel*, fields are depicted, corresponding to the coding regions that are mapped on the query sequence. The upper fields are related to the coding regions found on the forward orientation of the query sequence, the lower fields to the coding regions detected on the reverse orientation. The header of the fields is displayed in pink, indicating that the annotation project has not been calculated yet.

The zoom slider, located on top of the *Project sequences panel* in default configuration, allows zooming from full-length sequence view up to base level view. The red vertical line indicates the cursor position on the query sequence.

6 Calculating an annotation project

1. For the current annotation project, call the *Project settings* dialog box with ***File*** > ***Run calculation...*** ( ).

The sequence of calculation steps within an annotation run is the following:

1. In a first step, the six frames of the query sequence are screened for all possible coding regions.
2. Secondly, the possible query coding regions are screened against the coding regions that are mapped on the template sequences. Template coding regions showing any homology with query coding regions are retained and linked to the corresponding query coding region in a ranking based on *feature identity* and/or *chromosome synteny* scores.

3. Finally, query coding regions, for which template coding regions have been found showing a homology level which is above a threshold level set by the user, are annotated based on the descriptions available from those template coding regions. Which descriptions should be used can be defined by the user.

The settings comprised in these three steps are grouped in the *Project settings* dialog box.

2. In our example, do not change the settings and press <***OK***> to start the calculations.

During the calculations, the program shows the progress in the bottom of the window as a percentage and there is a green progress bar that proceeds from left to right.

Upon a finished calculation, the *Sequence display* panel is updated (see Figure 4).

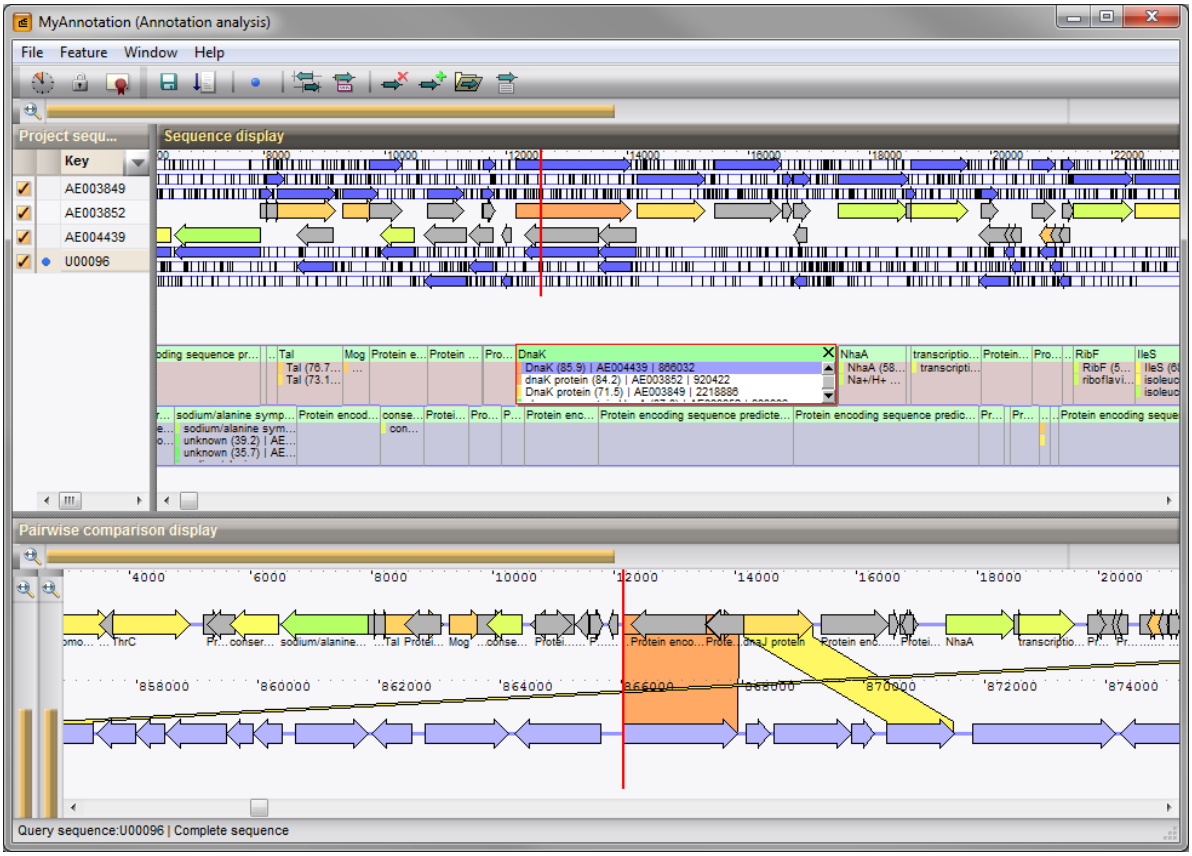

Figure 4: The *Annotation* window after finishing the calculations.

7 General functions

1. Zoom in or zoom out the *Sequence display* panel with the zoom slider located on top of this panel in default configuration. The zoom keeps the cursor position (red line) in focus.

In the lower part of the *Sequence display* panel, the query coding regions and template hits that passed the annotation project criteria are displayed. The header of each coding region is colored in green, displaying the product description of the hit that is used for annotating the query sequence (default this is the description of the best scoring hit, but this can be changed by the use).

2. To view all hits of an open reading frame, select the open reading frame with the left mouse pointer: the field changes into a pop-up chart (see Figure 5).

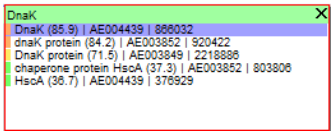

Figure 5: Hits of an open reading frame.

Within an information chart, hits found for that particular open reading frame are listed according their identification score. 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%). Next to the color code stands the product description of the template feature. The identity score of the full query protein with the template protein is displayed next to the product description, followed by the ***Key*** of the template sequence, and ending with the start position of the template sequence.

When a hit is selected from an information chart with the mouse pointer, the query sequence (top) and template sequence (bottom) are plotted in the *Pairwise comparison display panel*, with focus around the selected template and query sequences (see Figure 4). Chromosome parallelism is mapped around the hit. The color of the blocks, depicting the parallel hits, represents the identification score of the hit (red (100%) going over green (50%) and ending by blue (0%)). The alignment in the *Pairwise comparison display panel* can be zoomed in and zoomed out with the zoom slider, located on top of this panel in default configuration. Two additional sliders "Distance cutoff" and "Identity cutoff" are present at the left of the *Pairwise comparison display panel* in default configuration, allowing you to change the layout of the pairwise alignment.

In the upper part of the *Sequence display* panel, all open reading frames that passed the annotation project criteria are plotted on the query sequence and on the six reading frames. The open reading frames are plotted in blue on the reading frames. On the query sequence, the open reading frames that show a hit with the template features, are colored using the color scale used for representing the identity score. The color is obtained from the hit that is used to annotate the query feature (default this is the best scoring hit, but this can be changed by the user). Open reading frames which did not show any hit with any of the features from the template sequences (at least for the given annotation settings) are plotted in gray on the forward and reverse query sequence.

3. A detailed view of a query coding sequence with its respective hits, can be called by selecting the query coding sequence in the *Sequence display* panel and selecting the menu item ***Feature*** > ***Identification details...***. This brings up the *Annotation Detail* window (see Figure 6).

4. Close the *Annotation Detail* window with ***File*** > ***Exit***.

5. Save the annotation project with ***File*** > ***Save project*** (, **Ctrl+S**). All calculations are stored along.

6. Select ***File*** > ***Export annotation table...*** to export the annotations to a csv file.

7. Select ***File*** > ***Save annotated sequence as...*** to save the annotated sequence in the database.

A dialog pops up prompting for the entry key and sequence type. The default suggested settings can be changed if desired.

8. Save the changes to the existing entry U00096 and confirm the overwrite action.

9. Minimize the *Annotation* window and click on the colored dot of the U00096 entry in the *Experiment presence* panel of the *Main* window.

This opens the *Sequence editor* window. The annotations are displayed in the *Annotation* panel (see Figure 7).

10. Close the *Sequence editor* window.

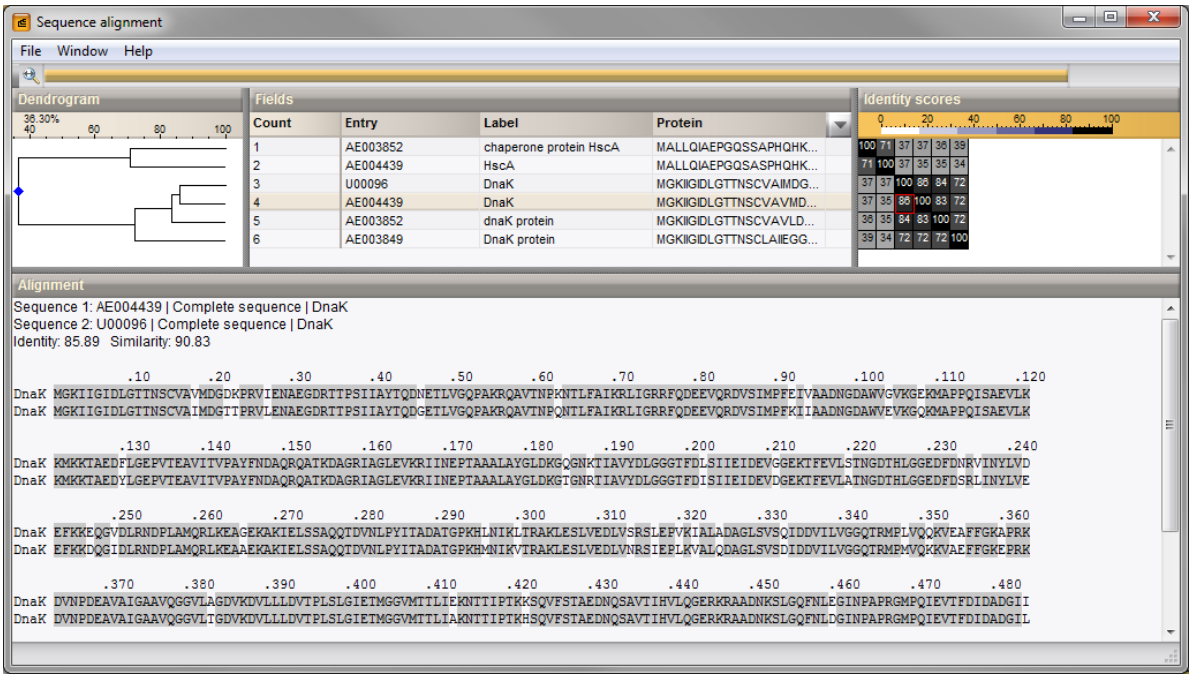

Figure 6: Detailed comparison of a single annotated CDS.

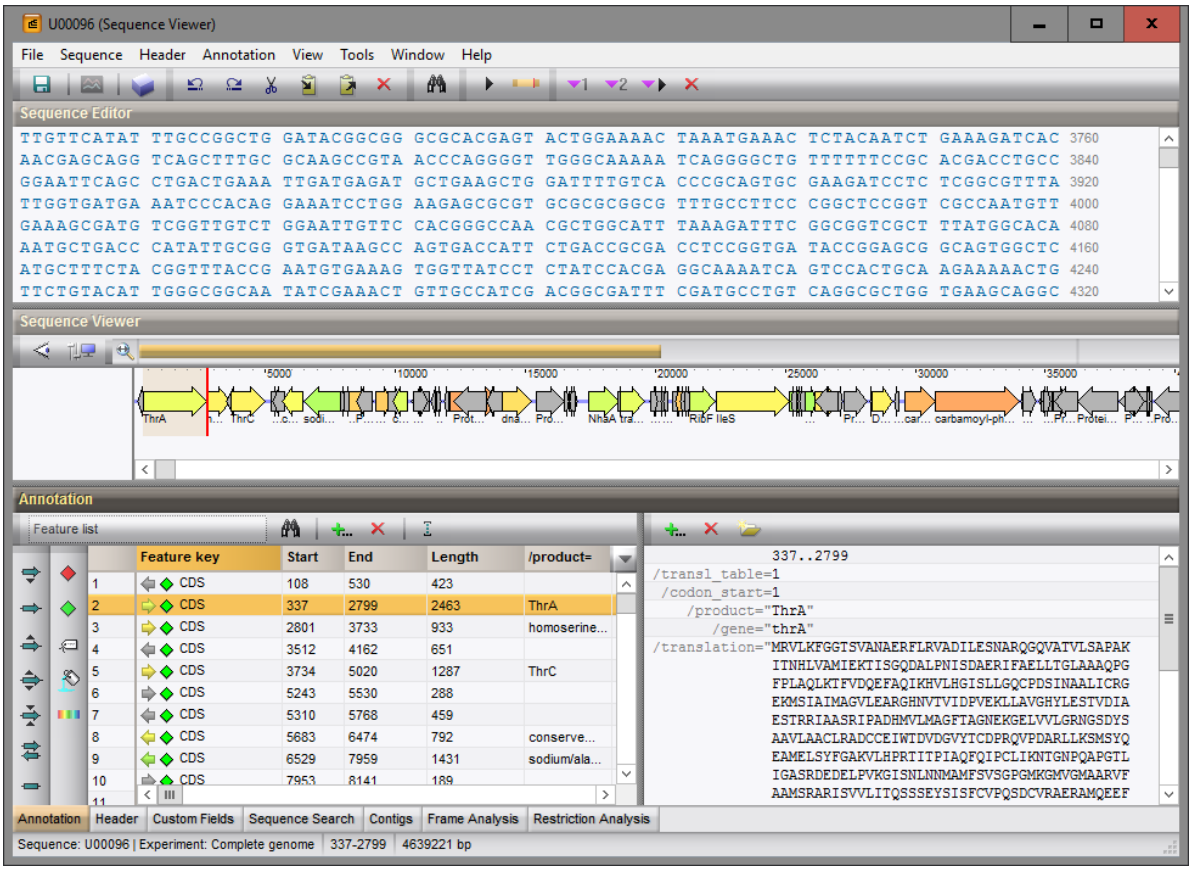

Figure 7: The *Annotation* panel in the *Sequence editor* window.

8 Editing functions

1. To remove a mapped coding region, select the feature in the *Sequence display* panel and select ***Feature*** > ***Delete***.
2. To map an open reading frame region which is not mapped on the query sequence, select the open reading frame in the reading frame in the upper part of the *Sequence display* panel and select ***Feature*** > ***Add from selection...***. Manually mapped features are displayed with a turquoise color on the query sequence.

Standard, the annotation description of the best scoring hit found within the template sequences is used for annotating the query sequence. If desired, next best scoring hits can also be selected for annotation:

3. Select an individual hit within the information chart and choose ***Feature*** > ***Annotate*** > ***From selected hit***.

The query feature is now annotated based on the selected hit and the product description of the selected hit is displayed in the header of the information chart. The product description of the selected hit is displayed with a blue color in the chart.

4. To restore the original annotation of a query sequence, select the feature in the *Sequence display* panel and choose ***Feature*** > ***Annotate*** > ***From best hit***.

The hit with highest identity score is again used to annotate the query sequence and the header information of the information chart is updated. The product description of the best scoring hit is displayed with a blue color in the chart.

9 Blast analysis

Within the *Annotation* window, two different types of BLAST searches can be launched:

- The BLAST functionality under ***Feature*** > ***Annotate*** allows the user to perform a BLAST screening of query open reading frame regions against the EMBL-GENBANK public databases. The results of the BLAST are imported into the annotation project and act as normal template hits.
- Alternatively, the BLAST functionality under ***Feature*** > ***BLAST ...*** refers to *real* BLAST projects and should be seen as a totally independent analysis, starting from the sequence selection in the *Sequence display* panel of the *Annotation* window. When selecting an ORF in one of the six reading frames, the sequence spanning the ORF is selected and the BLAST project can be launched for the selected subsequence with ***Feature*** > ***BLAST selected sequence (DNA)...*** or ***Feature*** > ***BLAST selected sequence (protein)...***.

For detailed information on each of these BLAST searches we refer to the chapter in the BioNumerics manual covering the BLAST functionality.

1. Select a feature and select ***Feature*** > ***Annotate*** > ***Automatic annotation of selected feature...*** in the *Annotation* window.

This action calls the *BLAST search settings* dialog box (see Figure 8).

2. Choose the ***BLAST program type***, and ***BLAST screening database***. Press <***OK***> to launch the BLAST search.

The function runs in background and the submission progress is displayed in the status bar of the *Annotation* window. The imported BLAST hits are marked with a colored diamond in the information charts, whereas template hits, which have been found on the template sequences included within the project, are marked with colored squares. The color code of the signs indicates the quality of identification (red (100%), going over green (50%) and ending by blue (0%). When a query open reading frame region has been screened

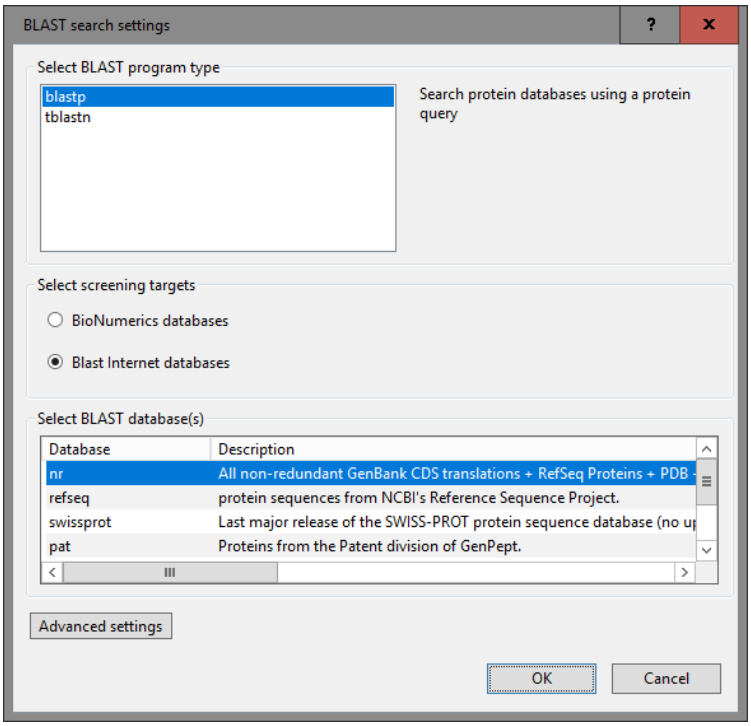

Figure 8: The *BLAST search settings* dialog box.

against public databases via the BLAST-functions, the feature is marked with a purple attachment appearing at the bottom of the information chart (see Figure 9).

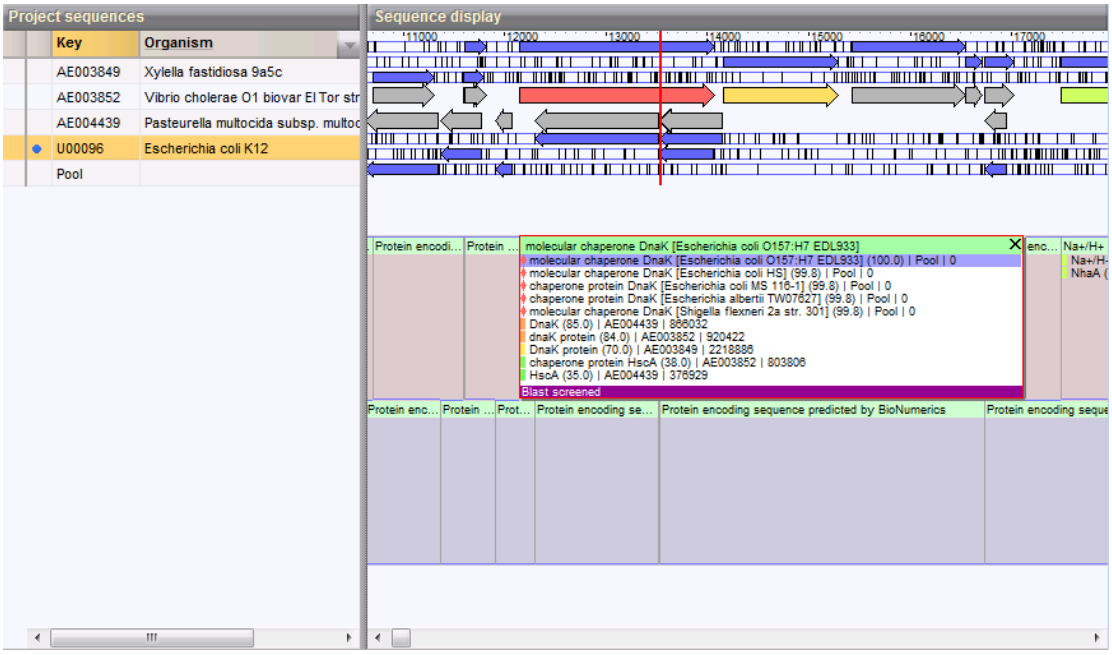

Figure 9: Blast screening of a selected feature.

3. With ***Feature*** > ***Annotate*** > ***Automatic annotation of all unknown features...***, all mapped features which did not show any hit with the template features included within the project can be blasted against a database.

This action also calls the *BLAST search settings* dialog box. For large sequences, it is recommended to run this function overnight.

4. Close the annotation project with ***File*** > ***Exit***.

If unsaved data is present in the annotation project, a dialog box pops up, prompting to save the changes for the annotation project.