BIONUMERICS Tutorial:

Comparing whole genomes

1 Aim

The *Chromosome Comparison* window in BIONUMERICS has been designed for large-scale comparison of sequences of unlimited length. In this tutorial you will learn how to create and calculate a chromosome comparison project and alignment in BIONUMERICS.

2 Preparing the database

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

In order to create a new chromosome comparison, we first have to import whole genome sequences from e.g. FASTA files, text files or online from EBI or NCBI using the BIONUMERICS import routines present.

2. In the *Main* window, select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.

3. Choose ***Download sequences from internet*** under ***Sequence type data*** and click <***Import***>.

4. Enter the accession codes **AE005174**, **BA000007** and **U00096** in the ***Accession codes*** input field, separated by the separation character “,” (see Figure 1).

5. Specify “,” as the ***Separation character*** and choose ***EBI*** as download site.

6. With the option ***Preview sequences*** checked, press <***Next***>.

The import routine fetches the sequences from the selected database and shows detailed information in the next step (see Figure 2).

7. Press <***Next***> and click <***Create new***> to create a new import template.

Each header tag (e.g. ID, AC, . . .) corresponds to a row in the grid panel (see Figure 3).

8. Select ”AC - ACCESSION” in the list and click <***Edit destination***> or double-click on ”AC - ACCESSION”. Select ***Key*** and press <***OK***>.

The grid is updated (see Figure 3).

9. Click <***Next***> and press <***Finish***>.

10. Specify a template name (e.g. **EBI**) and optionally enter a description.

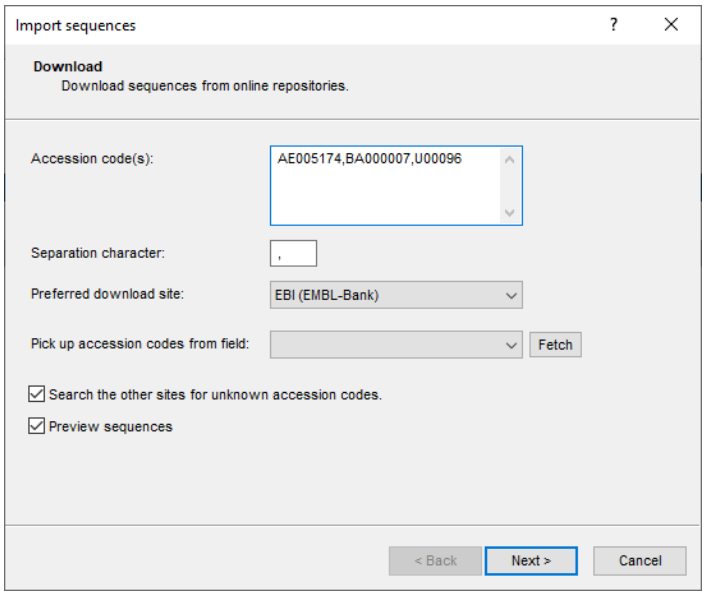

Figure 1: Download sequences from EBI.

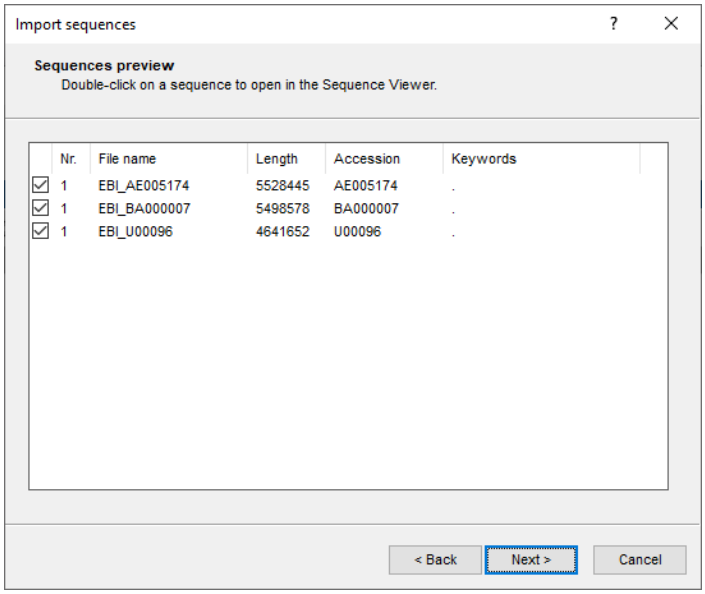

Figure 2: Fetched information.

11. Highlight the newly created template and select ***Create new*** as ***Experiment type*** (see Figure 4).

12. Press <***Next***>.

13. Specify a sequence type name (e.g. **genseq**) and press <***OK***> and confirm the action.

14. Press <***Finish***>.

The three sequences are imported in the database and are automatically selected (Figure 5).

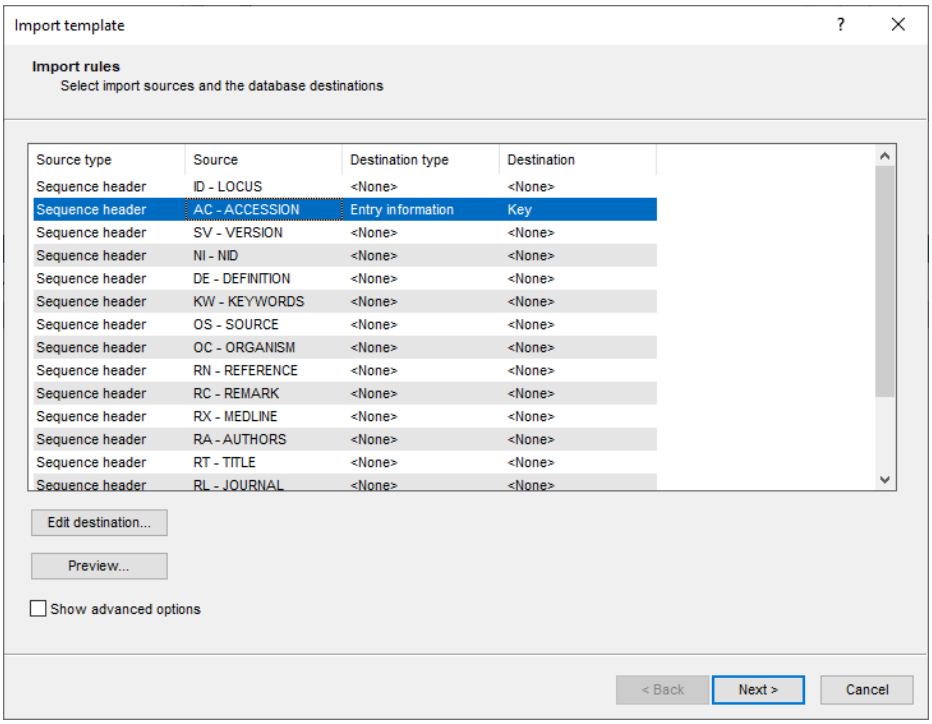

Figure 3: Create a new import template.

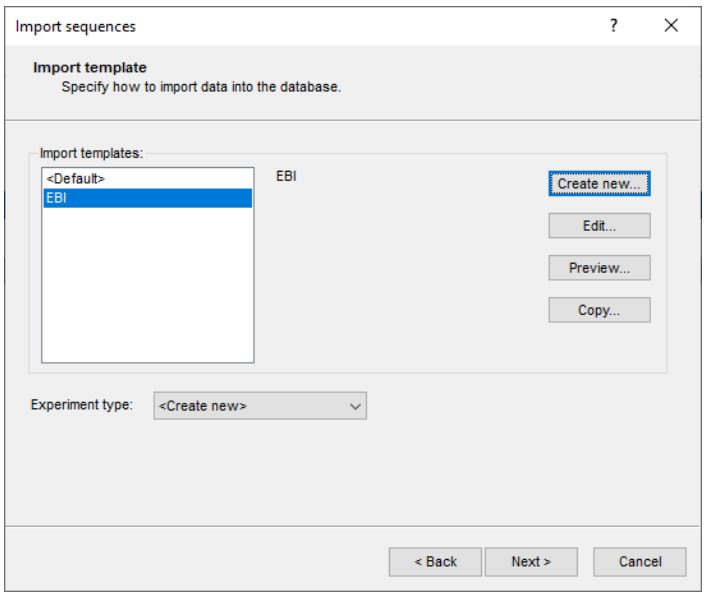

Figure 4: Select import template.

3 Creating a new chromosome comparison project

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

1. Select the three entries in the database using the check boxes or the **Ctrl-** key.

2. To create a new chromosome comparison project, select the *Chromosome comparisons tab* in the *Main* window and select ***Edit*** > ***Create new object...*** (+).

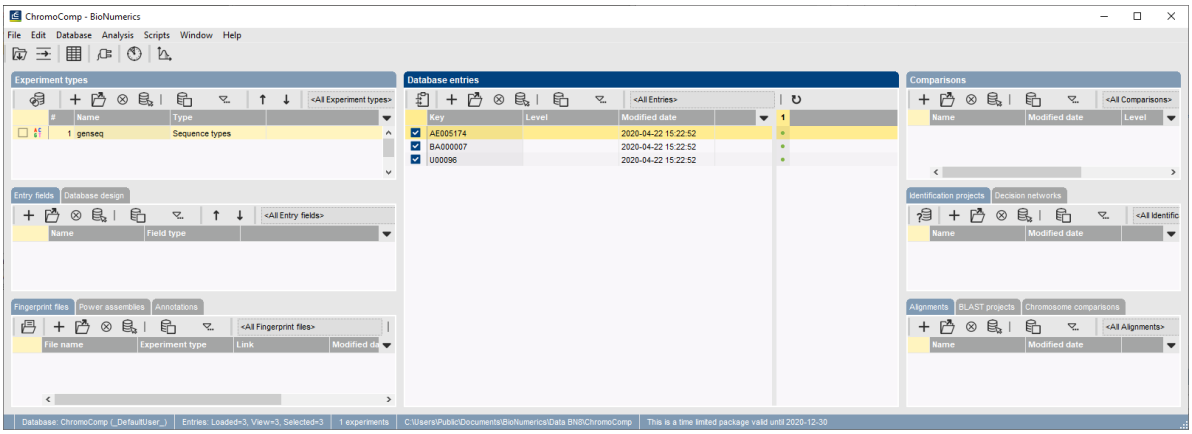

Figure 5: The *Main* window after import of the sequences.

A name for the new project is prompted for.

3. Specify a name for example **DNA chromosome clustering** and press <***OK***>.

The new project is added to the *Chromosome comparisons* 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 project.

4. Leave the **genseq** type selected in the list and press <***OK***> to open the *Chromosome Comparison* window (see Figure 6).

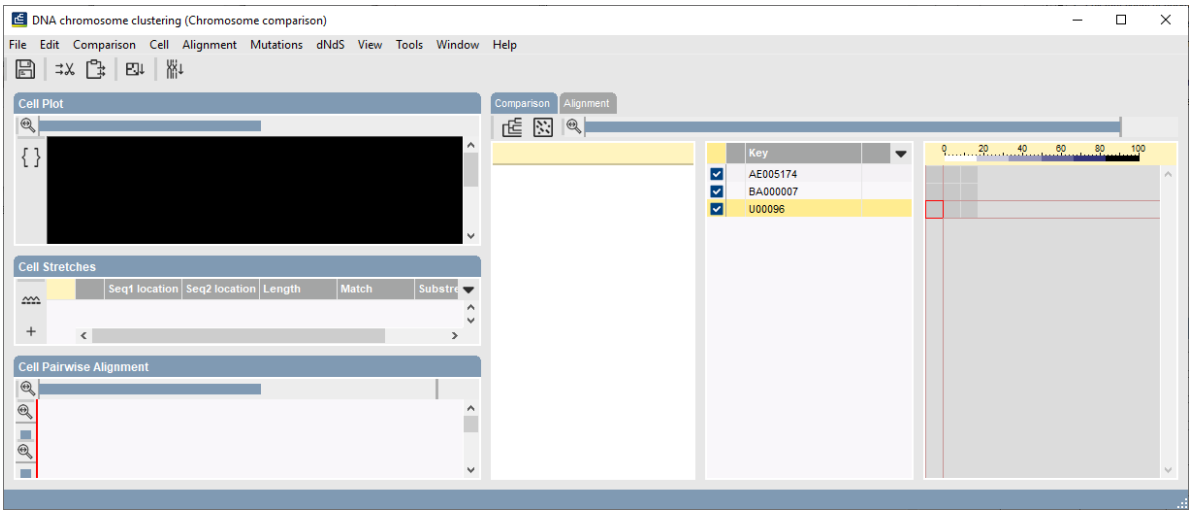

Figure 6: The *Chromosome Comparison* window.

The *Chromosome Comparison* window displays several view panels (see Figure 6): their function and arrangement is directed by the fact that a multiple chromosome comparison is an $n \times n$ matrix built up out of $n \times n$ pairwise comparisons. This means that the general overview of a chromosome comparison will be an $n \times n$ matrix (*Comparison* panel), with more detailed views being a selected single cell out of the matrix (a pairwise comparison) or a multiple alignment of $n-1$ sequences against a template (*Alignment* panel).

Related to the single cell view are three panels displaying pairwise comparison information: a dot plot view of the two selected sequences (*Cell Plot* panel), a listing of all stretches of homology

found between the two sequences (*Cell Stretches* panel), and a graphical synteny representation of the two sequences for selected regions of homology (*Cell Pairwise Alignment* panel).

The color of each specific cell in the *Comparison* panel indicates its calculation stage. A cell is gray if no calculation has been performed for this cell. If a cell is fully calculated, it will be depicted in green. Initially all cells have a gray color.

4 Pairwise comparing chromosomes

When creating a new chromosome comparison (either a DNA based chromosome comparison or a CDS based chromosome comparison), the user will have to define one or more seeds, which will direct the screening method. The seeds will specify the sensitivity of the screening (shorter seeds are more sensitive) and the type of screening (DNA based or amino acid based).

1. Choose ***Comparison*** > ***Calculate matrix...*** ( ).

2. Click on the *Seeds tab* in the *Project settings* dialog box.

New projects are initially defined with the standard seed ”11111” for amino acid-based screening.

3. Click on the *Comparison tab* in the *Project settings* dialog box where the settings can be specified to direct the screening.

4. For the current chromosome comparison, select a ***Full sequence based*** project, applying ***Matrix*** comparison (for both **Direct** and **Inverted** sequence orientation).

5. Press <***OK***> to start the calculations.

The calculation progress is indicated on the gray rectangles in the matrix, turning green when calculated. When all cells in the matrix are processed, each cell displays an identity score, with a corresponding color scale. The scale goes from black, corresponding with 100% identity, over blue towards white (0% identity) (see Figure 7).

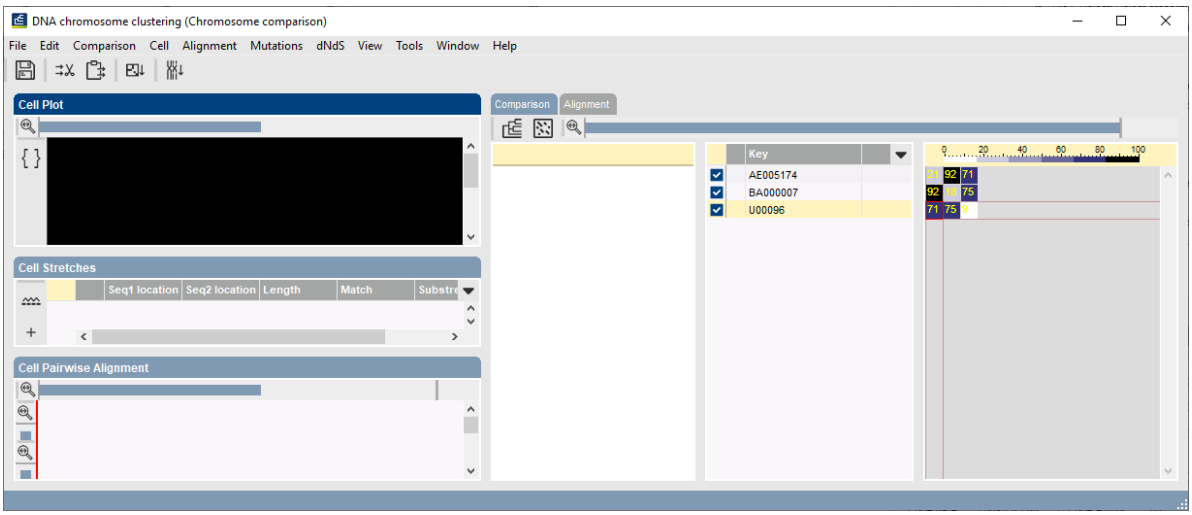

Figure 7: Matrix with processed cells.

6. The matrix of dot plots can be called with (see Figure 8).

This view shows each cell from the matrix as a reduced dot plot, where the first sequence is plotted along the X-axis against the second sequence along the Y-axis. If a dot plot is selected in

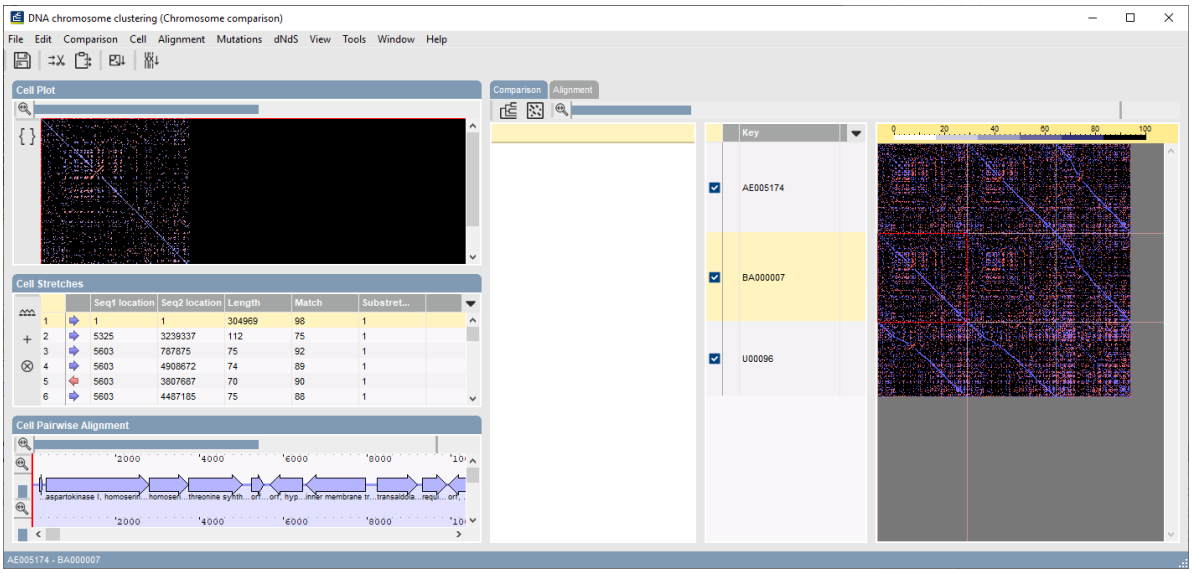

Figure 8: Matrix of dot plots.

the similarity matrix, the detailed matrix of dot plots is simultaneously displayed in the cell viewing panels.

Within all dot plots, blue dots represent stretches of homology between both sequences in the direct orientation, whereas red dots represent stretches of homology between the first sequence (normal direction) and the second sequence in inverse orientation.

7. A clustering of the entries, based upon the pairwise identity scores, can be calculated by selecting ***Comparison*** > ***Cluster matrix*** ( ) (see Figure 8).

8. To switch back to the identity score matrix representation, select again.

If a cell of the similarity matrix is selected either in the identity score matrix or in the dot plot matrix, the corresponding pairwise sequence alignment information is displayed in the cell view panels (see Figure 8).

A spot within the dot plot panel can be selected by clicking with the mouse pointer on the spot. The selector will automatically jump to the nearest spot.

Dots on the dot plot lying adjacent to each other are an indication of longer regions of homology between both sequences. They indicate regions of discontinuous parallelism, i.e. stretches of continuous homology interrupted by gaps in one of the two sequences. If such regions of discontinuous parallelism occur, one can try to link the individual stretches into one block, i.e into one stretch of discontinuous homology. Such blocks of discontinuous parallelism created by joining stretches of continuous homology are denoted as *super stretches*.

9. Selecting ***Cell*** > ***Cell Stretches*** > ***Switch stretch/superstretches listing*** ( ) maps the super stretches.

10. Selecting ***Cell*** > ***Cell Stretches*** > ***Switch stretch/superstretches listing*** ( ) again removes the super stretches from the plot.

Within the *Cell Stretches* panel, stretches mapping on the second sequence in direct orientation are indicated with a blue arrow, those mapping on the second sequence in inverted orientation are marked with a red arrow. The *Cell Pairwise Alignment* panel inverts the second sequence if a stretch is selected which maps on the second sequence in inverted orientation.

5 Calculating a chromosome alignment

The detailed information about pairwise homologies between sequences obtained within a multiple chromosome comparison can be used for calculating alignments between these sequences. To deal with the problem of rearrangements and to visualize the syntenies and parallelisms of the genomes in a comprehensible way, pointing out a sequence as guiding reference and aligning the other sequences against this reference is the most obvious approach.

1. Click on the *Alignment tab*.

2. To start a new alignment, select ***Alignment*** > ***Create new alignment...*** ( ).

3. Choose **U00096** as *template entry* (= guiding reference) and accept the other default settings and press <***OK***> to start the calculations.

The *Alignment overview panel* shows a listing of the aligned query sequences, plotted as horizontal graphs. On top is a scaling of the template sequence plotted together with a gray graph which represents the number of aligned sequences (consensus plot) along the Y-axis in function of the alignment position on the template sequence (X-axis). In the listing, graphs plot the identity score (Y-axis) of the aligned blocks of the query sequence in function of alignment position on the template sequence (X-axis). The red dots (standard color definition) indicate blocks of the aligned sequence which are inverted in the alignment, blue dots represent blocks of the aligned sequence in their original orientation in the alignment.

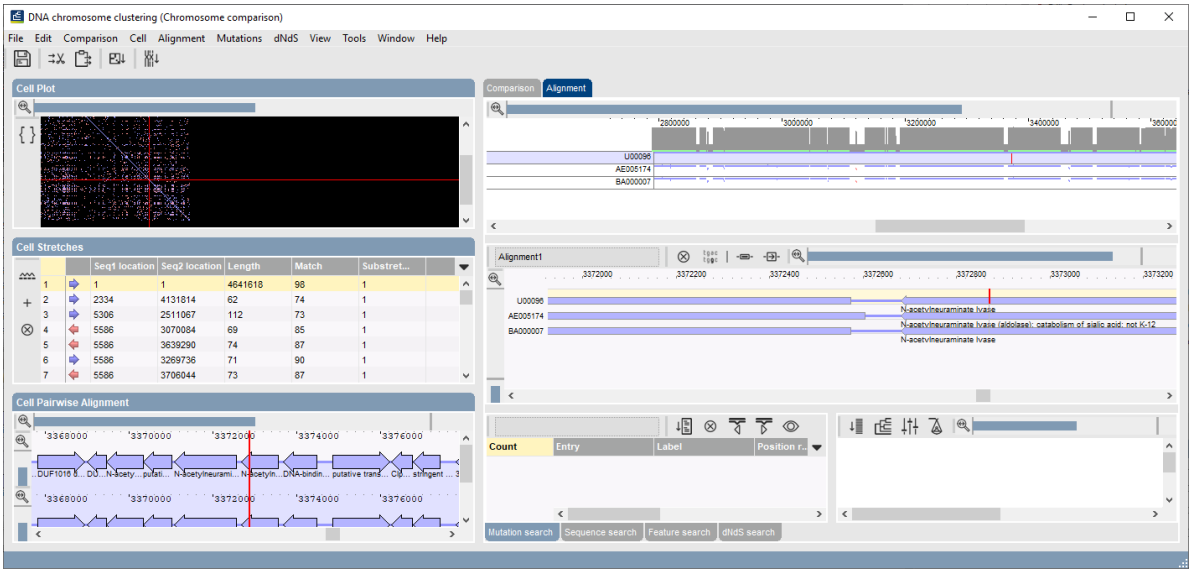

Figure 9: Alignment overview and detail.

4. Use the zoom slider at the top of the panel to zoom in on the overview plot.

Both in the *Alignment overview* and *Alignment detail panel*, a change in cursor position and a selection of sequences can be made with the mouse pointer. The selection and cursor position is updated in both panels.

In the *Alignment detail panel*, the cursor indication has a double function: the gray line indicates the alignment position whereas the red line indicates the current pairwise sequence selection (template plus one of the query sequences).

5. Change the alignment detail view from feature view (graphical) to text view (nucleic and amino acid information), with ***Alignment*** > ***Show text view*** ( ).

6. To switch back from text view to feature view, click the button again.

6 Mutation analysis

The mutation search function in the *Chromosome Comparison* window allows the detection of the following mutation types:

- A **silent mutation** is a nucleotide change which does not lead to an amino acid change. A silent mutation localized within a non coding sequence is called an **intergenic mutation**.
- A **missense mutation** or **non-synonymous mutation** is a translation change which leads to an amino acid change within the coding sequence located at this position.
- An **indel mutation**, which is either a deletion or insertion of a subsequence (or base).

1. Select ***Mutations*** > ***Search...*** ( ) to start a mutation search.

2. Keep the defaults selected and press the <***Find***> button.

The search function progress status is shown and the results for the mutation search function are listed in the *Mutation analysis* panel (see Figure 10).

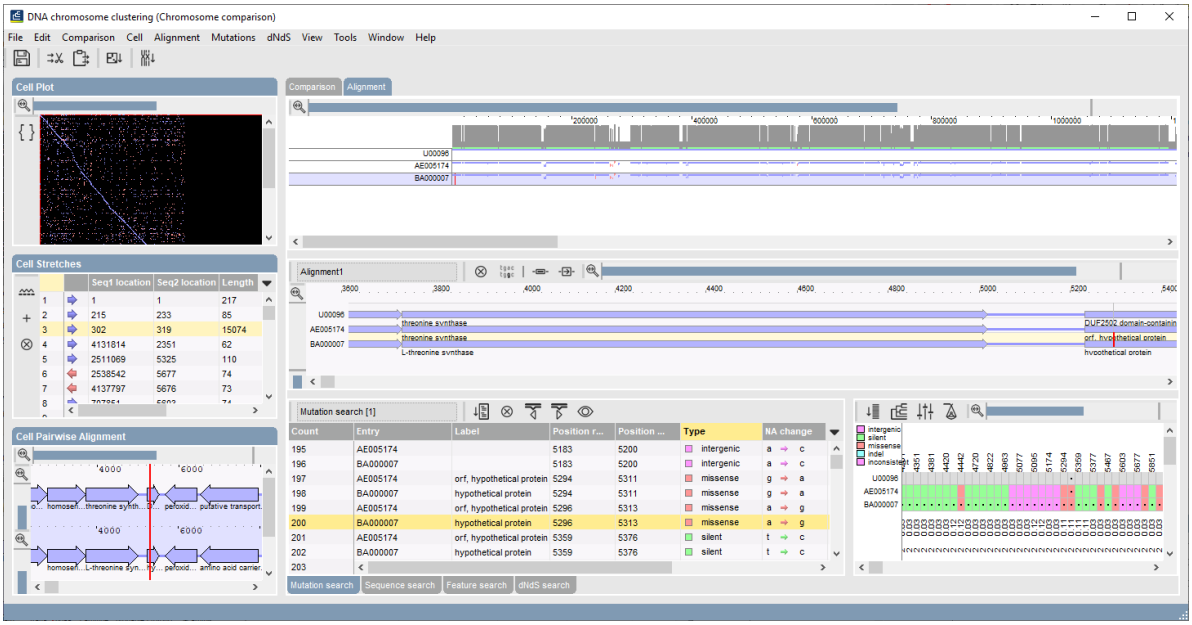

Figure 10: The *Chromosome Comparison* window after performing a mutation search.

At the left side of the *Mutation analysis* panel, the mutations that were found are listed with following information:

- ”Entry” indicates the entry key of the sequence on which the mutation is located.
- In the column ”Label” the preferential qualifier used for feature labeling, as defined in the alignment display settings is displayed.

- ”Position reference” gives the alignment position of the mutation on the reference (= template) sequence.
- ”Position sequence” gives the alignment position of the mutation on the sequence.
- ”Type” states the type of mutation: intergenic (silent) mutation, synonymous (silent) mutation, non-synonymous (missense) mutation or indels (insertions and deletions).
- ”NA change” gives the nucleotide base change: the first letter is the nucleotide base on the template (reference) sequence, the second is the base on the query sequence.
- ”AA change” gives the change at translation level as results from the nucleotide change. The first letter is the amino acid located within the translation product on the template sequence, the second letter is the amino acid on the query sequence.
- The quality score displayed in the field ”Quality” is a measure of confidence.

The positions of the mutations in the *Alignment detail panel* are color-labeled according to the mutation type.

3. If the mutations are not visualized in the *Detailed alignment panel* choose ***Mutations*** > ***View mutations on alignment*** ( ).

4. Choose ***Mutations*** > ***Jump to previous*** ( ) to jump to the next mutation upstream from the current cursor position.

5. Select ***Mutations*** > ***Jump to next*** ( ) to look for the next mutation downstream from the current cursor position.

The right panel of the *Mutation analysis* panel depicts a tabular view of alignment positions that resolve to mutations in one of the query sequences. The rows within the matrix correspond to the query sequences, whereas the columns are the mutation-affected alignment positions.