Find Substrings at Sequences

This workflow describes how to find substrings in sequences and group these sequences by different parameters.

First, the workflow reads sequences and text strings (patterns) from files. Then, these data sets are multiplexed using this rule: every sequence is united with every pattern. After multiplexing these united data sets are transported to the find patterns element. The results of patterns searching are grouped by id of a sequence: original and find patterns annotations are united into two new grouped annotations sets. And finally, the grouped data are written into file, specified by a user.

By default, sequence multiplexed using the rule "1 to 1". You can configure this value in the *Multiplexer* element parameters. Also, you can configure the *Pattern* element parameters and *Grouper* element parameters.
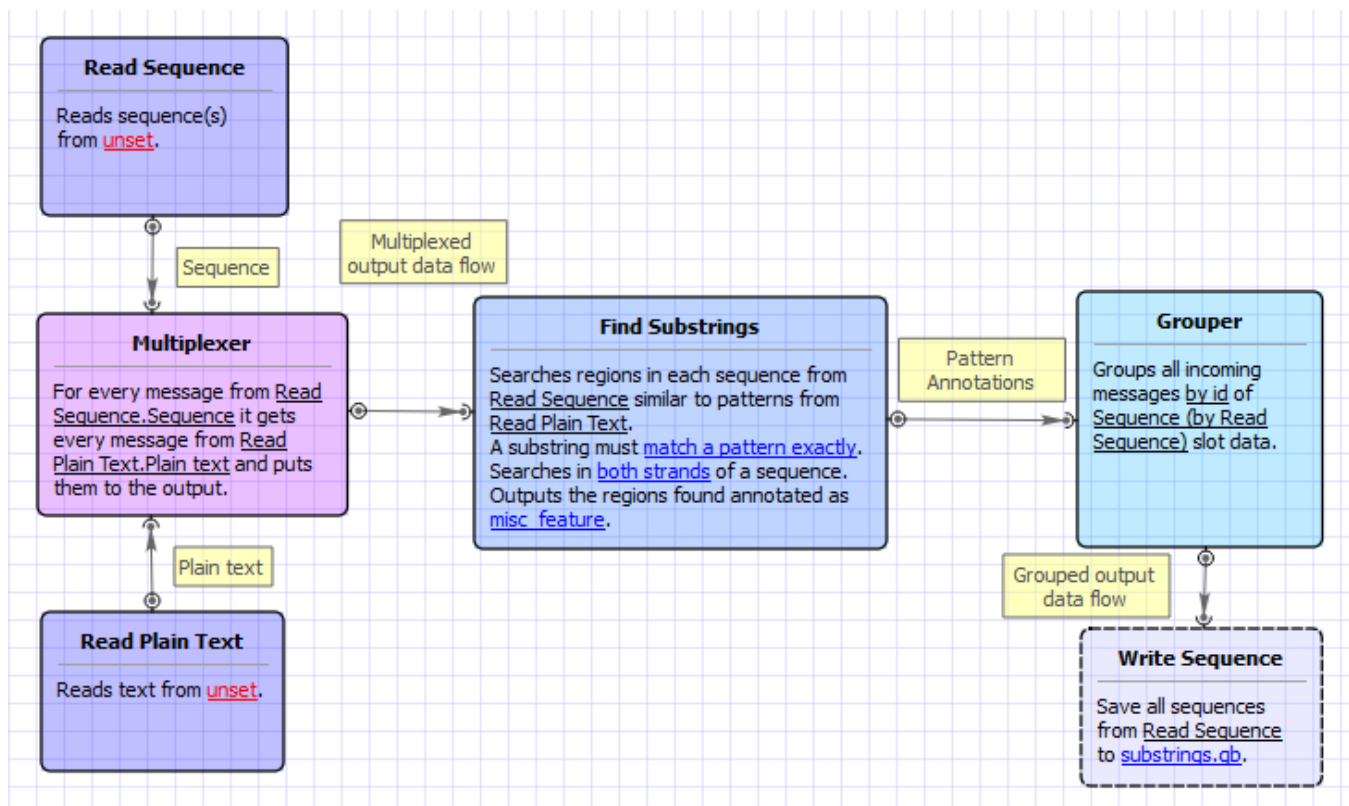
> ✓ **How to Use This Sample**
>
> If you haven't used the workflow samples in UGENE before, look at the "How to Use Sample Workflows" section of the documentation.

Workflow Sample Location

The workflow sample "Find Substrings at Sequences" can be found in the "Data Merging" section of the Workflow Designer samples.
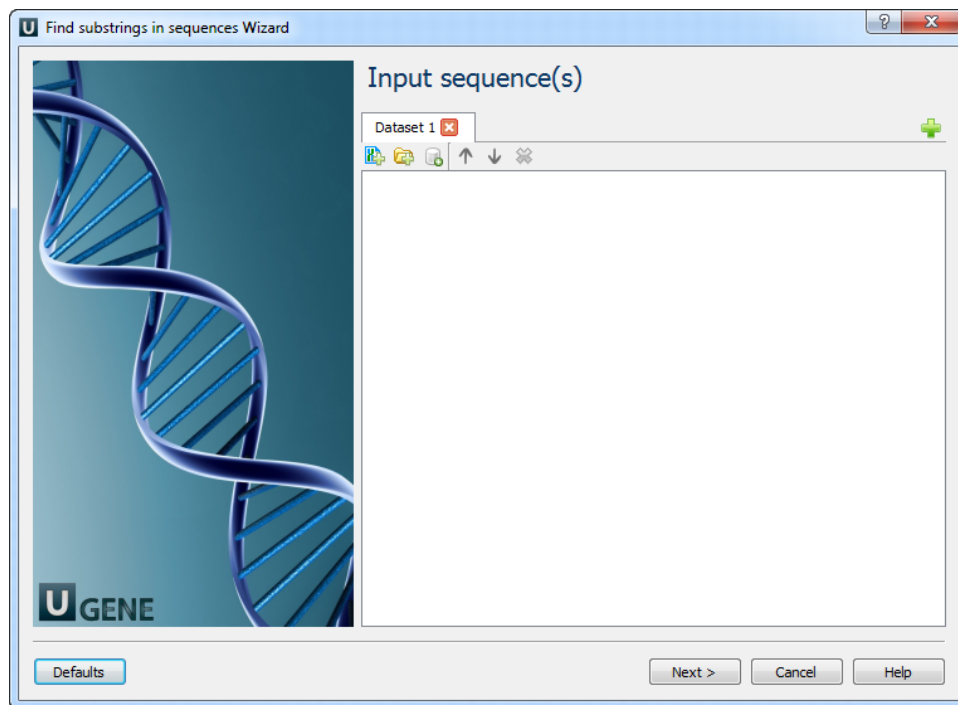
Workflow Image
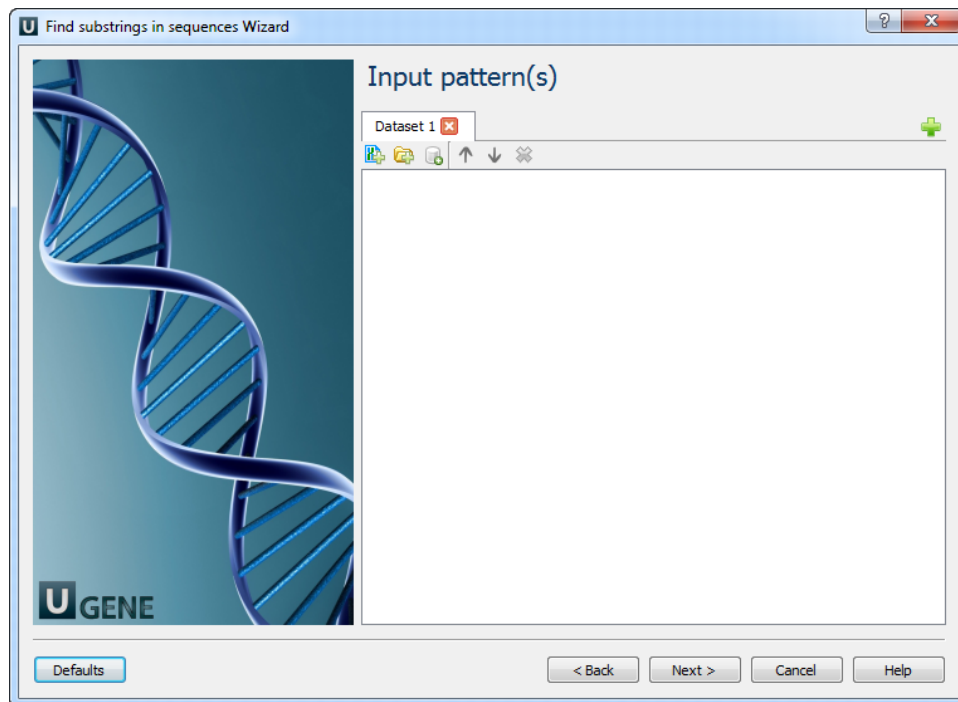
The workflow looks as follows:

Workflow Wizard
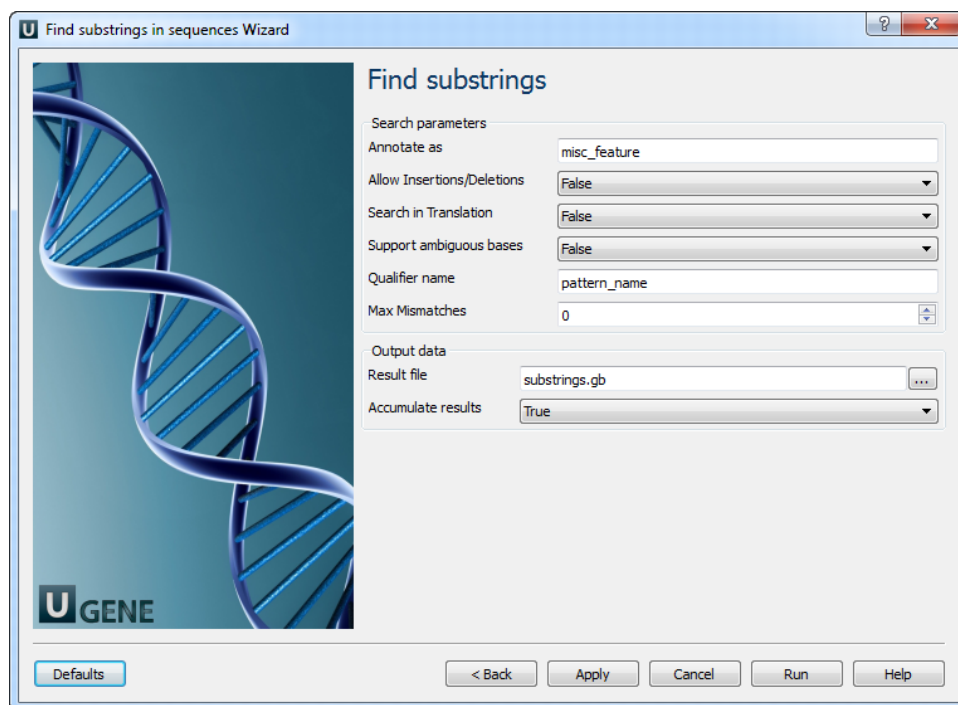
The wizard has 3 pages.

1. <u>Input sequence(s):</u> On this page you must input sequence(s).

2. Input pattern(s): On this page you must input pattern(s).

3. Find substrings: On this page you can modify search and output parameters.

The following parameters are available:

Annotate as	Name of the result annotations.
Allow Insertions /Deletions	Takes into account possibility of insertions/deletions when searching. By default substitutions are only considered.
Search in Translation	Translates a supplied nucleotide sequence to protein and searches in the translated sequence.
Support ambiguous bases	Performs correct handling of ambiguous bases. When this option is activated insertions and deletions are not considered.
Qualifier name	Name of qualifier in result annotations which is containing a pattern name.
Max Mismatches	Maximum number of mismatches between a substring and a pattern.
Result file	Location of output data file. If this attribute is set, slot "Location" in port will not be used.
Accumulate results	Accumulate all incoming data in one file or create separate files for each input.In the latter case, an incremental numerical suffix is added to the file name.