

DNA Sequence Search

User manual

A fast and easy to handle MAC OS application to search and detect user defined nucleotide sequences in DNA strings.

Typical users of this app are scientists and researcher at universities, pharmaceutical institutes, and research-oriented companies.

Applications

- General purpose nucleotide pattern search in DNA files.
- Sequence finder for detection of sequences e.g. ORF - open reading frame, initiation codon, stop codon, ...
- Detect recognition sequences of enzymes such as methylase and restriction enzymes.
- Detect sequences motives that are bipartite such as promoter sequences.
- Codon usage (search for usage of codons that code for the same amino acid).

Contents

| | |
|--------------------------------|---|
| Open DNA File | 2 |
| Define a search pattern | 3 |
| Calculate and execute a search | 4 |
| Save search results to disk | 5 |
| Configuration | 6 |
| Usage conditions | 7 |

Open DNA File

Pushing the **Open** button, displays a file selection menu to select a DNA file, stored on the disk.

Fasta file format

While loading the DNA file, the app checks if it complies with FASTA file format. In case of incompatibility the file load is stopped and an error message informs the user.

Fasta file example

```
>MZ501053.1 Escherichia phage BrunoManser, complete genome
ATGTTAAACAGCGAGGACGAATCTATGAAACAGGAAAAAGCGCCCGTCGTTTCAGGGTGGCAACTTTAAAG
AGCTATAACCGGAAAGAGTTCGGAACCGTGTTAGGCAAAACGCGACAGACCACGCCGAAGCAGTTGTTTGA
TCTGGCGGTTAAGTATTTTCGAATGGGCGGAAGACAACGCAATCAAAGCGGCTGAGACTGCCAGTTTTTCAG
GGCGATGTAGAAGAGTCGTTAGTGCATAAGCCGCGAGTGTTTACCGTTACCGGGTTTAAGCTTTTCTGCA
GCCTGAGCGACGGCACTATCGCGCGATACCGATCCGAACCGGACTATGCCGAAGTTATGGAATTCGTCTGA
TTCCGTTATCAATGAACAGAAATTCCAGCT...
```

File information display

Once the file is successfully loaded, the header information and file length are displayed in the DNA File line on the screen.

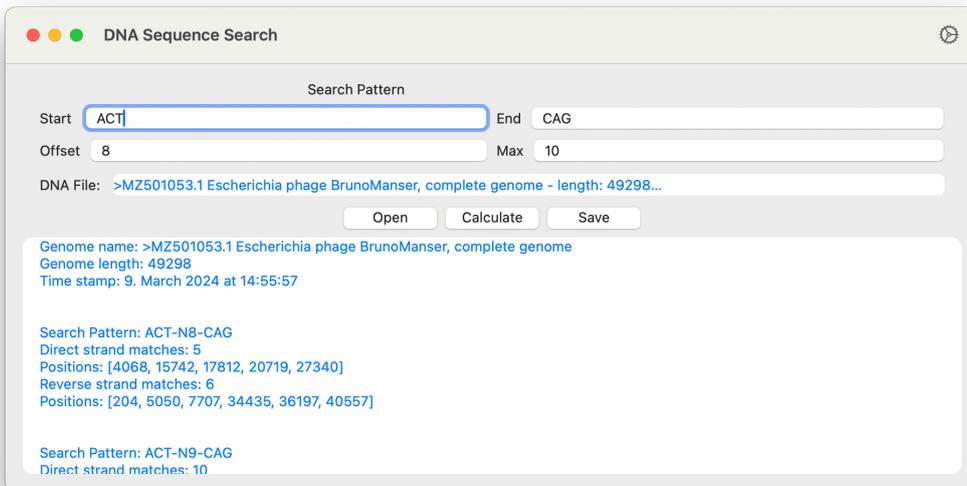
Header line example

```
>MZ501053.1 Escherichia phage BrunoManser, complete genome - length 49298
```

DNA file source

- National Institutes of Health, <https://www.ncbi.nlm.nih.gov/tools/sviewer/seqtrackdata/>

Define a search pattern



Four input fields are available to define the search pattern:

- **Start - End**
2 alphanumeric fields to enter an unlimited sequence of nucleotides (Adenin, Cytossin, Guanin, Thyamin; A-C-G-T) in upper case characters.
- **Offset - Max**
2 numeric fields to define a nucleotide distance between start and end pattern. Offset defines the minimum and Max the maximum distance.
Defining Offset and Max in one search, allows the user to calculate a range of patterns, p.e. Start - Offset - End until Start - Max - End.

Valid search examples

| Start | End | Offset | Max | Results in |
|-------|------|--------|-----|--|
| CGT | | | | Searches defined pattern and displays the results in the scroll view area. |
| GTAC | TAG | | | Searches both patterns and displays the results in the scroll view area. |
| CGTA | TAGC | 2 | | Searches CGTA-N2-TAG and displays the results in the scroll view area. |
| GC | ACG | 2 | 4 | Searches CG-N2-ACG, CG-N3-ACG, CG-N4-ACG and displays the results in the scroll view area. |

Invalid search examples

| Start | End | Offset | Max | Results in |
|-------|-----|--------|-----|--|
| | CGA | | | Invalid search definition: No end pattern without start pattern allowed. |
| CGT | ACC | | 2 | Invalid search definition: No max value without defined offset. |
| AC | CG | 4 | 2 | Invalid search definition: Max must be greater then offset |
| cat | | | | Will not show any result as nucleotides in FASTA are upper cased only. |

Calculate and execute a search

By pushing the **Calculate** button, *DNA Sequence Search* starts various checks of the input parameters and informs the user if any invalidities were detected. If the input is valid the search is started.

Method

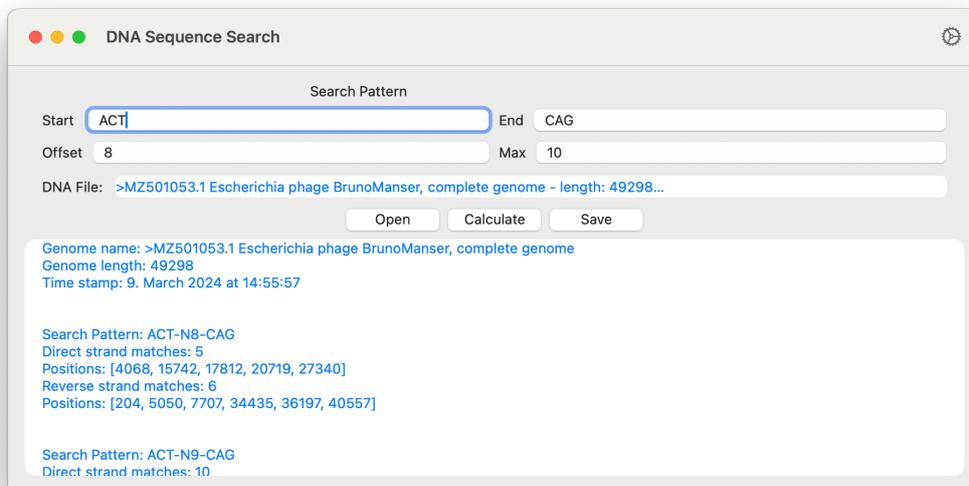
The loaded FASTA formatted DNA file consists only of the DNA's direct strand information. **DNA Sequence Search** completes it with the reverse strand data automatically. The app searches the user-defined string in direct and reverse direction. See example below.

Example pattern: GAAT

| Direction | DNA string |
|-----------|----------------------------------|
| direct → | ...GGTCAG GAAT TCGCTGA... |
| reverse ← | ...CCAGTCT TAAG CGACT... |

Results

Results of all searches are displayed in the scroll view area on the bottom of the app screen. The user can easily scroll through the results of the different searches. Each result block is headed by the search session header. CAUTION: Pushing the Save button stores the latest calculation only!



- **Search session header** - consists of
 - Genome name
 - Genome length
 - Time stamp

- **Positions and number of matches**

DNA Sequence Search finds the user-defined patterns and counts the number of matches. The search is done in direct and reverse strand and the results are shown in blocks for both strands separately. The position of a match is always the first position of the pattern in reading direction. Hence, in direct strand from left to right, in reverse strand from right to left. The results are headed by the search pattern, number of matches in direct / reverse strand and the positions of the matches.

Save search results to disk

By pushing the **Save** button, a file selection menu is opened which allows the user to store the result of the latest calculation to a file on the disk.

The results can be exported as:

- **Details** - consisting of
 - Search session header
 - Search pattern
 - Number of matches in direct strand
 - Positions in direct strand
 - Number of matches in reverse strand
 - Positions in reverse strand
- **Summary** - consisting of
 - Search session header
 - Search pattern
 - Number of matches in direct strand
 - Number of matches in reverse strand

The standard file extension is .txt even if the user selected CSV output format. In this case, the user has to enter the file extension .csv manually.

Configuration

In the configuration screen the user can choose the following options:

Format - export file

- **Text**
The results, either *details* or *summary* are exported as text file. The file extension is .txt.
- **CSV**
The results are exported as CSV table. The search session header occupies the first columns of the top three lines. The content of the table depends on the selection (*details* or *summary*).
 - Details - 3 columns:
search pattern - match position direct strand - match position reverse strand
 - Summary - 3 columns:
search pattern - # of matches direct strand - # of matches reverse strand

When saving the file to disk, the user must take care to change the file extension to .csv, otherwise it uses the standard .txt extension.

Content

Allows the user to decide between a detailed listing or a result summary.

- **Details** - consisting of
 - Search session header
 - Search pattern
 - Number of matches in direct strand
 - Positions in direct strand
 - Number of matches in reverse strand
 - Positions in reverse strand
- **Summary** - consisting of
 - Search session header
 - Search pattern
 - Number of matches in direct strand
 - Number of matches in reverse strand

Usage conditions

The following list shows conditions that have to be considered. Some of them will result in warnings or error messages.

Conditions

- A FASTA file must be loaded.
- A start sequence must be defined.
- Offset can't be greater than the file length.
- Range calculations (start - offset - max - end) are limited to chunks of a maximum of 1000. Greater search requests must be divided into multiple sessions.
- An execution time warning will be displayed if the expected calculation time exceeds 10 seconds. The user can ignore the warning or divide the search request in several sessions.