
DNA Sequence Search application guide

What is the app used for?

The purpose of this app is to identify the position of a user defined nucleotide sequences within a DNA genome.

What is a DNA genome?

Deoxyribonucleic acid, usually referred to briefly as DNA, is a nucleic acid composed of different deoxyribonucleotides. It carries the genetic information in all living beings and the DNA viruses. The long-chain polynucleotide contains special sequences of its nucleotides in sections of genes. The basic building blocks of DNA strands are four different nucleotides, each consisting of a phosphate residue, the sugar deoxyribose and one of four nucleic bases (adenine, thymine, guanine and cytosine; often abbreviated as **A, T, G and C**). The sequence of bases (nucleotide sequence) in certain DNA strand sections contains information.

Where to get DNA information?

There is a common accessible database available, which researcher all over the world use provided by the National Institutes of Health.

Link: <https://www.ncbi.nlm.nih.gov/tools/sviewer/seqtrackdata/>

You can also download a phage DNA file *TestPhage.fasta.zip* in the download page for app test purposes.

Data format of DNA data?

The common data format for DNA data is called FASTA, which is basically a text format with some internal organizational rules. The DNA Sequence Search app checks imported files to comply with this format and reject's if not.

Who are the target customers of this app?

The main target group for this app are people involved or interested in research and education.

What are potential questions a user wants to got answered?

Example 1:

How many restriction sites of the enzyme (X) can be found in the phage DNA (Y)?

(X) can be any restriction enzyme or sequence e.g. EcoKI which has the recognition sequence AAC[N6]GTGC.

(Y) can be any DNA string e.g. the phage *Test Phage* (find Fasta file in the download section)

The novelty is that with other DNA search tools, one cannot define N and only search for single sequences, not bipartite ones.

N ... representing any nucleotide ACG or T

Phage... virus infecting bacteria

Enzyme.. proteins that act as biological catalysts

How to configure the app to calculate this request?

1. Load the DNA file by clicking the *Open* button and select the appropriate file (*Test Phage.fasta*) on the disk. If the file format is correct the app shows the fasta file id line as well as the length of the DNA string in the *DNA File* field.
2. Enter the start sequence **AAC** in the *Start* field.
3. Enter the end sequence **GTGC** in the *End* field.
4. Enter **6** in the *Offset* field.
5. Click *Calculate*.

DNA Sequence Search displays a header for this calculation in the *data window* which consists of:

Genome name: >MZ123456.1 *Test phage, complete genome* (= fasta file id)
Genome length: 49298 (= DNA length)
Time stamp: date and time (when it was calculated)

Following the header it displays the calculation results in the *data window*.

Search Pattern: AAC-N6-GTGC
Direct strand matches: 2
Positions: [8990, 16338]
Reverse strand matches: 0
Positions: []

Example 2:

How many restriction sites of the enzyme (X) can be found in the phage DNA (Y)?

In this case (X) should be a sequence AAC[N6]-[N9]GTGC.

This means, the app should find a series of sequence starting AAC[N6]GTGC until AAC[N9]GTGC.

How to configure the app to calculate this request?

Given, we loaded and calculated the example above:

1. Override the *Offset* field with **7**.
2. Enter **9** in the *Max* field.
3. Click *Calculate*.

DNA Sequence Search displays again the header for this calculation in the *data window* which consists of:

Genome name: >MZ123456.1 *Test phage, complete genome* (= fasta file id)
Genome length: 49298 (= DNA length)
Time stamp: date and time (when it was calculated)

Following the header it displays the calculation results in the *data window*.

Search Pattern: AAC-N7-GTGC
Direct strand matches: 2
Positions: [44252, 45131]
Reverse strand matches: 4
Positions: [22783, 29714, 42506, 47455]

Search Pattern: AAC-N8-GTGC
Direct strand matches: 0
Positions: []
Reverse strand matches: 1
Positions: [27670]

Search Pattern: AAC-N9-GTGC
Direct strand matches: 1
Positions: [15108]
Reverse strand matches: 3
Positions: [12423, 41962, 48104]

In this case the speciality is that the user can easily define distances given by the numbers in the *Offset* and the *Max* field to calculate repeated sequences with increasing [Nx] instead of starting the calculation manually for each N.

Example 3:

Single instances of nucleotide combinations can of course be calculated simple by entering the sequence in the *Start* field e.g. **ACG**. Before you start this calculation make sure that the distance fields (*Offset*, *Max*) contain the number 0.

The result will consist again of the header and the detail section. In the mentioned example (ACG) the direct strand matches are 1144 and the reverse strand matches are 884.

Example 4:

Find two nucleotide combinations in one calculation cycle p.e. **ACT**, **GCT**.
In this case enter **ACT** in the *Start* field and **GCT** in the *End* field and click *Calculate*.

The result will consist again of the header and the detail section.

The result for **ACT** is 554 in direct strand and 512 in reverse strand, for **GCT** it finds 703 matches in direct strand and 785 in reverse strand. Of course it displays the positions of each match in the *data window*.

What is the speciality of the data window?

DNA Sequence Search displays the calculation results as described above in the *data window*. The results of all calculations are available until the program will be closed. By scrolling up and down the user can see all results of the complete session.

How can I export the results?

Clicking the *Save* button opens a file dialogue that allows to store a result file to your disk. The representation of the results can be defined by the options in the *Configuration* page. Standard configuration for *Format* is *Text* and for *Content* is *Details*.

What can be configured in the app?

The *Configuration* page offers 2 entries.

Format	Text / CSV; defines the format the results are represented in the <i>output file</i> .
Content	Details / Summary defines how the results are presented in the <i>data window</i> , as well as stored in the <i>output file</i> . <i>Details</i> Calculation header, Search pattern, number of matches, positions <i>Summary</i> Calculation header, Search pattern, number of matches

For more details see **DNA Sequence Search User Manual**, at https://www.piccosnext.com/?page_id=28