PERCON User Manual v2.5 May 19, 2015

PERCON DISTANCE VERS.4.35.1s R2

# Introduction

PERCON stands for Probabilistic Evaluation of Relation by Counting of OctaNucleotides.

This program is used to search, extract and classify alpha satellite (AS) monomers in nucleotide sequences. AS is composed of monomers of about 171 bp tandemly repeated in centromeric and pericentromeric regions. These monomers are classed by PERCON into 12 standard classes belonging to 5 suprachromosomal families (SFs). Every monomer is also independently classed by the A/B box in position 35-51 of 171 bp AS consensus. The box has two configurations and can be of A- or B-type. Some subclasses of AS monomers within class M1 can have an extra nucleotide in position 21 of the monomer.

# Authors

PERCON program was developed by V.A. Shepelev (spl@img.ras.ru). Alpha-satellite block was developed in collaboration with I.A. Alexandrov (ivanalx@hotmail.com).

# Software requirements

The program is 16-bit MS-DOS application and requires NTVDM subsystem (NT Virtual DOS Machine) to run under Windows NT, 2000, XP, or Windows 7. NTVDM is not supported in 64-bit versions of Windows and the only way to run program is to use virtualization software.

# Input

PERCON supports input sequences in GenBank format with gaps expanded. Files obtained from repositories should be translated to ASCII CR+LF (DOS/Windows) newlines.

# Limitations

1. This program was designed to aid research of a narrow circle of AS experts and was not intended for the broad audience. It is mainly the toolbox to assist in AS research and has a lot of parameters and modes, which are not always compatible.
2. This is a limited simplified version capable of production of SF-annotated AS maps. Some menu items are intended for more advanced versions of the program, which have functions disabled in this version. Likewise some steps in execution of the program may not make sense in this limited version, but have to be executed nevertheless.
3. PERCON is not completely user-friendly and not totally foolproof program in its current form. Please, be careful! Not all parameters are applicable in the entire range as proposed. Please do not use extreme parameters.
4. Program does not support batch processing, sequences from different files should be selected manually. You can combine multiple sequences in a single file, but depending on the size, the processing may take a long time.

# General consideration of the methods

### Use of Terms

DB — is a plain text file in DOS format that contains entries (nucleotide sequences) in GenBank format.

Window:

1. the fragment of the sequence analyzed at a given step, long sequences are analyzed in steps as 5000 bp windows;
2. window (WI) parameter in dot-matrix options.

The search of the AS monomer in the window is organized in three steps:

1. Fast search of AS-similar sequences by comparison of the octanucleotide dictionaries of the window and AS consensus. Selected windows go to the next step.
2. Refined analysis of AS in the window by construction of dot-matrices and their examination. Selected windows go to the next step.
3. Repeated alignment of AS consensus to the window and the extraction of monomers.

First step is very fast (near file reading speed). Second step is a little bit slower. The time-limiting stage is repeated alignment.

Every monomer is classed into one of 13 monomer classes by simple naive Bayesian classifier or remains unclassed (Um).

12 standard classes are characteristic of the five SFs (J1, J2, D1, D2, W1, W2, W3, W4, W5, R1, R2 and M1) and Xm stands for random sequence. Independently every monomer is classed into A or B type or random (X) or unclassed (U). SF1 is formed by J1 and J2, SF2 is formed by D1 and D2, SF3 is formed by W1, W2, W3, W4 and W5, SF4 is formed by M1 and SF5 is formed by R1 and R2 classes. According to the presence of the monomer classes the window is classed into one or more SFs. Classes J1, D2, W4-5, R2 and M1 belong to the A type; classes J2, D1, W1-3 and R2 belong to the B type.

# Package contents

dist80s.exe — main application;

ASC — file that describes AS consensus;

CONS\_ALP.DNG — AS consensus file;

cmxx-txt.txt — auxiliary file;

\_SQfile.tmp — auxiliary file to temporary store the sequence;

1.dng, 2.dng, 3a.dng, 4.dng, hor11mer12.txt — blank spaces for test sequences, not used in this version;

AP001464.gb — example of sequence;

M001464 — output example.

# Installation

Extract the distribution archive to the directory you wish to install PERCON (for example: C:\PERCON).

Reminder: this should be directory separate from OS system files. All files included in the archive are needed to successfully run the program.

Once the program files are unzipped from their archive and into their own folder, the user starts the program by double-clicking on the dist80s.exe file.

# Quick configuration and usage

1. First, user should find and load file with AS consensus. Press **F8**. Current directory name appears. Enter directory name, or press **Enter**. Select file ASC. Press **Enter**. Press **Esc**. This step should be fulfilled only once in every session.
2. Press **F2** to load AS consensus.
3. Press **F4**. Press **Enter**. Press **Esc**.
4. Press **F5**. Press **Enter**. Choose between **F7** for long sequences (longer than 450 Kbp) and **F8** for short sequences. Press **Enter**. And press **Enter** once more\*. Find your input file in directory and select it. Press **Enter**. Press **Esc**.
5. Press **Esc**. See if the output goes.  
   If no output appears in 10-20s something is wrong. Press **Ctrl-Break** to break the program. Close the window or shell and repeat from the beginning.
6. After calculation stops you can see the result. Default output filename is MALUX3.

\*At this point the program asks: “starting seq name? (Please specify valid case) !!!!!!!!”. If the user wants to start his analysis from the first sequence in DB (file) he presses Enter. If the user wants to start his analysis not from the first sequence he should enter the name for the starting sequence instead of '!!!!!!!!', which represents the first 8 symbols of the sequence name.

# Output

Output is stored as plain text and can be viewable in any text editor such as Windows Notepad.

Output can be divided into several parts. First, the header which lists various parameters and options that the program uses during processing.

|  |
| --- |
| PERCON DISTANCE: VERS.4.35.1s R2 8-5-2015  INPUT FILE: C:\PERCON\ASC  OUTPUT DIR: malux3  DB : C:\PERCON\AP001464.TXT  left\_lim= 0.00 right\_lim= 0.60  WINDOW=5000 STEP=4800  th/wi= 31/ 63  LenX\_left\_lim= 140  dev\_SAVE\_left\_lim= 30  max\_str\_SAVE\_left\_lim= 30  dev\_MARK\_left\_lim= 50  max\_str\_MARK\_left\_lim= 50  Sequence positions: 1-450000  MAPPING param: min stretch len= 15 min seq number= 1 min region len= 50  ALIGNMENT parameters:  match value=9  open deletion value=30  deletion value=3  mismatch value=5  multiregion alignment  max\_contig\_stop= 20  non\_lost\_res\_stop= 20  align value stop= 205  align\_len\_stop= 40  rs\_cut\_off\_val=0.290  delay to prevent overhit 0  SQ load: to 450KB memory buffer  REVN to do  alignment: sn\_t |

Second is the main part which displays detailed information for each classed monomer.

First line under the title includes:

1. Name of the sequence (AP001464);
2. Position of window in sequence (1-5000);
3. C if AS is found in complementary strand;
4. Three values of PERCON distances. The second roX is best for AS consensus, others are shown for compatibility;
5. The max length of a diagonal on dot-matrix (168);
6. Standard deviation for the expected number of points on dot-matrix (1187.9),
7. Asterisk marks the fact that the good diagonal is present on dot-matrix;
8. M indicates that dot-matrix diagonal cover well the sequence in question.

Next line is the real map for A/B-boxes obtained by independent classification tool (i.e. independently of classification into monomer classes).

Next line is monomer class map. This map may have extra symbols:

tl — for M1 monomer with rs\_cut\_off\_value ≥ 0.6;

qo — for M1 monomer with rs\_cut\_off\_value < 0.6;

= — for long non-AS sequence;

~ — for short non-AS sequence.

Next line shows the expected A/B map as prescribed by monomer class. The real map may sometimes deviate from the expected one.

Next line shows suprachromosomal family (SF) classification determined according to monomeric classes in the sequence.

|  |
| --- |
| seq ID from to ro1 roX roY max\_str dev  AP001464 1- 5000 C 0.392 0.018 0.453 168 1187.9 \* M  q\_A\_A\_B\_A\_A\_B\_B\_A\_B\_A\_B\_A\_B\_B\_B\_A\_B\_A\_B\_A\_A\_B\_A\_B\_A\_A\_B\_B\_A\_q  umR2R2UmR2R2R1R1R2R1R2R1R2R1R1R1R2R1R2R1R2R1R1UmUmR2R2R1R1R2um  box by monomer:  U\_A\_A\_U\_A\_A\_B\_B\_A\_B\_A\_B\_A\_B\_B\_B\_A\_B\_A\_B\_A\_B\_B\_U\_U\_A\_A\_B\_B\_A\_U  SF classification: SF5 impured;  type from to len ident score rs 171/2  um q 1 19 19 C 78.95 115 0.67 u  R2 A 21 190 170 C 88.24 1250 0.82 171  R1 B 192 362 171 C 88.30 1259 0.82 171  R1 B 363 530 168 C 89.88 1246 0.82 171  R2 A 535 704 170 C 88.24 1250 0.82 171  R2 A 706 875 170 C 85.88 1194 0.78 171  Um B 877 1047 171 C 85.38 1189 0.77 171  Um A 1048 1194 147 C 89.80 1113 0.84 u  R1 B 1197 1367 171 C 84.80 1175 0.76 171  R1 A 1370 1529 160 C 84.43 1123 0.75 171  R2 A 1531 1701 171 C 88.89 1273 0.83 171  R1 B 1702 1868 167 C 88.30 1237 0.80 171  R2 A 1869 2037 169 C 88.82 1236 0.81 171  R1 B 2039 2196 158 C 80.70 1073 0.70 171  R2 A 2198 2367 170 C 87.65 1236 0.81 171  R1 B 2372 2532 161 C 84.24 1099 0.74 171  R1 B 2533 2703 171 C 86.55 1217 0.79 171  R1 B 2705 2874 170 C 90.00 1292 0.84 171  R2 A 2875 3045 171 C 88.30 1259 0.82 171  R1 B 3046 3215 170 C 88.24 1250 0.82 171  R2 A 3217 3388 172 C 86.63 1198 0.77 171  R1 B 3389 3559 171 C 87.13 1231 0.80 171  R2 A 3560 3730 171 C 84.80 1175 0.76 171  R1 B 3731 3899 169 C 89.35 1269 0.83 171  R1 B 3906 4072 167 C 88.62 1237 0.82 171  R2 A 4073 4240 168 C 90.00 1266 0.83 171  R2 A 4242 4409 168 C 91.76 1308 0.85 171  Um B 4411 4581 171 C 85.96 1203 0.78 171  R2 A 4582 4752 171 C 87.72 1245 0.81 171  R2 A 4754 4923 170 C 88.82 1264 0.83 171  um q 4924 5000 77 C 93.51 623 0.90 u |

After subtitle, there is a description of every classed monomer:

1. Class of monomer;
2. A/B type of monomer, as determined by independent tool;
3. Start and end position of the monomer in the sequence;
4. Monomer length captured by alignment (real monomer may be longer due to

alignment of the ends problem);

1. C for complementary strand;
2. Identity to consensus;
3. Alignment score;
4. Relative similarity (rs) calculated as score divided by max score possible under current scoring rule;
5. 171 bp monomer type vs 172 bp monomer type (as determined by a special tool which evaluates the deletion in pos. 21).

Then, if it was defined in the settings, the sequences of classed monomers in FASTA format will be printed.

|  |
| --- |
| >um\_1\_AP001464  GATCTGTAAGTGGAAACATG  >R2\_17\_AP001464  TATCTGCAAGTGGACATTTGGACAGCTTTGAGGCTTATAGAGAAAATGGAAATATCTTCA  CATAAAAACTAGACAGAAGCATTCTCAGAAACTTTTTTTTGATGCTTTTATTTAACTCAC  AGAGTTGAACATTCCTTTTCATAGAGCAGTTTTGAAACACTCTTTTTGAAGGATC  >R1\_188\_AP001464  AATTTGCAAGTGGATATTTTGAAAGCTTTGAAGCCTTCGCTGGACACGGAATTATCTTCC  CATAAAAACTAGACAGAAGCATTCTCAGAAACTTATTTGTGATGTTTGCATTTAACTCAC  AGAGCTGAACATTCCTTTTCATTGAGCACTTTTGAAACACTCTTTTTGTAGTATC  >R1\_359\_AP001464  CACTTACAAGTGGATATTTGGACAGCTTCGAGGCTTTCGTTGGAAAAGGGAATATCTTCA  CATAAAAACCAGACTGAAGGCATTCTCAGAAACTTCTTTGTGATGTGCACATTCAACTCA  CAGAGTTGAAACTTTCTTTTGATAAAGCAGCTTTGAAACACTCTTTTTGTAGAATT |

The monomers in the outprint are at least 175 bp long. This includes the length of alignment as indicated in “len” column, extended if needed to match a complete monomer plus an overhang of extra 4 bp at the end. The overhang comes from the start of the next monomer, if AS goes on direct strand, or from the start of the previous monomer, if AS goes on reverse strand. After multiple alignment the overhangs usually have to be trimmed.

At the end of report there is some statistics.

|  |
| --- |
| invalid format doks=0  total OK format doks=1  short OK format doks=0  DISTANCE distribution:  0.0 up to 0.1: 0 7 0  0.1 up to 0.2: 0 0 0  0.2 up to 0.3: 0 0 0  0.3 up to 0.4: 4 0 0  0.4 up to 0.5: 3 0 6  0.5 up to 0.6: 0 0 1  0.6 up to 0.7: 0 4 0  0.7 up to 0.8: 2 13 2  0.8 up to 0.9: 25 18 17  0.9 up to 1.0: 15 3 23  1.0 up to 1.1: 1 3 1  1.1 up to 1.2: 0 2 0  1.2 up to 1.3: 0 0 0  1.3 up to 1.4: 0 0 0  more then 1.4 0 0 0  total: 50 50 50  mean: 0.822 0.726 0.840  std: 0.177 0.302 0.160  total calculation time= 45.70 |

# Menu structure

## Load

The command loads a table that describes AS consensus which before that needs to be chosen from InpDir (F8).

Consensus file needs an additional file (table) that references on it and defines start and end position which is used. Table file ASC contains line: “CONS\_ALP.dng 1 171”.

CONS\_ALP.dng — is the consensus sequence in GenBank format. The comment line begins with an asterisk.

## SF4

Is used to load matrices for M1 subclasses and should be chosen after command Load (F2) despite the fact that these matrices are not used in this release.

## More

### Mode & Parameters of filters

### Distance value filter

### No filter

Uses the mildest possible filter for similarity to consensus (ro).

left limit = 0.00

right limit = 999999.00

### Set filter

Sets the filter for sequence similarity to consensus.

This regulates the fast homology search using octanucleotide dictionaries. Ro=0 for identical sequences (left limit). The sequences with ro less than the right limit are no longer analyzed. If the filter is used, the most diverged AS sequences could be lost. If no filter option is used more sequences have to be evaluated at the following steps and more time is used.

*default:*

left limit = 0.00

right limit = 0.60

### Set RS cut-off value

Sets cut-off value for relative similarity to AS consensus. This value affects the most divergent AS sequences which could be included or excluded from the downstream analysis. For instance, the use of rs=0.2 would allow to analyze more divergent AS sequences, but will increase the background of false AS taken into analysis.

default: 0.29

### Calculation delay in milliseconds

Calculation delay in milliseconds to prevent overheating of processor. Not needed in modern computers.

default: 0

### dev Mark lower limit

These are the parameter used in automatic dot matrix analysis. The minimum value of standard deviation of the observed number of points on dot-matrix from expected, which be marked by \*. (see dev Save lower limit)

default: 30

### str\_max Mark lower limit

These are the parameter used in automatic dot matrix analysis. Lower limit for maximal diagonal on dot-matrix, which be marked by \*. (see str\_max Save lower limit)

default: 30

### Semi-width of the corridor

These are the parameter used in automatic dot matrix analysis.

default: 5

### Mapping parameters

These are the parameters used in automatic dot matrix analysis.

### Mapping: min stretch length to consider

The minimum length of the diagonal stretch on dot-matrix that should be considered.

default: 15

### Mapping: min sequence number in pile

Minimum number of diagonal stretches on dot-matrix to analyzed fragment.

default: 1

### Mapping: min region length

The minimum length of the analyzed fragment formed by dot-matrix diagonals.

default: 50

### vs DB via DISK special file

For long sequences more than 450 kb use this option and choose file with sequences to scan.

### vs. DB with short (<450kb) sq.

For sequences shorter than 450 kb use this option and choose file with sequences to scan.

## Parameters

### Pace Window parameter

The size of sequence analyzed in one step (max 6000 bp). Longer sequences are divided into windows of this size. The shorter sequences will be processed in one piece.

default: 5000

### Step parameter

Step of the overlap for the windows into which the long sequences are divided. The overlap should be longer than monomer length (171 bp). The default overlap is 200 bp, which makes the step value 4800 (5000-200).

default: 4800

### Lower sequence length limit

The minimal length of sequence (bp) which can be analyzed. The minimal possible value of this parameter is 50.

default: 140

### Threshold/Window parameters

Settings used in dot-matrix analysis. Selected fragments must satisfy the conditions: inside the window length WI there must be at least TH number of dots. The pair WI4/TH4 is used. Other parameters are out of use.

default:

TH1 = 63

TH2 = 63

TH3 = 63

TH4 = 31

WI1 = 63

WI2 = 63

WI3 = 63

WI4 = 63

### Probability

Dot probability on dot-matrix

default: 0.0000001

### dev Save lower limit

The minimum value of standard deviation of the observed number of points on dot-matrix from expected due to TH4/WI4 or Probability. The range of recommended values: 50-100.

default: 30

### str\_max Save lower limit

Lower limit for maximal diagonal on dot-matrix. The range of recommended values: 50-100.

default: 30

### echo\_step value

During processing the program will display every Nth sequence on the screen to control processing.

default: 100

## Options

### max\_str Distribution Save style

Print max\_str distribution values to output (see str\_max Save lower limit)

### Not save

default

### Save

### dev Distribution Save style

Print dev distribution values to output (see dev Save lower limit)

### Not save

default

### Save

### Save Fragment Style

#### Put SQ fragment (monomers) to output? <Y/N>

Print sequences of the classed monomers in the report.

default: no

#### Print fragment name prefix:

Prefix in the header of FASTA sequence.

default: >8

#### Put SQ fragment name to output? <Y/N>

Print the sequence name after class of monomer.

default: yes

### Save File Name

Specify path and name for output file.

default: malux3

### Extra SQ Lines To Skip

Sets number of sequence lines to skip at the beginning of every analyzed sequence in the file. Every line contains 60 bp. By the use of this option a sequence or sequences may be analyzed not from the start, but from any given site.

default: 0

## Input Directory

Selects ASC consensus file. This is the first step to get started.