Program NsiteH

ACTION: Search for statistically nonrandom and conservative motifs of human/animal and plant transcription Regulatory Elements (REs) available in both of aligned orthologous/homologous DNA sequences.

SEARCH CONDITIONS: (1) Expected Mean Numbers of any regulatory motif found must be less than a given number (default: 0.05).
(2) Homology Level of any motif in one sequence with the corresponding area of another sequence (in relation to ALIGNMENT) must be higher than a given level (default: 80%).

AUTHORS: I.A.Shahmuradov & V.V.Solovyev
LAST UPDATE: 15 Novembef 2013
VERSION: 5.2013
ACCESS: http://softberry.com

Program description:

As **NsiteH** and **NsiteM** programs, **Nsite** is based on the previously proposed real site and/or their IUPAC consensus based probabilistic approach of searching for putative REs in nucleotide sequences and statistically estimating motifs found (*Shahmuradov et al., Genetika (Russ.), 1986, 22, 357-367; Solovyev,V.V. and Kolchanov,N.A. In: "Computer analysis of genetic macromolecules. Structure, Function and Evolution", World Scientific, 1993,16-20*). The main features of the approach are the follows:

(i) RE may consist of a single box (a continuous DNA segment) or two boxes, spaced by some DNA sequence, where only length, but not nucleotide content, of this spacer is important for functioning of such a composite site.
(ii) A real RE or its IUPAC consensus contains both variable positions (where the presence of a certain group of nucleotides is permissible), and strictly conserved positions (where a strong identity between real site/consensus and predicted motif is required).
(iii) The nonequivalence of these positions should be taken into account, i.e., complete homology at conserved positions is needed, and a violation of homology in the variable positions should be permissible.
(iv) The homology between the RE and motif on query DNA sequence may a chance happening, therefore, estimation of its statistical significance is of major importance. A conclusion on the functional significance of the revealed homology can be reached, only if the homology is significantly nonrandom, i.e., the homology is not a chance event.
(v) Characteristics such as nucleotide frequencies should not be used when describing the consensus because of its small size. Instead, one should use estimates based on the number of nucleotides of various types in the consensus.
(vi) Although all available databases on REs usually annotate a fixed distance between two boxes of composite elements, some variability of the spacer length, seems, to have a place. Therefore, a search algorithm for composite REs should allow some limited flexibility in the sense of spacer length, relaying on both the known experimental data and theoretical assumption.

One of the important components of this approach is preparation of RE data set. **Nsite** uses the following 3 sets of human/animal and plant REs.

[1]: Set of human/animal REs prepared by merging and verification (excluding longer than 50 bp REs; limination of redundancies) RE/TF information from **ooTFD** (*Ghosh,D. Nucleic Acids Res, 2000, 28, 308- 310*). It includes **8030** real/consensus REs.

[2]: Set of human/animal REs prepared by merging and verification RE/TF information from **RegsiteAN DB** developed Softberry Inc. It includes **...** real/consensus REs.

[3]: Set of plant REs prepared by merging and verification RE/TF information from RegsitePL **DB** developed by Softberry Inc. The current version of this set includes 2359 real/consensus REs, but it is regularly updated.

Moreover, user can perform a search for motifs of REs from his own dataset in a format described below.

Input query sequences of length 100000 nucleotides or less must be given in FASTA format.

RULES for creating USER's set of REs:

1. USER's set must include only sequences of real REs and/or their consensuses.

2. Every real RE/consensus is described in 3 lines:

    LINE 1: Name/description of RE/consensus
    LINE 2: Sequence of of RE/consensus
    LINE 3: <par1> <par2> <par3> <par4>

3. Sequence (LINE2) may include both standard nucleotides (A/a, T/t, G/g,C/c) and any combination of them in according to IUPAC abbreviations:R - A or G, Y - T or C, K - G or T, M - A or C, S - G or C, W - A or T, B - G or T or C, D - A or G or T, H - A or C or T, V – A or G or C, N - A or G or C or T.

   In the case of composite REs, 2 boxes are seperated by "-".

   Length of RE/consensus sequence must not exceed 80 symbols, including "-" in the case of composite elements.

   Capital letters indicate Conservative nucleotides (positions) where a mismatch is not allowed.

4. In LINE 3: <par1> - a maximal number of mismatches for the 1st box
   <par2> - a maximal number of mismatches for the 2nd box (for composite REs)
   If RE contains a single box, then <par2> = 0;
   If any mismatch is not allowed, then <par1> = <par2> = 0)

   <par3> - minimal distance between boxes of composite RE
   <par4> - maximal distance between boxes of composite RE (for a single-box REs <par3> = <par4> = 0)

All <par1>, <par2>, <par3> and <par3> are given as INTEGER in 4i5 format.

Example of USER's set of 3 REs:

RE 1
agTGGcgAggcg
2 0 0 0
RE2
caggccTGc-CCAGctgg
1 1 8 10
RE 3
RRTGTGGWWW
0 0 0 0

How run NsiteH program:

How run NsiteH program:

nsiteh -i:<parI> -d:<parD> -h:<parH> -a:<parA> [-c:<parC>] [-o:<parO>] [-p:<parP>] [-n:<parN>] [-m:<parM>] [-r:<parR>] [-v:<parV>] [-u:<parU>] [-s:<parS>]

Options/Arguments:

<parI> Input File with the first Query DNA sequence in FASTA format (max. length: 100 000 bp).
If this option/argument is absent, Program display HELP information and ends.
In Query sequences symbols besides of "a/A", "g/G", "c/C" and "t/T" are accepted as "N".

<parH> Input File with the second Query DNA sequence in FASTA format (max. length: 100 000 bp).
If this option/argument is absent, Program display HELP information and ends.
Program displays HELP information and ends.

<parA> File of Alignment of 2 homologous sequences by sbl program
If this option/argument is absent, Program display HELP information and ends.
To get a File of Alignment, run:

sbl <parI: SEQ 1> <parH: SEQ 2, hom> -o:weak.cfg -S:2 -D:0 > <parA: Align.>

In SBL command line the 1st SEQ file must be given <parI> and the 2nd SEQ file must be given <parH>.

<parC> Conservation Level (Integer, > 0, <= 100)
Default: 80 (%)

<parD> Set of REs
<parD> = RegsitePLDB.dat ... search for motifs of Plant REs (RegsitePL DB)
<parD> = ooTFD_NR.dat ... search for motifs of Human.Animal REs (ooTFD)
<parD> = RegsiteANDB.dat ... search for motifs of Human/Animal REs (RegsiteAN DB)
<parD> = <User Set of RES> ... search for motifs of user-provided REs (see below).

If this option/argument is absent OR DB namee is not correct, Program display HELP information and ends.

<parO> Output File

Default: NsiteH.res

<parP> Print (y) or not (n) Query sequence
Default: n

<parN> Positions of motifs found are given in relation to Right Boundaries of Upstream Sequences [if Query sequences include upstream sequences of genes] (y), or positions of motifs found are given as in Query sequences (n).
If <parN> = y, Data File with Right Boundaries positions Upstream Sequences must be given by <parU>.
Default: n

<parM> Mean Expected Number (Real, >= 0.).
Default: 0.05

<parR> Statistical Significance Level (Real, > 0. , <= 100.0).
Default: 0.95

<parV> To allow variation for the known distance (D) between 2 blocks of Composite Regularory Elements by VL% from (D-VL*D/100) to (D+VL*D/100).
Default: parV=20%

<parU> Data File with Right Boundaries positions of upstream sequences.
If <parN>=n, this, argument is ignored);
if <parN>=y, but <parU> is absent or empty, right boundary position is a position of the last nucleotide of Query sequence.

<parS> Minimal level of homology between Known RE/consensus and motif found.
Default: parS=80 (%)

NOTE: If number of arguments is more than 13 or no any argument is given, Program display HELP information and ends.

NsiteH output:

Every OUTPUT file begins with description of the Program's allocation, Search Parameters, as well as used abbreviations (in the case of using Data sets created by us). The next 2 includes name and length of the first query sequence. At last, name of REs, statistical estimation and sequences of motifs found are given.

For example:

```
Program   NsiteH | Version 5.2013
  Search for motifs of   2356 RegulatoryElements (REs) in a pair of Homologous Sequences
  SET of REs: RegsitePDB: 2359 Plant Transcription REs [Last update: 05.02.2012; Softberry Inc]
_____________________________________________________
Search PARAMETRS:
    Expected  Mean  Number                :  0.0500000
    Statistical Siginicance Level         :  0.9500000
    Minimal Conservative Level            :  80 %
    Level of homology between known RE and motif:   80%
    Variation of Distance between RE Blocks     :   20%
NOTE: RE - Regulatory Element/Consensus   | AC - Accession No of RE in a given DB
```

```
       OS - Organism/Species   | BF - Binding Factor or One of them
       Mism. - Mismatches   | Mean. Exp. Number - Mean Expected Number   | Up.Conf.Int. - Upper
Confidence Interval
============================================================
 QUERY: >gi|19674|emb|X12512.1| Nicotiana plumbaginifolia Cab-E gene 5'-flanking region
 Right Boundary of Upstream Sequence:    554
 Length of Query Sequence:        554 bp      | Nucleotide Frequencies:  A -  0.34   G -  0.18   T -
0.31   C -  0.16

      1  CATCCTTACT AACGGTAAAT AACTTAGAAG TTATTGTATA CGTATGATCG
     51  AGCTGTTGGA CTTGTAGTAT CAAACTTTCA ATGACGCATC AAAATTAATT
    101  ATGGTAGCTT CGCGTTGGGA CACTTGTACA TGCATTAACT TGATTTCAAT
    151  TTCTTTTTTA AAAATATTTG TCTATTGTCA ATTTACCACT CGTACTTGAA
    201  GTGGGCCTAT TTGACAGGTC AGCTAAATAC AGAAGTGTAT GAACAATGCG
    251  TGGCCAAGAG TAACTCTTAT GCTAAAGACA AGTGGATATT ATATTGCAAT
    301  AATCCACAAT CAGACGTGGC AAATTTGGAT TGGCTATAAG AGAGCAAATC
    351  TTCATTAGGT AAGTTTTTTA AACATAAAAA GTATCTAAAA AAATCTTGTC
    401  ATGTTTAACG GTGCTGAACT TTGCCAAATG GACAAGAATG CAAAAGGTTA
    451  AAATTGCAAT CCACCAATTG AAAAGTAGAT ATAGATACTC AAGGATAAGG
    501  GTCTTTGGGC CTGTAAAGCC ATTTATATAC ACTTAGTGCA AAGCCCATGA
    551  AACT
.............................................................
 RE:     2. AC: RSP00002//OS: Brassica napus /GENE: Oleosin/RE: ABRE-3 /BF: B.napus embryo protein
factor
 Motifs on "+" Strand: Mean Exp. Number   0.01834     Up.Conf.Int.  1     Found   1
    -243  AgACGTGGC     -235 (Mism.= 1; Cons.: 100 %)
.............................................................
 RE:    35. AC: RSP00035//OS: barley (Hordeum vulgare) /GENE: Al21/RE: D1 /BF: DOF
 Motifs on "+" Strand: Mean Exp. Number   0.04169     Up.Conf.Int.  1     Found   1
    -114  CAAAAGG     -108 (Mism.= 0; Cons.: 100 %)
.............................................................
 RE:    93. AC: RSP00093//OS: barley (Hordeum vulgare) /GENE: Amy pHV19/RE: TATCCAC box /BF: unknown
nuclear factor
 Motifs on "-" Strand: Mean Exp. Number   0.03964     Up.Conf.Int.  1     Found   1
    -267  TATCCAC     -273 (Mism.= 0; Cons.:  85 %)
.............................................................
 RE:   204. AC: RSP00204//OS: arabidopsis (Arabidopsis thaliana) /GENE: AtEm6/RE: ABRE/6.2 /BF: ABI5
 Motifs on "+" Strand: Mean Exp. Number   0.02794     Up.Conf.Int.  1     Found   1
    -244  cAgACGTGGC     -235 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   208. AC: RSP00208//OS: French bean (Phaseolus vulgaris) /GENE: DLEC2/RE: DLEC2,A /BF: MAT2
(ROM2)
 Motifs on "-" Strand: Mean Exp. Number   0.00260     Up.Conf.Int.  1     Found   1
    -235  GCCACGTCtGaT     -246 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   219. AC: RSP00219//OS: arabidopsis (Arabidopsis thaliana) /GENE: RBCS-1A/RE: G box-1 /BF:
HY5; Arabidopsis bZIP protein - transcriptionl factor
 Motifs on "+" Strand: Mean Exp. Number   0.00908     Up.Conf.Int.  1     Found   1
    -245  TcAgACGTGGCA     -234 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   248. AC: RSP00248//OS: rice (Oryza sativa), Oryza sativa /GENE: alpha-globulin/RE: REB2 /BF:
REB
 Motifs on "-" Strand: Mean Exp. Number   0.00345     Up.Conf.Int.  1     Found   1
    -235  GCCACGTCtG     -244 (Mism.= 1; Cons.: 100 %)
.............................................................
 RE:   301. AC: RSP00301//OS: rice (Oryza sativa) (Oryza sativa) /GENE: GluB-1/RE: PROL box /BF:
unknown nuclear factor
 Motifs on "+" Strand: Mean Exp. Number   0.03776     Up.Conf.Int.  1     Found   1
     -18  TGCAAAG      -12 (Mism.= 0; Cons.:  85 %)
.............................................................
 RE:   304. AC: RSP00304//OS: maize (Zea mays) /GENE: Synthetic oligonucleotides/RE: KN1/KIP BS /BF:
KN1/KIP
 Motifs on "+" Strand: Mean Exp. Number   0.01171     Up.Conf.Int.  1     Found   1
    -343  TGACAGGT     -336 (Mism.= 0; Cons.:  87 %)
.............................................................
 RE:   629. AC: RSP00629//OS: arabidopsis (Arabidopsis thaliana) /GENE: Lhcb1*3/RE: CCA1 BS2 /BF:
CCA1
 Motifs on "+" Strand: Mean Exp. Number   0.04127     Up.Conf.Int.  1     Found   1
    -166  AAAAATCT     -159 (Mism.= 0; Cons.: 100 %)
.............................................................
 RE:   683. AC: RSP00683//OS: arabidopsis (Arabidopsis thaliana) /GENE: Adh/RE: -190 half G-box
(core) /BF: GBF3
 Motifs on "+" Strand: Mean Exp. Number   0.00753     Up.Conf.Int.  1     Found   1
    -132  GCCAAaTGGA     -123 (Mism.= 1; Cons.: 100 %)
.............................................................
 RE:   738. AC: RSP00738//OS: arabidopsis (Arabidopsis thaliana) /GENE: CAB2/RE: CUF-1 BS /BF: CUF-1
 Motifs on "-" Strand: Mean Exp. Number   0.03677     Up.Conf.Int.  1     Found   1
```

```
    -235  GCCACGTctG     -244 (Mism.= 2; Cons.: 100 %)
.....................................................................................
.....................................................................................
.....................................................................................


.............................................................
 Totally       68 motifs of    66 different REs have been found
 Of them:      50 motifs of    49 different REs have Conservative Level >=  80 %
-------------------------------------------------------------
 QUERY: >gi|3036947|dbj|AB012638.1| Nicotiana sylvestris Lhcb1*5 genes for light harvesting
chlorophyll a/b-binding protein, complete cds
 Right Boundary of Upstream Sequence:    520
 Length of Query Sequence:        520 bp      | Nucleotide Frequencies:  A -  0.33   G -  0.18   T -
0.30   C -  0.18

      1  ATCCCGAACC CAAAGTTTGA AATCCTGGCT CCGCCTCTGA TCCGTGCCCC
     51  CTAAAGGCTT TTAGCATTAA GGGTGTCAAG TGATTTAAAT TAATCATTTT
    101  CAAGGTATAC ACATATACAT ATACAAGATT TCTGCTGAAG TTTACGGGTC
    151  CGCCACTGCA TACAAGTGGT CCTATTTTAC AGGTCAGCTA AATATACAGA
    201  AGTGTATGAA CAATGCATGG CCAGGAGTTA CTCTTATGCT CTGGCTAAGT
    251  TGATATTATA TTGCAATAAT CCACAATCAG ACGTGGCAAA TTTGGATTGG
    301  CTATAAGAAG GAAATCTTCA TTGGCTTAGA TTTTTTAAAC GTATAAAGTA
    351  TCTACAAAAA TCTAGTCATC TTTAACGGTG CAGAACTTTG CCAAATGGAA
    401  AAGAATGCAA AAGGTTACAA ATTGTCATCC ACCAATGGAA AAGCAGATAT
    451  AGATATTCAA GGATAAGGTA GTCTTTGGGC CTGTAAATTC ATTTATATAC
    501  ACTTAGTACA AAGCCCATAA
.............................................................
 RE:     2. AC: RSP00002//OS: Brassica napus /GENE: Oleosin/RE: ABRE-3 /BF: B.napus embryo protein
factor
 Motifs on "+" Strand: Mean Exp. Number   0.02276     Up.Conf.Int.  1     Found   1
    -242  AgACGTGGC     -234 (Mism.= 1; Cons.: 100 %)
.............................................................
 RE:     6. AC: RSP00006//OS: soybean (Glycine max) /GENE: GS15/RE: ATRE /BF: unknown nuclear factor
 Motifs on "+" Strand: Mean Exp. Number   0.02006     Up.Conf.Int.  1     Found   1
     -36  AAATTcaTTTATAT      -23 (Mism.= 2; Cons.:  85 %)
.............................................................
 RE:    35. AC: RSP00035//OS: barley (Hordeum vulgare) /GENE: Al21/RE: D1 /BF: DOF
 Motifs on "+" Strand: Mean Exp. Number   0.04055     Up.Conf.Int.  1     Found   1
    -113  CAAAAGG     -107 (Mism.= 0; Cons.: 100 %)
.............................................................
 RE:   204. AC: RSP00204//OS: arabidopsis (Arabidopsis thaliana) /GENE: AtEm6/RE: ABRE/6.2 /BF: ABI5
 Motifs on "+" Strand: Mean Exp. Number   0.03270     Up.Conf.Int.  1     Found   1
    -243  cAgACGTGGC     -234 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   208. AC: RSP00208//OS: French bean (Phaseolus vulgaris) /GENE: DLEC2/RE: DLEC2,A /BF: MAT2
(ROM2)
 Motifs on "-" Strand: Mean Exp. Number   0.00300     Up.Conf.Int.  1     Found   1
    -234  GCCACGTCtGaT     -245 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   219. AC: RSP00219//OS: arabidopsis (Arabidopsis thaliana) /GENE: RBCS-1A/RE: G box-1 /BF:
HY5; Arabidopsis bZIP protein - transcriptionl factor
 Motifs on "+" Strand: Mean Exp. Number   0.01014     Up.Conf.Int.  1     Found   1
    -244  TcAgACGTGGCA     -233 (Mism.= 2; Cons.: 100 %)
.............................................................
 RE:   248. AC: RSP00248//OS: rice (Oryza sativa), Oryza sativa /GENE: alpha-globulin/RE: REB2 /BF:
REB
 Motifs on "-" Strand: Mean Exp. Number   0.00424     Up.Conf.Int.  1     Found   1
    -234  GCCACGTCtG     -243 (Mism.= 1; Cons.: 100 %)
.............................................................
 RE:   350. AC: RSP00350//OS: maize (Zea mays) /GENE: C1/RE: VP1 RE /BF: unknown nuclear factor
 Motifs on "-" Strand: Mean Exp. Number   0.04384     Up.Conf.Int.  1     Found   1
    -298  tGgCCATGCAT     -308 (Mism.= 2; Cons.:  90 %)
.....................................................................
.....................................................................
.....................................................................
 .............................................................
 Totally       71 motifs of    67 different REs have been found
 Of them:      45 motifs of    44 different REs have Conservative Level >=  80 %
```