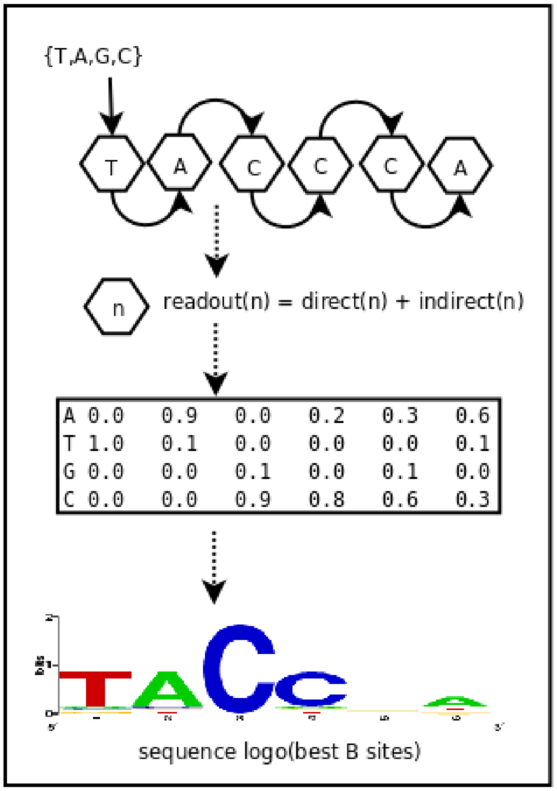
DNAPROT manual

Bruno Contreras-Moreira
Laboratory of Computational Biology
http://www.eead.csic.es/compbio
ARAID / Estación Experimental de Aula Dei / CSIC
Zaragoza, España

January 31, 2012

Contents


{T,A,G,C}
T
A
C
C
C
A
n
readout(n) = direct(n) + indirect(n)
sequence logo(best B sites)

A	0.0	0.9	0.0	0.2	0.3	0.6
T	1.0	0.1	0.0	0.0	0.0	0.1
G	0.0	0.0	0.1	0.0	0.1	0.0
C	0.0	0.0	0.9	0.8	0.6	0.3

Figure 1: DNAPROT simplified flow chart.

1 Description

This document describes the software DNAPROT and provides a few examples on how to use it. DNAPROT is an algorithm that essentially converts an array of atomic interactions, inferred in an input protein-DNA complex in PDB format, into a position weight matrix (PWM) which to some extent captures binding specificity.

2 Requirements and installation

DNAPROT has only been tested in Linux systems, and includes both Perl scripts and binary files for which source code is not available as it makes use of several bits originally produced by third parties. Therefore, a Perl interpreter is needed to run this software, which should be installed by default in all Linux systems.

In order to test this software please follow these steps:

1. Unpack the DNAPROT software package with `tar xvfz dnaprot_X.Y.tgz`
2. `cd dnaprot_X.Y`
3. Try the main Perl script, named `dnaprot.pl`, with the included sample input file (`1je8_AB.pdb`) in PDB format, by means of the instruction `./dnaprot.pl -i 1je8_AB.pdb` . You should get an output very similar to the contents of file `sample_output.txt`. In addition, the script should produce a PDB file with remediated DNA coordinates of the original input complex, which can be visualized using standard software such as RasMol, Jmol or PyMOL.

2.1 Perl modules

The Perl module FindBin, which is by default installed on most Linux distributions, is required by the DNAPROT scripts. The best place to obtain it otherwise is the Comprehensive Perl Archive Network (CPAN), which can be also browsed from the terminal with the command `cpan`.

3 User manual

This section describes the available options for the DNAPROT algorithm.

3.1 Input files

The input required to run DNAPROT consists of a text file, in PDB format, which contains the atomic coordinates of a complex between one or more protein chains and one or two DNA chains, which are expected to form a duplex. The DNAPROT package includes a sample input file named `1je8_AB.pdb`.

It's OK if they input file contains a single DNA chain, but the algorithm will find problems when two or more non-helical chains are included. For this reason the `dnaprot.pl` script attempts to remediate the DNA coordinates before calling the DNAPROT algorithm itself. The script used for this task is named `correct_dna.pl` and can be found in the `bin/` directory. Essentially, the script selects the longest DNA duplex found in the input coordinates, but will try to use the original coordinates if no duplexes are recognized. The remediated file, which will contain in most cases a duplex with DNA chains renamed as `a` and `b`, is part of the generated output:

```
# file with remediated DNA coordinates: complex_1je8_AB.pdb
```

3.2 Basic DNAPROT options

Typing `$ dnaprot.pl -h`, on the terminal will show the basic options:

```
$ ./dnaprot.pl -h
usage: ./dnaprot.pl [options]
```

```
 -h this message
 -i input complex file in PDB format
 -r use relaxed dnaprot                   (optional,...
 -d do not use hetatoms found in input    (optional)
 -p pass DNAPROT options                  (optional,...
 -L list all DNAPROT available options    (optional)
 -V show version and references and exit  (optional)
```

The only required option is `-i`, used to choose an input file.

Among all possible ways of running DNAPROT, by default this script produces three structure-based PWMs: a readout matrix, a contact matrix and a mean matrix which will up-weight the contact contribution when the number of observed interface interactions is below 5. This default behaviour amounts to using the `-p ' -P -1 -C -e -c -O '` arguments.

This takes us to the `-p` option, which can be used to pass arbitrary options to the DNAPROT binary, between quotes. The full list of available options can be retrieved by typing `$ dnaprot.pl -L` .

The `-d` flag can be passed to ignore any hetatoms found in the input PDB file, which can be useful to study the effect of removing existing water molecules.

Finally, the `-r` option calls a relaxed version of the DNAPROT binary that in our tests is useful to study the interface of homology models, since usually these modelled molecules contain conformational errors. In fact these relaxed behaviour is obtained by calling the `my_vwhbplus.pl` script (in the `bin/hbplus/` directory) with a relaxed maximum distance for Van der Waals contacts, that increases to 4.4Å.

3.3 Advanced DNAPROT options

As mentioned in the previous section, the full list of available options of DNAPROT can be retrieved by typing `$ dnaprot.pl -L` . The list looks as follows:

```
$ ./dnaprot.pl -L

-m number of start site                      (optional, continue runs)
-l number of last site                       (optional, stop runs)
-c apply residue-DNA base distance correction (optional)
-b apply B-factor column correction          (optional)
    Bfactor columns as user-defined weights
-w calculate nucleotide substitution weights  (optional)
    to correct recognition scores
-e evaluate protein-DNA complex and exit      (optional)
-S symmetry-correct scoring matrices         (optional)
-M explore only mutated contacts (Bfact>0.00) (optional, applies to -P)
    keeps conserved residues unchanged
-P produce PWM with this background %GC,      (optional)
    use -1 for an occurrence matrix or -9 for a SC count matrix
```

```
-C produce PWM by counting interface contacts (optional)
-t dissecT (in)direct interface contributions (optional)
-a get mPWM by averaging PWM and cPWM with    (optional)
    this [0-1] cPWM weight
-o output complex with consensus DNA          (optional, expects file name...
-W pseudo weight factor [0,1]                 (optional, default 0.01)
    affects PWM calculation with -P
-O get optimally weighted mPWM                (optional, up-weights cPWM...
-I use this PDB file with other molecules     (optional)
    present at the interface, usually HET water
-D use this [0-1] float to weight deformation (optional)
    1-D is the direct readout weight
-s file with DNA sequences to be parsed       (optional)
   Line format (max 2000 char): identifier \ ggaa...tagc \
```

The -P,-C,-e,-c,-O options have already been mentioned, they are now explained in more detail:

- -P is used to request a readout PWM.
- -C is used to request a contact PWM (cPWM), as done by Morozov and Siggia in their 2007 paper.
- -c is used to linearly correct the interaction weights extracted from the hydrogen bond and hydrophobic interaction matrices in terms of the physical distance of the interaction. Given a distance limit of 3.9Åbetween pairs of interacting atoms, based on the work of Brameld for hydrogen bonds, the following formula is applied:

$$score(dist) = rawscore\,(1.0 + \frac{3.9 - dist}{3.9})$$

- -e is used to request only an evaluation of the input protein-DNA complex, otherwise all possible oligonucleotides of the same length will be sampled and scored. In order to control oligonucleotide sampling users are advised to use options -m,-l,-s.
- -b is useful to freeze parts of the interface DNA, by adding Bfactors with 00.00 value in the input PDB file, or to up-weight parts of the DNA motif by adding arbitrary values (between 00.00 and 99.99) to the same Bfactor column of a PDB file:

```
                                                             Bfactor
                                                             |
ATOM   1369  N9   DG C  14      14.938 -11.067  92.134  1.00 29.65
```

- -M is similar to -b but for protein residues: it will change the -P behaviour in order to drive in silico mutagenesis only for nucleotides contacted by mutated protein residues, which are identified as having Bfactors > 00.00 .

- `-S` can be selected in order to correct for asymmetric weights in interaction matrices, giving equal values to atoms that in theory can play the same role in hydrogen bonds (A:N3 = G:N3 , T:O2 = C:O2 , ARG:NH1 = ARG:NH2 , ASP:OD1 = ASP:OD2 , GLU:OE1 = GLU:OE2).

- `-D` is used to change the contribution of DNA deformation (indirect readout) to the PWM generated with `-P`, as explained in reference 1. This parameter is important particularly for proteins that rely more on indirect readout mechanisms. The next equation describes the readout score for base b in position j of the input DNA motif:

$$score(b,j) = e^{-(1-D)\,direct(b,j)+D\,indirect(b,j)}$$

- `-W` is used to change the pseudo weight factor used during the computation of `-P` PWMs, as discussed in reference 1. In our tests this parameter had very little impact.

- `-o` creates an output PDB file, with the name passed by the user, including the consensus DNA sequence calculated by DNAPROT.

- `-a` gives control on how `-P` and `-C` PWMs should be averaged linearly. The closer to 1, the higher the contribution of cPWM.

- `-t` will produce four different PWMs, each one derived from indirect readout, hydrogen bonds, water-mediated hydrogen bonds and hydrophobic interactions respectively.

- `-w` produces a string which summarizes the weight of each nucleotide within the original DNA sequence contained in the input coordinates. The weight or relative importance of each nucleotide is scaled between 0 and 9. This option requires in silico mutating all nucleotides in order to estimate the standard deviation of each column of the corresponding PWM. Standard deviations are normalized with respect to the highest deviation. Positions (columns) with high values are those whose interface scores change most when the the nucleotide mutates. The generated string looks like this:

```
# Estimating nucleotide weights with first protein model ...
CGTACCCATTAATGGGTACG
00505905300350950500
```

4 A few examples of use

This section presents a few different ways of running DNAPROT with the provided sample input, which corresponds to the dimer NarL bound to DNA.

4.1 Checking the interface of a PDB complex

The output produced by command `$ ./dnaprot.pl -i 1je8_AB.pdb` should include:

- A list of atomic interactions found at the original interface

  ```
  # Original interface contacts:
  : w : LYS   NZ  A0188 <- 5.54 ->  DT   O4 b0013 : score 1.895
  : w : LYS   NZ  A0188 <- 5.40 ->  DA   N7 b0012 : score 1.356
  : H : LYS   NZ  A0192 <- 2.86 ->  DG   N7 b0015 : score 5.35927
  : H : LYS   NZ  A0192 <- 2.91 ->  DG   O6 b0015 : score 5.39154
  : H : LYS   NZ  A0192 <- 3.05 ->  DG   O6 b0016 : score 5.23718
  : w : LYS   NZ  B0188 <- 5.74 ->  DT   O4 a0033 : score 1.895
  : w : LYS   NZ  B0188 <- 5.41 ->  DA   N7 a0032 : score 1.356
  : H : LYS   NZ  B0192 <- 3.22 ->  DG   N7 a0035 : score 4.96871
  : H : LYS   NZ  B0192 <- 3.28 ->  DG   O6 a0036 : score 4.98359
  : V : VAL  CG1  A0189 <- 3.65 ->  DT   C7 a0023 : score 4.95446
  : V : LYS   CE  B0188 <- 3.85 ->  DT   C7 a0033 : score 2.39937
  : V : VAL  CG1  B0189 <- 3.80 ->  DT   C7 b0003 : score 4.77538
  : V : LYS   CE  A0188 <- 3.83 ->  DT   C7 b0013 : score 2.41152
  ```

  where the first field classifies interactions as hydrogen bonds (H), water-mediated hydrogen bonds (w) or hydrophobic interactions (V). On the right side details are provided for the atom in the protein side, while nucleotide atoms are shown on the left. The Cartesian distance between both atoms is reported among `<- arrows ->`.

- A summary string showing the original DNA motif with bases involved in atomic interactions, with the total log-odd and several DNA deformation energies (step, base pair, stability) reported:

  ```
  seq orig PDB 0 cgTaCCcATtaATgGGtAcg 46.983 13 total      \\
  46.983 step 15.19 bp 4.15 dGDNA 23.22
  ```

- A readout PWM:

  ```
  # Estimating occurrence matrix with first protein model (onlymutated=0)...
  > PW = (PS * (1-w)) + (PD * weight) | weight = 0.50 | PMscale = 100
  A |  ..  4  29  15   0  21  73  12  25  29  29   5  26  13   5  17  83  ..
  C |  ..  6  24  71  94  27   7  42  19  23  11   7  17   7   3  27   5  ..
  G |  ..  5  27   5   1  23   7  10  21  18  42   7  34  76  69  18   5  ..
  T |  .. 81  16   5   1  25   9  32  31  26  14  77  19   0  19  34   3  ..
  ```

- A summary string reporting bases involved in indirect readout:

  ```
  # indirect readout mask (38) : 00000000000011001100
  ```

- Two string showing the PWM score of the original and the best (consensus) possible DNA sequence:

  ```
  seq orig weight: CGTACCCATTAATGGGTACG 990.00
  seq cons weight: CGTACCCACTAGTGGGTACG 1013.00
  ```

- A contact PWM and a report of the contacts observed in both DNA strands:

  ```
  # Estimating PWM by counting interface contacts with first protein model  \\
  (min_contacts_cons=20)...
  > PW = (PS * (1-w)) + (PD * weight) | weight = -1.00 | PMscale = 100
  A |  .. 13  56   9  14  23  38  22  24  24  30  19  19  14  10  14  60  ..
  C |  .. 14  14  69  52  31  20  22  24  24  22  19  19  16  10  14  12  ..
  G |  .. 13  13   9  14  21  19  25  24  24  22  19  39  52  66  14  12  ..
  T |  .. 56  13   9  16  21  19  27  24  24  22  39  19  14  10  54  12  ..

  # contact count :     \\
  .. 9/0 9/0 6/6 0/8 0/2 0/4 0/1 0/0 0/0 1/0 4/0 4/0 8/0 5/6 0/8 0/10 ..
  ```

- A mean PWM:

  ```
  # mean PWM (weight=0.500000) (optimal, interactions=13)
  > PW = (PS * (1-w)) + (PD * weight) | weight = 0.50 | PMscale = 100
  A |  ..  8  42  12   7  22  56  17  24  26  30  12  22  14   7  15  71  ..
  C |  .. 11  19  70  73  29  13  33  23  24  16  13  18  11   8  20  10  ..
  G |  ..  9  21   7   8  22  13  17  22  21  32  13  36  64  67  17   8  ..
  T |  .. 68  14   7   8  23  14  29  27  25  18  58  20   7  14  44   7  ..
  ```

4.2 Dissecting the recognition mechanisms of a given interface

The output produced by command `$ ./dnaprot.pl -i 1je8_AB.pdb -p ’-e -t’` should include:

- An indirect readout PWM and a binary string marking those nucleotides involved:

  ```
  # Dissecting PWM contributions (fixed %GC=50)...
  : indirect readout interface 00000000000011001100
  > PW = (PS * (1-w)) + (PD * weight) | weight = 1.00
  A |  .. 12.00 33.00 16.00 17.00 18.00 31.00 25.00 25.00 34.00 32.00 ..
  C |  .. 32.00 19.00 35.00 37.00 30.00 24.00 18.00 17.00 19.00 18.00 ..
  G |  .. 21.00 31.00 15.00 27.00 21.00 21.00 18.00 18.00 16.00 15.00 ..
  T |  .. 31.00 13.00 30.00 15.00 27.00 20.00 35.00 36.00 27.00 31.00 ..
  ```

- A hydrogen bonds PWM.

- A water-mediated hydrogen bonds PWM.
- A hydrophobic interface PWM.

4.3 Checking the binding scores (affinity) along known DNA sequences

We can create a text file (named for instance `NarL.sites`) with DNA sequences, such as promoters or binding sites, that we wish to scan with the DNAPROT algorithm and the NarL PDB complex. The accepted format is:

```
1 \ cgtacccattaatgagtaag \
2 \ cgtacccattaatgggcaag \
3 \ cgtacccattaagaggtatg \
```

The scan can be done in two ways:

1) Scoring each DNA sequence directly in terms of the readout scores, with the command `./dnaprot.pl -i 1je8_AB.pdb -p '-s NarL.sites'`
2) Scoring each site with a calculated PWM, such as a readout PWM or a contact PWM, with a command such as
`./dnaprot.pl -i 1je8_AB.pdb -p '-s NarL.sites -P -1'` or
`./dnaprot.pl -i 1je8_AB.pdb -p '-s NarL.sites -C'`

In either case, the output will look like, showing the scores of matching the PWM in direct and reverse complement orientations:

```
# Parsing the DNA sequences file...
seq 0 1 0 CGTACCCATTAATGAGTAAG 906.00 284.00 dGDNA 20.96
seq 1 2 0 CGTACCCATTAATGGGCAAG 956.00 295.00 dGDNA 23.51
seq 2 3 0 CGTACCCATTAAGAGGTATG 889.00 288.00 dGDNA 21.71
```

5 Known limitations of DNAPROT

The algorithms presented here have several limitations, which are now briefly described.

- Quality of coordinates at the interface. Often the input coordinates contain very few atomic interactions at the interface and therefore the process of in silico mutagenesis fails to provide a comprehensive report of specific DNA binding. This can happen both with experimental structures and homology models. In our 3D-footprint benchmark we found that complexes that yield less than 5 contacts produce bad PWMs.

- Geometric description of atomic interactions. It is common to find protein residues at DNA interfaces which have multiple short-distance contacts with purines or pyrimidines that the algorithm cannot label as hydrogen bonds or hydrophobic contacts. While these cases provide good quality cPWMs, again the quality of PWMs are negatively affected by these situations. Future work will have to consider enhancing the repertoire of possible interactions at the interface. Our 2009 paper shows that hydrogen bonds are the most common interactions, and that different superfamilies employ different indirect readout mechanisms.

- Quality of produced PWMs. The PWMs generated by DNAPROT have not been derived from a collection of known DNA sequences, and therefore are not necessarily suited to carry out pattern matching straight away. In our 3D-footprint tests we found it useful to refine them by piling the best B sites scored by any given PWM, usually setting B to 50, and then calculating the resulting refined PWM by calculating the observed nucleotide frequencies in the B aligned sites.

6 Credits

DNAPROT is designed, created and maintained at the Laboratory of Computational Biology at Estación Experimental de Aula Dei in Zaragoza, Spain, although the first prototypes of the software were originally developed at the Center for Genomic Sciences of Universidad Nacional Autónoma de México (CCG/UNAM).

So far the code has been written mostly by Bruno Contreras-Moreira, but it also includes contributions from my colleagues Vladimir Espinosa and Paul W.Fitzjohn, and also from Marc Parisien, who guided me in measuring DNA deformation.

Moreover the DNAPROT algorithm relies on a modified version of the HBPLUS program, originally written by IK McDonald, and on 3DNA programs written by KJ Lu.

7 References

The key literature references describing the algorithm are:

1 Espinosa Angarica,V., González Pérez,A., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics, 9:436. http://www.biomedcentral.com/1471-2105/9/436

2 Contreras-Moreira,B. (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucleic Acids Research, 38: D91-D97. http://nar.oxfordjournals.org/cgi/content/abstract/38/suppl_1/D91

The original references describing HBPLUS and X3DNA are:

3 McDonald IK and Thornton JM (1994) Satisfying Hydrogen Bonding Potential in Proteins. Journal of Molecular Biology 238:777-793.
http://www.ncbi.nlm.nih.gov/pubmed/8182748

4 Lu XJ and Olson WK (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Research,31:5108-5121.
http://nar.oxfordjournals.org/cgi/content/short/31/17/5108