Table of Contents:

1. The Purpose of the Scripts

2. Contact Information

3. List and Description of Scripts

4. Installation and Setup

5. Usage

6. Changelog

1. The Purpose of the Scripts

Computational redesign of native or synthetic promoters for altered nucleosome affinity.

2. Contact Information

Hal Alper

The University of Texas at Austin

200 E. Dean Keeton Street, Stop C0400

Austin, TX 78712-1589

CPE 5.408 / phone: (512) 471-4417 / fax: (512) 471-7060

E-mail: halper@che.utexas.edu

3. List and Description of Scripts

affinity.m

Takes a DNA sequence and computes nucleosome affinity values for each nucleotide.

containsforbidden.m

This script looks for instances of user-defined DNA motifs in a DNA sequence. Motifs can include degenerate bases.

gccontent.m

Computes the GC content of a sequence

gcprofile.m

Calculates the GC contents of each 100bp sliding window of input DNA sequence.

maxprom.m

This program will take a promoter and iteratively decrease predicted nucleosome occupancy in user-defined basepair increments (rounds) until the occupancy can no longer be decreased.

nucleomin.m

Nucleomin takes a sequence as input and searches all n-nucleotide variants of the starting sequence to find the candidate with the minimum predicted nucleosome affinity, with the requirement that the sequence is also synthesizable and contains no additional or fewer transcription factor binding sites. n is user-defined.

problemrank.m

Notes the positions of the input DNA sequence which contain particular DNA motifs and ranks them from lowest nucleotide to highest nucleotide.

randprom.m

Initializes a random DNA sequence for a synthetic promoter and generates a list of sequences within the promoter which must be conserved during the design process based on user specifications.

randseq.m

Makes a random DNA sequence of the length and GC content specified

remforbidden.m

Tries to remove as many matches to a set of DNA motifs as possible from an input sequence. Users may also specify the locations of bases which may not be changed during this process.

seqarea.m

Computes the cumulative affinity score for a DNA sequence.

seqcheck.m

The sole purpose of this program is to make sure that a sequence can be synthesized by IDT's gblocks. It was sufficient at the time of writing but some features of it may no longer be necessary as synthesis technology improves.

synthprom.m

This function takes a general outline for a promoter and makes a synthetic nucleosome optimized promoter.

4. Installation and Setup

Setup instructions provided for Windows systems.

1) Obtain a copy of MATLAB (tested on r2011b) with the bioinformatics toolbox (tested on r2013b) installed

2) Copy the scripts listed above into the MATLAB working directory

3) Download a copy of the FORTRAN code for NuPoP (as of 6/28/2013 it was located at http://nucleosome.stats.northwestern.edu/ as "NuPoP\_F")

insert Download PERL and MinGW for Gfortran compiler

4) Edit the FORTRAN code for NuPoP\_F as follows:

Replace the following in npred.f90:

REPLACE:

implicit none

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

integer i,lfn,mlL,rep,species,order; character\*80 fileName; character\*3 tpc

real\*8 freqL1(4),tranL1(4,4),tranL2(16,4),tranL3(64,4),tranL4(256,4),Pd(500,11)

real\*8 freqN4(64,4),tranN4((147-4)\*256,4),freqN1(147,4),tranN1(584,4)

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

open(1,file='yourpath/NuPoP\_F/profile/freqL.txt')

read(1,\*) freqL1; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL.txt')

do i=1,4; read(1,\*) tranL1(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL2.txt')

do i=1,16; read(1,\*) tranL2(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL3.txt')

do i=1,64; read(1,\*) tranL3(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL4.txt')

do i=1,256; read(1,\*) tranL4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/147freqN.txt')

do i=1,147; read(1,\*) freqN1(i,:); end do

close(1)

open(1,file='yourpath/NuPoP\_F/profile/147tranN.txt')

do i=1,584; read(1,\*) tranN1(i,:); end do

close(1)

open(1,file='yourpath/NuPoP\_F/profile/146-149freqN4.txt')

do i=1,64; read(1,\*) freqN4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/146-149tranN4.txt')

do i=1,(147-4)\*256; read(1,\*) tranN4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/Pd.txt')

do i=1,500; read(1,\*) Pd(i,1:11); end do; close(1)

write(\*,'(a)') 'Please input'

write(\*,'(a)',advance='no') ' File name of DNA sequence (FASTA) : '; read\*,fileName

mlL=500

write(\*,'(a)',advance='no') ' Order of Markov model (1 or 4) : '; read\*,order

if(order/=1.and.order/=4) then; print\*,'1 or 4 should be inputed! stop.'; stop; end if

rep=1

print\*,' '

write(\*,'(a)') 'Select the species from the following list:'

print\*,'1=Human 2=Mouse 3=Rat'

print\*,'4=Zebrafish 5=D. melanogaster 6=C. elegans'

print\*,'7=S. cerevisiae 8=C. albicans 9=S. pombe'

print\*,'10=A. thaliana 11=Maize 0=Other'

print\*,' '

write(\*,'(a)',advance='no') 'Input the lable of selected species : '; read\*,species

print\*,' '

write(\*,'(a)') 'Predicting......'

lfn=len\_trim(fileName)

if(order==1) then

call vtbfb(lfn,trim(fileName),freqL1,tranL1,freqN1,tranN1,mlL,rep,species,Pd)

else if(order==4) then

call vtbfbNL4(lfn,trim(fileName),freqL1,tranL1,tranL2,tranL3,tranL4,freqN4,tranN4,mlL,rep,species,Pd)

end if

write(\*,'(a)') ' Done.'

end

WITH:

implicit none

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

integer i,lfn,mlL,rep,species,order; character\*80 fileName,stringorder,stringspecies; character\*3 tpc

real\*8 freqL1(4),tranL1(4,4),tranL2(16,4),tranL3(64,4),tranL4(256,4),Pd(500,11)

real\*8 freqN4(64,4),tranN4((147-4)\*256,4),freqN1(147,4),tranN1(584,4)

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

open(1,file='yourpath/NuPoP\_F/profile/freqL.txt')

read(1,\*) freqL1; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL.txt')

do i=1,4; read(1,\*) tranL1(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL2.txt')

do i=1,16; read(1,\*) tranL2(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL3.txt')

do i=1,64; read(1,\*) tranL3(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/tranL4.txt')

do i=1,256; read(1,\*) tranL4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/147freqN.txt')

do i=1,147; read(1,\*) freqN1(i,:); end do

close(1)

open(1,file='yourpath/NuPoP\_F/profile/147tranN.txt')

do i=1,584; read(1,\*) tranN1(i,:); end do

close(1)

open(1,file='yourpath/NuPoP\_F/profile/146-149freqN4.txt')

do i=1,64; read(1,\*) freqN4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/146-149tranN4.txt')

do i=1,(147-4)\*256; read(1,\*) tranN4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP\_F/profile/Pd.txt')

do i=1,500; read(1,\*) Pd(i,1:11); end do; close(1)

CALL GETARG(1,fileName)

CALL GETARG(2,stringorder)

CALL GETARG(3,stringspecies)

read(stringorder,\*) order

read(stringspecies,\*) species

mlL=500

if(order/=1.and.order/=4) then; print\*,'1 or 4 should be inputed! stop.'; stop; end if

rep=1

lfn=len\_trim(fileName)

if(order==1) then

call vtbfb(lfn,trim(fileName),freqL1,tranL1,freqN1,tranN1,mlL,rep,species,Pd)

else if(order==4) then

call vtbfbNL4(lfn,trim(fileName),freqL1,tranL1,tranL2,tranL3,tranL4,freqN4,tranN4,mlL,rep,species,Pd)

end if

end

5) Replace the string "yourpath" with the directory in which NuPoP\_F is located (Use the full path, e.g. "C:/Users/..."

6) Rename the file to "Npred2.f90" and compile Npred2.f90 as Npred2.exe using the instructions provided in the manual included with NuPoP\_F. Replace the INSTALL file provided by NuPop with the included INSTALL file. See the NuPoP\_F manual for more detailed information as to the installation of NuPoP.

7) Add the directory containing Npred2.exe to your system's path. This can be found by clicking Start, right-clicking "Computer", selecting "Properties", selecting "Advanced System Settings", and on the "Advanced" tab, clicking "Environment Variables", and editing the "Path" variable in the "System Variables" category by placing a semicolon after the last directory and adding the one you want to add.

You're now ready to begin designing promoters!

5. Usage

1) Pick a promoter. Promoters must be designed including 200bp upstream and 100bp downstream of its genomic or plasmid context. This will ensure that the nucleosome affinity values calculated for promoter variants will be comparable to one another. Note the nucleotide positions of the start and the end of the promoter.

2) Annotate the transcription factor binding sites and note the nucleotides covered by the binding sites. A particularly user-friendly repository is the Yeast Promoter Atlas http://ypa.ee.ncku.edu.tw/

3) Annotate any sequences you would not want introduced into the designed promoter. These sequences, if present in the wild-type promoter, will not be altered.

4) Build input files. For the TEF promoter, we enter the DNA sequence of the promoter itself plus 200bp upstream and 100bp downstream as follows:

TEF='GGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTACCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCCTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAAAAGCTGGAGCTCATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGCGCATCGCCGTACCACTTCAAAACACCCAAGCACAGCATACTAAATTTCCCCTCTTTCTTCCTCTAGGGTGTCGTTAATTACCCGTACTAAAGGTTTGGAAAAGAAAAAAGAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAGGCAATAAAAATTTTTATCACGTTTCTTTTTCTTGAAAATTTTTTTTTTGATTTTTTTCTCTTTCGATGACCTCCCATTGATATTTAAGTTAATAAACGGTCTTCAATTTCTCAAGTTTCAGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAGAAAGAAAGCATAGCAATCTAATCTAAGTTTTCTAGAACTAGTATGTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATGGTGATGTTAATGGTCACAAATTTTCTGTCT';

For its nucleotides covered by transcription factor binding sites, we enter:

TEFforbidden=[281:291 334:343 377:383 443:484];

For the sequences which will not be introduced or removed from the designed promoter, make a cell array containing the relevant motifs. We used the TF consensus list found at yeastract.com, in addition to the start codon and TATA box for our studies.

These input files are included in Sample Data.mat

Example MATLAB Commands:

Optimize TEF in 1bp steps:

[TEFproms,TEFareas,TEFcurves]=maxprom(TEF,TEFstart,TEFend,1,TEFforbidden,forbiddenseqs);

Design Psynth1 in 1bp steps:

[psynth1proms,psynth1areas,psynth1curves]=synthprom(psynth1params,psynth1start,psynth1end,1,forbiddenseqs);

For each command, the first output is a list of nucleosome-optimized promoters, starting from the wild-type (or seed) sequence, and proceeding in 1bp steps toward a variant with reduced predicted nucleosome affinity. The second output is the corresponding cumulative affinity score for each promoter, and the third output is the nucleosome affinity curves used to compute the cumulative affinity score for each promoter.

As the programs are running, they will periodically display a progress indicator which describes how far along the program is in computing the current mutation.

6. Changelog

4/18/2014: Updated Install info. Thanks to Joseph Cheng