MCB 372

PSI BLAST, scalars

J. Peter Gogarten
Office: *BPB 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

Assignment from Wednesday

1) Read through the Perl scripts extract_lines.pl and extract_lines_mod.pl
2) Why does the first of these get along without `chomp ($line);`
   DISCUSS
3) Write a short Perl script that calculates the circumference of a circle given a radius provided by the user (see exercises 1-4 chapter 2 in Learning Perl). (One set of answers is given in Appendix A of the book)
   GO OVER EXAMPLES

From Lab exercises:

Which option turns off the low complexity filter? **-F F**
Which option, and which setting, sets the word size to 2? **–W 2**
Which option allows to use two processors? **–a 2**

Exercises from Wednesday:

```
#!/usr/bin/perl #-w
my $i='';
print "\$i= $i\n";
$i = 1;
print "\$i= $i\n";
$i++;
print "\$i= $i\n";
$i *= $i;
print "\$i= $i\n";
$i .= $i;
print "\$i= $i\n";
$i = $i/11;
print "\$i= $i\n";
$i = $i . "score and" . $i+3 ;
print "\$i= $i\n";
$i = $i+3 . "score and" . $i;
print "\$i= $i\n";
```

$i=
$i= 1
$i= 2
$i= 4
$i= 44
$i= 4
$i= 7
$i= 10score and7

Exercises from Wednesday:

```
$a=1;
$b=2;
$c = $a +$b;
print "\$c= $c\n";
$c = $a / $b;
print "\$c= $c\n";
$c = "$a + $b";
print "\$c= $c\n";
$c = '$a + $b';
print "\$c= $c\n";
$c = $a + $b++; # better use parenthesis $b is 3 at the end of this line
print "\$c= $c\n";
$c += $a ; #add the value of $a to $c and stores the ressult in $c
print "\$c= $c\n";
```

$c= 3
$c= 0.5
$c= 1 + 2
$c= $a + $b
$c= 3
$c= 4

Exercises from Wednesday:

```
#!/usr/bin/perl -w
print "\n\n";
@myArray = ('A', 'B', 'C', 'D', 'E');
print $#myArray; # returns highest number of field in array
print "\n";
print length($myArray[0]); # returns lenght of scalar - no idea what it does with an arr
print "\n";
print $myArray[1]; #returns value in slot 1 (the 2nd entry - perl starts a 0)
print "\n";
print $n=@myArray; #one way to get the number of elements in an array
print "\n";
print reverse (@myArray); #comes in handy for DNA sequences.
print "\n";

~
~
```

4
1
B
5
EDCBA

Psi-Blast: Detecting structural homologs

Psi-Blast was designed to detect homology for highly divergent amino acid sequences

Psi = position-specific iterated

Psi-Blast is a good technique to find "potential candidate" genes

Example: Search for Olfactory Receptor genes in Mosquito genome
Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ (2002) G protein-coupled receptors in Anopheles gambiae. Science 298:176-8

by Bob Friedman

Psi-Blast Model

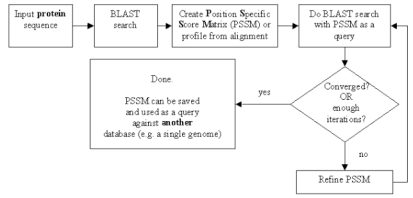
Model of Psi-Blast:
1. Use results of gapped BlastP query to construct a multiple sequence alignment
2. Construct a position-specific scoring matrix from the alignment
3. Search database with alignment instead of query sequence
4. Add matches to alignment and repeat

Similar to Blast, the E-value in Psi-Blast is important in establishing matches
E-value defaults to 0.001 & Blosom62

Psi-Blast can use existing multiple alignment - particularly powerful when the gene functions are known (prior knowledge) or use RPS-Blast database

by Bob Friedman

PSI BLAST scheme



© Olga Zhaxybayeva

Position-specific Matrix

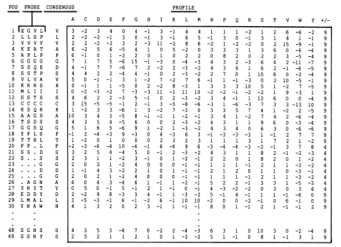
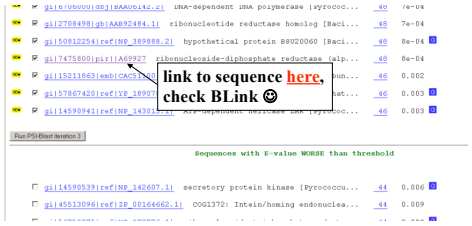


Fig. 1. The concept of a profile. (a) A flow diagram of profile analysis. (b) A 49-residue sample profile for the immunoglobulin variable region domain, generated from the four-probe sequences shown at the left (see Fig. 2b for details). The profile is shown in the box. The rightmost column of the profile gives the penalty for insertion/deletion (+/−). Positions 31–47 of the profile are omitted from the figure for clarity. Notice that where gaps appear in some of the probe sequences, the insertion/deletion penalty is lower than elsewhere.

M Gribskov, A D McLachlan, and D Eisenberg (1987) Profile analysis: detection of distantly related proteins. PNAS 84:4355-8.

by Bob Friedman

Psi-Blast Results Query: 55670331 (intein)



PSI BLAST and E-values!

Psi-Blast is for finding matches among divergent sequences (position-specific information)
WARNING: For the nth iteration of a PSI BLAST search, the E-value gives the number of matches to the profile NOT to the initial query sequence! The danger is that the profile was corrupted in an earlier iteration.

PSI Blast from the command line

Often you want to run a PSIBLAST search with two different databanks - one to create the PSSM, the other to get sequences:
To create the PSSM:

blastpgp -d nr -i subI -j 5 -C subI.ckp -a 2 -o subI.out -h 0.00001 -F f

blastpgp -d swissprot -i gamma -j 5 -C gamma.ckp -a 2 -o gamma.out -h 0.00001 -F f

Runs a 4 iterations of a PSIblast
the -h option tells the program to use matches with E <10^-5 for the next iteration, (the default is 10^{-3})
-C creates a checkpoint (called subI.ckp),
-o writes the output to subI.out,
-i option specifies input as using subI as input (a fasta formated aa sequence).
The nr databank used is stored `in /common/data/`
-a 2 use two processors
-h e-value threshold for inclusion in multipass model [Real]
default = 0.002 THIS IS A RATHER HIGH NUMBER!!!

(It might help to use the node with more memory (017)
(command is `ssh node017`)

To use the PSSM:

```
blastpgp -d /Users/jpgogarten/genomes/msb8.faa -i subI -a 2 -R
subI.ckp -o subI.out3 -F f

blastpgp -d /Users/jpgogarten/genomes/msb8.faa -i gamma -a 2 -R
gamma.ckp -o gamma.out3 -F f
```

Runs another iteration of the same blast search, but uses the databank `/Users/jpgogarten/genomes/msb8.faa`

-R tells the program where to resume
-d specifies a different databank
-i input file - same sequence as before
-o output_filename
-a 2 use two processors
-h e-value threshold for inclusion in multipass model [Real]
default = 0.002. This is a rather high number, but might be ok for the last iteration.

More on blastall:

available at safari books online
http://proquestcombo.safaribooksonline.com/

Installation instructions and info on parameters at the NCBI:
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/
ftp://ftp.ncbi.nlm.nih.gov/blast/documents/formatdb.html
ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast.html
ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastpgp.html
ftp://ftp.ncbi.nlm.nih.gov/blast/documents/fastacmd.html
ftp://ftp.ncbi.nlm.nih.gov/blast/documents/

http://www.bioinformatics.ubc.ca/resources/tools/blastall

http://en.wikipedia.org/wiki/BLAST

PSI Blast and finding gene families within genomes

PSSMs can be useful to find gene family members in a genome.
1st step: Get PSSM

A) do PSI blast search with one or several seed sequences using nr as target database

```
blastpgp -d nr -i query.name -j 5 -C query.ckp -a 2 -o query.out
    -h 0.00001 -F f
```

A) Use CDD. Problem is that the PSSMs are not easily obtained. You can download the CDD PSSMs from the NCBI's FTP server, but these are not in the correct checkpoint format to act as seeds for a databank search. According to Eric Sayers from the NCBI help desk:

Yes, indeed. The problem is that we produce two "flavors" of scoremats: one with intermediate data (frequencies) and one with final data (integer scores). Blastpgp can only use the intermediate data scoremats, and unfortunately the scoremats on the ftp side are final data scoremats. We are in the process of trying to make this easier, perhaps by placing the intermediate scoremats on the ftp site as well. In the meantime, you can use Cn3D 4.2 to convert the final data scoremat into an intermediate one as follows:

1) download Cn3D 4.2 from the CD-Tree release (http://www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml)
2) Load the cd of interest into Cn3D 4.2 (find the cd on the web and click structure view to view it in cn3d 4.2
3) In the sequence window of cn3d 4.2, choose View/Export/PSSM – this will produce an intermediate scoremat

Note: Cn3D 4.2 only runs under windows …. ^%*&^^$%$

PSI Blast and finding gene families within genomes

2nd step: use PSSM to search genome:

A) Use protein sequences encoded in genome as target:

```
blastpgp -d target_genome.faa -i query.name -a 2 -R query.ckp -o
    query.out3 -F f
```

B) Use nucleotide sequence and tblastn. This is an advantage if you are also interested in pseudogenes, and/or if you don't trust the genome annotation:

```
blastall -i query.name -d target_genome_nucl.ffn -p psitblastn -R
    query.ckp
```

Assignment for Wednesday

1) Review PSIblast
2) Write a 3 sentence outline for your student project
3) Re-read chapter 2 p32 - p34 on control structures and page 142 -146 on for, foreach, and while loops

For next week:

```
Background: @a=(0..50);
```

#assigns numbers from 0 to 50 to an array, so that $a[0] =0; $a[1] =1; $a[50] =50

4) Write perlscripts that add all numbers from 1 to 50. Try to do this using at least to different control structures.