Sequence Analysis Spring 2000 Dr. Richard Friedman
(212)305-6901 (76901) friedman@cuccfa.ccc.columbia.edu 130BB

Lesson 9 Profile Methods

Assignment:

1. Read GCG Program manual:

A. Profile Analysis
B. ProfileScan
C. ProfileGap
D. ProfileMake
E. ProfileSearch
F. ProfileSegments
G. Meme
H. MotifSearch

2. The following file on cuccfa:

/usr2/seq/doc/blast2.doc

Theory:

A profile is a matrix in which a row corresponds to a position in a multiple sequence alignment and a column corresponds to a type of residue (amino acid or nucleotide). An element of the profile matrix gives the weight associated with the residue in question. A more positive weight indicates that the residue is likely to occur in the position. A less positive weight indicates that the residue is unlikely to occur in the position.

```
{
GAP_CREATE 12
GAP_EXTEND 4
}
```

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	2	0
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-3
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-2	1
L	-1	-4	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-3
M	-1	-3	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-2
N	-2	1	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

Table 1 A sample scoring matrix

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len
W	-2	-2	-3	-3	-3	4	-3	0	-1	0	1	-1	-1	-2	-1	4	1	-2	-2	4	3	-2	21	21
F	0	0	2	0	-1	3	1	0	0	-2	0	0	0	-2	-1	-2	0	0	0	1	3	-1	21	21
N	3	5	-1	3	3	-4	3	2	-1	3	-2	-1	6	1	2	1	2	2	-1	-2	-2	2	21	21
Q	4	7	-1	8	8	-6	4	11	-3	4	-3	-2	7	4	11	3	1	2	-1	-6	-2	9	21	21
L	1	-1	1	-5	-4	8	-2	0	7	-3	9	6	1	-1	-4	-3	0	1	6	1	7	-4	21	21
V	4	4	-6	7	5	-2	3	2	8	1	6	6	2	4	6	-2	0	4	9	-11	-5	6	21	21
Y	4	1	4	-1	0	9	3	2	3	-5	5	0	3	-2	-4	-6	5	2	2	3	10	-2	100	100
N	6	14	-4	11	9	-6	11	5	0	6	-3	-1	15	0	5	2	7	4	0	-4	-3	7	100	100
D	4	12	-6	14	13	-9	8	7	3	5	-2	0	9	4	8	4	5	4	3	-9	-8	11	100	100
S	14	10	3	9	8	-16	16	0	-4	13	-10	-2	8	13	5	8	17	15	1	-7	-17	7	100	100

Table2 A Sample Profile

How profile matrices are derived can be understood in terms of the following simple example:

	Position:			
Sequence	1	2	3	Weight
1	G	T	A	0.15
2	A	A	A	0.15
3	T	C	C	0.15
4	G	G	G	0.15
5	C	C	T	0.4

A. A profile based on numbers of residues.

Position	A	T	G	C
1	1	1	2	1
2	1	1	1	2
3	2	1	1	1

A= profile matrix.
i= position in multiple sequence alignment.
r= letter of the alphabet representing a kind of residue.
$A_{i,r}$ = element of the profile matrix corresponding to the ith position and the rth residue.
$N_{i,r}$ = Number of residues of kind r in position i.

$$A_{i,r} = N_{i,r}$$

The profile matrix entry is equal to the number of residues in a given position that are of a given kind.

B. A profile based on fractions of residues:

Position	A	T	G	C
1	0.20	0.20	0.40	0.20
2	0.20	0.20	0.20	0.40
3	0.40	0.20	0.20	0.20

N= The total number of sequences.

$f_{i,r} = \frac{N_{i,r}}{N}$ = the fraction of each kind of residue in each position.

$A_{i,r} = f_{i,r}$

The profile matrix entry is equal to the fraction of residues in a given position that are of a given kind.

C. A profile based of weighted fractions of residues:

Position	A	T	G	C
1	0.15	0.15	0.3	0.4
2	0.15	0.15	0.15	0.55
3	0.3	0.4	0.15	0.15

j = the jth sequence in the alignment.
w_j = the weight of the jth sequence in the alignment.
$N_{i,r}^{j}$ = the number of residues of kind r in position i in sequence j. $N_{i,r}^{j} = 0$ or 1.

$$A_{i,r} = \frac{\sum_{j=1,N} w_j N_{i,r}^{j}}{\sum_{j=1,N} w_j}$$

The profile matrix entry is equal to the weighted fraction of residues in a given position that are of a given kind.

D. A profile based of the log weighted fractions of residues:

Position	A	T	G	C
1	-1.90	-1.90	-1.20	-0.92
2	-1.90	-1.90	-1.90	-0.60
3	-1.20	-0.92	-1.90	-1.90

$$A_{i,r} = \ln\left(\frac{\sum_{j=1,N} w_j N_{i,r}^{j}}{\sum_{j=1,N} w_j}\right)$$

The profile matrix entry is equal to the natural logarithm of the weighted fraction of residues in a given position that are of a given kind.

E. A profile that takes relations between residues into account:

The above method fails for residues that do not occur at a particular position. The solution to this problem is to sum over all residues as weighted by their similarity matrix entry.

$$A_{i,r} = \sum_{r'=1,k} S_{r,r'} \ln\left(\frac{\sum_{j=1,N} w_j N^j_{i,r'}}{\sum_{j=1,N} w_j}\right)$$

$S_{r,r'}$ = The similarity matrix element between residue r and residue r'.

k = number of different kinds of residues.
for nucleic acids k=4.
for proteins k=20.

Please don't ask me to give a detailed worked example.

Individual profile-generating programs may use slightly different equations than the ones given above. However the above equations illustrate the basic principles of deriving profile matrices.

Summary of Commands:

Note: In this document different fonts have different meanings:

Times is used to explain commands.

`Courier is used to indicate commands and command options.`

`Courier italics are used to indicate command parameters, for example, filenames.`

`Courier bold is used to indicate commands that are not displayed.`

`Courier bold italics are used to indicate computer-generated output.`

Helvetica is used to indicate menu items.

`profilescan` Compares a sequence to a library of profiles at high stringency. Finds the profiles that best fit the sequence.

`profilescan -inter` Compares a sequence to a library of profiles at low stringency. Finds the profiles that best fit the sequence.

`to` *`profiledir`* Goes to the directory containing the Prosite profile library.

`profilegap` Compares a sequence and a profile, producing a sequence-profile alignment.

`profilemake` Makes a profile from a multiple sequence alignment. The aligned sequences must be in a GCG *.msf file. The profile is in a *.prf file.

Specify the sequences by typing “name.msf{*}”.

`profilesearch` Compares a profile to a sequence database. Finds the sequences that best fit the profile. The list of sequences are placed in a *.pfs file.

`profilesearch -minl=`*`cutoff`* Compares a profile to a sequence database. Finds the sequences that best fit the profile. The list of sequences with a minimum Z-score of *`cutoff`* are placed in a *.pfs file.

`profilesegments` Aligns a profile to the sequences found by profilesearch. Uses a *.pfs as an input.

`blastall -p` *`blastprogramname`* `-i` *`protein.tfa`* `-d` *`databasename`* `-h` *`cutoff`* `-o` *`outputfile`*

Runs a gapped BLAST search on cuccfa. This program can be used to search files that are updated weekly, such as the NCBI nonredundant protein database.

`blastprogramname` kind of blast program being run.

BLAST Programs:

`blastp` Protein query sequence compared to protein database.

`blastn` Nucleic acid query sequence compared to nucleic acid database.

blastx Nucleic acid query sequence translated in all six frames and compared to protein database.

tblastn Protein query sequence compared to nucleic acid query sequence translated in all six frames.

tblastx Nucleic acid query sequence translated in all six frames compared to nucleic acid database translated in all six frames.

protein.tfa Name of protein file in Fasta format.

databasename Name of database.

cutoff Probability cutoff above which sequences are not reported.

outputfile Name of outputfile.

datalist Lists the Blast2-searchable databases installed on cuccfa.

chmod +x *doblastall* Makes the doblastall command file executable.

blastpgp -i *protein.tfa* -d *databasename* -h *matrixcutoff* -j *#iterations* -o *outputfile*

Runs a PSIBLAST search on cuccfa. Protein-protein searches only. parameters have the same meaning as for Blastall except:

marixcutoff Probability cutoff above which sequences are not included in the matrix.

#iterations Number of iterations.

Example:

blastpgp -i *src_human.fd.tfa* -d *swhuman* -h *.0005* -j *10* -o *src_human.fd.j10.swhuman.psiblast*

Runs a PSIBLAST search with *src_human.fd.tfa* as a query sequence of the database *swhuman* with a cutoff of *.0005* for inlcusion in the matrix for a maximum of 10 iterations with the output going into a file named *src_human.fd.j10.swhuman.psiblast*.

blastpgp -B *matrix.phy* -i *protein.tfa* -d *databasename* -h *matrixcutoff* -j *#iterations* -o *outputfile*

Runs a special version of PSIBLAST that takes a multiple sequence alignment as input for the first iteration. The multiple sequence alignment should be in the same format as the file */usr2/seq/seqclass/lab9/5rad9.ul.phy* . Residues from which a profile is made (resides similar to the consensus) should be in upper case. Other

residues should be in lower case. The sequence in file *protein.tfa* must also occur in the mutiple sequence alignment. If the multiple sequence alignment contains a truncated form of the protein, the single protein file must contain the same form.

Example:
`blastpgp -B 5rad9.ul.phy -i human.trunc.tfa -d nr -e .0005 -j 2 -o human.trunc.5rad9.ul.0005.j2.nr.psiblast`

Runs a PSIBLAST search with `human.trunc.tfa` as a query sequence and `5rad9.ul.phy` as the query alignment of the nonredundant database with a cutoff of `.0005` for maximum of `2` iterations with the output going into a file named `human.trunc.5rad9.ul.0005.j2.nr.psiblast` .

`impala -i src_human.fd.tfa -P /usr2/seq/wolf1187/wolf1187 -o src_human.fd.wolf1187.imp`

Compares the sequence `src_human.tfa` to a library of PSIBLAST profiles, in this case one that contains a matrix for every fold in the SCOP (Structural Classification Of Proteins).database and outputs the file to `src_human.fd.wolf1187.imp`.

CONVERGED Appears in PSIBLAST output file to indicate that no new database sequences have been added to the profile matrix, so that the run has converged.

`meme` Generates a number of short nonoverlapping profiles from a series of unaligned sequences in a GCG list file. Outputs a *.prf file.

`motifsearch` Uses the short profiles generated by meme (in a *prf file) to search a protein database.

http://pfam.wustl.edu/ PFAM database (effectively another profile database).

http://www.blocks.fhcrc.org/ Search a sequence against a database of short sequence alignment profiles.

Note: In PSIBLAST searches, I recommend a cutoff of E=.0005, both for inclusion of a sequence in the matrix and for identifying homologs. It is sometimes useful to check sequences with E ≤ .0005 at each step for low complexity regions with Seg. The reliablity of the statistical estimates generated by Motifsearch have not yet been established to the best of my knowledge.

Lab (or homework):

Make a directory entitled `lab9` and go to it.

1. Identifying a sequence

   A. Copy the related sequences, `unknown1.pep` and `unknown2.pep` from `/usr2/seq/seqclass/lab9`.

B. Search the GCG default database for a profile corresponding to *unknown1.pep.* What kind of protein is it?

C. Repeat the same search with a less stringent cutoff.

D. Align *unknown2.pep* to the profile that fits *unknown1.pep* best.

2. Making a profile from a multiple sequence alignment and using it to identify members of a family.

A. Make a profile from the best sequence alignment **of SH2 domains only** from the previous lab. If you don't have such an alignment from the previous lesson I have placed one in called *sh2.best.msf* in */usr2/seq/seqclass/lab9* .

B. Search the Swissprot database for proteins from humans that contain SH2 domains. Hint: Your sequence specification should be `sw:*human.`

C. Determine a cutoff for your SH2 profile. Use the contents of the file */usr2/seq/seqclass/lab10/sh2gibson.msf* as a reference.

Hint: A good way to check proteins whose names do not correspond to proteins in *sh2gibson.msf* is to use the command

`typedata` *`sw:proteiname`* `| grep SH2`

What is the cutoff?

How many human proteins containing SH2 domains are in the Swissprot database, according to your profile?

D. Align your SH2 profile to each of the human sequences detected by it.

3. Searches done with Ssearch3 and PSIBLAST.

A. Obtain the file sw:src_human.

B. Prepare a gcg formatted file consisting of the protein's SH2 domain.

C. Filter the sequence of the file prepared in part C for low complexity regions.

D. Perform a full Smith-Waterman search of a database consisting of the human proteins in Swiss-Prot with the sequence prepared in part C as a query sequence.

E. Perform a PSIBLAST search with the same query sequence and database (swhuman) with a stringency of .001 until convergence.

F. Compare the sensitivity of Profilesearch, Ssearch3, and PSIBLAST.

4. Search done with the version of PSIBLAST that takes an alignment as input.

The alignment of the 5 known rad9 proteins is: 5rad9.ul.phy. The human protein with its sequence shortened to the region included in the multiple alignment is human.trunc.tfa . Based on this sequence and alignment, try to determine the biochemical function of RAD9.

5. Search done with meme and motifscan:

Use Meme and Motifscan to attempt to determine the biochemical function of the RAD9 proteins.

6. Use impala with the Wolff1187 database, Blocks, and PFAM to identify the domains in src_human.fd.tfa.

Web sites:

http://www.ncbi.nlm.nih.gov/
PSIBLAST

http://ulrec3.unil.ch/software/profilescan.html
Prosite Profiles.

http://www.sdsc.edu/MEME/meme/website/
Meme and a Motifsearch-like program called Mast

http://pfam.wustl.edu/
PFAM Database

BLOCKS Database (a database of short profiles)
http://www.blocks.fhcrc.org/

Bibliography:

1. Gribskov, M. McLachlan, A.D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins, PNAS, 84, 43355-4358.
2. Gribskov, M., Luethy, R., Eisenberg D. (1989) Profile analysis. in Methods in enzymology, 183, 146-159, Academic Press, San Diego, California, USA.
3. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J, (1997) Nucleic Acids Res. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nuc. Acid. Res. 25,3389-3402.
4. Biological Sequence Analysis, R. Durbin, S. Eddy, A. Krough, and G. Mitchison, Cambridge, 1998.,Chaps. 1-5.
5. http://blocks.fhcrc.org/blocks/help/about_blocks.html
6. J.G. Henikoff, S. Henikoff & S. Pietrokovski, New features of the Blocks Database servers, Nucl. Acids Res. 27:226-228 (1999).
7. S. Henikoff, J.G. Henikoff and S. Pietrokovski, Blocks+: A non-redundant database of protein alignment blocks dervied from multiple compilations, Bioinformatics 15(6):471- 479 (1999).
8. S. Henikoff and J.G. Henikoff Protein family classification based on searching a database of blocks, Genomics 19:97-107 (1994).
9, Schaffer, A.A., Wolf, Y. I., Ponting, C.P., Koonin, E.V., Aravind, L., and Altschul, S.F. (2000) IMPALA: Matching a Protein Sequence Against a Collection of PSI-BLAST-Constructed Position-Specific Score Matrices, Bioinformatics, in press.

Answer key to lab 8

(I thank Dr. Murad Nayal for checking the exercises).

Clustalw aligned the full protein sequences better than Pileup. By the criterion of producing the most columns with a consensus sequence the order was msa > Pileup > Clustalw, however this might not actually correspond to structure. Comparison to structural tests has shown that msa generally gives the best alignment.