PSI (position-specific iterated) **BLAST**

The NCBI page described PSI blast as follows:

"*Position-Specific Iterated BLAST (PSI-BLAST) provides an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues.*

The program first performs a gapped BLAST database search. The PSI-BLAST program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching.

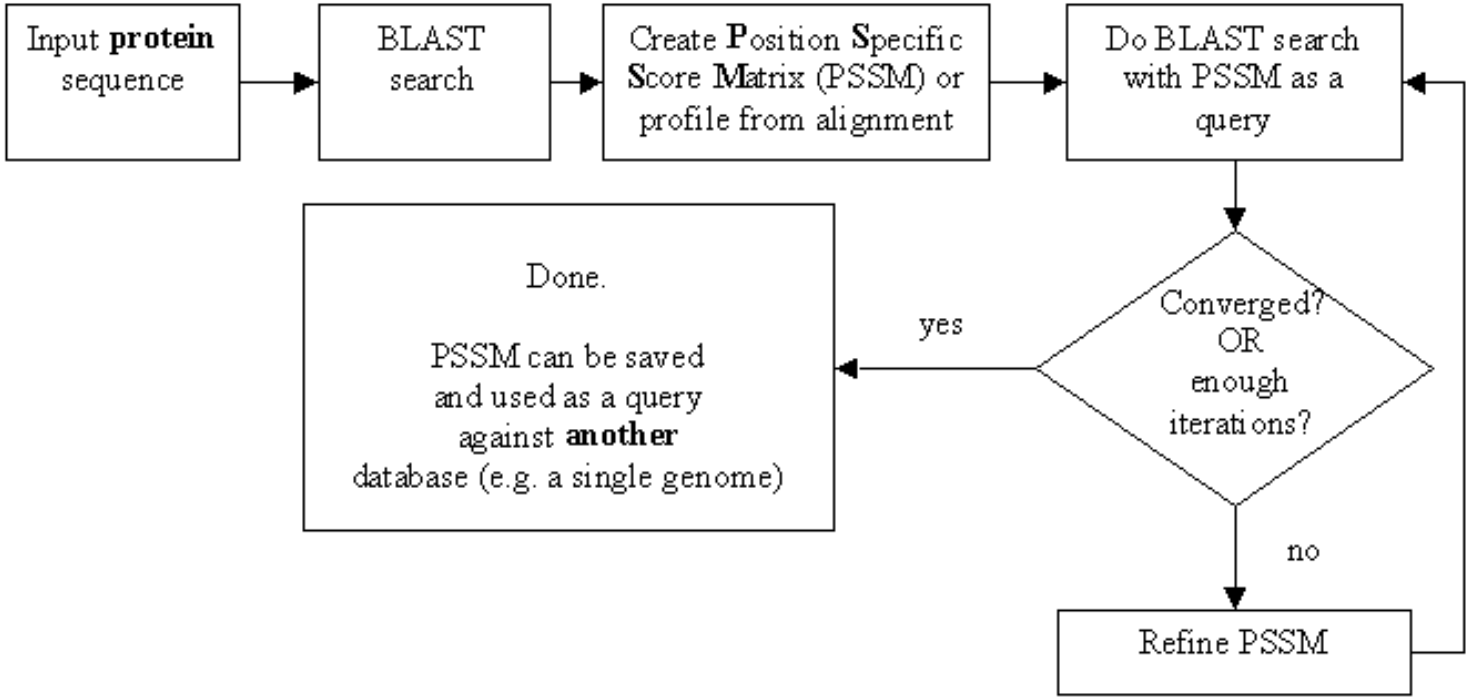
PSI-BLAST may be iterated until no new significant alignments are found. At this time PSI-BLAST may be used only for comparing protein queries with protein databases."

The Psi-Blast Approach

1. Use results of 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

Psi-Blast can use existing multiple alignment, or
use RPS-Blast to search a database of PSSMs

PSI BLAST scheme



Position-specific Matrix

POS	PROBE	CONSENSUS	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
			PROFILE																				
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.												.										
.	.												.										
.	.												.										
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

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. 2*b* 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)

NEW	☑	gi|6706000|dbj|BAA06142.2|	DNA-dependent DNA polymerase [Pyrococ...	48	7e-04
NEW	☑	gi|2708498|gb|AAB92484.1|	ribonucleotide reductase homolog [Baci...	48	7e-04
NEW	☑	gi|50812254|ref|NP_389888.2|	hypothetical protein BSU20060 [Baci...	48	8e-04 G
NEW	☑	gi|7475800|pir||A69927	ribonucleoside-diphosphate reductase (alp...	48	8e-04
NEW	☑	gi|15211863|emb|CAC51100	bun...	46	0.002
NEW	☑	gi|57867420|ref|YP_18907	hat...	46	0.003 G
NEW	☑	gi|14590941|ref|NP_143015.1|	ATP-dependent helicase LHR [Pyrococ...	46	0.003 G

link to sequence here, check BLink ☺

Run PSI-Blast iteration 3

Sequences with E-value WORSE than threshold

☐	gi|14590539|ref|NP_142607.1|	secretory protein kinase [Pyrococcu...	44	0.006 G
☐	gi|45513096|ref|ZP_00164662.1|	COG1372: Intein/homing endonuclea...	44	0.009

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 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.

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
```

Psi-Blast finds homologs 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.