PSI-BLAST

Position-Specific Iterated BLAST

Stephen F. Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Protein and DNA “Profiles”

The idea behind DNA and protein “profiles” has a long history. These structures are also called “position-specific score matrices” or “PSSMs”, and they are closely related to “Hidden Markov Models” or “HMMs”.

McLachlan, A.D. (1977) “Analysis of periodic patterns in amino acid sequences: collagen.” *Biopolymers* **16**:1271-1297.

Stormo, G.D., *et al.* (1982) “Use of the ‘perceptron’ algorithm to distinguish translational sites in *E. coli*.” *Nucl. Acids Res.* **10**:2997-3011.

McLachlan, A.D. (1983) “Analysis of gene duplication repeats in the myosin rod.” *J. Mol. Biol.* **169**:15-30.

Staden, R. (1984) “Computer methods to locate signals in nucleic acid sequences.” *Nucl. Acids Res.* **12**:505-19.

Schneider, T.S., *et al.* (1986) “Information content of binding sites on nucleotide sequences.” *J. Mol. Biol.* **188**:415-431.

Taylor, W.R. (1986) “Identification of protein sequence homology by consensus template alignment.” *J. Mol. Biol.* **188**:233-258.

Berg, O.G. & von Hippel, P.H. (1987) “Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.” *J. Mol. Biol.* **193**:723-750.

Dodd, I.B. & Egan, J.B. (1987) “Systematic method for the detection of potential lambda cro-like DNA-binding regions in proteins.” *J. Mol. Biol.* **194**:557-564.

Gribskov, M., *et al.* (1987) “Profile analysis: detection of distantly related proteins.” *Proc. Natl. Acad. Sci. USA* **84**:4355-4358.

Patthy, L. (1987) “Detecting homology of distantly related proteins with consensus sequences.” *J. Mol. Biol.* **198**:567-577.

Structure of a Profile

A profile of length L is an $L \times 20$ (for proteins) or $L \times 4$ (for DNA) array $s_{i,j}$. An element $s_{i,j}$ of this array represents the score for aligning letter j at position i.

A profile can be aligned to an individual sequence in exactly the same way that a sequence can be, using the Needleman-Wunsch or Smith-Waterman algorithm.

The scores of a profile may be derived from a multiple sequence alignment in a variety of ways.

Steps in Profile Analysis

1) Select a set of related sequences, often by running a database search with a query sequence.

2) Construct a multiple sequence alignment of the sequences.

3) Derive a profile from the multiple sequence alignment.

4) Compare the profile to a database of sequences.

5) Iterate, by returning to step 1).

In the mid-1990s, this process could involve running as many as four separate programs, some of which were fairly slow (i.e. steps 2 and 4 above). It generally required a fair amount of expertise, and was not accessible to most biologists.

Can Profile Analysis be Automated?

Requirements:

A way to collect a set of sequences. Use the output of a BLAST search.

A way to define the length of a profile.

A *fast* way to construct a multiple alignment.

A way to derive profile scores from the multiple alignment.

A *fast* way to search the database with a profile. Generalize BLAST.

For iteration, a way to assess the significance of profile-sequence alignments.

PSI-BLAST strategy:

Keep approach simple at first; consider refinements later.

Altschul, S.F., *et al.* (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.* **25**:3389-3402.

Profile Length

Here is one alignment returned by a BLAST protein database search:

```
>sp|Q99728.2|BARD1_HUMAN
Length=777

 Score = 53.1 bits (126),  Expect = 3e-07, Method: Composition-based stats.
 Identities = 32/111 (29%), Positives = 55/111 (50%), Gaps = 15/111 (14%)

Query  24   THVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGR  83
            THVV+  DA    + TLK  LGI  G W++ + WV   ++ +    E  +E+
Sbjct  605  THVVVPGDA---VQSTLKCMLGILNGCWILKFEWVKACLRRKVCEQEEKYEIP-------  654

Query  84   NHQGPKRARESQDR---KIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVV  131
              +GP+R+R ++++   K+F G     +G F + P D L  +V   G  ++
Sbjct  655  --EGPRRSRLNREQLLPKLFDGCYFYLWGTFKHHPKDNLIKLVTAGGGQIL  703
```

Some alignments will cover essentially the whole query sequence; some just small regions. Different alignments will have insertions and deletions in different places.

How long should the profile be, and what should each “column” correspond to?

PSI-BLAST Profile Length

The profile constructed by PSI-BLAST has exactly the same length as the query sequence. Insertions with respect to the query are simply ignored.

```
>sp|Q99728.2|BARD1_HUMAN
Length=777

 Score = 53.1 bits (126),  Expect = 3e-07, Method: Composition-based stats.
 Identities = 32/111 (29%), Positives = 55/111 (50%), Gaps = 15/111 (14%)

Query  24   THVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGR  83
            THVV+  DA    + TLK  LGI  G W++ + WV   ++ +    E  +E+
Sbjct  605  THVVVPGDA---VQSTLKCMLGILNGCWILKFEWVKACLRRKVCEQEEKYEIP-------  654

Query  84   NHQGPKRARESQDR---KIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVV  131
              +GP+R+R ++++   K+F G     +G F + P D L  +V   G  ++
Sbjct  655  --EGPRRSRLNREQLLPKLFDGCYFYLWGTFKHHPKDNLIKLVTAGGGQIL  703
```

These aligned letters are ignored.

Multiple Alignment

“True” multiple alignment algorithms generally try to take account of all sequences simultaneously when constructing a multiple alignment.

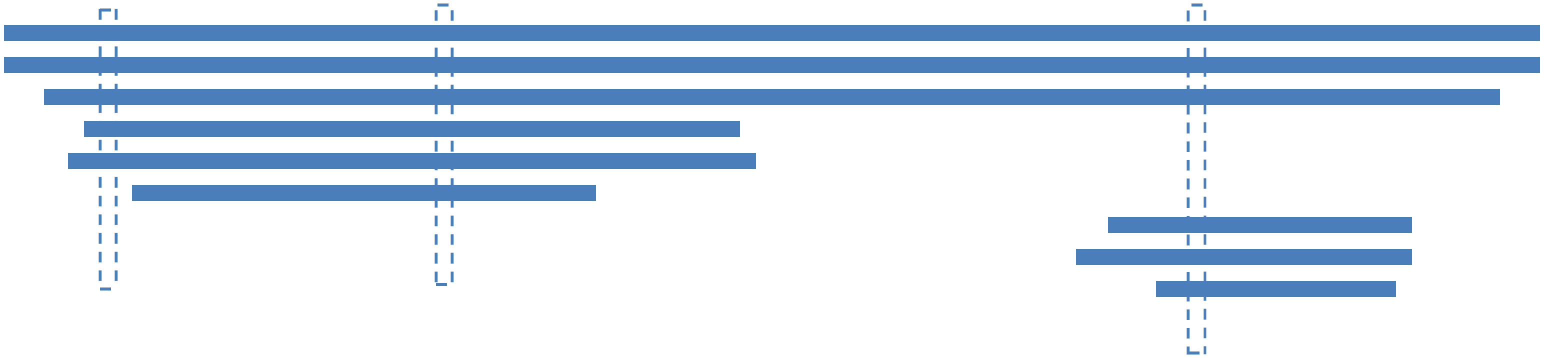
For speed and simplicity, PSI-BLAST simply collapses the pairwise alignments produced by BLAST into a multiple alignment, with each profile column corresponding to a single letter from the query sequence.

```
Query:    ...THVVMKTDAEFVCERTLKYFL...
          ...THVVVPGDA---VQSTLKCML...

Query:    ...THVVMKTDAEFVCERTLKYFL...
          ...TRVIVPGEG--VQSTTKCMLL...
```

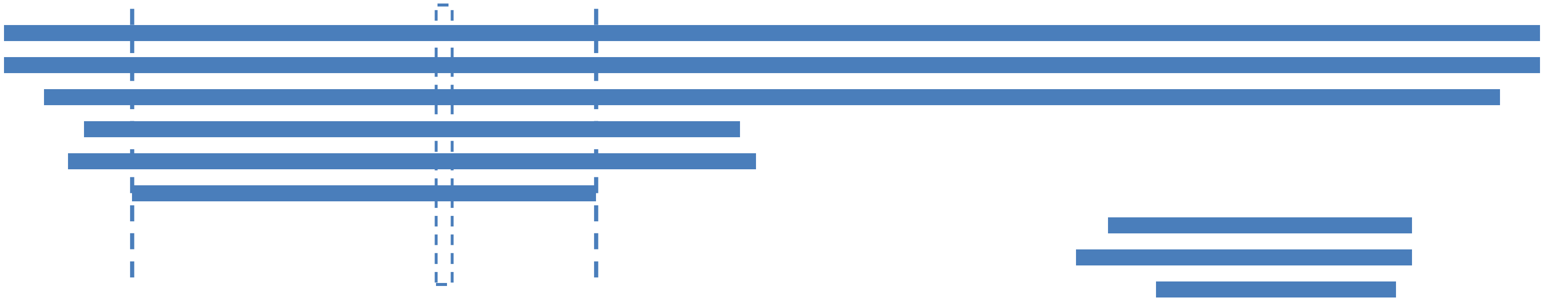
⇓

```
Query:    ...THVVMKTDAEFVCERTLKYFL...
          ...THVVVPGDA---VQSTLKCML...
          ...TRVIVPGEG--VQSTTKCMLL...
```

Sequence Weights

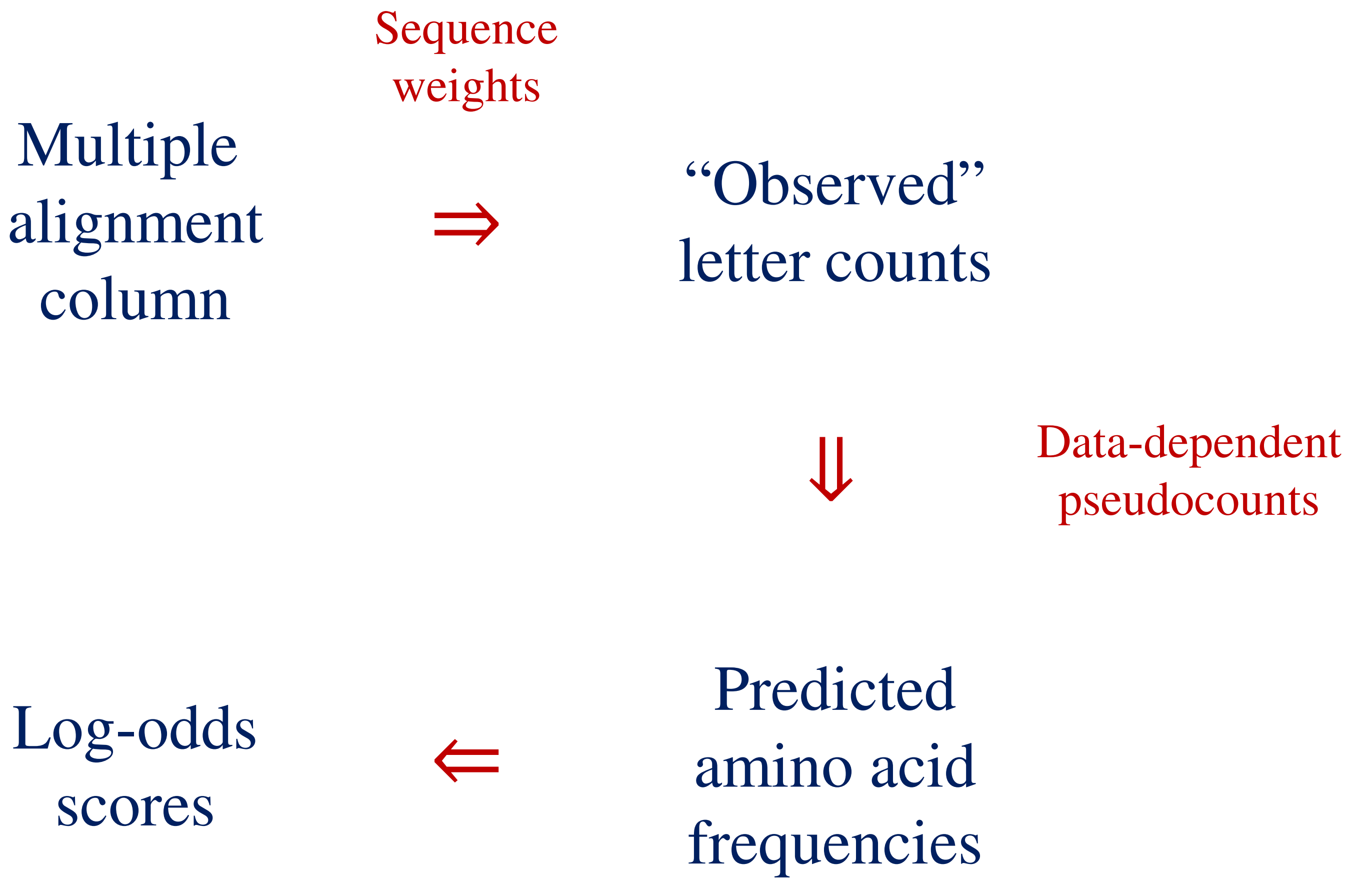


Different columns involve different sets of sequences. Thus the sequence weights may very from one profile column to another.



For a given column ***C***, consider only those sequences that participate in ***C***, and calculate sequence weights using the maximal extent of the subalignment containing all those sequences.

The Construction of Profile Scores



Multiple alignment column

Sequence weights ⇒

“Observed” letter counts

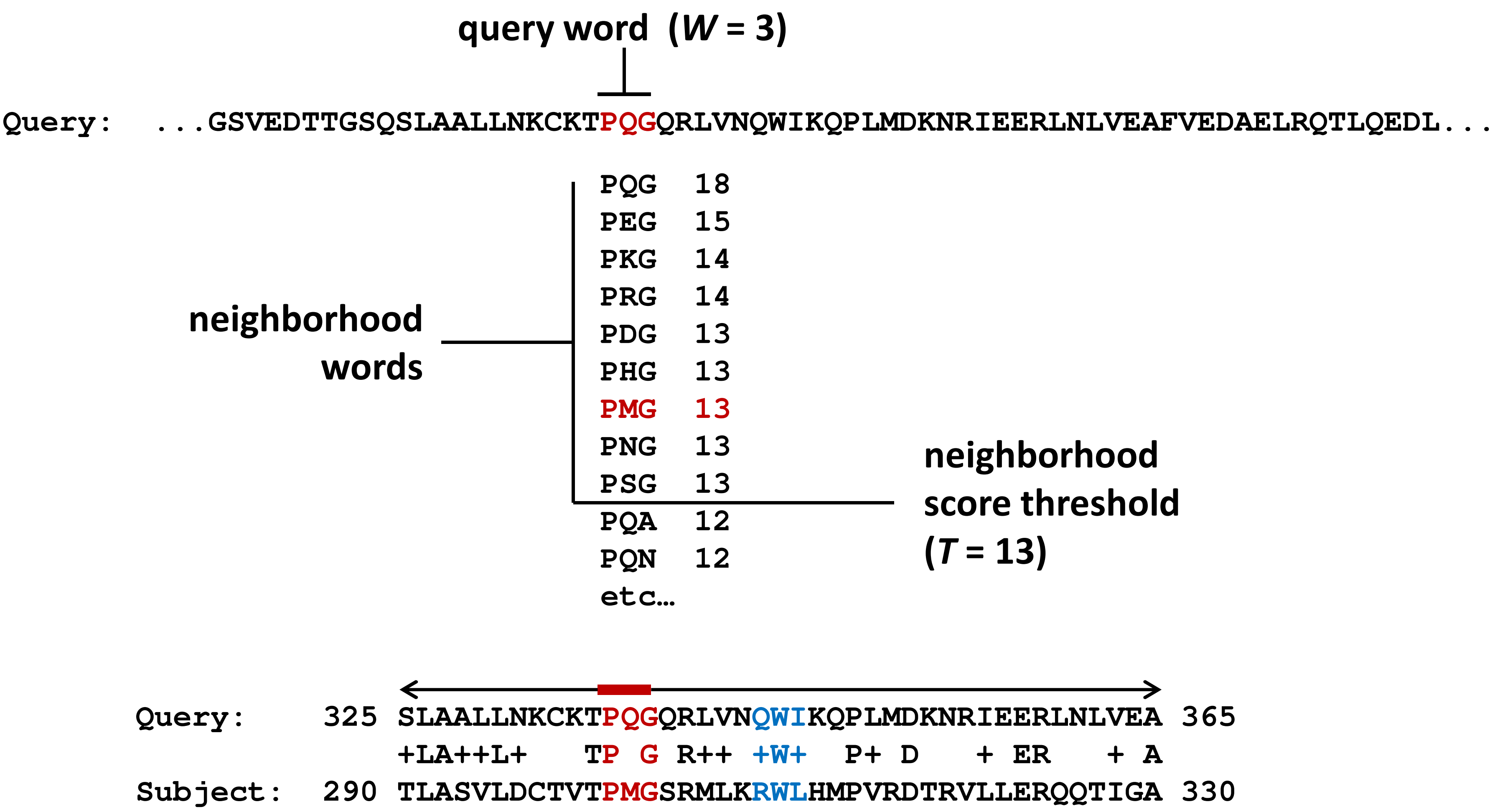
⇓ Data-dependent pseudocounts

Predicted amino acid frequencies

⇐

Log-odds scores

The BLAST Algorithm Applied to Profiles



query word (*W* = 3)

```
Query:   ...GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL...
```

neighborhood words

PQG	18
PEG	15
PKG	14
PRG	14
PDG	13
PHG	13
PMG	13
PNG	13
PSG	13
PQA	12
PQN	12
etc…	

neighborhood score threshold (*T* = 13)

```
Query:     325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
               +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Subject:   290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

With the query sequence replaced by a profile, one can still construct a list of neighborhood words, and proceed exactly as before.

PSI-BLAST Statistics

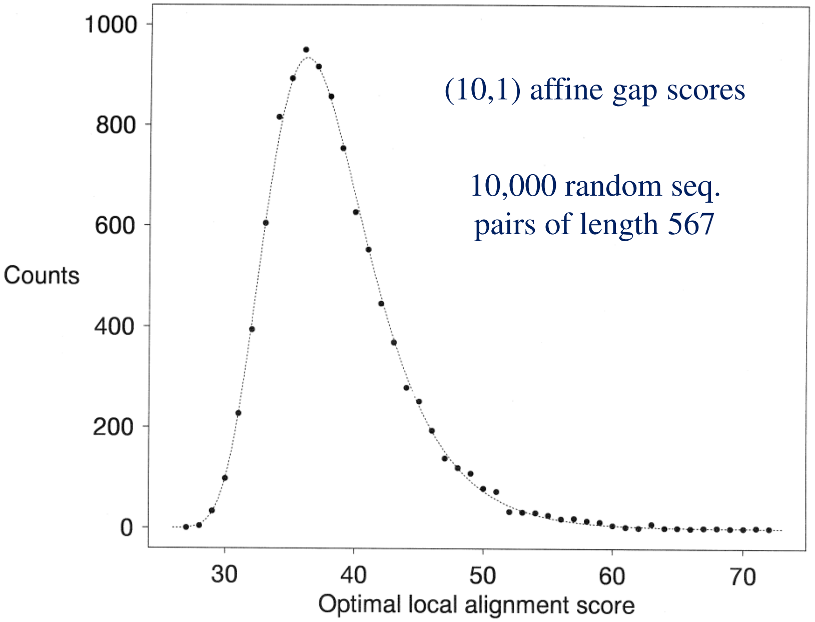
Problem: By random simulation, we can estimate the statistical parameters λ_g and K_g for gapped profile-sequence alignment, but it takes too much time to do this for the new profile generated by each PSI-BLAST iteration.

However: By extending Karlin's theory to profiles, we can calculate analytically, and rapidly, the ungapped statistical parameters λ_u and K_u for profile-sequence comparison.

Hypothesis: If particular gap costs transform a specific λ_u to λ_g for sequence-sequence comparison, they will do approximately the same for profile-sequence comparison.

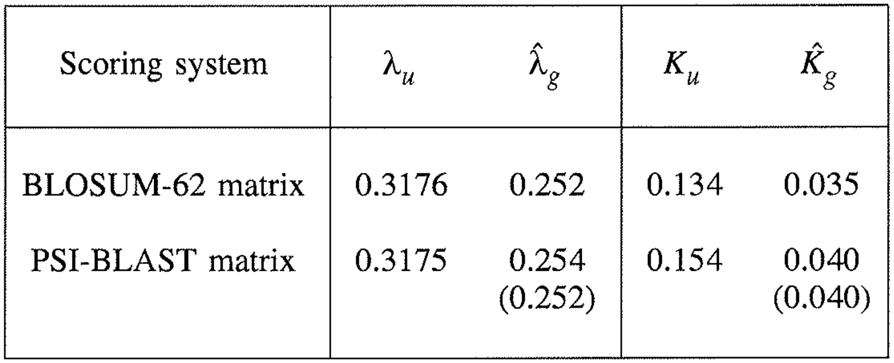
Solution: Scale each new profile so that it has the same λ_u as a standard substitution matrix, such as BLOSUM-62. (This is fast, by Karlin's theory.) Assume the precomputed λ_g for sequence-sequence comparison is valid for profile-sequence comparison.

PSI-BLAST profile derived from 128 significant local alignments from the comparison of inflenza A virus hemagglutinin precursor to SWISS-PROT.



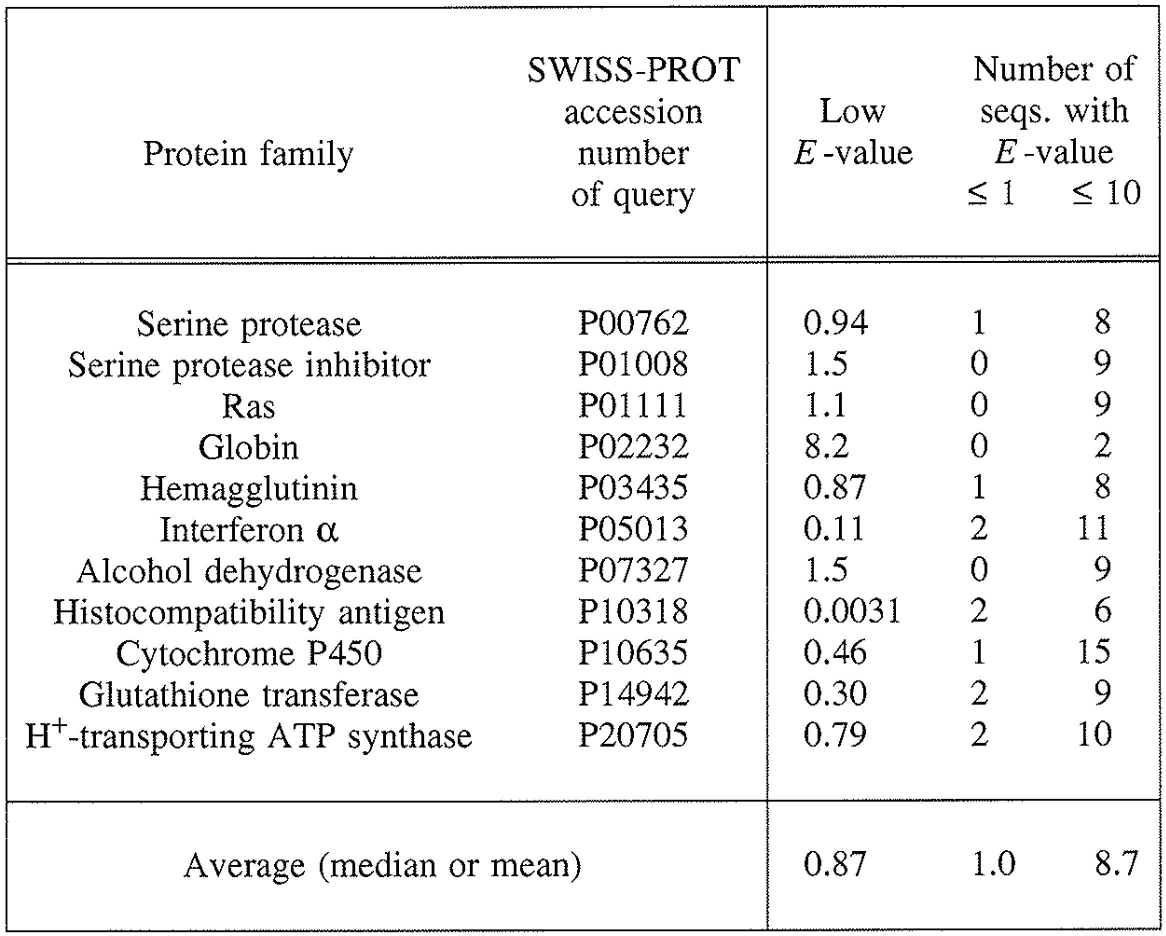
(10,1) affine gap scores

10,000 random seq. pairs of length 567



Scoring system	λ_u	$\hat{\lambda}_g$	K_u	$\hat{K}_g$
BLOSUM-62 matrix	0.3176	0.252	0.134	0.035
PSI-BLAST matrix	0.3175	0.254 (0.252)	0.154	0.040 (0.040)

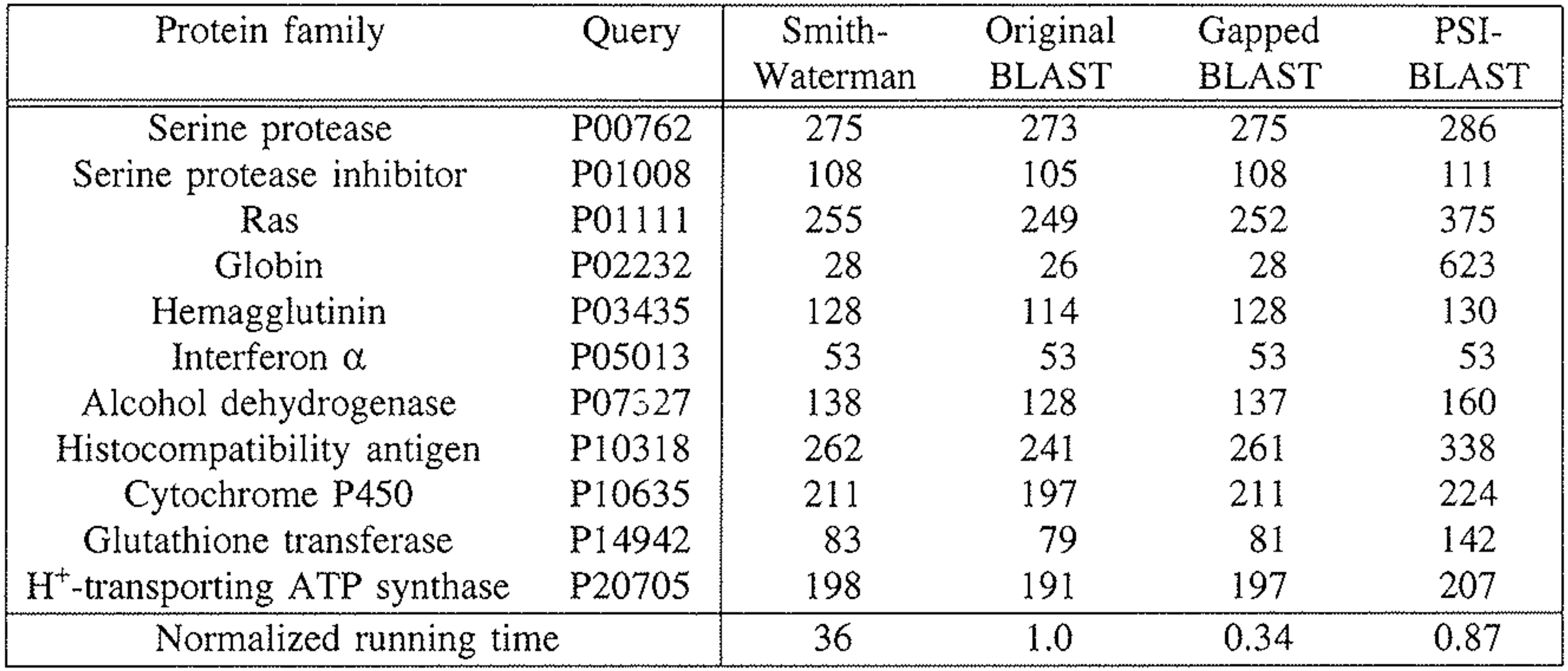
Accuracy of PSI-BLAST Statistics



Protein family	SWISS-PROT accession number of query	Low *E*-value	Number of seqs. with *E*-value	
			≤ 1	≤ 10
Serine protease	P00762	0.94	1	8
Serine protease inhibitor	P01008	1.5	0	9
Ras	P01111	1.1	0	9
Globin	P02232	8.2	0	2
Hemagglutinin	P03435	0.87	1	8
Interferon α	P05013	0.11	2	11
Alcohol dehydrogenase	P07327	1.5	0	9
Histocompatibility antigen	P10318	0.0031	2	6
Cytochrome P450	P10635	0.46	1	15
Glutathione transferase	P14942	0.30	2	9
H^+-transporting ATP synthase	P20705	0.79	2	10
Average (median or mean)		0.87	1.0	8.7

Sequences compared to a shuffled version of SWISS-PROT

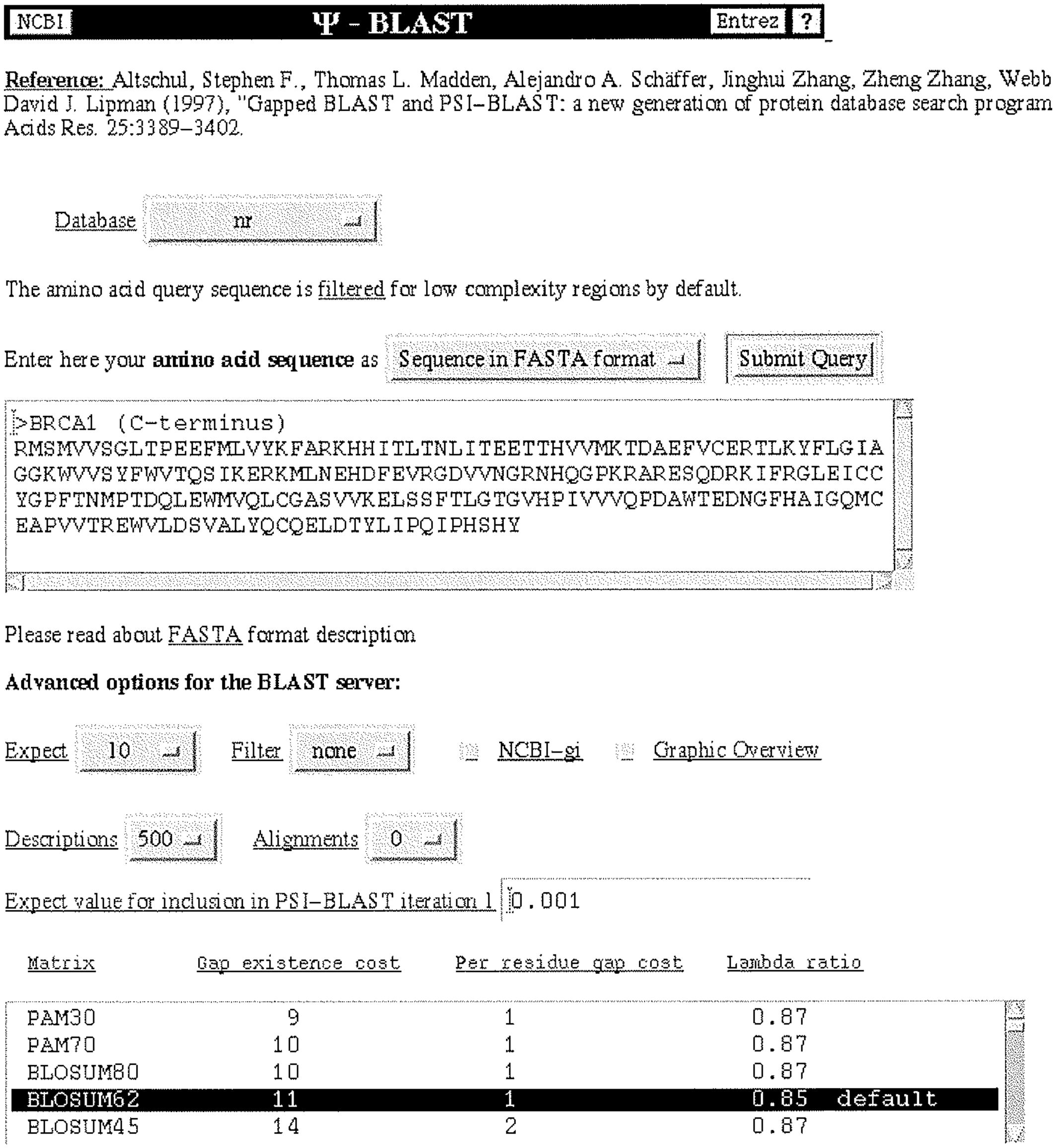
PSI-BLAST Search Results



Protein family	Query	Smith-Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon α	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H^+-transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

Number of SWISS-PROT sequences with an E-value ≤ 0.01.
By SWISS-PROT annotation, all but one are true positives.

Running PSI-BLAST



Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: nr

The amino acid query sequence is filtered for low complexity regions by default.

Enter here your **amino acid sequence** as: Sequence in FASTA format

```
>BRCA1 (C-terminus)
RMSMVVSGLTPEEFMLVYKFARKHHITLTNLITEETTHVVMKTDAEFVCERTLKYFLGIA
GGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICC
YGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNGFHAIGQMC
EAPVVTREWVLDSVALYQCQELDTYLIPQIPHSHY
```

Please read about FASTA format description

Advanced options for the BLAST server:

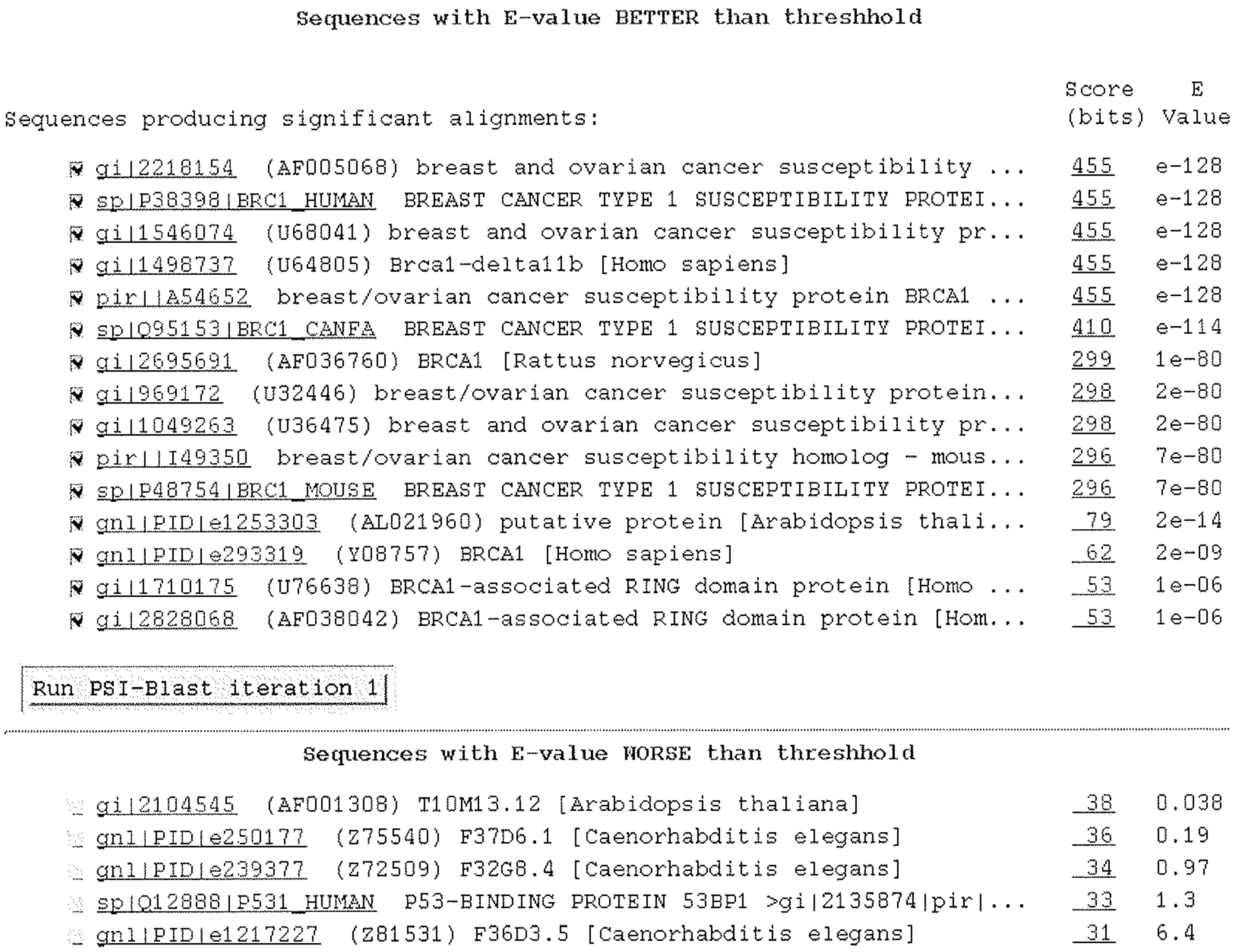
Expect: 10 Filter: none □ NCBI-gi □ Graphic Overview

Descriptions: 500 Alignments: 0

Expect value for inclusion in PSI-BLAST iteration 1: 0.001

Matrix	Gap existence cost	Per residue gap cost	Lambda ratio	
PAM30	9	1	0.87	
PAM70	10	1	0.87	
BLOSUM80	10	1	0.87	
BLOSUM62	11	1	0.85	default
BLOSUM45	14	2	0.87	

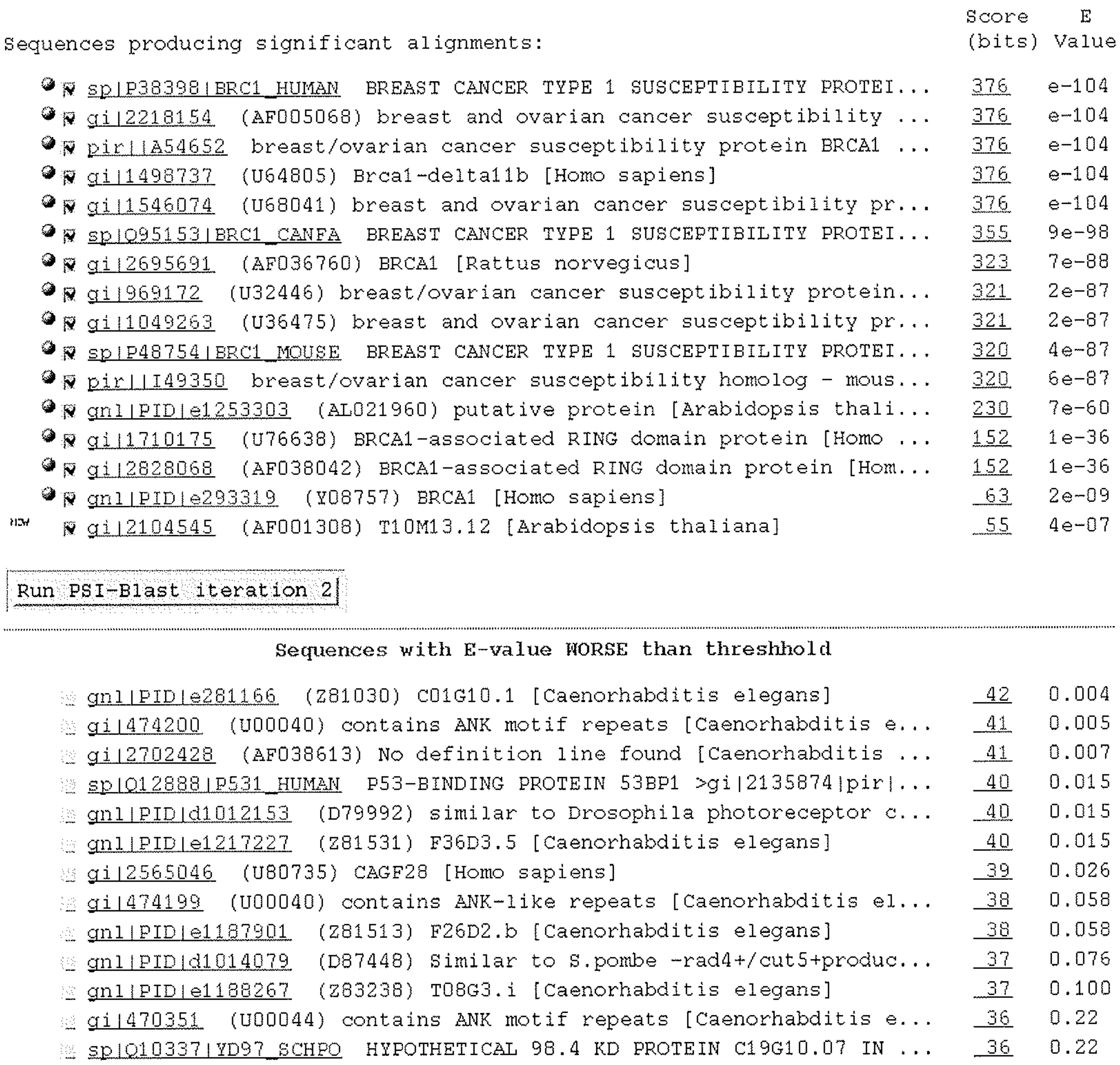
Results of Initial BLAST Run



```
                    Sequences with E-value BETTER than threshhold

                                                                          Score    E
Sequences producing significant alignments:                               (bits) Value

    ☑ gi|2218154  (AF005068) breast and ovarian cancer susceptibility ...   455   e-128
    ☑ sp|P38398|BRC1_HUMAN  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   455   e-128
    ☑ gi|1546074  (U68041) breast and ovarian cancer susceptibility pr...   455   e-128
    ☑ gi|1498737  (U64805) Brca1-delta11b [Homo sapiens]                    455   e-128
    ☑ pir||A54652  breast/ovarian cancer susceptibility protein BRCA1 ...   455   e-128
    ☑ sp|Q95153|BRC1_CANFA  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   410   e-114
    ☑ gi|2695691  (AF036760) BRCA1 [Rattus norvegicus]                      299   1e-80
    ☑ gi|969172  (U32446) breast/ovarian cancer susceptibility protein...   298   2e-80
    ☑ gi|1049263  (U36475) breast and ovarian cancer susceptibility pr...   298   2e-80
    ☑ pir||I49350  breast/ovarian cancer susceptibility homolog - mous...   296   7e-80
    ☑ sp|P48754|BRC1_MOUSE  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   296   7e-80
    ☑ gnl|PID|e1253303  (AL021960) putative protein [Arabidopsis thali...    79   2e-14
    ☑ gnl|PID|e293319  (Y08757) BRCA1 [Homo sapiens]                         62   2e-09
    ☑ gi|1710175  (U76638) BRCA1-associated RING domain protein [Homo ...    53   1e-06
    ☑ gi|2828068  (AF038042) BRCA1-associated RING domain protein [Hom...    53   1e-06

 Run PSI-Blast iteration 1
```

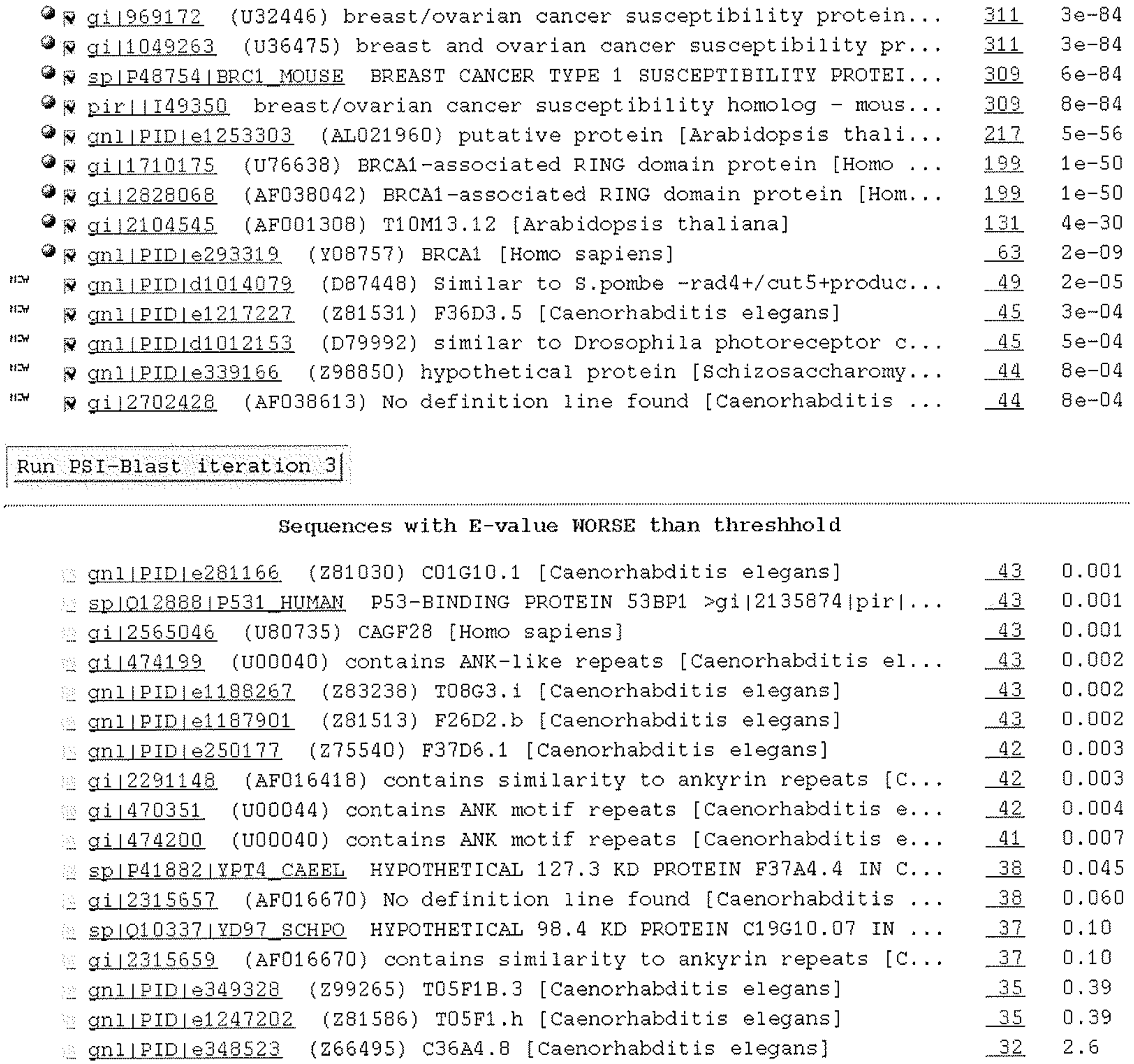
```
                    Sequences with E-value WORSE than threshhold

    ☐ gi|2104545  (AF001308) T10M13.12 [Arabidopsis thaliana]                38   0.038
    ☐ gnl|PID|e250177  (Z75540) F37D6.1 [Caenorhabditis elegans]             36   0.19
    ☐ gnl|PID|e239377  (Z72509) F32G8.4 [Caenorhabditis elegans]             34   0.97
    ☐ sp|Q12888|P531_HUMAN  P53-BINDING PROTEIN 53BP1 >gi|2135874|pir|...    33   1.3
    ☐ gnl|PID|e1217227  (Z81531) F36D3.5 [Caenorhabditis elegans]            31   6.4
```

Results of First PSI-BLAST Iteration



```
                                                                          Score    E
Sequences producing significant alignments:                               (bits) Value

     sp|P38398|BRC1_HUMAN  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   376   e-104
     gi|2218154  (AF005068) breast and ovarian cancer susceptibility ...   376   e-104
     pir||A54652  breast/ovarian cancer susceptibility protein BRCA1 ...   376   e-104
     gi|1498737  (U64805) Brca1-delta11b [Homo sapiens]                    376   e-104
     gi|1546074  (U68041) breast and ovarian cancer susceptibility pr...   376   e-104
     sp|Q95153|BRC1_CANFA  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   355   9e-98
     gi|2695691  (AF036760) BRCA1 [Rattus norvegicus]                      323   7e-88
     gi|969172  (U32446) breast/ovarian cancer susceptibility protein...   321   2e-87
     gi|1049263  (U36475) breast and ovarian cancer susceptibility pr...   321   2e-87
     sp|P48754|BRC1_MOUSE  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   320   4e-87
     pir||I49350  breast/ovarian cancer susceptibility homolog - mous...   320   6e-87
     gnl|PID|e1253303  (AL021960) putative protein [Arabidopsis thali...   230   7e-60
     gi|1710175  (U76638) BRCA1-associated RING domain protein [Homo ...   152   1e-36
     gi|2828068  (AF038042) BRCA1-associated RING domain protein [Hom...   152   1e-36
     gnl|PID|e293319  (Y08757) BRCA1 [Homo sapiens]                         63   2e-09
NEW  gi|2104545  (AF001308) T10M13.12 [Arabidopsis thaliana]                55   4e-07


                     Sequences with E-value WORSE than threshhold

     gnl|PID|e281166  (Z81030) C01G10.1 [Caenorhabditis elegans]            42   0.004
     gi|474200  (U00040) contains ANK motif repeats [Caenorhabditis e...    41   0.005
     gi|2702428  (AF038613) No definition line found [Caenorhabditis ...    41   0.007
     sp|Q12888|P531_HUMAN  P53-BINDING PROTEIN 53BP1 >gi|2135874|pir|...    40   0.015
     gnl|PID|d1012153  (D79992) similar to Drosophila photoreceptor c...    40   0.015
     gnl|PID|e1217227  (Z81531) F36D3.5 [Caenorhabditis elegans]            40   0.015
     gi|2565046  (U80735) CAGF28 [Homo sapiens]                             39   0.026
     gi|474199  (U00040) contains ANK-like repeats [Caenorhabditis el...    38   0.058
     gnl|PID|e1187901  (Z81513) F26D2.b [Caenorhabditis elegans]            38   0.058
     gnl|PID|d1014079  (D87448) Similar to S.pombe -rad4+/cut5+produc...    37   0.076
     gnl|PID|e1188267  (Z83238) T08G3.i [Caenorhabditis elegans]            37   0.100
     gi|470351  (U00044) contains ANK motif repeats [Caenorhabditis e...    36   0.22
     sp|Q10337|YD97_SCHPO  HYPOTHETICAL 98.4 KD PROTEIN C19G10.07 IN ...    36   0.22
```

Results of Second PSI-BLAST Iteration



```
    ● ☑ gi|969172  (U32446) breast/ovarian cancer susceptibility protein...   311   3e-84
    ● ☑ gi|1049263  (U36475) breast and ovarian cancer susceptibility pr...   311   3e-84
    ● ☑ sp|P48754|BRC1_MOUSE  BREAST CANCER TYPE 1 SUSCEPTIBILITY PROTEI...   309   6e-84
    ● ☑ pir||I49350  breast/ovarian cancer susceptibility homolog - mous...   309   8e-84
    ● ☑ gnl|PID|e1253303  (AL021960) putative protein [Arabidopsis thali...   217   5e-56
    ● ☑ gi|1710175  (U76638) BRCA1-associated RING domain protein [Homo ...   199   1e-50
    ● ☑ gi|2828068  (AF038042) BRCA1-associated RING domain protein [Hom...   199   1e-50
    ● ☑ gi|2104545  (AF001308) T10M13.12 [Arabidopsis thaliana]               131   4e-30
    ● ☑ gnl|PID|e293319  (Y08757) BRCA1 [Homo sapiens]                         63   2e-09
NEW   ☑ gnl|PID|d1014079  (D87448) Similar to S.pombe -rad4+/cut5+produc...    49   2e-05
NEW   ☑ gnl|PID|e1217227  (Z81531) F36D3.5 [Caenorhabditis elegans]            45   3e-04
NEW   ☑ gnl|PID|d1012153  (D79992) similar to Drosophila photoreceptor c...    45   5e-04
NEW   ☑ gnl|PID|e339166  (Z98850) hypothetical protein [Schizosaccharomy...    44   8e-04
NEW   ☑ gi|2702428  (AF038613) No definition line found [Caenorhabditis ...    44   8e-04

Run PSI-Blast iteration 3

                    Sequences with E-value WORSE than threshhold

      □ gnl|PID|e281166  (Z81030) C01G10.1 [Caenorhabditis elegans]            43   0.001
      □ sp|Q12888|P531_HUMAN  P53-BINDING PROTEIN 53BP1 >gi|2135874|pir|...    43   0.001
      □ gi|2565046  (U80735) CAGF28 [Homo sapiens]                             43   0.001
      □ gi|474199  (U00040) contains ANK-like repeats [Caenorhabditis el...    43   0.002
      □ gnl|PID|e1188267  (Z83238) T08G3.i [Caenorhabditis elegans]            43   0.002
      □ gnl|PID|e1187901  (Z81513) F26D2.b [Caenorhabditis elegans]            43   0.002
      □ gnl|PID|e250177  (Z75540) F37D6.1 [Caenorhabditis elegans]             42   0.003
      □ gi|2291148  (AF016418) contains similarity to ankyrin repeats [C...    42   0.003
      □ gi|470351  (U00044) contains ANK motif repeats [Caenorhabditis e...    42   0.004
      □ gi|474200  (U00040) contains ANK motif repeats [Caenorhabditis e...    41   0.007
      □ sp|P41882|YPT4_CAEEL  HYPOTHETICAL 127.3 KD PROTEIN F37A4.4 IN C...    38   0.045
      □ gi|2315657  (AF016670) No definition line found [Caenorhabditis ...    38   0.060
      □ sp|Q10337|YD97_SCHPO  HYPOTHETICAL 98.4 KD PROTEIN C19G10.07 IN ...    37   0.10
      □ gi|2315659  (AF016670) contains similarity to ankyrin repeats [C...    37   0.10
      □ gnl|PID|e349328  (Z99265) T05F1B.3 [Caenorhabditis elegans]            35   0.39
      □ gnl|PID|e1247202  (Z81586) T05F1.h [Caenorhabditis elegans]            35   0.39
      □ gnl|PID|e348523  (Z66495) C36A4.8 [Caenorhabditis elegans]             32   2.6
```

The Corruption of Profiles

PSI-BLAST E-values are calculated for the profiles PSI-BLAST produces, and can not be interpreted as referring to the original query sequence.

Once a sequence unrelated to the query is included in a PSI-BLAST multiple alignment, and thus in the construction of PSI-BLAST's profile, it will bring in many of its "friends" on the next iteration, and this process can snowball. Sequence weighting will exacerbate this process.

Profile corruption is a major problem for iterative approaches such as PSI-BLAST.

PSI-BLAST Refinements

Schäffer, A.A., *et al.* (2001) “Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.” *Nucleic Acids Res.* **29**:2994-3005.

> Composition-based statistics;
> Different treatment of indels in calculating predicted amino acid frequencies;
> Optional use of Smith-Waterman algorithm in final stage;
> Filter database sequences for low-complexity segments;
> etc.

Altschul, S.F., *et al.* (2005) "Protein database searches using compositionally adjusted substitution matrices." *FEBS J.* **272**:5101-5109.

> In the initial BLAST search, the adjustment of substitution matrices for use with sequences having non-standard amino composition.

Altschul, S.F., *et al.* (2009) "PSI-BLAST pseudocounts and the minimum description length principle.“ *Nucleic Acids Res.* **37**:815-824.

> Refined calculation of the effective number of independent observations in a column;
> Number of pseudocounts dependent on column entropy.

<u>Possible future development</u>:

> Sequence trimming to avoid the over-extension of true alignments;
> the problem was described in:

Gonzalez, M.W. & Pearson, W.R. (2010) “Homologous over-extension: a challenge for iterative similarity searches.” *Nucleic Acids Res.* **38**:2177-2179.