

Bioinformatics for Biologists

Sequence Analysis: Part II. Pattern Searching

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute

Topics to Cover

- Pattern searching
 - PSI-BLAST
 - PHI-BLAST
 - Finding patterns

PSI-BLAST

- **Position Specific Iterative BLAST** uses a profile (or position specific scoring matrix, PSSM) that is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search.
- The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero.
- The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" is used to refine the profile. This iterative searching strategy results in increased sensitivity.

Start with a BLASTP search

Sequences with E-value BETTER than threshold

		Score	E
			(bits) Value
NEW	<input checked="" type="checkbox"/> gi 2501594 sp Q57997 Y577_METJA	Protein MJ0577	244 5e-65
NEW	<input checked="" type="checkbox"/> gi 2501593 sp Q57951 Y531_METJA	Hypothetical protein MJ0531	75 8e-14
NEW	<input checked="" type="checkbox"/> gi 1177001 sp P42297 YXIE_BACSU	Hypothetical protein yxiE precursor	65 6e-11
NEW	<input checked="" type="checkbox"/> gi 2501590 sp P73475 YC30_SYNY3	Hypothetical protein slr1230	59 3e-09
NEW	<input checked="" type="checkbox"/> gi 2501596 sp Q50777 YB54_METTM	Hypothetical 16.1 kDa protein in...	54 2e-07
NEW	<input checked="" type="checkbox"/> gi 2501591 sp P74148 YD88_SYNY3	Hypothetical protein sl11388	51 8e-07
NEW	<input checked="" type="checkbox"/> gi 2507517 sp P39177 UP12_ECOLI	Unknown protein from 2D-page (Sp...	49 3e-06
NEW	<input checked="" type="checkbox"/> gi 3334425 sp O27222 YB54_METTH	Hypothetical protein MTH1154	49 4e-06
NEW	<input checked="" type="checkbox"/> gi 1176031 sp P45680 YJ16_COXBU	Hypothetical protein CBU1916	44 1e-04
NEW	<input checked="" type="checkbox"/> gi 2501592 sp P72817 YG54_SYNY3	Hypothetical protein sl11654	44 1e-04
NEW	<input checked="" type="checkbox"/> gi 2501595 sp P74897 YQA3_THEAQ	Hypothetical 14.6 kDa protein in...	44 2e-04
NEW	<input checked="" type="checkbox"/> gi 33518627 sp O07552 NHAX_BACSU	Stress response protein nhaX	44 2e-04
NEW	<input checked="" type="checkbox"/> gi 12231054 sp P87132 YFK5_SCHPO	Hypothetical protein Cl67.05 in...	41 0.001
NEW	<input checked="" type="checkbox"/> gi 1731241 sp Q10851 YK05_MYCTU	Hypothetical protein Rv2005c/MT2...	40 0.003
NEW	<input checked="" type="checkbox"/> gi 2501589 sp P72745 YB01_SYNY3	Hypothetical protein slr1101	39 0.005

Run PSI-Blast iteration 2

PSI-BLAST - Iteration 1

<input checked="" type="checkbox"/>	gi 2501594 sp Q57997 Y577_METJA	Protein MJ0577	192	3e-49
<input checked="" type="checkbox"/>	gi 1177001 sp P42297 YXIE_BACSU	Hypothetical protein yxiE precursor	160	1e-39
<input checked="" type="checkbox"/>	gi 2501591 sp P74148 YD88_SYNY3	Hypothetical protein sl11388	159	2e-39
<input checked="" type="checkbox"/>	gi 2501593 sp Q57951 Y531_METJA	Hypothetical protein MJ0531	157	7e-39
<input checked="" type="checkbox"/>	gi 2501592 sp P72817 YG54_SYNY3	Hypothetical protein sl11654	149	2e-36
<input checked="" type="checkbox"/>	gi 3334425 sp O27222 YB54_METTH	Hypothetical protein MTH1154	137	9e-33
<input checked="" type="checkbox"/>	gi 2501596 sp Q50777 YB54_METTM	Hypothetical 16.1 kDa protein in...	134	6e-32
<input checked="" type="checkbox"/>	gi 2507517 sp P39177 UP12_ECOLI	Unknown protein from 2D-page (Sp...	133	1e-31
<input checked="" type="checkbox"/>	gi 1731241 sp Q10851 YK05_MYCTU	Hypothetical protein Rv2005c/MT2...	124	1e-28
<input checked="" type="checkbox"/>	gi 2501589 sp P72745 YB01_SYNY3	Hypothetical protein slr1101	111	5e-25
<input checked="" type="checkbox"/>	gi 1176031 sp P45680 YJ16_COXBU	Hypothetical protein CBU1916	110	1e-24
<input checked="" type="checkbox"/>	gi 2501595 sp P74897 YQA3_THEAQ	Hypothetical 14.6 kDa protein in...	108	4e-24
<input checked="" type="checkbox"/>	gi 12231054 sp P87132 YFK5_SCHPO	Hypothetical protein C167.05 in...	107	1e-23
<input checked="" type="checkbox"/>	gi 33518627 sp O07552 NHAX_BACSU	Stress response protein nhaX	95	8e-20
<input checked="" type="checkbox"/>	gi 2501590 sp P73475 YC30_SYNY3	Hypothetical protein slr1230	92	4e-19
NEW	<input checked="" type="checkbox"/> gi 2507516 sp P37903 UP03_ECOLI	Unknown protein 2D_000B3L from 2...	88	8e-18
NEW	<input checked="" type="checkbox"/> gi 1731252 sp Q10862 YJ96_MYCTU	Hypothetical protein Rv1996/MT20...	82	4e-16
NEW	<input checked="" type="checkbox"/> gi 2507515 sp P44195 YDAA_HAEIN	Protein HI1426	55	1e-07
NEW	<input checked="" type="checkbox"/> gi 2507514 sp P03807 YDAA_ECOLI	Protein ydaA	52	4e-07
NEW	<input checked="" type="checkbox"/> gi 1174913 sp P44880 USPA_HAEIN	Universal stress protein A homolog	47	1e-05
NEW	<input checked="" type="checkbox"/> gi 2829581 sp P71893 YN19_MYCTU	Hypothetical protein Rv2319c/MT2...	41	7e-04
NEW	<input checked="" type="checkbox"/> gi 17380539 sp P28242 USPA_ECOLI	Universal stress protein A	40	0.002
NEW	<input checked="" type="checkbox"/> gi 1175845 sp P46888 YECG_ECOLI	Hypothetical protein yecG	40	0.003

Amino acids

PSSM from PSI-BLAST

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

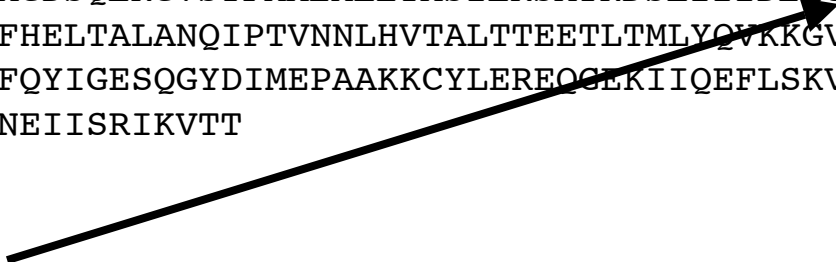
POSITIONS

Pattern Hit Initiated (PHI)- BLAST

>HUMAN MSH2

MAVQPKETLQLESAAEVGFVRRFFQGMPEKPTTTVRLFDRGDFYTAHGEDALLAAREVFKTQGVIKYMGPA
GAKNLQSVVLSKMNFEFVKDLLLVQRVEVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNND
MSASIGVVGVMKSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKCEVLPGETAGDM
GKLRQIIQRGGILITERKKADFSTKDIYQDLNRLKGGKGEQMNSAVLPEMENQVAVSSLSAVIKFLELL
SDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPL
MDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAANLQDCYRLYQGINQLPNVIQA
LEKHEGKHQKLLLAVFVTPLTDLRSDFSKFQEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEK
KMOSTLISAARDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLN
EEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAILEKGQG
RIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA
EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYI
ATKIGAFCMFATHFHELTALANQIPTVNNLHVLTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHV
FQYIGESQGYDIMEPAAKKCYLEREQGLKIIQEFLSKVKQMPFTEMSEENITIKLQ
NEIISRIKVT

DNA mismatch
repair proteins mutS
family signature



PHI-BLAST

```
>gi|4099512|gb|AAD00647.1| (U87911) MutS homolog 2 [Arabidopsis thaliana]
      Length = 117

      Score = 136 bits (364), Expect = 1e-40
      Identities = 88/117 (75%), Positives = 98/117 (83%)

Query: 668 TGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSI VDCILARVGAGDSQLKGVSTFMA 727
          TGPNMGGKST+IRQ GVIVLMAQ+G FVPC+ A +SI DCI ARVGAGD QL+GVSTFM
Sbjct: 1   TGPNMGGKSTFIROVGVIVLMAOVGSFVPCDKASISIRDCIFARVGAGDCQLRGVSTFMQ 60

Query: 728 EMLETASILRATTKDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHF 784
      pattern 743 *****
          EMLETASIL+ AT SLIIIDELGRGTSTYDGFGLAWAI E++ A +FATH+
Sbjct: 61 EMLETASILRATTKDSLIIIDELGRGTSTYDGFGLAWAICEHLVQVKRAPTLFATHY 117
```


Pattern Searching

RRRRYYYY	4 purines followed by 4 pyrimidines
TATAA[1,0,0]	TATAA, allowing 1 mismatch
p1=6...8 GAGA ~p1	a hairpin with GAGA as the loop
p1=6...6 3...8 p1	exact 6 character repeat separated by up to 8
p1=6...6 3..8 p1[1,1,1]	allow one mismatch, deletion and insertion

Pattern Searching Programs

Patscan `scan_for_matches patfile < inputfile`

fuzznuc, EMBOSS programs; web and Unix
fuzzprot,
fuzztrans,
dreg

Demo

- Readseq
- Entrez
- NCBI
- WU-BLAST2
- FASTA
- Smith-Waterman
- BLAT
- PatScan