

#6

Fast Algorithms for Homology Search

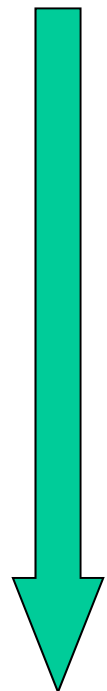
Topics:

- **FASTA**
 - Lookup Table, k-tuple
 - WDG (Weighted Directed Graph)
- **BLAST**
 - neighborhood words table
 - finite automaton
- **PSI-BLAST**
 - BLAST using “profile”
 - Iterative refinement of profile

Homology search from databases

Dynamic programming method (Smith-Waterman algorithm) is too slow for database search, because of rapid increase of database entries.

Approximated search methods are required.

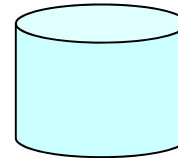


SSEARCH (Smith-Waterman)

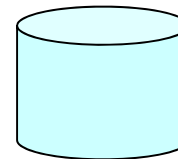
FASTA

BLAST

faster



nr-nt (DNA) database
105,901,840 entries
(Release 09-05-17, May 09)



nr-aa (Protein) database
8,243,496 entries
(Release 07-05-17, May 09)

FASTA

D. Lipman and W. Pearson: “Improved tools for biological sequence comparison”, Proc. Natl. Acad. Sci. USA, 85:2444-2448 (1988)



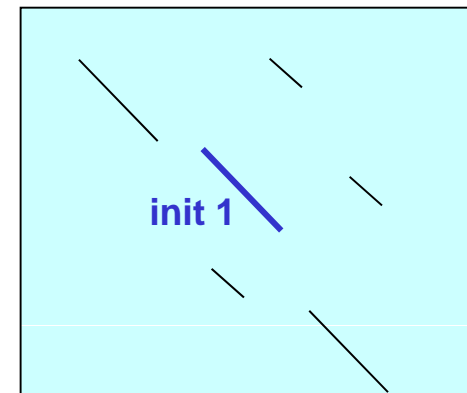
Fast heuristic method to compare query sequence (DNA, or Protein) against a sequence (in a database).

Step1: Find substring matches.

Speed-up by “lookup table” technique.

Step2: Find 10 best diagonal runs.

Get “init 1” score, as the best diagonal score.



Step3: Attempt to join several good diagonal runs.

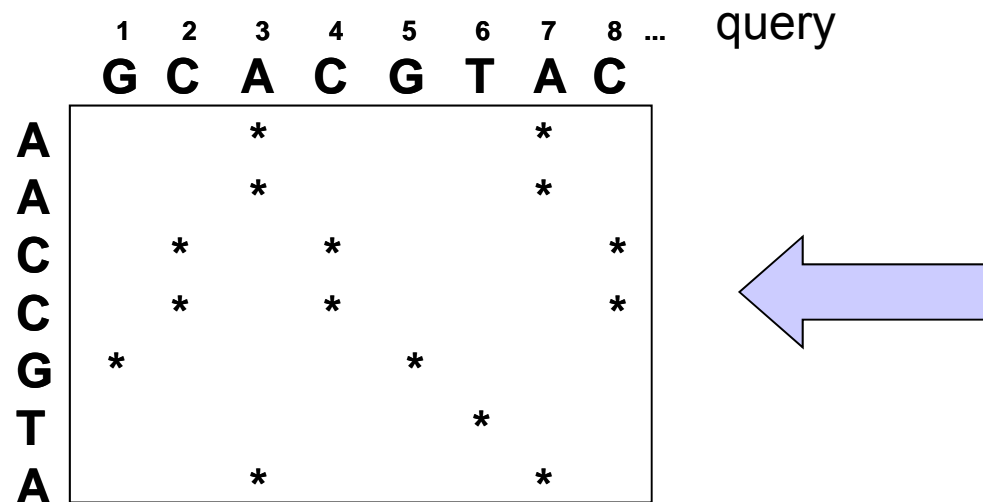
Get “init n” score, as the best path score from “WDG” approach.

Step4: Also perform dynamic programming with only narrow band around the “init 1” diagonal. Get “opt” score, as the DP alignment score.

(when used for DB search, database sequences are ranked by “init n” or “opt” scores.

FASTA

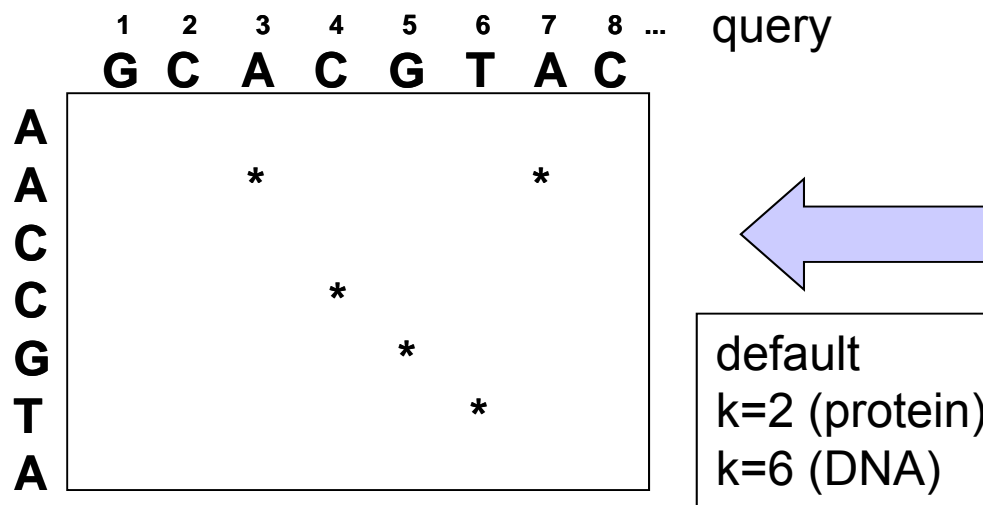
Step1: Find substring matches.



k-tuple = 1
Lookup Table

average length
is $L / 4$

A: 3, 7, ...
C: 2, 4, 8, ...
G: 1, 5, ...
T: 6



k-tuple = 2
Lookup Table

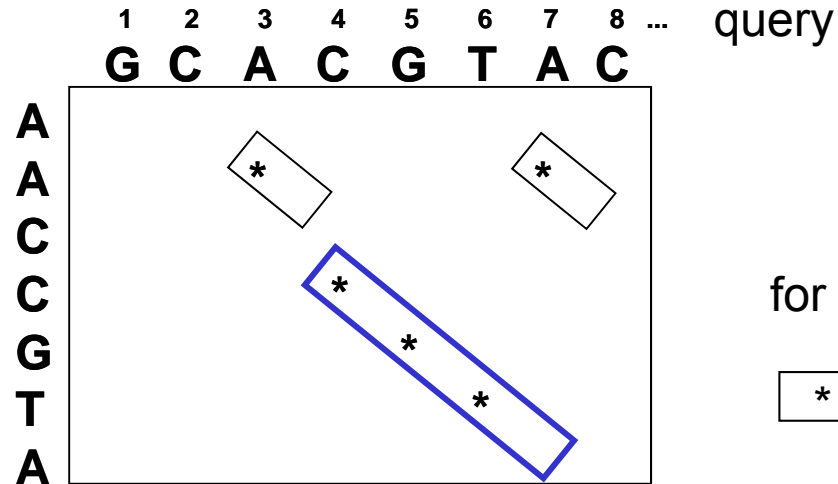
average length
is $L / 4^2$

AA: ... GA: ...
AC: 3, 7, ... GC: 1, ...
AG: ... GG: ...
AT: ... GT: 5, ...
CA: 2 TA: 6, ...
CC: ... TC: ...
CG: 4, ... TG: ...
CT: ... TT: ...

default
k=2 (protein)
k=6 (DNA)

FASTA

Step2: Find 10 best diagonal runs.



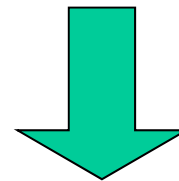
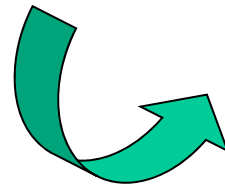
Scoring

for DNA

$$\boxed{* * * _ * _ *} = a \times (\# \text{ of } k\text{-tuple match}) - b \times (\# \text{ of gap})$$

for Protein

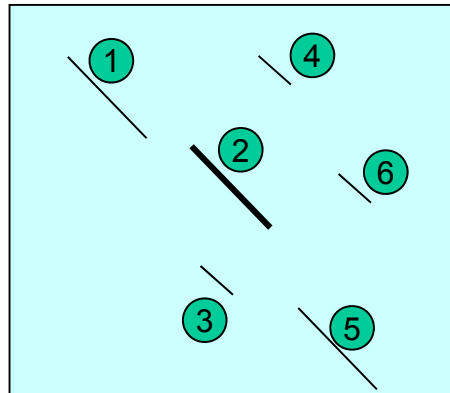
$$\boxed{* * * _ * _ *} = \sum \text{Matrix-score (match)} - b \times (\# \text{ of gap})$$



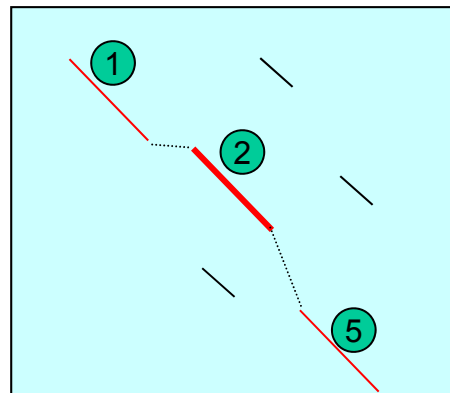
10 best “diagonal” regions are recorded.
The best score = **“init 1” score**.

FASTA

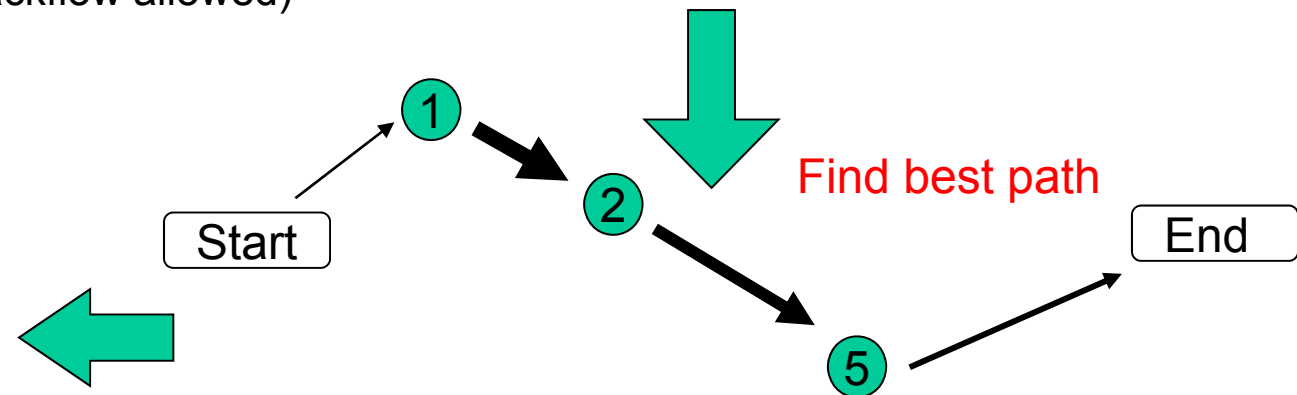
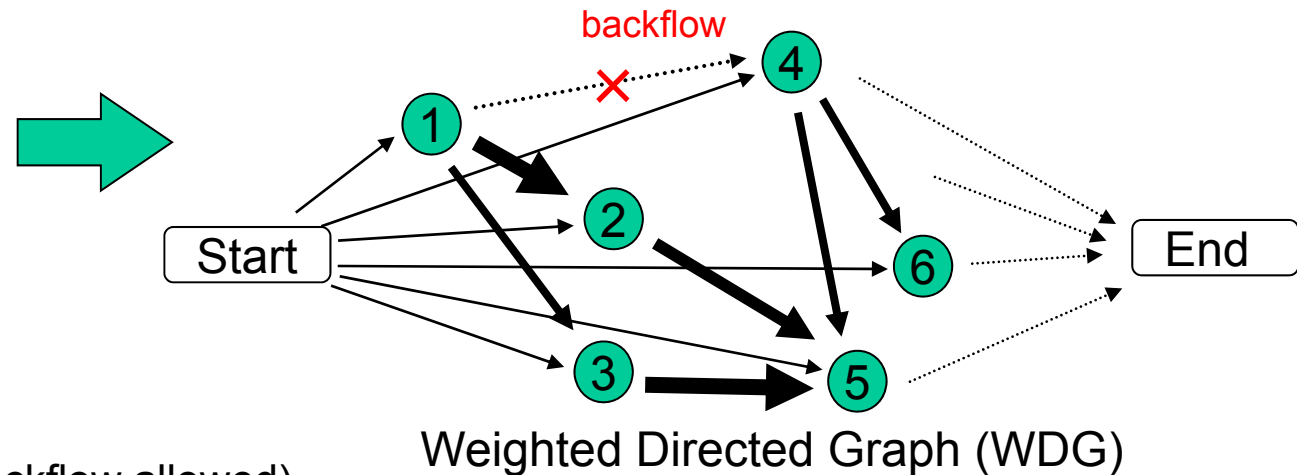
Step3: Attempt to join several good diagonal runs.



up left to down right (no backflow allowed)

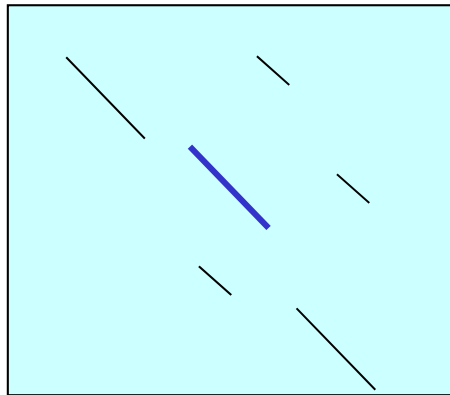


Best alignment score = **“init n” score**.

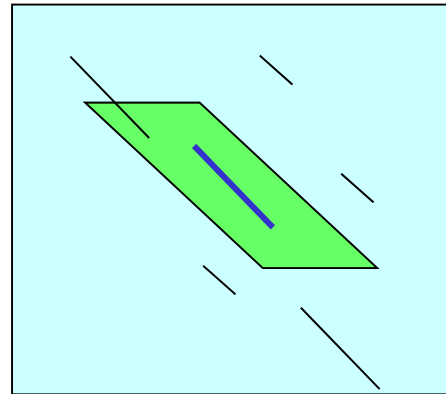
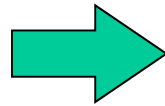


FASTA

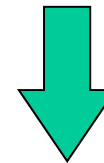
Step4: Perform usual DP with a limited band area



“init 1” diagonal



narrow band around **init 1** diagonal



Smith-waterman method
only within narrow band

Alignment score = **“opt”** score.

use
“init n” score
or
“opt” score

BLAST

BLAST = (Basic Local Alignment Search Tool)

Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers,
David Lipman:

“Basic local alignment search tool”, J. Mol. Biol., 215:403-410 (1990).



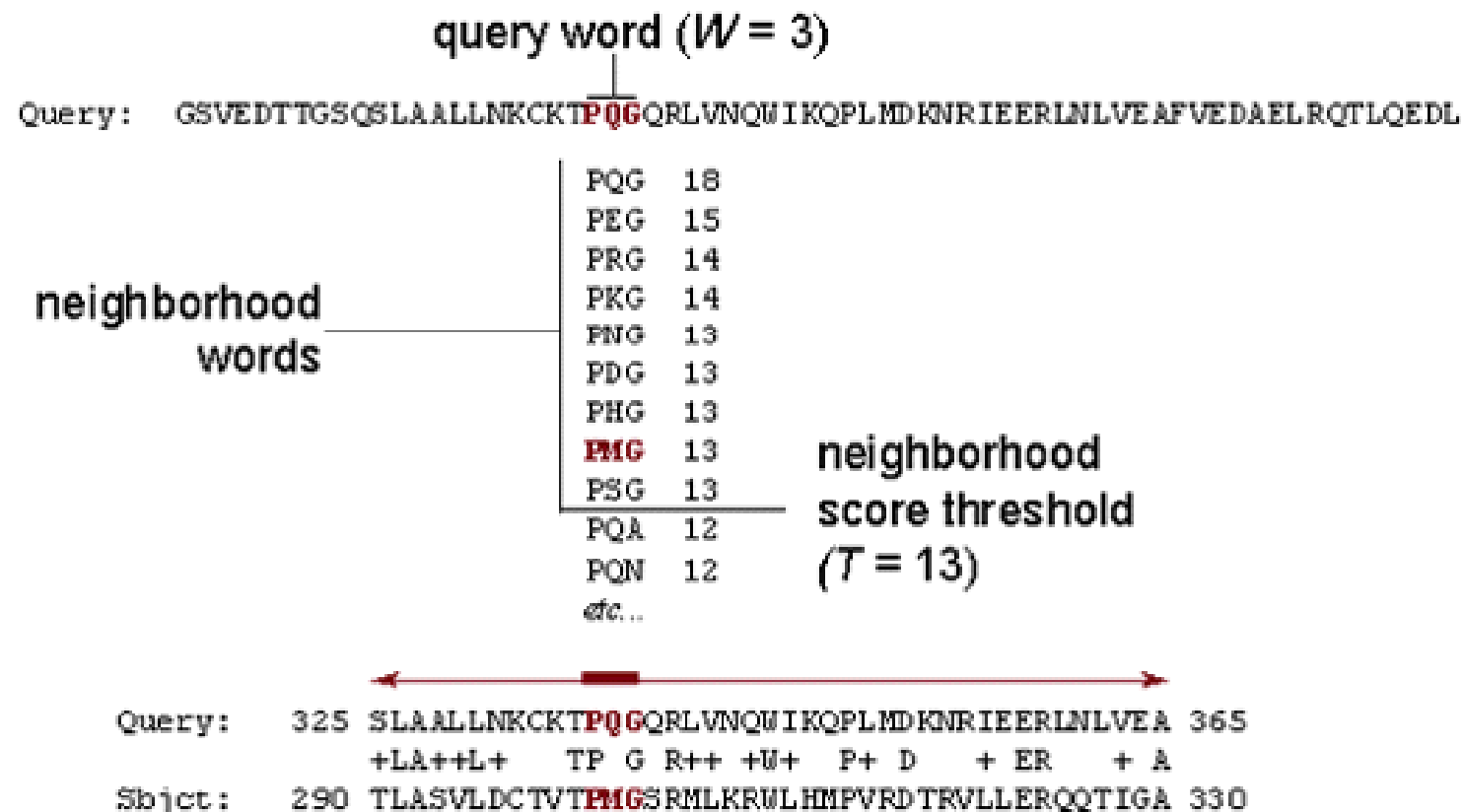
Step1: Prepare “neighborhood words” table with word length “W”.
default: W=3 (protein), W=11 (DNA)

Step2: Search database sequences with neighborhood words.
Speed-up by “finite automaton” technique.

Step3: Extend hits and find “HSP” (High-scoring Segment Pair)
which has at least Score “S”.
Report the “MSP” (Maximal Segment Pair) which has
the maximum score.

BLAST

The BLAST Search Algorithm

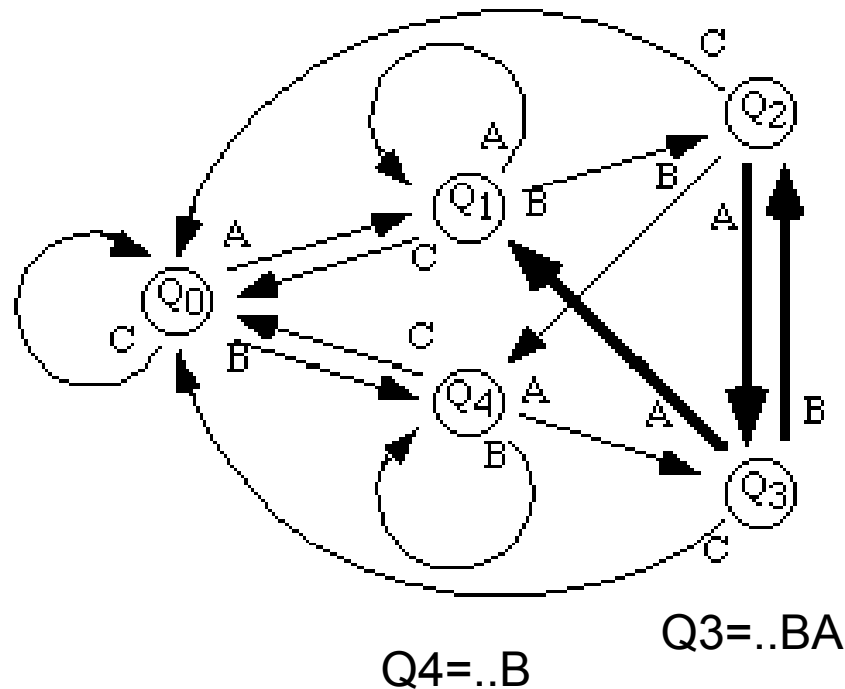


High-scoring Segment Pair (HSP)

BLAST

Substring search by finite automaton

Finding {ABA, BAA, BAB}



Q2->Q3 = ABA
 Q3->Q1 = BAA
 Q3->Q2 = BAB

BLAST

Threshold Score

MSP score “S” follows “Extreme distribution”.

$$P(S \geq x) = 1 - \exp(-\exp(-\lambda(x - \mu)))$$

Characteristic value μ is obtained as

$\mu = \log(KMN) / \lambda$, λ is a constant, and

M, N are length of query and database.

$$P(S \geq x) = 1 - \exp(-KMN \exp(-\lambda x))$$

$$\doteq KMN \exp(-\lambda x) \quad \text{.. Poisson distribution}$$

Therefore, for any probability P (e.g. $P=0.05$),

$$P = KMN \exp(-\lambda x)$$

Threshold score x can be calculated as

$$x = 1/\lambda \{ \log(K/P) + \log(MN) \}$$

BLAST programs

Basic BLAST

Choose a BLAST program to run.

nucleotide blast

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

protein blast

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast

blastx

Search **protein** database using a **translated nucleotide** query

tblastn

Search **translated nucleotide** database using a **protein** query

tblastx

Search **translated nucleotide** database using a **translated nucleotide** query

program	target DB	query (input) sequence
blastn	DNA	DNA
blastp	amino acid	amino acid
blastx	amino acid	DNA (translated to a.a.)
tblastn	DNA (translated to a.a.)	amino acid
tblastp	(<i>obsolete</i>)	
tblastx	DNA (translated to a.a.)	DNA (translated to a.a.)

BLAST search example

Query = human ALDH2 gene DNA sequence (first 50 bases)

Sequences producing significant alignments:
(Click headers to sort columns)

		Score		Expectation		Match%	
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000690.2	Homo sapiens aldehyde dehydrogenase 2 family (mitochondrial) mRNA	81.8	81.8	82%	2e-14	100%	U E G
XM_509379.2	PREDICTED: Pan troglodytes mitochondrial aldehyde dehydrogenase 2 family (mitochondrial) mRNA	81.8	81.8	82%	2e-14	100%	G
XR_012809.1	PREDICTED: Macaca mulatta mitochondrial aldehyde dehydrogenase 2 family (mitochondrial) mRNA	81.8	81.8	82%	2e-14	100%	U G
XM_001490910.1	PREDICTED: Equus caballus similar to aldehyde dehydrogenase 2 family (mitochondrial) mRNA	52.0	52.0	68%	1e-05	94%	U G
NM_001075367.1	Bos taurus similar to Aldehyde dehydrogenase, mitochondrial	52.0	52.0	68%	1e-05	94%	U G
XM_848535.1	PREDICTED: Canis familiaris similar to Aldehyde dehydrogenase, mitochondrial	52.0	52.0	68%	1e-05	94%	U E G
XM_849411.1	PREDICTED: Canis familiaris similar to Aldehyde dehydrogenase, mitochondrial	52.0	52.0	68%	1e-05	94%	G
NM_009656.3	Mus musculus aldehyde dehydrogenase 2, mitochondrial	46.1	46.1	62%	9e-04	93%	U E G
NM_032416.1	Rattus norvegicus aldehyde dehydrogenase 2, mitochondrial	46.1	46.1	62%	9e-04	93%	U E G
XM_845808.1	PREDICTED: Canis familiaris similar to Aldehyde dehydrogenase, mitochondrial	42.1	42.1	66%	0.013	90%	U G
XM_001257227.1	PREDICTED: Bos taurus similar to Aldehyde dehydrogenase, mitochondrial	40.1	40.1	56%	0.053	92%	G
NM_001093553.1	Xenopus laevis MGC80785 protein (MGC80785)	38.2	38.2	46%	0.21	95%	U G
NM_001004907.1	Xenopus tropicalis aldehyde dehydrogenase 2 family (mitochondrial) mRNA	38.2	38.2	46%	0.21	95%	U G
XM_001666588.1	Caenorhabditis briggsae AF16 hypothetical protein (CBG2)	36.2	36.2	44%	0.82	95%	G
XM_001643835.1	Vanderwaltozyma polyspora DSM 70294 hypothetical protein (Kpo1_4)	36.2	36.2	44%	0.82	95%	G
XM_001264785.1	Neosartorya fischeri NRRL 181 GTP-binding protein YchF (NFIA_01582)	34.2	34.2	34%	3.3	100%	G
XM_001233054.1	PREDICTED: Gallus gallus armadillo repeat containing protein	34.2	34.2	34%	3.3	100%	U G
XM_418230.2	PREDICTED: Gallus gallus armadillo repeat containing protein	34.2	34.2	34%	3.3	100%	U G
XM_415171.2	PREDICTED: Gallus gallus aldehyde dehydrogenase 2 family (mitochondrial) mRNA	34.2	34.2	50%	3.3	92%	U G
XM_766467.1	Giardia lamblia ATCC 50803 inositol 5-phosphatase (GLP_630_47132)	34.2	34.2	34%	3.3	100%	G
XM_671948.1	Plasmodium berghei strain ANKA hypothetical protein (PB000013.02.0)	34.2	34.2	34%	3.3	100%	G
NM_173915.2	Bos taurus gastrin (GAS), mRNA	34.2	34.2	34%	3.3	100%	U G
XM_001492144.1	PREDICTED: Equus caballus similar to KIAA1432, (LOC100059612), mRNA	32.2	32.2	32%	13	100%	U G
NM_129488.3	Arabidopsis thaliana jacalin lectin family protein (AT2G39310) mRNA	32.2	32.2	32%	13	100%	U E G
NM_116485.3	Arabidopsis thaliana TOC159 (translocon outer membrane complex 159) mRNA	32.2	32.2	48%	13	91%	U E G

NCBI **blastn**: <http://www.ncbi.nlm.nih.gov/BLAST/>

BLAST search example (cont'd)

human

```
> ref|NM\_000690.2| UEG Homo sapiens aldehyde dehydrogenase 2 family (mitochondrial)
(ALDH2), nuclear gene encoding mitochondrial protein, mRNA
Length=2445
```

```
Score = 81.8 bits (41), Expect = 2e-14
Identities = 41/41 (100%), Gaps = 0/41 (0%)
Strand=Plus/Plus
```

41 bp perfect match

Expectation: 2×10^{-14}

```
Query 10 GGGTCAACTGCTATGATGTGTTTGGAGCCCAGTCACCCTTT 50
        |||
Sbjct 1847 GGGTCAACTGCTATGATGTGTTTGGAGCCCAGTCACCCTTT 1887
```

chimpanzee

```
> ref|XM\_509379.2| G PREDICTED: Pan troglodytes mitochondrial aldehyde dehydrogenase
2 (ALDH2), mRNA
Length=1967
```

```
Score = 81.8 bits (41), Expect = 2e-14
Identities = 41/41 (100%), Gaps = 0/41 (0%)
Strand=Plus/Plus
```

41 bp perfect match

Expectation : 2×10^{-14}

```
Query 10 GGGTCAACTGCTATGATGTGTTTGGAGCCCAGTCACCCTTT 50
        |||
Sbjct 1406 GGGTCAACTGCTATGATGTGTTTGGAGCCCAGTCACCCTTT 1446
```

thale cress (シロイヌナズナ)

```
> ref|NM\_129488.3| UEG Arabidopsis thaliana jacalin lectin family protein (AT2G39310)
mRNA, complete cds
Length=1610
```

```
Score = 32.2 bits (16), Expect = 13
Identities = 16/16 (100%), Gaps = 0/16 (0%)
Strand=Plus/Minus
```

```
Query 16 ACTGCTATGATGTGTT 31
        |||
Sbjct 283 ACTGCTATGATGTGTT 268
```



jackfruit

16 bp perfect match

Expectation : 13

not significant

PSI-BLAST

PSI (Position Specific Iterated) -BLAST

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”,
Nucleic Acids Res., 25(17), 3389-3402 (1997).



Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a **profile (or position specific scoring matrix, PSSM) 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" used to refine the profile. This iterative searching strategy results in increased sensitivity.

(from “PSI-BLAST tutorial”, NCBI)

PSI-BLAST

