今期から始まった遠隔授業の受講者の方々にも対応するため、早めにファイルを置いていますが、授業直前に内容を修正する可能性があります。2010年4月19日

#5
Homology Search from Databases

Topics:

Homology Search from Databases

- Amino Acid Substitution Matrix (PAM, BLOSUM)
- Similarity vs. Homology
- E-value, Bit Score
- P-value

Homology search from databases

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

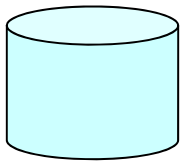
Approximated search methods are required.

SSEARCH (Smith-Waterman)

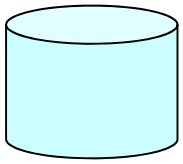
FASTA

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)

BLAST programs

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous 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	*(obsolete)*	
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 (mitoch human	81.8	81.8	82%	2e-14	100%	U E G
XM_509379.2	PREDICTED: Pan troglodytes mito chimpangee lehydrogenase	81.8	81.8	82%	2e-14	100%	G
XR_012809.1	PREDICTED: Macaca mulatta mitochondrial aldehyde de Macaca monkey	81.8	81.8	82%	2e-14	100%	U G
XM_001490910.1	PREDICTED: Equus caballus similar to horse dehydr	52.0	52.0	68%	1e-05	94%	U G
NM_001075367.1	Bos taurus similar to Aldehyde dehydrogenase, mitochondria cow	52.0	52.0	68%	1e-05	94%	U G
XM_848535.1	PREDICTED: Canis familiaris similar to Alde dog ydrogenase, mit	52.0	52.0	68%	1e-05	94%	U E G
XM_849411.1	PREDICTED: Canis familiaris similar to Aldehyde dehydrogenase, mit	52.0	52.0	68%	1e-05	94%	G
NM_009656.3	Mus musculus aldehyde dehydrogenase 2, mitochondrial (mouse	46.1	46.1	62%	9e-04	93%	U E G
NM_032416.1	Rattus norvegicus aldehyde dehydrogenase rat RNA >emb|X	46.1	46.1	62%	9e-04	93%	U E G
XM_845808.1	PREDICTED: Canis familiaris similar to Aldehyde dehydrogenase, mit	42.1	42.1	66%	0.013	90%	U G
XM_001257227.1	PREDICTED: Bos taurus similar to Aldehyde dehydrogenase, mitocho	40.1	40.1	56%	0.053	92%	G
NM_001093553.1	Xenopus laevis MGC80785 protein (MGC80 African frog	38.2	38.2	46%	0.21	95%	U G
NM_001004907.1	Xenopus tropicalis aldehyde dehydrogenase 2 family (mitochondrial)	38.2	38.2	46%	0.21	95%	U G
XM_001666588.1	Caenorhabditis briggsae AF16 hypothetical protein (CBG2 worm	36.2	36.2	44%	0.82	95%	G
XM_001643835.1	Vanderwaltozyma polyspora DSM 70294 hypothetical protein (Kpol 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 hen	34.2	34.2	34%	3.3	100%	U G
XM_418230.2	PREDICTED: Gallus gallus armadillo repeat containing 6, transcript va	34.2	34.2	34%	3.3	100%	U G
XM_415171.2	PREDICTED: Gallus gallus aldehyde dehydrogenase 2 family (mitoch	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 >emb|X16581.1|BTPPGAST Bovine	34.2	34.2	34%	3.3	100%	U G
XM_001492144.1	PREDICTED: Equus caballus similar to KIAA1432, (LOC100059612),	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 1	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

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

41 bp perfect match
Expectation: 2×10^{-14}

chimpangee

```
>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

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

41 bp perfect match
Expectation : 2×10^{-14}

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

Similarity vs. Homology

Similarity:

The extent to which sequences are related.
Similarity makes no statement about descent from a common ancestor.
Similarity is a direct measurement and can be discussed in terms of percentages.

Homology:

Sequence similarity that can be attributed to descent from a common ancestor.
Homology is a matter of opinion, not directly measurable or observable.

(advanced)
Homologous = Orthologous or Paralogous
Orthologous - Homologous sequences in different species.
Paralogous - Homologous sequences in the same species that arose by means of gene duplication.

cited from http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Similarity/simsrch3ahomology.html

Amino Acid Substitution Matrix

- PAM (Point / Percent Accepted Mutation)

  M. Dayhoff and R. Schwartz: “Matrices for detecting distant relationship”, Atlas of Protein Sequences, 353-358 (1979).

Obtained from global alignment of 71 groups of closely related proteins.
PAM = a unit of evolution (1 PAM = 1 point mutation/100 amino acids).
Different PAM matrices are derived from the PAM 1 matrix by matrix multiplication.

$$\text{log-odds} = \log_2 \frac{\text{(freq. of changes in homologous proteins)}}{\text{(freq. of pairs aligned by chance)}}$$

PAM250 > PAM120 > PAM90 >> PAM1

used for detecting remote sequences

Amino Acid Substitution Matrix

- BLOSUM (BLOck SUbstitution Matrix)

S. Henikoff and J. G. Henikoff: "Amino acid substitution matrices from protein blocks", Proc. Nat'l Acad. of Sci. USA, 89(22):10915-10919 (1992).

Obtained from local alignment of conserved sequence blocks (from Blocks database)
BLOSUM Matrices derived from Blocks whose alignment corresponds to the BLOSUM
Matrix number *e.g.* BLOSUM 62 is derived from Blocks containing $\leq$62% identity.

$$\text{log-odds} = \log_2 \frac{\text{(freq. of substitution in a block)}}{\text{(freq. of pairs aligned by chance)}}$$

BLOSUM30 > BLOSUM62 > BLOSUM80

used for detecting remote sequences

Substitution Matrix and Probability

- Odds score (Odds ratio オッズ比)

$$\text{Odds}(A) = \frac{P(A)}{P(\text{not } A)} = \frac{P(A)}{1 - P(A)}$$

$$P(A) = \frac{\text{Odds}(A)}{1 + \text{Odds}(A)}$$

- Log-odds score

$$\text{Log-odds}(A) = \log_e \frac{P(A)}{P(\text{not } A)}$$

Substitution Matrix and Probability

- Log-odds score

"bidding" for homology

"bidding" for no homology

$$S = \log_e \frac{P(x, y \mid \text{Match})}{P(x, y \mid \text{Random})} = \log_e \frac{\Pi_i \; q(x_i, y_i)}{\Pi_i \; p(x_i) \times \Pi_i \; p(y_i)}$$

assuming position independence

$$= \log_e \Pi_i \frac{q(x_i, y_i)}{p(x_i)\, p(y_i)}$$

$$= \Sigma_i \; s(x_i, y_i)$$

i

x: G K D — D …

y : G — E V D …

where $s(a, b) = \log_e \dfrac{q(a, b)}{p(a) \times p(b)}$ → **substitution score matrix**

Substitution Matrix and Probability

- arbitrary scaling factor λ

$$s(a, b) = (1/\lambda) \cdot \boldsymbol{log}_e \frac{q(a, b)}{p(a) \times p(b)}$$

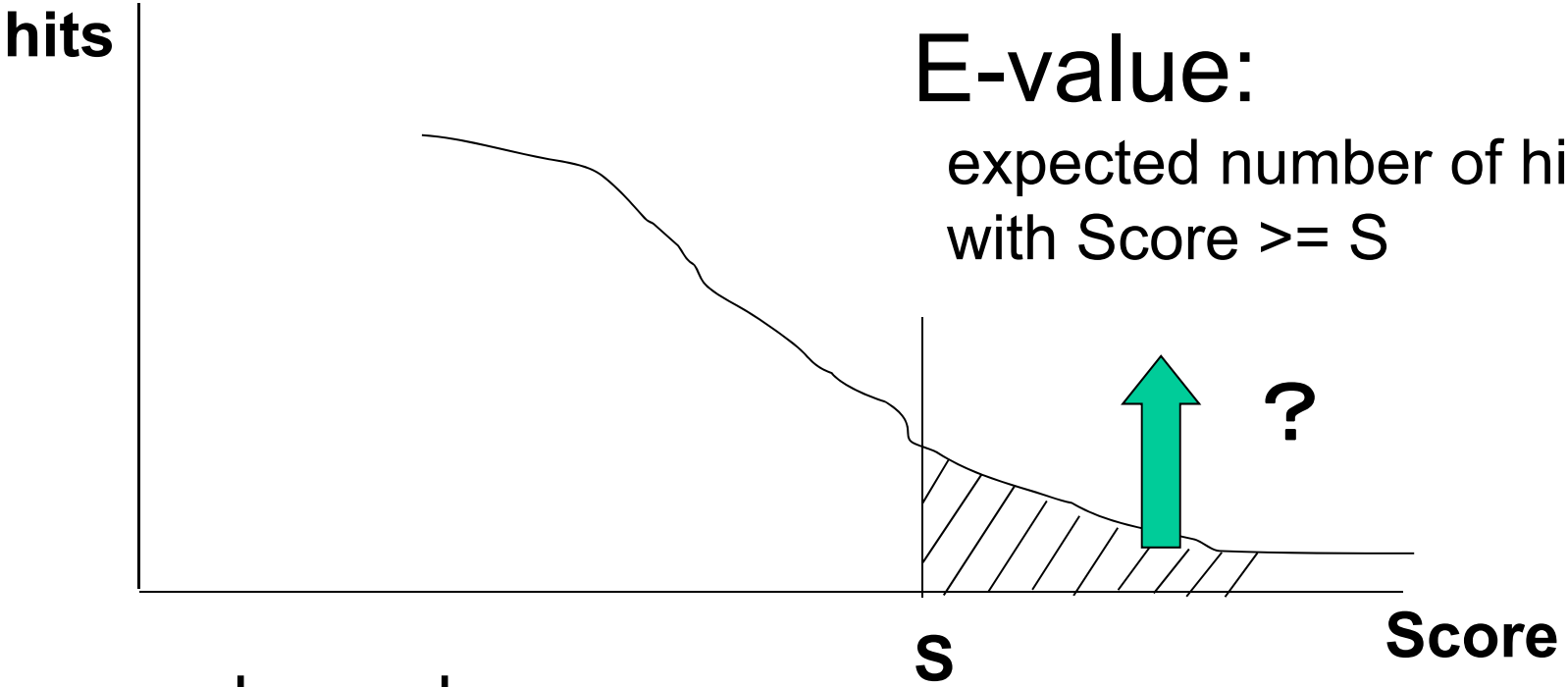
an arbitrary scaling factor (λ) can be introduced, without less of generality

$$q(a, b) = p(a)\, p(b)\ \boldsymbol{exp}\,(\lambda \cdot s(a, b))$$

Because q(a,b) is a probability,

$$\Sigma_{a,b}\ q(a, b) = \Sigma_{a,b}\ p(a)\, p(b)\ \boldsymbol{exp}\,(\lambda \cdot s(a, b)) = 1$$

E-value for Global Alignment

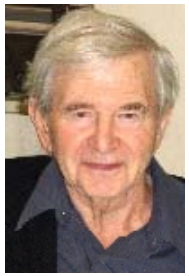


depends on

1) sequence compositions
2) scoring systems (matrix, gap penalty)
3) random model

E-value for (Ungapped) Local Alignment

Karlin, S. & Altschul, S.F. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes."
Proc. Natl. Acad. Sci. USA, 87:2264-2268 (1990).



Karlin

Altschul

$$E = K\,m\,n\,\exp(-\lambda S)$$

Maximum (minimum) of independent identically distributed (i.i.d.) random variables

→ → **Extreme Value Distribution** （極値分布）

cf.

Sum of independent identically distributed (i.i.d.) random variables

→ → Normal Distribution

E-value for (Ungapped) Local Alignment

E-value:
expected number of hits
with Score >= S

Extreme Distribution

$$E = 1 - \exp(-K m n \exp(-\lambda S))$$

$$\fallingdotseq K m n \exp(-\lambda S)$$

for x<<1
$\exp(-x) \fallingdotseq 1 - x$

m: length of query sequence
n: total length of database sequences
K: a constant determined by sequence composition
λ: score scaling factor

Memo:

1) **Expected number E increases proportional to the m × n product.**
2) **The number of high score hits decreases exponentially with S.**

Bit Score (Normalized Score)

Bit Score: $S' = -\log_2 (K \exp(-\lambda S))$

$$= \frac{\lambda S - \log_e K}{\log_e 2}$$

E-value can be simply represented as:

$$E = m\,n\,2^{-S'}$$

Raw Score: S

⇕

Bit Score : S'

memo:

$S' = 10$ (bits) $\rightarrow E \fallingdotseq m\,n\,10^{-3}$

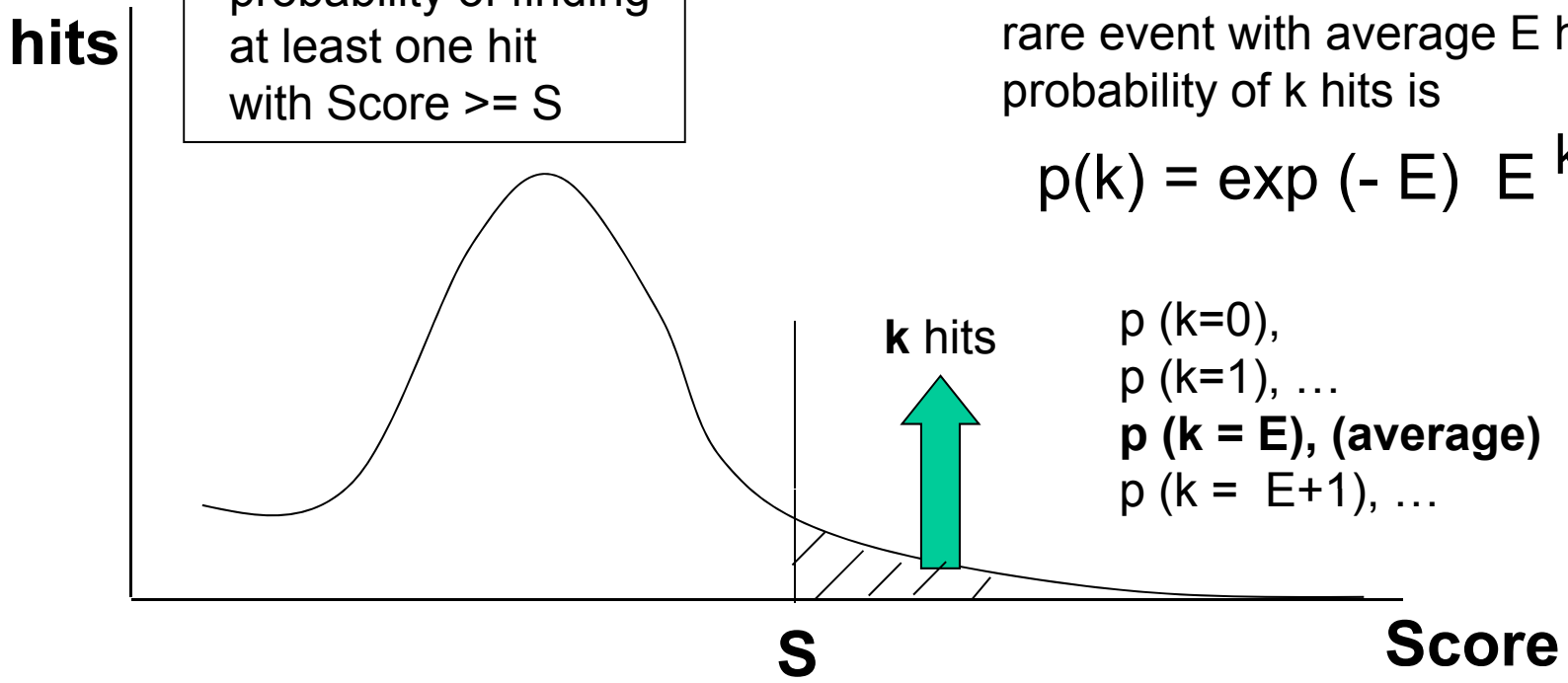
$S' = 20$ (bits) $\rightarrow E \fallingdotseq m\,n\,10^{-6}$

$S' = 30$ (bits) $\rightarrow E \fallingdotseq m\,n\,10^{-9}$

$S' = 100$ (bits) $\rightarrow E \fallingdotseq m\,n\,10^{-30}$

P-value for (Ungapped) Local Alignment

P-value:
probability of finding
at least one hit
with Score >= S



Poisson distribution

rare event with average E hits.
probability of k hits is

$$p(k) = \exp(-E)\ E^{k} / k!$$

p (k=0),
p (k=1), …
p (k = E), (average)
p (k = E+1), …

× sum

$$P = \text{Prob}(k > 0; S)$$
$$= 1 - p(k=0)$$
$$= 1 - \exp(-E)\ E^{0} / 0!$$
$$= 1 - \exp(-E)$$

memo:

E = 3 → P = 1 - exp (-3) ≒ 0.95
E = 0.01 → P = 1 - exp (-0.01) ≒ 0.01
E < 0.01 → P ≒ E