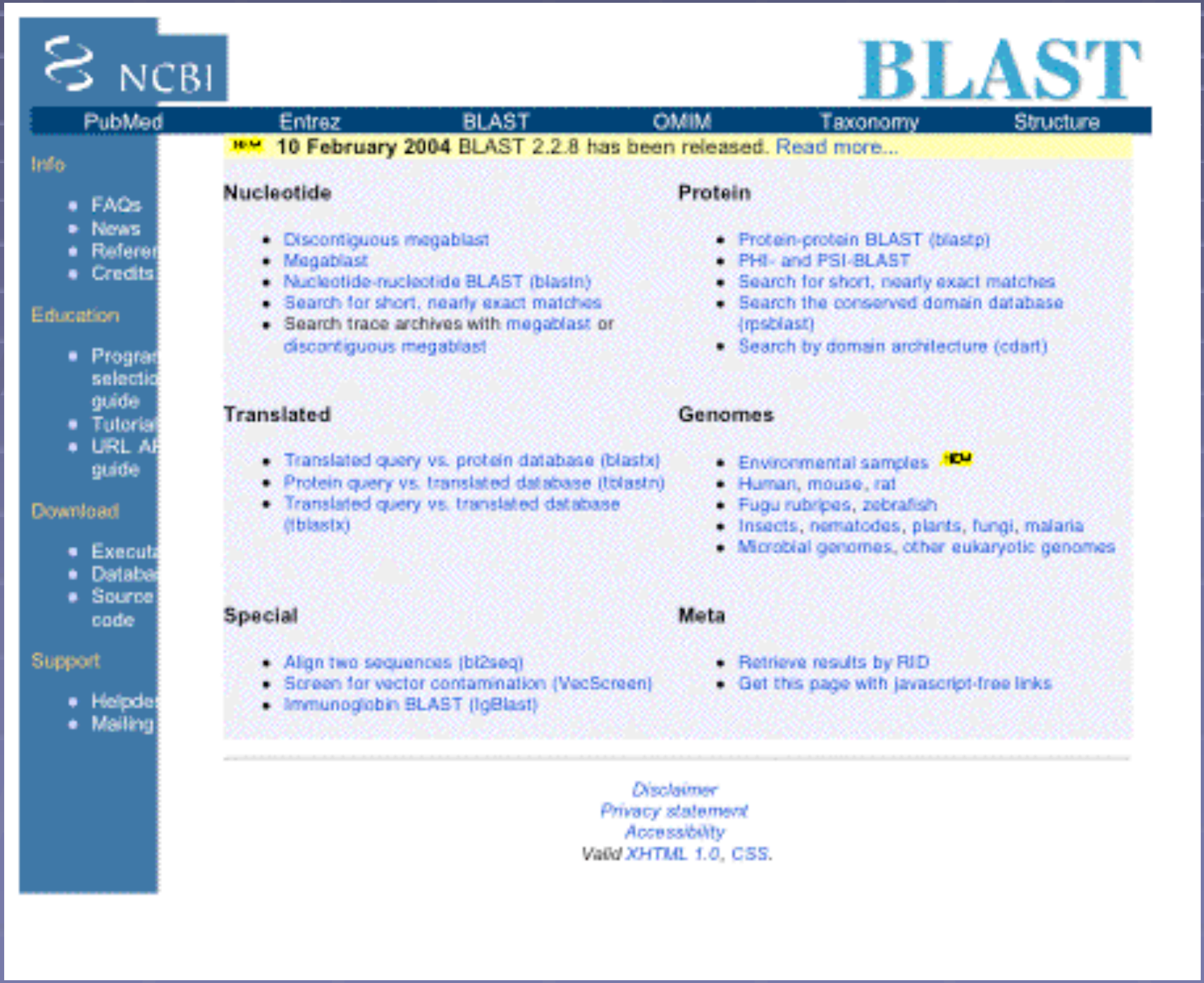
Searching Sequence databases 1: Blast



10 February 2004 BLAST 2.2.8 has been released. Read more...

Info

- FAQs
- News
- Referer
- Credits

Education

- Progra selectio guide
- Tutoria
- URL AF guide

Download

- Executa
- Databa
- Source code

Support

- Helpde
- Mailing

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Environmental samples NEW
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.

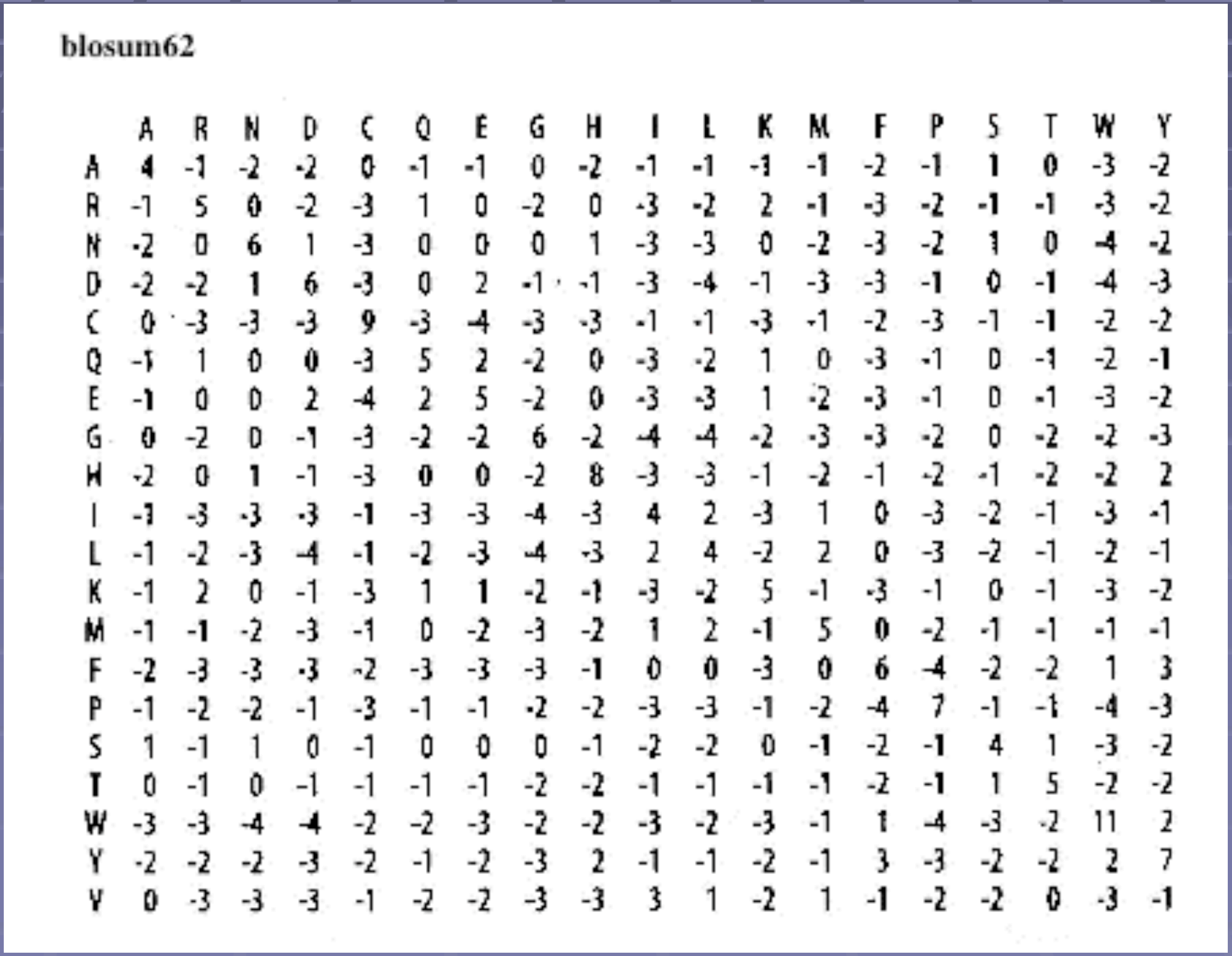
Quiz

- Expectation:
  - Discrete variable X takes values 1,2,3
    - Pr[X=1]=0.2
    - Pr[X=2]=0.6
    - Pr[X=3]=0.2
    - E(X)?
  - X is one of n values $X_1 \ldots X_{n,}$ and they are equi-probable.
    - E(X)?
  - How is a scoring matrix used?

Blosum62 (PAM)

ARSTW
AASTD

Score=8



blosum62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1

Matrix Multiplication

- Consider 3 nXn matrices A_1, A_2, A_3
- Let $A_3 = A_1A_2$

$$A_3\left[i,j\right] = \sum_{k=1}^{n} A_1[i,k]A_2[k,j]$$

PAM again

- Two sequences are 1 PAM apart if they differ in 1% of residues
- Two sequences s and t are k PAMs apart if
  - There exists sequence s’ such that
    - s and s’ are 1 PAM apart
    - s’ and t are k-1 PAMs apart

$$PAM_2['A','L'] = \sum_{X='A'}^{'Y'} PAM_1['A',X]PAM_1[X,'L']$$

$$PAM_3['A','L'] = \sum_{X='A'}^{'Y'} PAM_1['A',X]PAM_2[X,'L']$$

$$PAM_{250}['A','L'] = \sum_{X='A'}^{'Y'} PAM_1['A',X]PAM_{249}[X,'L']$$

$$PAM_2 = PAM_1^2$$

$$PAM_3 = PAM_1 \cdot PAM_2 = PAM_1^3$$

$$PAM_{250} = PAM_1 \cdot PAM_{249} = PAM_1^{250}$$

P-value computation

- How significant is a score? What happens to significance when you change the score function
- A simple empirical method:
  - Compute a distribution of scores against a random database.
  - Use an estimate of the area under the curve to get the probability.
  - OR, fit the distribution to one of the standard distributions.

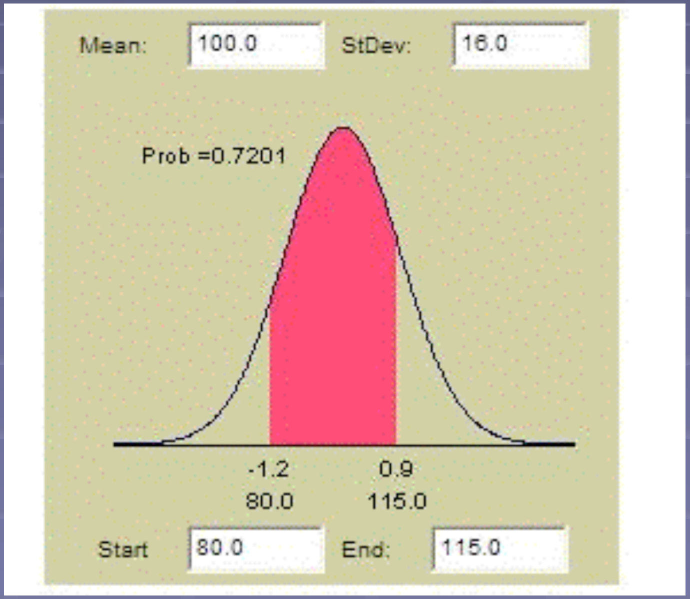
Z-scores for alignment

- Initial assumption was that the scores followed a normal distribution.
- Z-score computation:
  - For any alignment, score S, shuffle one of the sequences many times, and recompute alignment. Get mean and standard deviation

$$Z_S = \frac{S \square \square}{\square}$$

  - Look up a table to get a P-value

Normal Distribution



Mean: 100.0 StDev: 16.0

Start 80.0 End: 115.0

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$

Blast E-value

- 1990, Karlin and Altschul showed that ungapped local alignment scores follow an exponential distribution
- Practical consequence:
  - Longer tail.
  - Previously significant hits now not so significant

Exponential distribution

- Random Database, Pr(1) = p
- What is the expected number of hits to a sequence of k 1's

$$(n-k)p^k \approx ne^{k\ln p} = ne^{-k\ln\left(\frac{1}{p}\right)}$$

- Instead, consider a random binary Matrix. Expected # of diagonals of k 1s

$$E = (n-k)(m-k)p^k \approx nme^{k\ln p} = nme^{-k\ln\left(\frac{1}{p}\right)}$$

- As you increase k, the number decreases exponentially.
- The number of diagonals of k runs can be approximated by a Poisson process

$$\Pr[u \text{ hits}] = \frac{\square^u e^{\square\square}}{u!}$$

$$\Pr[u > 0] = 1 \square e^{\square\square}$$

- In ungapped alignments, we replace the coin tosses by column scores, but the behaviour does not change (Karlin & Altschul).
- As the score increases, the number of alignments that achieve the score decreases exponentially

Blast E-value

- Choose a score such that the expected score between a pair of residues < 0
- Expected number of alignments with a particular score

$$E = Kmne^{-\lambda S} = mn2^{-\left(\frac{\lambda S - \ln K}{\ln 2}\right)}$$

$$\Pr(\#\,\mathrm{hsp} > 0) = 1 - e^{-Kmne^{-\lambda S}}$$

- For small values, E-value and P-value are the same

Keyword Search