

Wright State University

CORE Scholar

Computer Science and Engineering Faculty
Publications

Computer Science & Engineering

2003

Searching Sequence Databases

Dan E. Krane

Wright State University - Main Campus, dan.krane@wright.edu

Michael L. Raymer

Wright State University - Main Campus, michael.raymer@wright.edu

Follow this and additional works at: <https://corescholar.libraries.wright.edu/cse>



Part of the [Computer Sciences Commons](#), and the [Engineering Commons](#)

Repository Citation

Krane, D. E., & Raymer, M. L. (2003). Searching Sequence Databases. .

<https://corescholar.libraries.wright.edu/cse/383>

This Presentation is brought to you for free and open access by Wright State University's CORE Scholar. It has been accepted for inclusion in Computer Science and Engineering Faculty Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

BIO/CS 471 – Algorithms for Bioinformatics

Searching Sequence Databases

Database Searching

- How can we find a particular short sequence in a database of sequences (or one HUGE sequence)?
- Problem is identical to local sequence alignment, but on a much larger scale.
- We must also have some idea of the *significance* of a database hit.
 - Databases always return some kind of hit, how much attention should be paid to the result?

BLAST

- BLAST – Basic Local Alignment Search Tool
- An approximation of the Needleman & Wunsch algorithm
- Sacrifices some search sensitivity for speed

Scoring Matrices

- DNA

- Identity
- Transition/Transversion

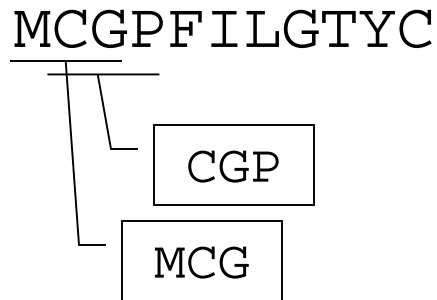
- Proteins

- PAM
- BLOSUM

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

The BLAST algorithm

- Break the search sequence into *words*
 - $W = 3$ for proteins, $W = 12$ for DNA



MCG, CGP, GPF, PFI, FIL,
ILG, LGT, GTY, TYC

- Include in the search all words that score above a certain value (T) for any search word

<u>MCG</u>	<u>CGP</u>	
MCT	MGP	...
MCN	CTP	
...	...	

**This list can be
computed in linear
time**

The Blast Algorithm (2)

- Search for the words in the database
 - Word locations can be precomputed and indexed
 - Searching for a short string in a long string
 - Regular expression matching: FSA
- HSP (High Scoring Pair) = A match between a query word and the database
- Find a “hit”: Two non-overlapping HSP’s on a diagonal within distance A
- Extend the hit until the score falls below a threshold value, X

The BLAST Search Algorithm

query word ($W = 3$)

Query: GSVEDTTGSQSLAALLNKCKT**TPQG**QRLVNQWIKQPLMDKNRIEERLNLVEAFVE DAELRQTLQEDL

neighborhood
words

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

neighborhood
score threshold
($T = 13$)

Query: 325 SLAALLNKCKT**TPQG**QRLVNQWIKQPLMDKNRIEERLNLVEA 365
 +LA++L+ TP G R++ +U+ P+ D + ER + A
 Sbjct: 290 TLASVLDCTV**PMG**SRLMKRULHMPVRDTRVLLERQQTIGA 330

High-scoring Segment Pair (HSP)

Results from a BLAST search

NCBI CD-Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?RID=1066068742-2377-951781.BLASTQ3> Go

gnl|CDD|5811, LOAD_USPA, USPA, An ATP binding domain seen as a stand alone in USPA.

CD-Length = 135 residues, 100.0% aligned
Score = 90.5 bits (224), Expect = 8e-20

Query: 6 **KKILYPTDFSETAEIALKHVKAFKTLKAEVILLHVIDEREIKKRDIFSLLLGVAGLNKS** 65
Sbjct: 1 **KKILVAIDGSPSEKALRWAVDLAKRRGAE LILLHVI PPSVSTAASPALDL**----- 51

Query: 66 **VEEFENELKNKLTEEAKNKMNIKKELEDVGFVKVDIIVVGIPHEEIVKIAEDEGV DIII** 125
Sbjct: 52 -----ALLLEEALKLLLEEALLEEEAGVKIDVEVEEGSPAEAILLEAESNADLIV 102

Query: 126 **MGSHGKTNLKEILLGSVTENVIKKSNKPVLVVK** 158
Sbjct: 103 **VGSRGRGGLRRLLLGSVSEKVLKAPCPVLVVR** 135

gnl|CDD|10459, COG0589, UspA, Universal stress protein UspA and related nucleotide-binding proteins [Signal transduction mechanisms]

CD-Length = 154 residues, 100.0% aligned
Score = 84.2 bits (207), Expect = 6e-18

Query: 1 **MSVMYKKILYPTDF-SETAEIALKHVKAFKTLKAEVILLHVIDEREIKKRDIFSLLLGV** 59
Sbjct: 1 **MPAMYKKILVAVDVGSEAAEKAL EAVALAKRLGAPLILLVVIDPLEPT**-----A 50

Query: 60 **AGLNKSV EEFENELKNKLTEEAKNKMNIKKELEDVG-FKVDIIVVGIPH-EEIVKIAE** 117
Sbjct: 51 **LVSVALADAPIPLSEEELEEEAEELLAEAALAEAGVPVVEVEVVEGSPSAEEILELAE** 110

Query: 118 **DEGV DIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKN** 161
Sbjct: 111 **EEDADLIVVGSRGRSGLSRLLLGSVAEKVLRHAPCPVLVVRSEG** 154

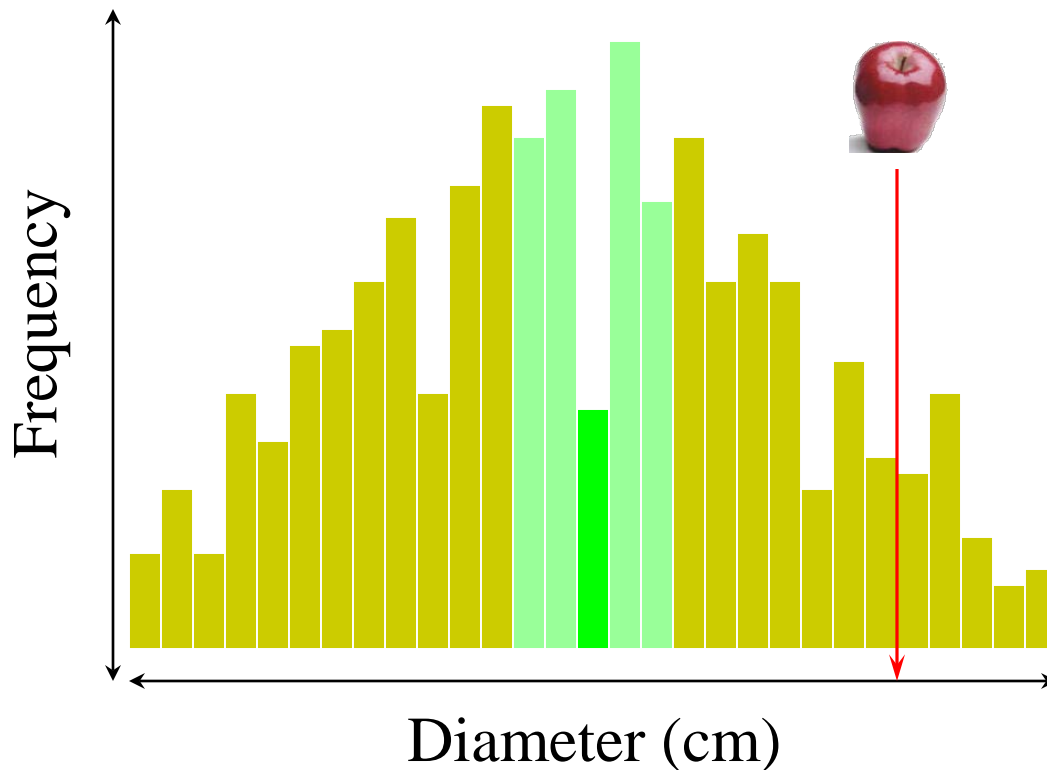
Internet

Search Significance Scores

- A search will *always* return some hits.
- How can we determine how “unusual” a particular alignment score is?
 - ORF’s
 - Assumptions

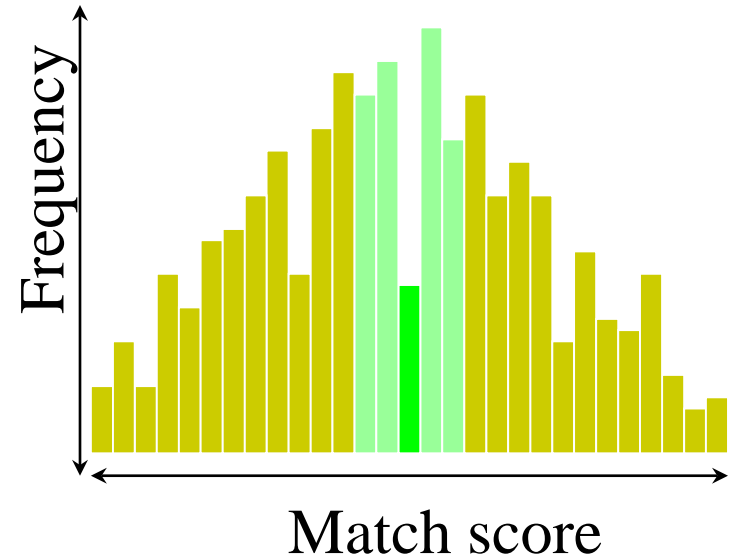
Assessing significance requires a *distribution*

- I have an apple of diameter 5". Is that unusual?



Is a match significant?

- Match scores for aligning my sequence with *random sequences*.
- Depends on:
 - Scoring system
 - Database
 - Sequence to search for
 - Length
 - Composition
- How do we determine the *random sequences*?



Generating “random” sequences

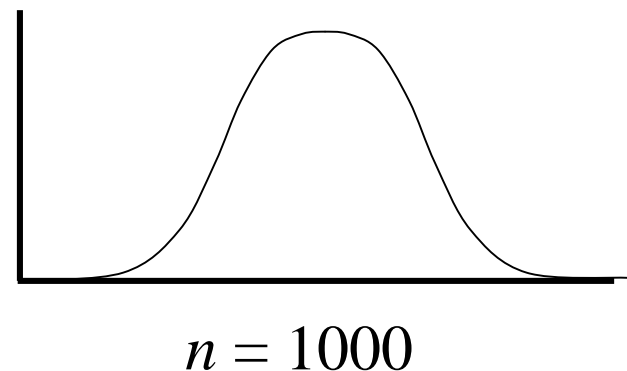
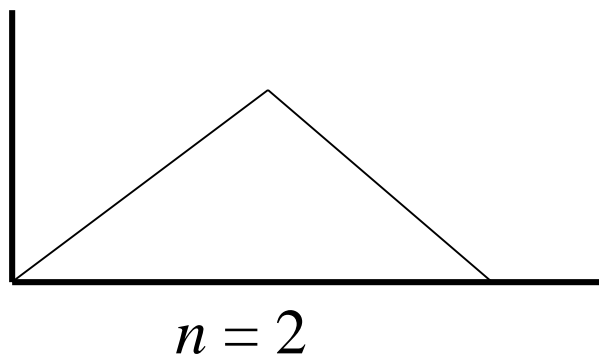
- Random uniform model:

$$P(G) = P(A) = P(C) = P(T) = 0.25$$

- Doesn't reflect nature
- Use sequences from a database
 - Might have genuine homology
 - We want unrelated sequences
- Random shuffling of sequences
 - Preserves composition
 - Removes true homology

What distribution do we expect to see?

- The mean of n random (i.i.d.) events tends towards a Gaussian distribution.
 - Example: Throw n dice and compute the mean.
 - Distribution of means:

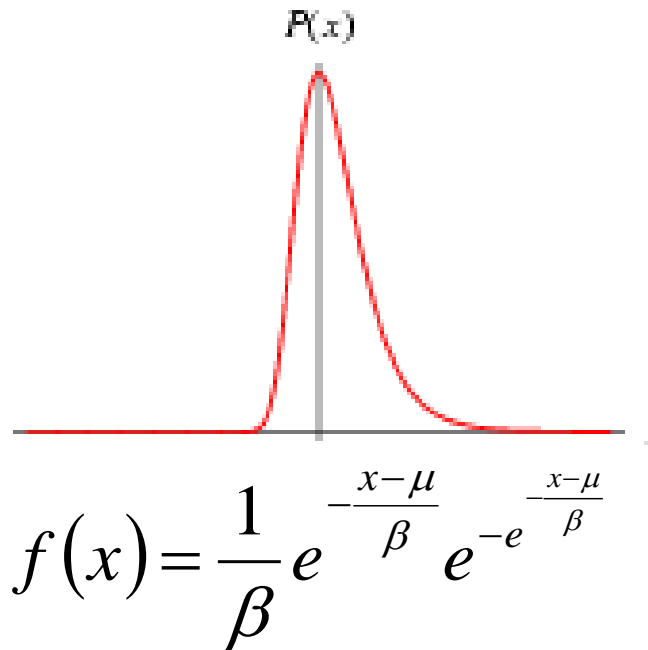


The extreme value distribution

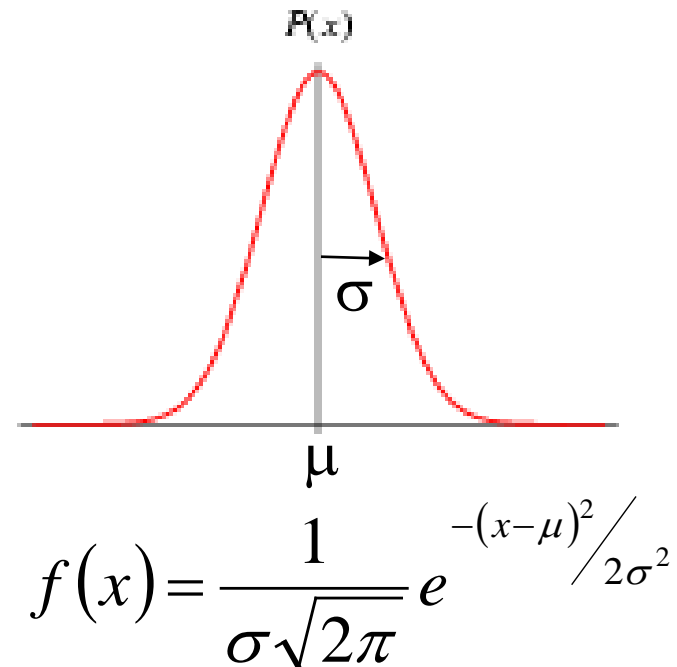
- This means that if we get the match scores for our sequence with n other sequences, the mean would follow a Gaussian distribution.
- The **maximum** of n (i.i.d.) random events tends towards the **extreme value distribution** as n grows large.

Comparing distributions

Extreme Value:



Gaussian:



Determining P-values

- If we can estimate β and μ , then we can determine, for a given match score x , the probability that a random match with score x or greater would have occurred in the database.
- For sequence matches, a scoring system and database can be parameterized by two parameters, K and λ , related to β and μ .
 - It would be nice if we could compare hit significance without regard to the database and scoring system used!

Bit Scores

- The expected number of hits with score $\geq S$ is:

$$E = Kmn e^{-\lambda S}$$

- Where m and n are the sequence lengths
- Normalize the raw score using:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- Obtains a “bit score” S' , with a *standard set of units*.
- The new E-value is: $E = mn 2^{-S'}$

P values and E values

- Blast reports *E*-values
- $E = 5$, $E = 10$ versus $P = 0.993$ and $P = 0.99995$
- When $E < 0.01$ *P*-values and *E*-values are nearly identical

BLAST parameters

- Lowering the neighborhood word threshold (T) allows more distantly related sequences to be found, at the expense of increased noise in the results set.
- Raising the segment extension cutoff (X) returns longer extensions for each hit.
- Changing the minimum E -value changes the threshold for reporting a hit.