

Bioinformatics: Life Sciences on Your Computer

E Value Mathematics

This is supplemental reading to the video on E Values.

Bits

BLAST developer Stephen Altschul at NCBI used bits to measure normalized scores. He could have used logs and powers of 10, but he chose to use powers of 2. For instance, the number 128 is equivalent to 2^7 or 7 bits.

Search space (bits)

The video lecture focused on the size of the database. The search space is also partially determined by the length of the query sequence. To calculate the search space, multiply query length by database length. If the query is 256 amino acids (8 bits) and the database has 30 million amino acids (~25 bits), the search space size, N, is 33 bits ($2^8 \times 2^{25}$). Exponents are additive when multiplied.

E value

The E value is the best way to determine homology. The video lecture covers E values and the discussion topics below in a little more depth.

Discussion topic: Suppose you played a big jackpot lottery. For convenience sake, let's say that the probability of winning with a single ticket is 1/10,000,000. Are you likely to win with one ticket? Now let's say that for one drawing, 90,000,000 tickets are sold. Now do you think it's likely that someone wins? On average, how many winners would you expect with that many tickets sold?

Discussion topic: Most large scale probability problems have answers very close to $p = 0$ or $p = 1$. One of the more famous problems is the monkey problem. It goes something like this. All of the monkeys in the universe (the # is given in the problem) are given keyboards. They type on the keyboards at a given rate (keystrokes per minute) and they work continuously for one year. What is the likelihood that a monkey would type the word "Hamlet?" What would you conclude about the monkey who typed that word? Now what is the likelihood that a monkey would type Act I, Scene I of the play, "Hamlet," even allowing for some misspellings? What would you conclude about that monkey? How do these details relate to BLAST?

Exercise/Discussion topic: If the search space is 36 bits, then calculate the E value if S' is: 30 bits? 35 bits? 40 bits? 50 bits? 75 bits? 100 bits? Please do not answer them all in one post.

Inferring homology

From BLAST or other alignments and searches, one can **never infer non-homology**. Unrelated sequences score similarly to random sequences. If the E value suggests that the similarity score is non-random, then sequences are NOT unrelated, Therefore, statistically significant similarity scores means that sequences are related!

If $E < 1 \times 10^{-4}$, then the sequences are homologous.

Confirming homology

Sometimes lab studies are used to confirm homology, especially of sequences with E values that are

"borderline." A function/structure study in the wet lab provides clues and context. On the computer side, one way to confirm homology is to "connect the dots." Sequences are not inherited from two parents. Human inheritance is based on receiving two sequences, one from each parent, but each **sequence** came from only one parent. Therefore, in the sequence world, if sequence A is related to sequence B, and sequence B is related to sequence C, then sequence A must be related to sequence C.

So is the alignment between A and C produces an E value of 2×10^{-3} , that is not strong enough to prove homology. However, if A & B align with a very strong E value, and B and C align with a very strong E value, then we can say that A and C are related sequences.

Another approach is to find a reasonable multiple alignment with sequences from a protein family. Looking at the genomic context also provides reasonable clues.

Discussion topic: Discuss wet lab and dry lab methods for assessing homology. Explain what "connect the dots" means."

RBP4, progesterone

An alignment between human RBP4 (retinol-binding protein) and human progesterone has an E value of 0.27. Not a great E value, but remember, we can never infer non-homology. The two proteins are about the same size, both have the lipocalin signature (GXW) and they can be multiply aligned. Both are soluble hydrophilic carrier proteins with similar 3-D structure. Although the E value does not show it, all of the other evidence suggests that these two sequences are related.

Score = 33.9 bits (76), Expect = 0.27, Method: Composition-based stats. Identities = 29/114 (25%), Positives = 50/114 (43%), Gaps = 8/114 (7%)									
Query	24	VSSPRVKENPDKARFSGTWYAMAKKDPEGLFLQDNIVAEPVSDETG-QMSATAKGRVRL	82						
		+S+ + K NPD +F+GTW +A + AE + Q +A A R L							
Sbjct	33	ISTIQPKANPDAQQFAGTWLLVAVGSACRFLQEQGHRAEATTLHVA PQGTAMAVSTPRKL	92						
Query	83	NNWDVCADMVGTFTDTEPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQY	136						
		+ +C + + DT +F ++ G +G + +TDY ++AV Y							
Sbjct	93	DG--ICWQVRQLYGD TGVLGRFLLQARGA-----RGAVHVVVAETDYQSPAVLY	139						

Discussion topic: Are these orthologs or paralogs? How does this fit into last week's discussion about orthologs and paralogs and which are the closer relatives?

Adjusting results

Too Many Results

- Smaller database
- Limit by organism
- Use portion of query
- Limit to one domain
- Adjust scoring matrix
- Adjust E value cutoff

Too Few Results

- Remove limits
- Try different PAM/BLOSUM
- Try FASTA or PSI-BLAST

