

Bioinformatics: Life Sciences on Your Computer

Week 2 Introduction

This reading is to introduce the Week 2 material.

BLAST - what is it?

When people think of bioinformatics, BLAST searches often come to mind quickly. I have a YouTube tutorial on BLAST--highly recommended. BLAST is the Basic Local Alignment Search Tool. Here's how it works:

1. Start with a DNA or protein sequence.
2. Search a large database of sequences
3. Output sequences in that database that are most similar in sequence to the sequence with which you started.

Sequences that are very similar are likely related, meaning that they were derived from the same ancestor sequence.

BLAST - why do it?

The primary reason to do a BLAST search is to find homologs, meaning related sequences. how and why is that important? Here are some reasons.

1. **Finding the closest relative in another organism.** If you are studying human disease-related genes, it could be useful to have an animal model for how that gene/protein works.
2. **Finding related sequences within an organism.** Gene duplication leads to related genes in the same species. Often those "paralogs" adopt different functions, but maintain some structural similarity, that could provide insight into how that structure is related to function.
3. **Identifying a DNA or RNA sequence.** With newly sequenced genomes, often a first step is finding the genes and assigning functions to those genes. BLAST searches help find some gene functions very quickly.
4. **Discovering new genes.** By starting with genomic DNA, translated BLAST searches may help find a protein that has not yet been discovered, although it may be related to known proteins.
5. **Finding sequence variants.** This is very useful with viruses such as HIV that mutate quickly.
6. **Discovering how organisms are related to each other.** We used to classify organisms by eyesight. It was easy to put plants and animals in different categories. Microscopy allowed scientists to discover two fundamentally different cell types: prokaryotic (bacteria and archaea) and eukaryotic. At the molecular level, we can use related sequences to study phylogeny, the study of how organisms are related.
7. **Finding amino acids important to function.** When you find a family of related proteins, often the entire protein family has some amino acids that are always the same in every sequence. Those amino acids are likely important to function, because nature does not tolerate mutation in those positions.

Discussion topic: What is your opinion on these and other reasons to find related sequences?

This week in Bioinformatics

We have several videos, readings and practice problems this week. They can be divided into three categories.

1. Pairwise Sequence Alignment
2. BLAST (and the related algorithm FASTA)
3. PSI-BLAST (and the related Delta-BLAST)

If BLAST is critical, why cover the other two topics?

Well, to find a related sequence in a database, it is absolutely necessarily to optimally align your starting (query) sequence with sequences in a database. Pairwise alignment is an essential building block of BLAST. And like most bioinformatics programs, BLAST is not 100% sensitive. It will not find all related sequences. For those harder to find related sequences, there is PSI-BLAST.

Goals for this week

Perform pairwise alignment.

Record and explain the identity and similarity between to aligned sequences.

Contrast global, local alignment

Describe the substitution matrix and gap penalties.

Run BLAST searches to find related sequences.

Use the E value to evaluate homology.

Calculate an E value based on search space and score, in bits.

Perform PSI-BLAST to find putative homologs.

That sounds like a lot, but lets work in small chunks and I think you will find it very doable. Good luck!