GETTING STARTED

Using the Website

Direct your browser on your phone, computer, or tablet to the following website:
http://www.kelleybioinfo.org

There you will see the homepage, as shown at right.

Touching or clicking an icon (e.g., “Alignment”) will take you to a new page that has tools related to the icon topic. The Alignment, Motifs, and Phylogeny buttons teach algorithms and tools for many types of sequence analysis with DNA, RNA, and proteins. The Protein and RNA buttons focus on algorithms for predicting structural features of the functional macromolecules, while the Probability button teaches how to generate substitution matrices.

Example: The Alignment Page

Clicking or touching the Alignment button will take you to the following page, which begins with the BLAST algorithm interactive tool. All the pages use this basic design.
General features

The “I” Button stands for “Interactive”. This is a tutorial on how to use the interactive algorithm learning tool.

Information on the interactive learning tool
Tutorials and test data for online bioinformatics software

Introductory information on the website, Bioinformatics, and Basic Biology.

Info on how to read different data formats.

Tutorials on more commonly used Bioinformatics websites.

Primers on getting started with Bioinformatics programming.

While most of the pages look like the Alignment page, the Basics page is organized differently and mostly contains information and tutorials.

How To Use This Book

I will assume you are familiar with how to read/use a book, but remember that the physical book is meant to be used in conjunction with the online component. Throughout the text you will be directed to online modules via URLs and QR codes. The online material is not supplemental, but is a critical portion of this hypertextbook.
Motivation

The purpose of this activity is to teach the basic concepts behind the BLAST algorithm and how to use a web-based implementation of this algorithm to analyze DNA and protein sequence data. BLAST (Basic Local Alignment Search Tool) is a fast computational method for making sequence alignments. Sequence alignments are a critical part of bioinformatics. Computational methods for making pairwise alignments of biological molecules (DNA, RNA, or protein) were some of the very first bioinformatics algorithms developed. Among other things, sequence alignments allow researchers to determine the organisms from which the molecule came (human, oyster, pine tree, bacterium, etc.) and predict the cellular function of biological molecules based only on their sequence. For example, BLAST can report with high confidence that the protein sequence YNFGSGSAYGGSFGGVDGLLAGGKATMONL is keratin from the domestic dog hair found on your sofa.

BLAST was created to speed up the process of making sequence alignments. Full pairwise sequence alignment methods (see Chapter 03) are too computationally intensive to handle the alignment of thousands or millions of sequences. BLAST speeds up this process by "chopping up" an input sequence into smaller bits and matching these smaller bits to millions of different sequences. The algorithm then attempts to extend the sequence alignment to make a full alignment. Then the algorithm ranks the sequence alignments, and the longest alignment with the fewest mismatches wins! In bioinformatics parlance we call the good matches "hits," and the best ones are "best hits" or "top hits." BLAST is a heuristic method, meaning that it is not guaranteed to find the optimal alignment, but it is much faster than more stringent approaches.

Learning Objectives

1. Know the basic purpose and utility of the BLAST computational method (Motivation).
2. Understand the concepts behind the BLAST algorithm (Concepts and Exercises).
3. Correctly solve sequence-matching problems based on the BLAST algorithm (Concepts and Exercises).
4. Learn how to use NCBI’s BLAST web-based sequence analysis website and be able to correctly interpret its output (Concepts and Exercises).

Concepts

As mentioned above, the purpose of the BLAST algorithm is to find the best hit (highest-scoring match) of an unknown DNA or protein sequence in a database. The method has been so successful because of its clever simplicity. In order to better grasp the method behind the algorithm, try the preparatory exercise. Using your brain and a pencil, try to find regions (local alignments) that best match the following DNA sequence and protein sequences. First, try to match the Query DNA sequence to the Sbjct (Subject) DNA sequence. Then do the same for the protein Query and Sbjct sequences. Find the regions of best alignment between the two, and keep in
mind that the match doesn’t have to be perfect and may even need spaces to help it line up. Circle or draw lines between the matching letters.

DNA MATCH

Query: AGCGAATATTATGAGTAGCAGAAGTCCTGGAGCCT

Sbjct: ACTACAGGGGAGTTTTGTTGAAGTTGCAAAGTCCTGGAGCCTCCAGAGGGC

PROTEIN MATCH

Query: MEMPATTALLNDRLAMLYFWKAEETCALEVCE

Sbjct: ETIRRAYPDANLLNDRLAMLYFWRKAEETCAPSVSRKIVATWMLLEVCE

Reflection

• How much of the query did you try to match at one time?
• How did you find a match? Can you describe it in words?
• Were there any mismatches for the best sequence?
• Were there ever multiple matches? Would breaking up the Query (introducing a gap) help?

Below is the answer. The vertical lines indicate a perfect match between the letters of the two sequences. Note that there are some mismatches and that a big gap must be inserted for the end of the protein Query sequence to match the Sbjct sequence (LEVCE).

DNA MATCH

Query: AGCGAATATTATGAGTAGCAGAAGTCCTGGAGCCT

Sbjct: ACTACAGGGGAGTTTTGTTGAAGTTGCAAAGTCCTGGAGCCTCCAGAGGGC

PROTEIN MATCH

Query: MEMPATTALLNDRLAMLYFWKAEETCA——— LEVCE

Sbjct: ETIRRAYPDANLLNDRLAMLYFWRKAEETCAPSVSRKIVATWMLLEVCE

Most students solve this problem by (i) sliding the Query sequence along the Sbjct sequence, (ii) finding a short region that matches well, and then (iii) extending the match as far as possible. This is essentially how the BLAST algorithm works. Figure 1.1.1 details the basic steps of the algorithm.
FIGURE 1.1.1. Principles behind the BLAST algorithm. (1) The first step of the algorithm is to break the Query sequence into smaller pieces called “words.” For DNA, the word size is usually 10 or 11 letters long, but the example uses four-letter words for simplicity. (Four-letter words. *snicker*) Protein sequence matches start with fewer letters. (2) Then the algorithm slides these smaller words across possible target sequences until it finds a perfect match with one of the small words. (3) Starting with this small alignment, BLAST then extends the alignment until it runs out of letters or the alignment becomes poor (lots of mismatches). The score of the alignment is determined by summing up the scores for matches and mismatches. For DNA, BLAST uses +5 for a match and −4 for a mismatch. The scores of protein sequence alignments are determined by using a special table of match/mismatch scores, such as the BLOSUM62 matrix (see Chapter 07).

Exercises

Interactive exercise (theory)
Use the online BLAST Interactive Link below to learn how the algorithm makes and scores BLAST sequence alignments. Click on the dark circle with the yellow letter I at the top of the page to learn how to use the BLAST Exercise teaching interactive. Once you learn how it works, solve the activity problem.

BLAST Interactive Link
Link: http://kelleybioinfo.org/algorithms/default.php?o=1
Problem

Practice with the BLAST Exercise Link, then solve the problem below.

1. Write the best alignment of the Query to each DNA sequence in the boxes and circle the first matching word from the Query.

2. Calculate scores and rank the three alignments.

<table>
<thead>
<tr>
<th>Length</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Size: 25</td>
<td>Match: 5</td>
</tr>
<tr>
<td>Query Size: 10</td>
<td>Mismatch: -.4</td>
</tr>
<tr>
<td>Word Size: 5</td>
<td></td>
</tr>
<tr>
<td>Query: GCACATG CCT</td>
<td></td>
</tr>
</tbody>
</table>

Query: G C A C A T G C C T

DNA 1: C A G A G A C C T A C T C T A C T G C C A

DNA 2: C A G A G A C C T A C T C T A G C A T C T T G

DNA 3: C A C A C A T G A C T T T G A C C A T G

<table>
<thead>
<tr>
<th>DNA 1:</th>
<th>DNA 2:</th>
<th>DNA 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score:</td>
<td>Rank:</td>
<td></td>
</tr>
</tbody>
</table>
Lab Exercises (Practice)

In this part of the exercise, you will learn how to analyze mystery DNA and protein sequences using the BLAST algorithm online at NCBI. You will also learn how to interpret the output from the program including what the values mean and how to find information about the best match in the database to your query sequence. You will also use a program, called ORF finder, that translates a DNA sequence into likely protein sequences.

NCBI BLAST Tutorial
Link:

Sample and lab exercise data:
http://kelleybioinfo.org/algorithms/data/DAli1.txt
Lab Exercise

Click on the sample and lab exercise data link for the sequence data used in this exercise.

Part 1
Use NCBI BLAST tools to analyze the following DNA that you just sequenced from a plasmid and answer the following questions:

>Part1_Plasmid_Derived_Sequence

CGTTTACGGCGTGGACTACCAGGGTATCTAATCCTGCTCCTCAGCGT
CAGTTACTGCCAGAGACCAGCCTCCACCACCCTGCCGATCTCCTGATATCTGCATTCCACCCTTA
CACCAGGAAATCCAGTCTCCCCCTGC

1. Use the NCBI BLAST tool to perform a sequence search with the above sequence.

   a. The highest-scoring BLAST hit is to what named organism? (Ignore the unknown/uncultured organism hits.)

   b. What is the gene name?

   c. What is the function of the gene, if known? (Don’t know? Try asking “Professor” Google or “Dr.” Wikipedia!)

   d. Who submitted the sequence?

   e. From what institution?

   f. Get the following data for this particular match.

      i. E-value:

      ii. Identities: