BINF
sights.

ISSUE08
2021-T2-W6

BLAST Basics.

Basic Local Alignment Search Tool

ACGT

P05 FEATURED ARTICLE
--
THE BIOINFORMATICIAN'S TOOLBOX:
HOW THE FOUR-LETTER CODE TELLS THE STORY OF LIFE, AND HOW YOU CAN READ IT

[INTERACTIVE WALKTHROUGH]

ATTGCCTTTAGATUNSW
ACTBIOINFORMATICS
GCCGATCGASOCIETY_

Society News

BINFSOC

R Workshop.

This week we held an informative session with our friends at MathSoc, presenting an 'introduction to R' workshop. R is a free and powerful statistical language, and is widely used in both uni courses and the workforce. The workshop covered what you can do with R, talked through the basic operations and syntax, and showed key functions in data handling and plotting.
If you missed out on the live event, a recording is up on our YoutTube channel. Use it as a revision session if you're a BINF3010 student this term, or doing computer science or engineering and want to brush up on your statistical analysis skills.

https://youtu.be/ze4AdZnCVBo

Merch.

Our first merch drop will occur next week. Stay tuned to our Facebook page for updates. Hoodies and t-shirts will be available, with posters coming out later this term.

Student Guide.

We are now in the planning stages of a Student Guide for first year bioinformatics students. We're aiming to include info about degree structure, where to find help and advice regarding your courses, and a general overview on what to expect from the bioinformatics world. If you're interested in giving us advice on what to include, as someone thinking of changing to bioinformatics or as a past student who wishes they had been told something about their journey beforehand, contact us at binfsights@unswbinfsoc.com

“

The decoding of the human genome tells us that we are indeed related to the animals, the insects, and the plants, and that, like it or not, Earth is where we belong.

— Ian McCallum

the swiss army knife of bioinformatics

AN INTRODUCTION TO NCBI’S BLAST, A USEFUL TOOL USED BY BIOLOGISTS AROUND THE WORLD TO IDENTIFY NEW PROTEINS, TRACE EVOLUTION BETWEEN SPECIES, AND RESEARCH THERAPEUTIC TARGETS

Writer *Anthony Nguyen, Cam McMenamie* ***Editor*** *Gabby Younes*

Basic Local Alignment Search Tool

Biological information is commonly represented as sequences. A sequence is simply a one-dimensional object that is made up of a selection of items, following one after the other. This very page you are reading is a sequence; a sequence of 26 possible letters in a particular order. You might be familiar with the binary code used by computers - each set of instructions is a sequence of just 2 'characters'. Every living organism on earth similarly has its own set of instructions contained in the DNA code. These DNA sequences can be understood as strings of four letters: A, G, C and T.

Every book you have ever read has told its story using 26 letters. Earth has told the story of life, the story of every living thing - with just four.

In most cases, these DNA sequences stretch over magnitudes of thousands, to billions. Over time, small changes (mutations) occur and can accumulate within these sequences, changing the biological information stored. This simple but incredibly effective phenomenon is responsible for driving the evolution of new biochemistry, new cell types, and ultimately, new species -- and is a fundamental principle to modern biology.

If you're a biologist, you might want a way to compare and rank how similar a group of different organisms are to one another. Alternatively, you might want to figure out which species a sample of unknown DNA sequence belongs to. Essentially, the goal is to extend the information that is already known about one sequence (such as the species to which it belongs), to another, potentially unknown sequence.

One of the most well-known tools used to quickly and efficiently do this is NCBI's Basic Local Alignment Search Tool, or BLAST. Cited over 12,000 times, BLAST can be used to compare pairs of sequences locally; section by section, and then assign a couple of overall metrics to quantify the similarities. The word we use when making a 'comparison' between two sequences is 'alignment'. Before we can determine how similar pairs of sequences are, we need to find the best way to align these sequences and then score how statistically significant each alignment is.

Now that we understand why we might want to use BLAST, let's step through how we can use the tool. We'll use the example of wanting to figure out which species an unknown sequence of DNA might belong to.

1 First, we navigate to the BLAST website.

> https://blast.ncbi.nlm.nih.gov/Blast.cgi

We have four algorithms to choose from, depending on the type of sequence that we have and what information we want from it.

2 Choose ‘Nucleotide BLAST’.

Since we have a sequence of DNA nucleotides and would like to compare the DNA sequence against others in a database to determine which species it might belong to, we will **choose 'Nucleotide BLAST’.**

(Note: it is also to compare **amino acid** sequences with each other. These are sequences of the building blocks for proteins; which are indirectly translated from the DNA.)

3 Paste in the DNA sequence.

Our sequences often come in a file format called FASTA. This will usually just consist of the sequence itself as well as a title/header containing information about that particular sequence. Our file looks like this:

```
>UnknownSequence
GTAATGTACATAACATTAATGTAATAAAGA
```

Lets paste all of this into the box, including the header. Notice that under ‘Choose Search Set’ > ‘Database’, that the database we are using is the nucleotide collection by default. We could change this in the future if we knew more about our sequence to narrow our search, or wanted to answer a specific question.

4 Scroll down and press BLAST.

Before we press BLAST, take some time to expand out some of the parameter options to see what we can fiddle with before running the program.

Note that one of the options to filter out ‘Low complexity regions’ has been turned on by default. This tells the BLAST algorithm to look out for any regions in your sequence that are made up of only a few elements, such as single nucleotide repeats. This makes sure that the final score is not made high by the presence of these regions, just by chance.

5 View your results.

The time the BLAST algorithm takes to run depends on how busy the server is at the moment as well as the length of your unknown sequence. If our sequence were much longer, we might want to save the results page so we can view it again later without needing to rerun BLAST. To do this, just save the Request ID (RID) that can be used to retrieve your previous searches (jobs).

6 Interpret your results.

The table is presented in descending order. So, the highest scoring results are collected at the top of the table. It is useful to look at the Expected (E) value of the result which measures how likely it was that the result was seen by pure chance. A good rule of thumb is that the smaller the number, the more significant the match.

The percentage identity is the percentage of nucleotides in the unknown region that were similar to the reference genome. Depending on the question we have, a low percentage is not necessarily unusable.

The query cover tells us how much of our unknown sequence was aligned to the reference sequence (sometimes, only part of our sequence will match another).

7 Explore.

From here, your question might already have an answer. From the low E value and high percentage identity and query cover of the top results, we have reason to believe that the DNA sequence we started with belongs to that of the species Bos Taurus, or cattle.
We can click on the Graphic Summary tab to get a picture of how well each result aligned with our sequence. It so happens that each line spans the full length, unbroken, which we expected given our coverage and identity values.

The Alignments tab will give us the chance to view the alignment, letter-by-letter. It is important to point out that while our Bos Taurus sequence (Query) was only 30 bases long, many of our results (Sbjct) aligned to positions much further downstream in the genome. Remember that BLAST performs local alignments. So we can align our sequence to the best scoring position anywhere along the reference genome.

--

While this was a simple example, we can’t always expect sequences to match one hundred percent. The BLAST algorithm has ways to deal with this by adding spaces or ‘gaps’ into our sequence to potentially produce a better alignment, at a cost to the score. This, along with the numerous other versions of the BLAST algorithm allows us to answer increasingly complex questions in biology -- given just a sequence of letters.

Open Position @ UNSW

PhD student in single cell computational biology

-

SUPERVISOR
--

Dr. Fabio Zanini
Data Driven Biomedicine lab @ UNSW Sydney

> fabilab.org

ABOUT THE PROJECT
--

While Physics has a standard model that explains most phenomena in the universe in terms of interactions between elementary particles, such a model is still missing in biology. To fill this gap, we are looking for a talented PhD student to create a standard model of cell biology.

-

PLANNED START: third/last quarter of 2021

MORE INFO > unswbinfsoc.com/phd-position

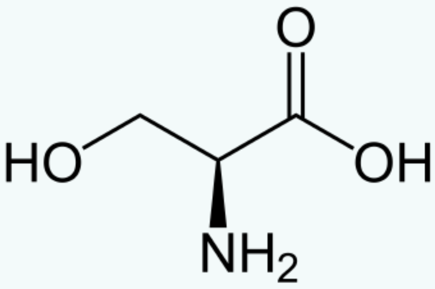
AMINO ACID OF THE WEEK

-

[SERINE]

CHEMICAL STRUCTURE

--



SERINE

S

Ser

105.093

RNA CODONS | U C N

A G Y

POLAR.

BUILDING BLOCK OF PROTEINS.

DISCOVERED IN 1865 BY GERMAN CHEMIST EMIL CRAMER. ISOLATED FROM SILK PROTEIN. STRUCTURE WAS IDENTIFIED IN 1902.

FOODS RICH IN SERINE INCLUDE EGGS, EDAMAME, LAMB, LIVER, PORK, SALMON, SARDINES, SEAWEED AND TOFU.

NON-ESSENTIAL; CAN BE SYNTHESISED UNDER PHYSIOLOGICAL CONDITIONS.

ACTS IN BRAIN AS A SIGNALLER. D-SERINE CAN ACT AT GLYCINE SITE ON NMDA RECEPTORS (IN NEURONS), ALLOWING THEM TO BE ACTIVATED BY GLUTAMATE.

-

BINF sights.

Contact us

IF YOU HAVE ANY COMMENTS or feedback regarding BINFsights, please write to us at binfsights@unswbinfsoc.com

We also encourage anyone to share with us anything you'd like us to take a look at, be it a bioinformatics tool that you have made or find useful; or news in the bioinformatics world that you'd like to see written about in future issues.

TO VIEW PAST AND PRESENT issues of BINFsights, check out our website at unswbinfsoc.com/binfsights

Stay tuned on our Facebook page for updates regarding events and society news.

-- The BINFSOC Team

BINF sights.