

Intro to NCBI Module

Purpose: To introduce students interested in the life sciences to bioinformatic and computational tools which are frequently used throughout many field of Biology

Goal: To develop a strong understanding of what NCBI is capable of and when it may be used

With the increasing dependence on computational power in the world of Biology, it is more important now than ever to understand the vast number of **bioinformatic** tools available. From annotating a **genome** to developing entire simulations of protein binding, the pool of opportunities for biological discoveries has grown infinitely large with the continued development of computational methods. Over the course of this module, you will explore one of the most common bioinformatic tools which is readily available to any scientist.

What is bioinformatics?

What is computational biology and how is it different from bioinformatics?

What is a genome?

National Center for Biotechnology Information (NCBI)

The NCBI houses a database full of biological data relevant for biomedical and genomic studies.

1. Navigate to the following webpage: <https://www.ncbi.nlm.nih.gov/>
2. Click on "About NCBI"
3. Click on "Our Mission"

What has led to the need for a database such as the NCBI?

Who sponsored the legislation that led to the creation of the NCBI?

What is the overall goal of the NCBI?

4. At the top of the page, type “NP_187663.1” into the search bar and click search
5. Click on “Gene” which can be found in the database block labeled “Genes”

At the top of the entry, you should see the following:

HB2 hemoglobin 2 [*Arabidopsis thaliana* (thale cress)]
Gene ID: 820216, updated on 6-May-2021

Download Datasets

From this title alone, we are able to learn some very important information about this gene. For example, we can see that this gene is named “hemoglobin 2” and that its gene symbol is HB2.

Why do you think it is important for a gene to have a shorthand gene symbol such as “HB2”?

We are also able to see which organism this specific gene is from. In this specific case, the HB2 gene was found in *Arabidopsis thaliana*, otherwise known as the thale cress.

+

Now, click on “gene” to see the full summary box:

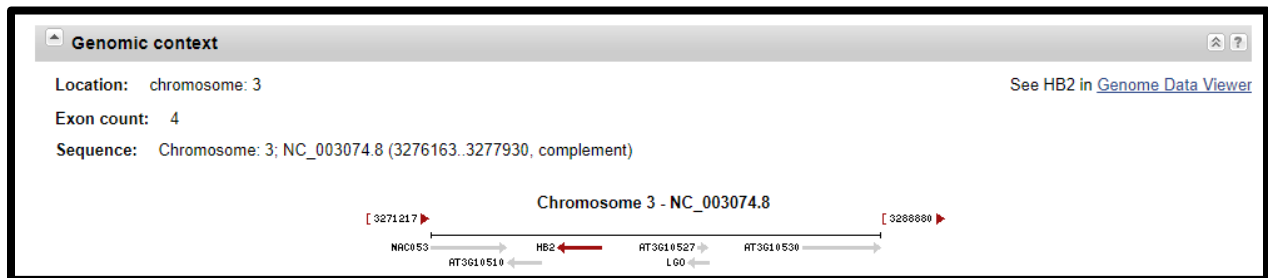
Summary

Gene symbol	HB2
Gene description	hemoglobin 2
Primary source	Araport:AT3G10520
Locus tag	AT3G10520
Gene type	protein coding
RNA name	hemoglobin 2
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetales; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	AHB2; ARABIDOPSIS HEMOGLOBIN 2; ARATH GLB2; ATGLB2; GLB2; haemoglobin 2; HEMOGLOBIN; HEMOGLOBIN 2; NON-SYMBIOTIC HAEMOGLOBIN 2; NSHB2
Summary	Encodes a class 2 non-symbiotic hemoglobin. Over-expression of AHb2 in seeds led to a 40% increase in the total fatty acid content of developing and mature seeds in three subsequent generations. This was mainly due to an increase in the poly-unsaturated C18:2 (omega-6) linoleic and C18:3 (omega-3) alpha-linolenic acids.

NEW Try the new [Gene table](#)
Try the new [Transcript table](#)

Again, we are provided with some very useful information about the HB2 gene. This is a protein coding gene (meaning that this gene carries the instructions for the production of a specific protein) as denoted by the “Gene type”. We are again shown the organism which this gene was isolated from along with the organism’s lineage (similar to a family tree showing which group the thale cress belongs to).

Finally, scroll down to the the genomic context box:



Here we are able to see the location of this gene on the genome of the organism. The HB2 gene is located on chromosome 3 of the thale cress genome. In addition, we are able to see other genes which are in close proximity to the gene being analyzed.

What is the gene code of the gene directly to the left of (upstream) HB2?

Now, click on the arrow for the gene directly upstream HB2.

What is the name of the gene directly upstream HB2?

So, we have learned a little bit about the HB2 gene in thale cress. What about the actual protein that HB2 encodes? Let's find some information about this protein.

6. At the top of the page, click the dropdown menu next to "Gene" and select "All Databases"

7. With “All Databases” selected, type “NP_187663.1” into the search bar and hit search
8. Click on “Protein” which can be found in the database block labeled “Proteins”

This page provides us with a trove of useful information. For now, we will focus on a few important details about this protein.

At the top of the page we are provided with the name of the protein (hemoglobin 2), the **gene locus** for the gene encoding this protein (NP_187663), the length of the protein (158 amino acids), and the genome type (linear).

hemoglobin 2 [Arabidopsis thaliana]

NCBI Reference Sequence: NP_187663.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

[Go to:](#) ☒

LOCUS NP_187663 158 aa linear PLN 14-FEB-2019

DEFINITION hemoglobin 2 [Arabidopsis thaliana].

What is a gene locus?

Now, click on “hemoglobin 2” and scroll down to the bottom of the page:

ORIGIN

```

1 mgeigfteq ealvkeswei lkqdpkysl hffsqileia paakglfsfl rdsdevphnn
61 pkikahavkv fkmctetaiq lreegkvvva dtlqylgsi hlksgrvidph fevvkeallr
121 tlkeglgeky neevagawsq aydhlaiaik temkqees
//

```

Here we see one of the most elegant facets of biology. The amino acid sequence for the hemoglobin 2 protein. Just as the combination of letters in each word written here conveys

specific meaning, the amino acid sequence of a protein (the primary structure of a protein) ultimately makes a protein what it is.

When looking at such amino acid sequences on any database, you will most likely see them written out in one letter codes as above. You will find a key below to help decipher which amino acids are in the hemoglobin 2 amino acid sequence.

Alanine - A	Glutamic Acid - E	Methionine - M	Tyrosine - Y
Arginine - R	Glycine - G	Phenylalanine - F	Valine - V
Asparagine - N	Histidine - H	Proline - P	
Aspartic Acid - D	Isoleucine - I	Serine - S	
Cysteine - C	Leucine - L	Threonine - T	
Glutamine - Q	Lysine - K	Tryptophan - W	

While the NCBI is home to an immense amount of biologically relevant data, here we focused on a few important pieces of information you can receive from this resource. We here at MoleculeGo hope you take some time on your own to look through the many parts of this website.

Now, let's put all that we have learned together:

9. Search "All Databases" for NP_179204.1

What organism is this gene found in?

What is the gene symbol for this gene?

What is the name of this gene?

What is the gene type of this gene?

Where is this gene located in the organism's genome?

What is the gene symbol for the gene directly upstream of this gene?

What is the name of the protein encoded by this gene?

How many amino acids is this protein?

What is the first amino acid in this protein's primary structure?

This gene seems to be very similar to the one we looked at previously. How can we see how similar the two really are to each other? This is where we will look at one final tool available on the NCBI website.

Below you will find the accession numbers for each of the proteins we have looked at:

NP_179204.1 (*Arabidopsis thaliana* hemoglobin 1)

NP_187663.1 (*Arabidopsis thaliana* hemoglobin 2)

10. Navigate back to the NCBI homepage (<https://www.ncbi.nlm.nih.gov/>)

11. On the right side of the page and under “Popular Resources” click “BLAST”

Popular Resources

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

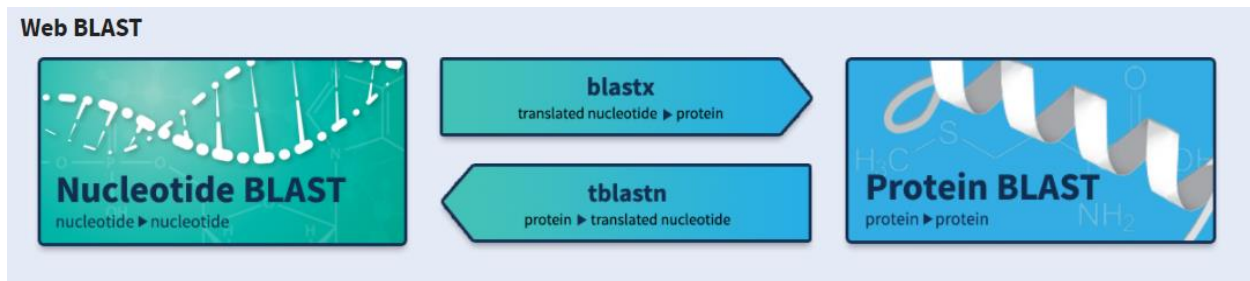
Gene

Protein

PubChem

BLAST stands for Basic Local Alignment Search Tool and is a resource that allows you to align two or more nucleotide or amino acid sequences for comparison. Using this tool we can determine how similar our two proteins are!

12. Select “Protein BLAST”



13. Check the box to the left of “Align two or more sequences”

The image shows the 'Standard Protein BLAST' form. The 'blastp' tab is selected. The 'Enter Query Sequence' section has a text box for the query sequence. Below it, the 'Align two or more sequences' checkbox is checked. The 'Job Title' field contains 'NP_179204:hemoglobin 1 [Arabidopsis thaliana]'. The 'Query subrange' section has 'From' and 'To' fields. A notification box on the right says 'New columns added to the Description Table'.

14. Type the accession number for *Arabidopsis thaliana* hemoglobin 1 into the first box and the accession number for *Arabidopsis thaliana* hemoglobin 2 into the second box

The image shows the 'Align Sequences Protein BLAST' form. The 'blastp' tab is selected. The 'Enter Query Sequence' section has a text box containing 'NP_179204.1'. Below it, the 'Align two or more sequences' checkbox is checked. The 'Job Title' field contains 'NP_179204:hemoglobin 1 [Arabidopsis thaliana]'. The 'Query subrange' section has 'From' and 'To' fields. A notification box on the right says 'New columns added to the Description Table'. The 'Enter Subject Sequence' section has a text box containing 'NP_187663.1'. Below it, the 'Subject subrange' section has 'From' and 'To' fields.

15. Scroll down the the bottom of the page and click the blue “BLAST” button
 - a. It will take a little bit for the program to run and the page will auto refresh until you see the following page

BLAST® » blastp suite-2sequences » results for RID-EH4JEWPY114

Home Recent Results Saved Strategies Help

[Edit Search](#) Save Search Search Summary ▾

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title	NP_179204:hemoglobin 1 [Arabidopsis thaliana]
RID	EH4JEWPY114 Search expires on 07-11 02:15 am Download All ▾
Program	Blast 2 sequences Citation ▾
Query ID	NP_179204.1 (amino acid)
Query Descr	hemoglobin 1 [Arabidopsis thaliana]
Query Length	160
Subject ID	NP_187663.1 (amino acid)
Subject Descr	hemoglobin 2 [Arabidopsis thaliana]
Subject Length	158

Other reports [Multiple alignment](#) [MSA viewer](#) ?

Filter Results

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

16. Scroll down and click on the “Alignments” tab

Descriptions Graphic Summary **Alignments** Dot Plot

Alignment view Pairwise [Restore defaults](#) Download ▾

1 sequences selected ?

[Download](#) ▾ [GenPept](#) [Graphics](#) [Next](#) [Previous](#) [Descriptions](#)

hemoglobin 2 [Arabidopsis thaliana]
 Sequence ID: NP_187663.1 Length: 158 Number of Matches: 1
[See 6 more title\(s\)](#) [See all Identical Proteins\(IPG\)](#)

Range 1: 2 to 153 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
171 bits(433)	3e-60	Compositional matrix adjust.	84/152(55%)	108/152(71%)	0/152(0%)

Query	5	GKIVFTEEQEALVVKSHSVNKKNSAELGLKLFKIFETAPTCKKMFSLRDSPIAEQNP	64
Sbjct	2	G+I FTE+QEALV +SN ++K++ + L F +I EIAP K +FSFLRDS NP	61
Query	65	KLKPHAMSVFVMCCESAVQLRKTKVTVRETTLLKRLGASHSKYGVVDEHFEVAKYALLET	124
Sbjct	62	KLK HA+ VF M CE+A+QLR+ GKV V +TTL+ LG+ H K GV+D HFEV K ALL T	121
Query	125	IKEAVPEMMSPEMKVAMGQAYDHLVAATKAEM	156
Sbjct	122	+KE + E ++ E++ AW QAYDHL AIK EM	153

Related Information
[Gene](#) - associated gene details
[Genome Data Viewer](#) - aligned genomic context
[Identical Proteins](#) - Identical proteins to NP_187663.1

This final piece of data we are going to look at is a great example of why in Biology there is always more to a protein than just a name. The lines labeled “Query” are the amino acid sequence for *Arabidopsis thaliana* hemoglobin 1 and the lines labeled “Sbjct” (or subject) are the amino acid sequence for *Arabidopsis thaliana* hemoglobin 2. Under the label “Identities” we see that 84 amino acids out of the 152 that were compared are identical in these two proteins. Under the label “Positives” we see that 108 amino acids out of the 152 that were compared are similar or identical to each other.

We can now confidently conclude that *Arabidopsis thaliana* hemoglobin 1 and *Arabidopsis thaliana* hemoglobin 2 are 55% identical and 71% similar. Even though the

only difference between the two protein names is a single number, they are 45% different from one another.

We now know to never judge a protein by its label.

Challenge Questions:

Q: What is the Expect (E) value?

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance. For more details please see the calculations in the [BLAST Course](#).

The Expect value can also be used as a convenient way to create a [significance threshold](#) for reporting results. You can change the Expect value threshold on most BLAST search pages. When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits can be reported.

What do you think is the difference between identity and similarity?

What would lead to some amino acids being labeled as similar and others not? (Think about what groups we can put amino acids in)

What is the "expect value" or "e-value" for our alignment in the previous exercise?

What does the "expect value" describe?

If a match between two sequences is to be considered significant, would we see an “e-value” that is very large or very small?
