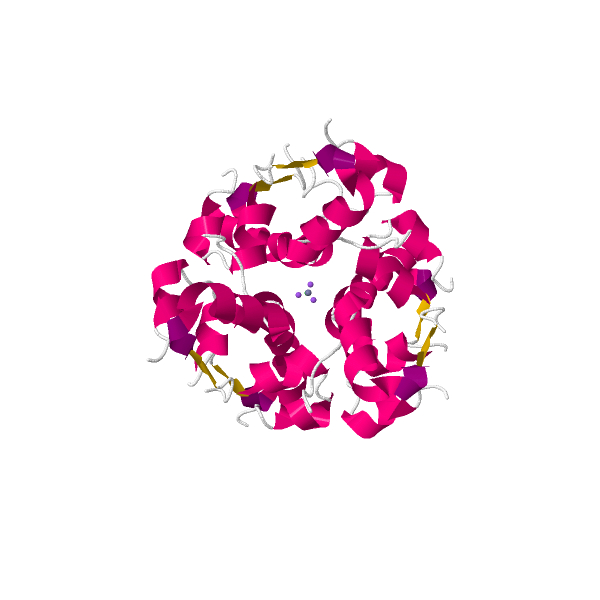


**info@rcsb.org • www.rcsb.org**

**Bioinformatics of Insulin**



This bioinformatics tutorial explores the relationship between the sequence, structure, and biological function of the protein hormone insulin. In this tutorial you will learn about the following:

1. Find and explore the structure of insulin using search tools on the RCSB PDB website;

Use molecular visualization tools to explore the insulin structure and function;

1. Find gene and protein sequences to figure out key steps in its biosynthesis.

The PDB archive is the primary repository of experimentally determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](http://www.wwpdb.org/" \t "_blank), the RCSB PDB (a) curates and annotates structural data from researchers around the world. The RCSB PDB also provides a variety of tools and resources for searching, visualizing, downloading, and analyzing biomolecular structures.

**Learning Goals:**

1. Query the RCSB PDB website to find a specific structure.
2. Learn about the structure from the structure summary page and by visualization of the structure
3. Explore the gene and protein sequences, and the protein structure of the molecule of interest to figure out key steps in its biosynthesis.

**Exercise:**

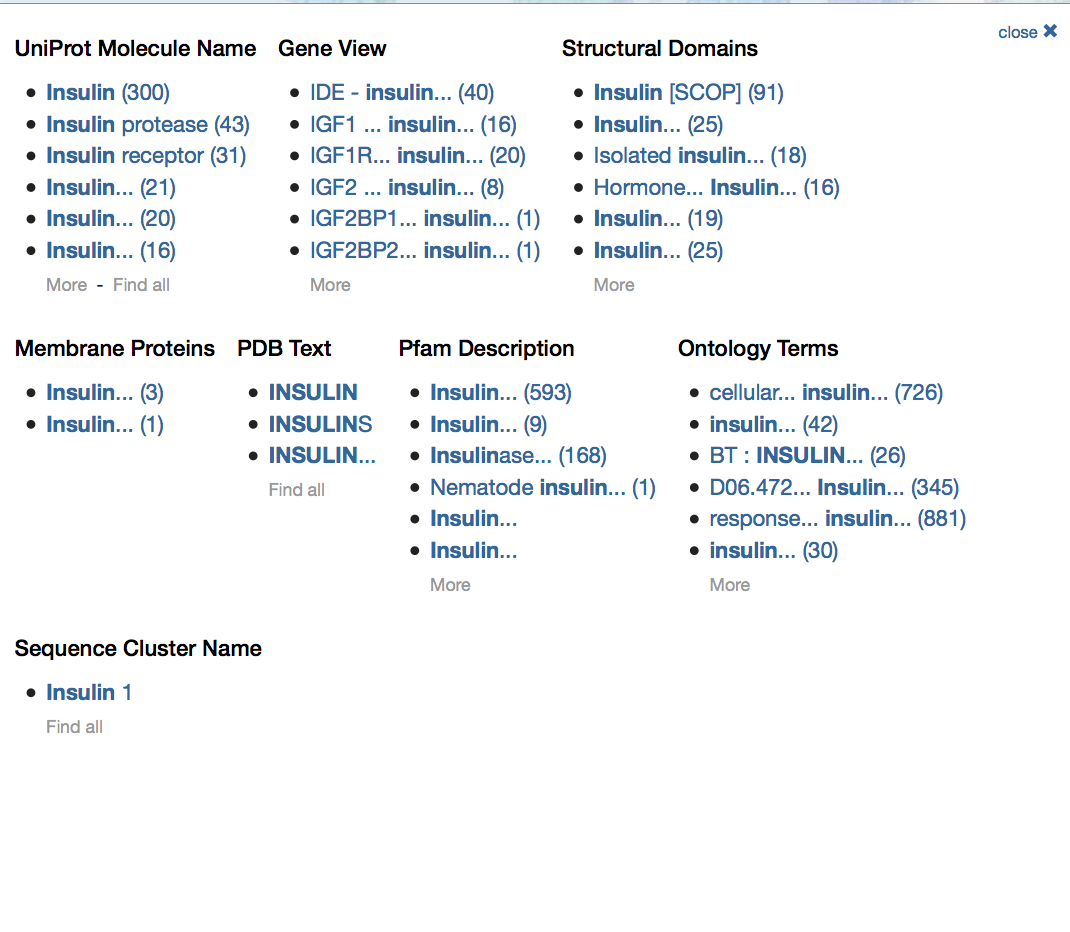
**Find structures of insulin at the RCSB PDB website.**

In this first part, we will find a structure of insulin in the RCSB PDB, and use several tools to explore its structure and function.

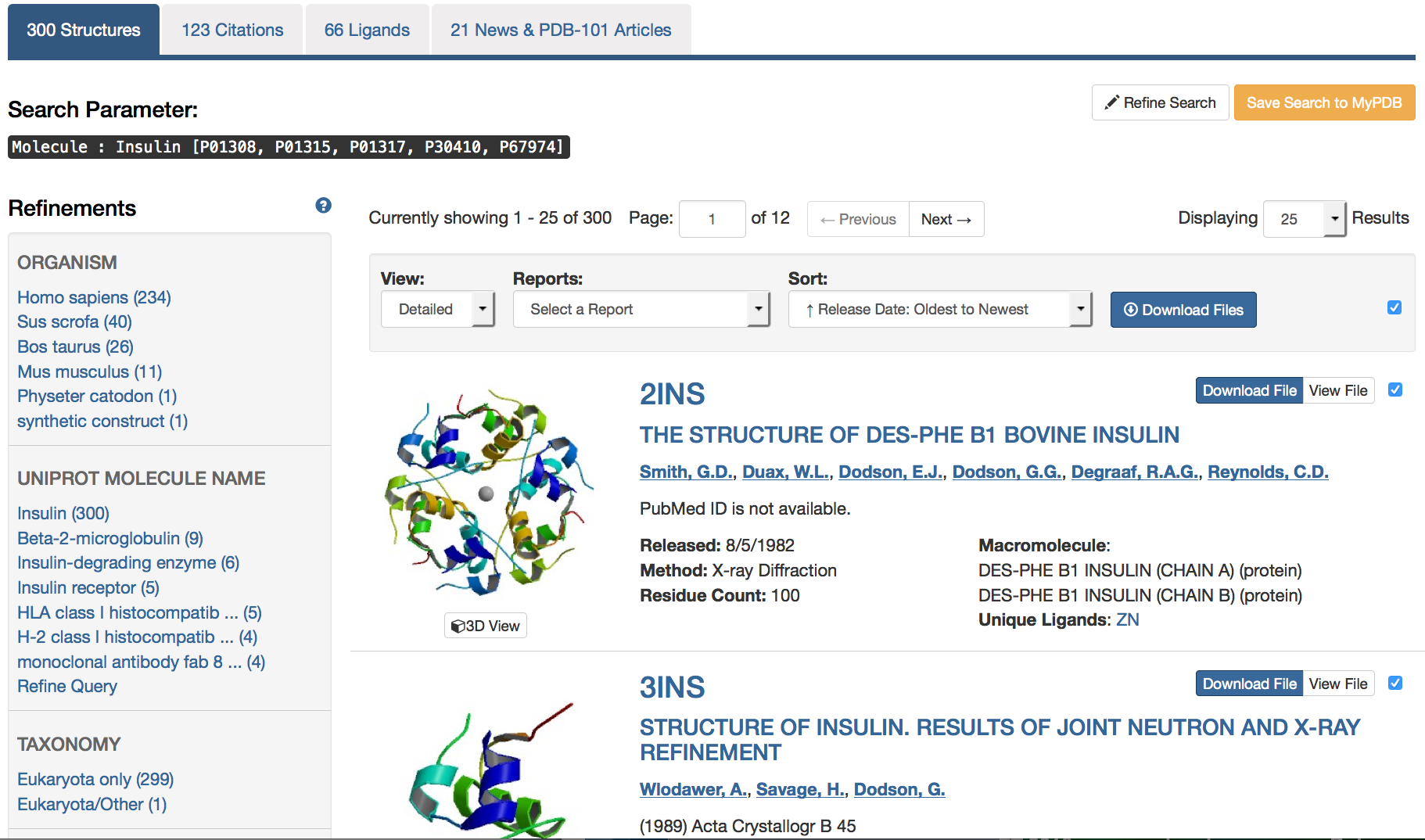
1. Go to the RCSB PDB website at <http://www.rcsb.org>
2. Perform a search by typing the protein name “insulin” in the search box on top of the home page:



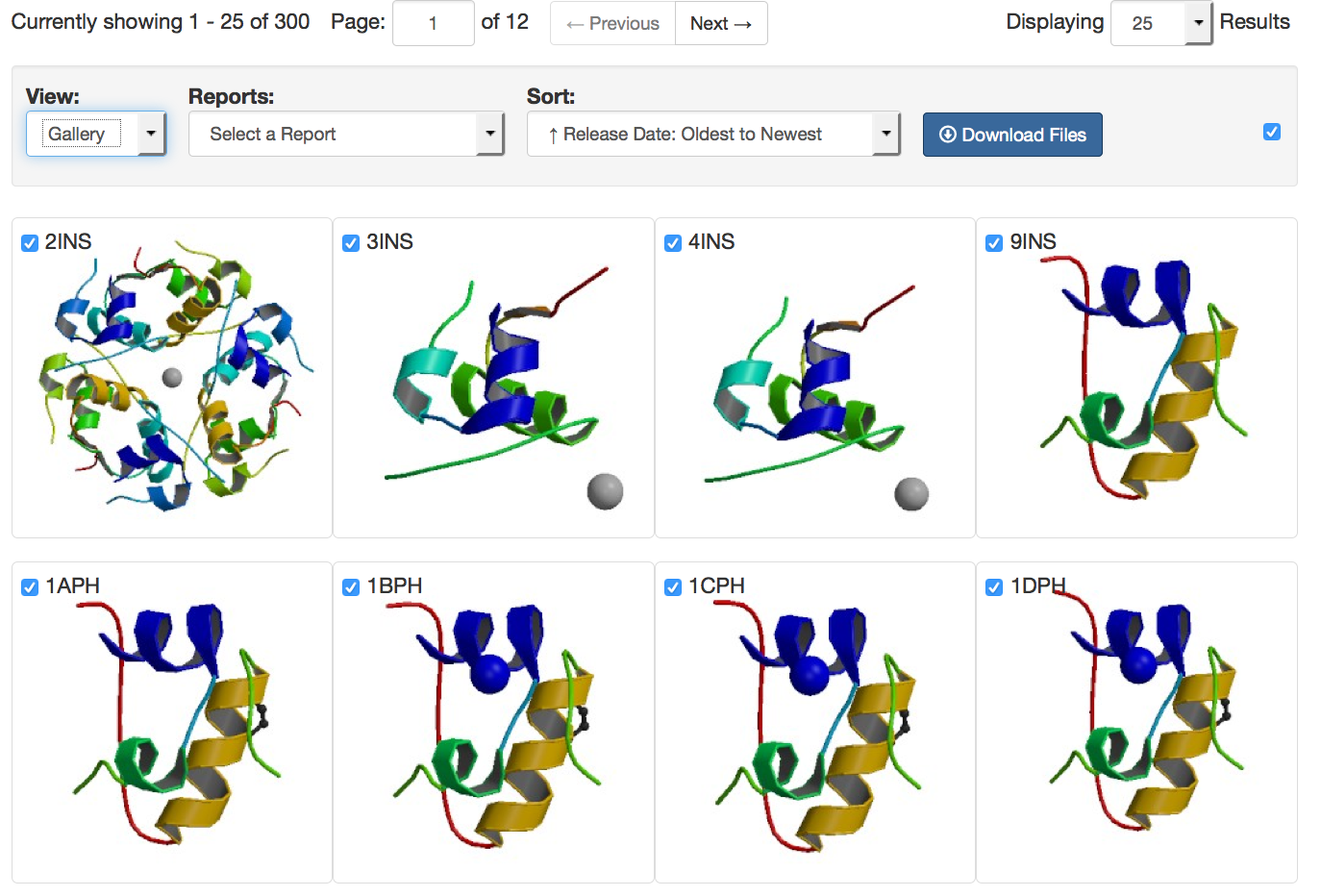
1. Click the *Go* button.
2. You should see a list of autocomplete suggestions to select from.



1. Click on the *UniProt Molecule Name* “Insulin” to see the results page, which contains a list of all Insulin structures in the PDB on the day that this search was done (September 2016).

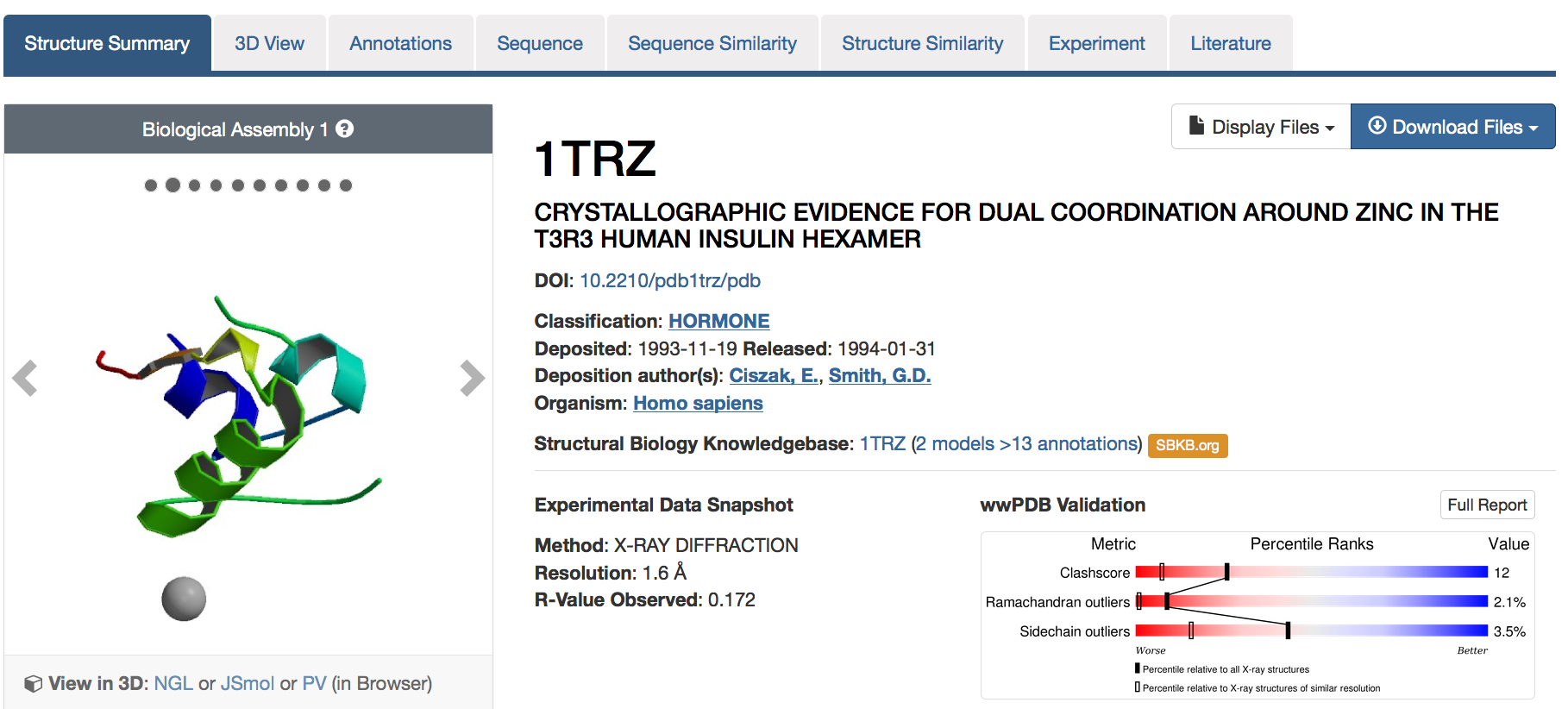


You can explore all of these different structures by clicking on different examples, creating reports, or generating a gallery image.

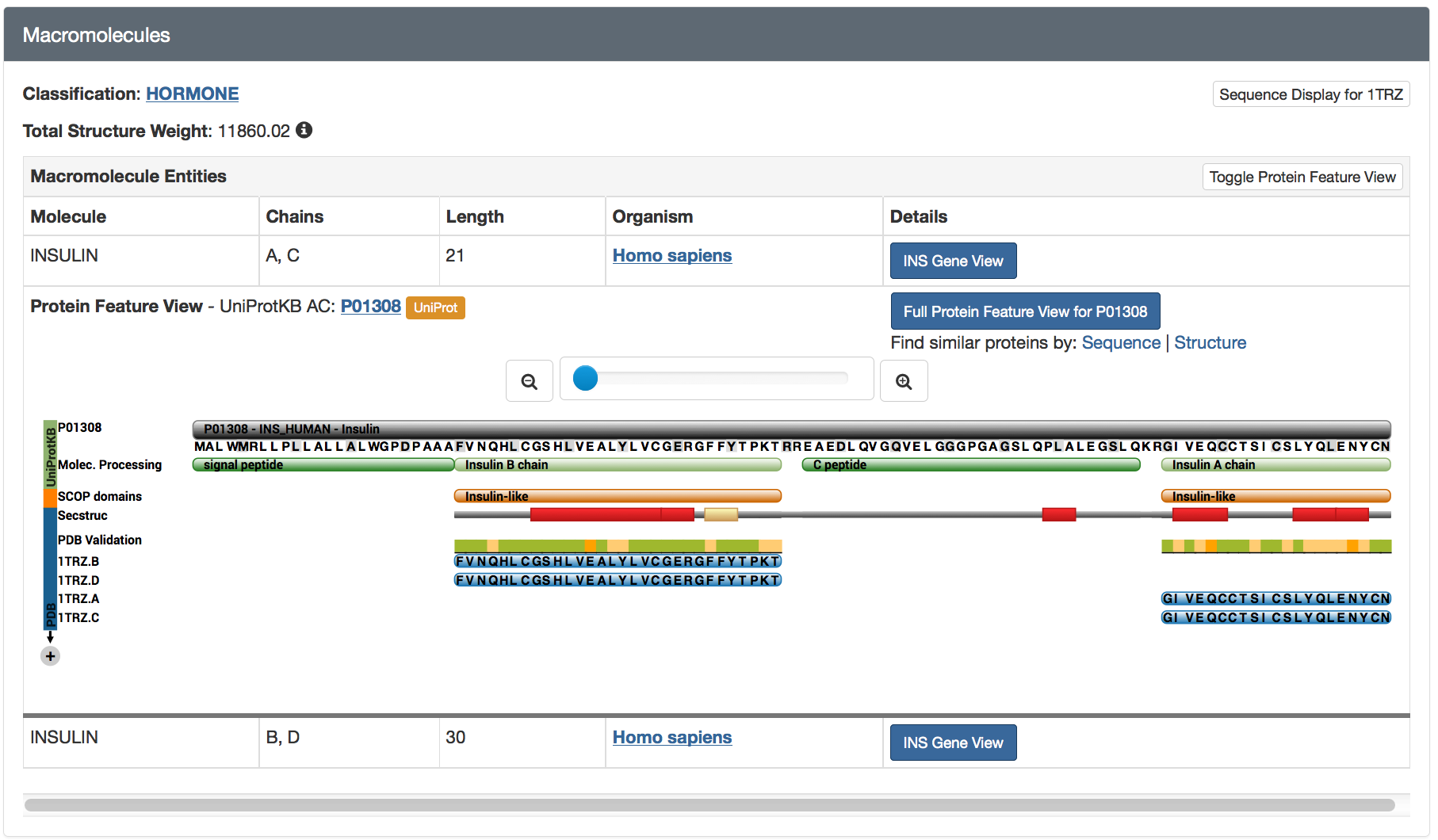


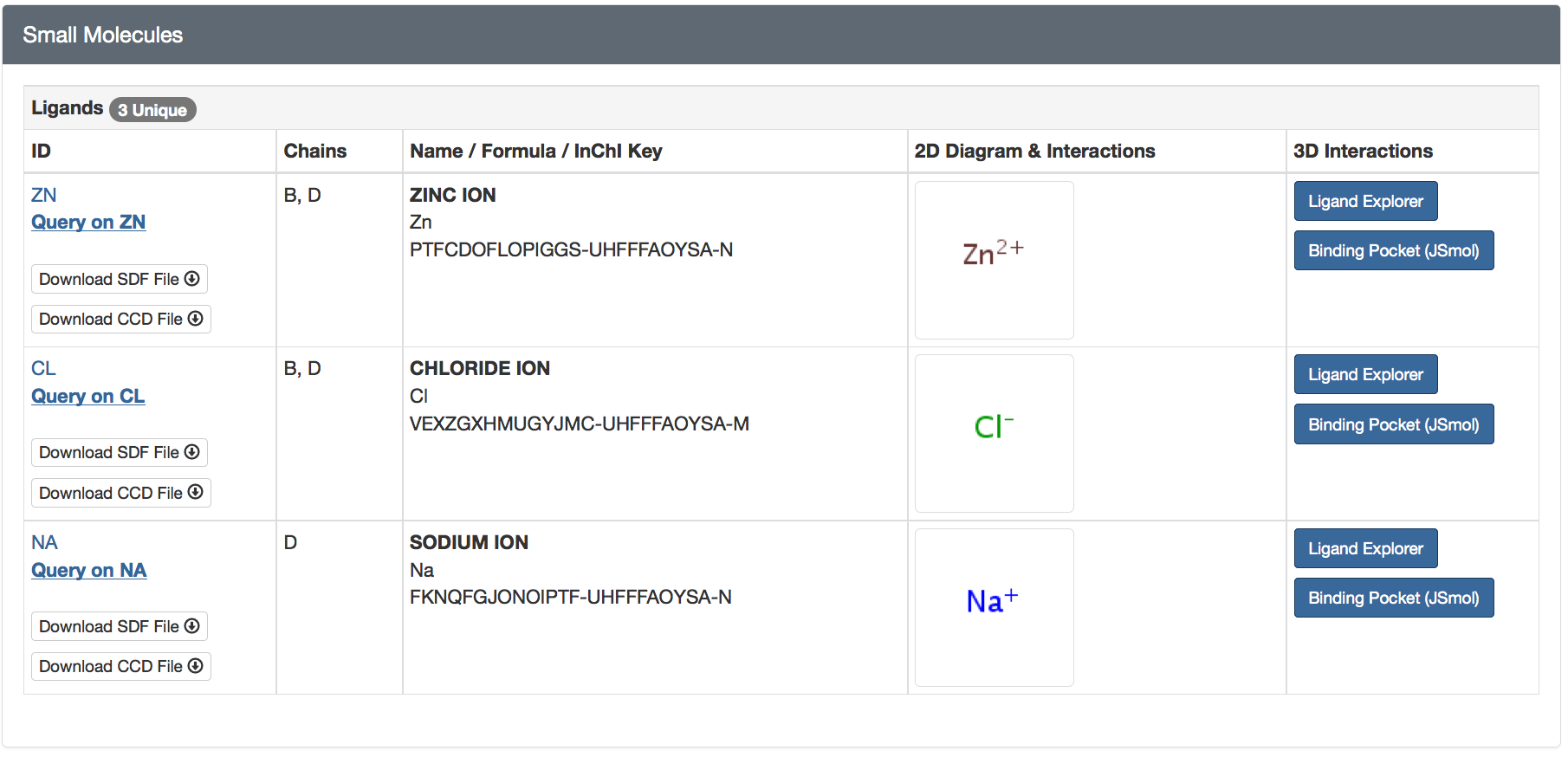
Example image collage for some of the structures found when searching for **Insulin.**

**Q1.** How many human insulin structures did you find in the archive? How did you figure this out?For the following sections of the exercise, we will use the structure of human insulin with PDB ID 1trz[[1]](#footnote-1). You can easily find this protein by entering the PDB ID 1trz in the search bar at the top of the page and clicking *’Go’*. This will take you to the *Structure Summary* page for 1trz.



Review the contents of this structure by scrolling down the page. Information about all components in this structure is presented in sections titled *Macromolecules* and *Small Molecules*, shown below.





To learn more about the overall insulin protein sequence and its various annotations, click on the small *+* sign in the panel showing various colored tracks. Alternatively click on the *Full Protein Feature View …* button under *Details*.

To learn more about the *Small Molecules* associated with the protein click on the hyperlinks and buttons shown above.

**Q2.** List 2 things that you can learn from the *Protein Feature View* page?

**Q3.** How many macromolecules (protein polymer chains) are included in this structure? (Hint: check the molecule names and chain identifiers in the *Macromolecules* box)

**Q4**. How many small molecules are included in this structure? (Hint: check the ligand names and chain identifiers in the *Small Molecules* box)

**Q5**. Explain why the structure title describes this structure as an insulin hexamer? (Hint: Read the abstract in the *Literature* box on the Structure Summary Page)

**Visualization of Insulin to understand its functions**

**a. Basic protein visualization using Protein Workshop**

Now let’s take a look at the insulin structure in close detail.

1. On the right side of the Structure Summary page, you will see a box containing an image and links to 3-D molecular viewers such as *NGL*, *JSmol*, *PV*, *Protein Workshop* etc.
2. Under the image click on the link that says *Protein Workshop*
3. This will download the viewer. In the process it will ask you if you want to *trust this application*. This is part of the Java security mechanisms; you simply accept/trust each one and click *run* when prompted.

Once the structure is loaded you should see something that looks like the following:

|  |  |
| --- | --- |
| Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 2.40.15 PM.png | The *viewer* is to the left and the *control panels* are to the right. If you click and drag the mouse in the viewer you will see the structure rotate. You can also zoom the structure by clicking the middle button and dragging (or, shift+click with a one-button mouse), and translate the structure using the right button (or ctrl+click with a one-button mouse). Take a minute to get familiar with these controls. |

Protein Workshop automatically displays a *ribbon representation* of the protein structure. This representation represents the polypeptide chain of the protein, but uses flat arrows to show beta strands and curly ribbons to show alpha helices. The chain is also colored in rainbow colors from blue to red from one end of the chain to the other, so you can follow the chain as it folds into this complex structure.

You can learn more about Insulin in the Molecule of the Month feature with the same title at: <http://pdb101.rcsb.org/motm/14>. If you want to try building your own model of Insulin, there is a paper cut-out form at <http://cdn.rcsb.org/pdb101/learn/resources/insulin-activity-assets/insulin-activity.pdf>. Alternatively select Learn >> Paper Models from the PDB101 home page (<http://pdb101.rcsb.org/>).

**b. Different ways of looking at Insulin in Protein Workshop**

The *control panel* of Protein Workshop is designed for quick and simple editing in a four-step process. Notice the boxes numbered 1-4 on the *Tools* panel. These are to help you go through the steps of using this tool. Other panels provide advanced options, described in more detail below. Panels 1-3 change when the Tools, Shortcuts, or Options buttons are selected:

|  |  |  |
| --- | --- | --- |
| Tools Menu | Shortcuts Menu | Options Menu |
| Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 2.52.44 PM.png | Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 2.52.53 PM.png | Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 2.53.07 PM.png |

To switch to a view that shows all atoms, you need to go through several steps. First, in the *Control Panel* under the Tools menu, turn off the ribbon representation, by performing one action in each of the four boxes:

1. Click on the *Visibility* Button

2. Select *Ribbons*

3. (No options in the Visibility Tool.)

4. Click on *1trz* (the PDB ID) in the bottom tree viewer.

At this point the viewer should be blank, because you have essentially chosen to toggle OFF the ribbon representation. If you click on *1trz* again, you can toggle the ribbon back ON. With the ribbon OFF, display atoms using a ball-and-stick representation:

1. Click on the *Visibility* Button

2. Select *Atoms and Bonds* this time.

3. (No options in the Visibility Tool.)

4. Click on *1trz* (the PDB ID) in the bottom tree viewer.

You should see the following:

|  |  |
| --- | --- |
| Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 3.01.13 PM.png | Each sphere represents an atom, and the lines between these atoms represent covalent bonds.  Carbon atoms are green  Nitrogen atoms are blue  Oxygen atoms are red  Sulfur atoms are yellow |

To get a better look at the protein structure with all atoms and bonds shown, zoom into the protein.

Hold the shift key, and click and drag the mouse downward.

As you can see, this representation shows a lot of information that makes it difficult to find specific structural features with all atoms shown. Features that you might be able to identify include:

* yellow sulfur atoms in cysteine
* single oxygen atoms in red are water molecules

To return to the default view of the structure, we’ll follow our steps in reverse.

1. Click on 1ema in the bottom tree viewer. This will make the molecule disappear.

2. Select on *Ribbons* in box #2.

3. Click on 1ema in the bottom tree viewer. This will make the molecule appear, but now in ribbon representation. You can zoom the image to show the whole protein by pressing the shift key and dragging the mouse up.

**c. Exploring the disulfide bridges using a combination of ribbon and atom views**

Often, we want to explore some parts of the protein in atomic detail (such as the disulfide bridges), but use a simple ribbon representation for the rest of the protein.

1. Still in Protein Workshop, select the Visibility tool

2. Select *Atoms and Bonds* for what you want the tool to affect – This means that when we select all the Cys residues in chains A and B, it will appear as atoms bonds.

3. (No options in the Visibility Tool)

4. In box 4, open up Chain A. Scroll down and select the position of the Cys. Selecting and deselecting will these residues cause the disulfide bonds to appear (in atom representation) and disappear. You can rotate the structure to see how this piece fits in the overall shape of the protein.

|  |  |
| --- | --- |
| Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 3.12.55 PM.png | Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 3.12.28 PM.png |
| Here, the disulfide bonds are not shown | Here, the disulfide bonds are shown |

## Q6. List the positions in chains A and B where you found a Cys residue. How many disulfide bonds do they form? Which of the Cys residues interact with each other to form these bonds?

## d. Visualize hydrogen bonds in proteins

The secondary structures of proteins, like the alpha helices seen in insulin, are stabilized by hydrogen bonds. The most important hydrogen bonds in proteins are formed between C=O hydrogen atom and the N-H oxygen atom in the protein backbone—these hydrogen bonds link different portions of the chain and stabilize the folded structure. However, it can be tricky to see these bonds using many of the structures in the PDB because crystallographic structures typically do not include the hydrogen atoms. Instead, we often draw the hydrogen bond between the nitrogen atom in the N-H group (which is included in the PDB file) and the oxygen atom of the C=O group.

Protein Workshop does not display hydrogen bonds, so we’ll look at hydrogen bonds in insulin using *JSMol*.

1. Go again to the *Structure Summary* page browser page for PDB entry 1trz.
2. Underneath the image on the right, select *3D View* to launch the *JSMol* viewer.

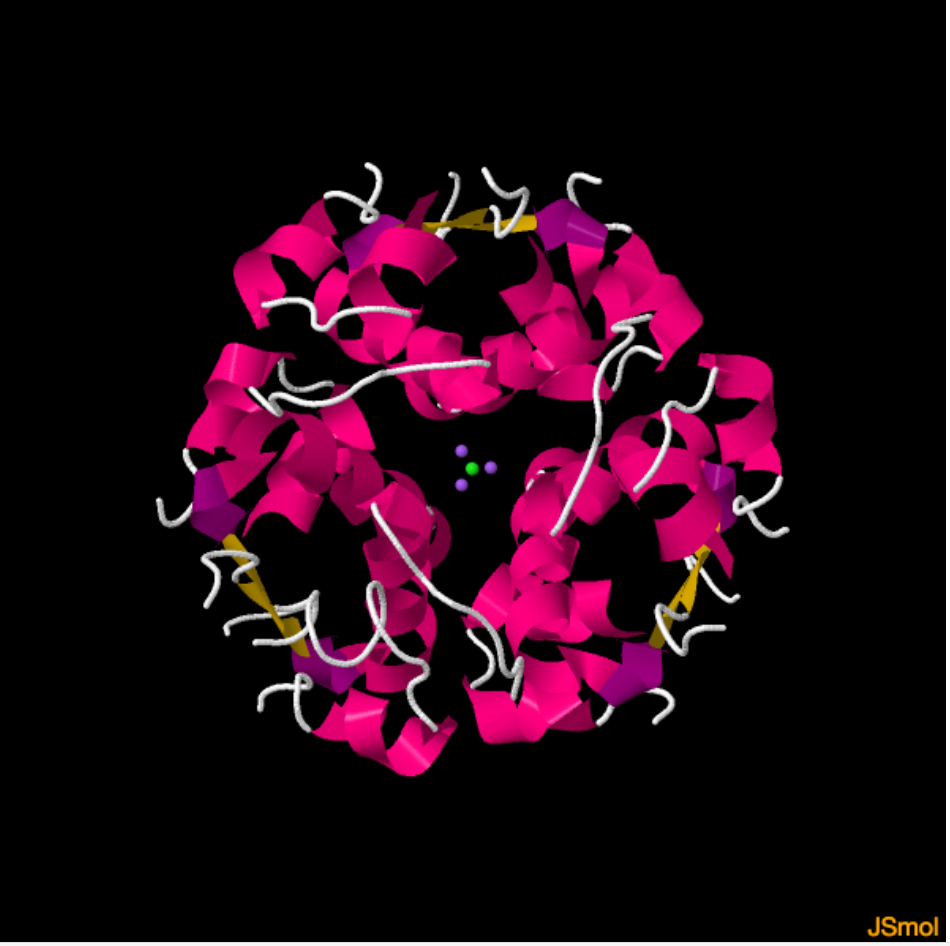
|  |  |
| --- | --- |
| 3. Commands can be typed in the ‘Scripting Options’ section (to see the text box, you will have to expand it by clicking the title) beneath the image viewer to change the representation of the molecule. | Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-27 at 3.39.51 PM.png |
| 4. To show the hydrogen bonds in the protein backbone:  a) Type **select protein** in the *Input* box, and select the *Submit* button.  b) Type **cartoon off** and select the *Submit* button.  c) Enter **select backbone** and select the *Submit* button.  d) Enter **wireframe 100** and select the *Submit* button.  e) Enter **calculate hbonds** and select the *Submit* button. |  |  |

You can rotate the Insulin to see all of the hydrogen bonds.

**Q7**. How manybackbonehydrogen bonds can you see in the N-terminal helix of chain A. (Hint: examine the residues carefully. If necessary, you may color the residues in the chains by rainbow under the Display options).

**d. Visualize the biological assembly of insulin**

Single copies of chains A and B are seen in the molecular images in the above steps. In its storage form insulin has been shown to form hexameric assemblies. Select the Biological Assembly 3 from the Structure *pull-down* menu in the *Structure Details* section to show this hexameric structure of insulin. It should appear like the following:



Examine the structure closely.

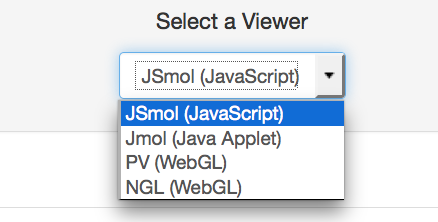
**Q8**. What are the purple and green dots in the center of the structure?

**Q9**. The title of the structure suggests that there is zinc in the structure. Where is it located?

**Q10**. What is the function of the zinc ion(s)?

**e. Using the other visualization tools**

At the bottom of the JSmol graphics window select other viewers to see the same structure (PDB ID 1trz).



The insulin molecule as visualized by *PV* and *NGL* are shown below:

|  |  |
| --- | --- |
| View in PV | View in NGL |
| Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-28 at 9.23.19 AM.png | Macintosh HD:Users:SDutta:Desktop:Screen Shot 2016-09-28 at 9.23.40 AM.png |

Both the visualization software shown above have their own advantages. For example, PV allows viewing the molecule based on symmetry, while NGL displays superior speed even when displaying very large molecules.

**Complete the following questions for HW:**

**Exploring Gene and Protein Sequences of Insulin to understand its biosynthesis**

In this part, we will use several online bioinformatics resources to find the sequence of insulin gene, translate this into a protein sequence, and analyze the translation product. The goal here is to figure out how the insulin gene product is processed during biosynthesis of the molecule.

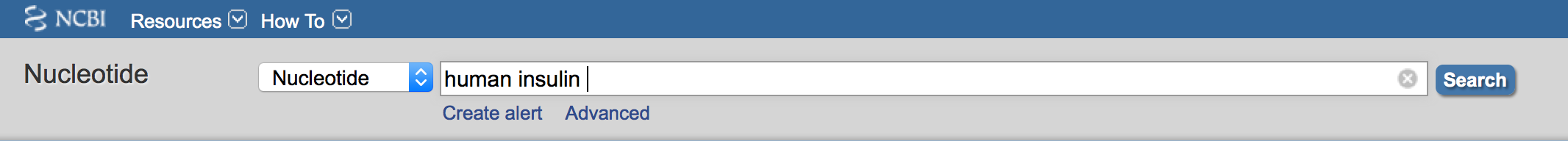
**a. Find the DNA sequence for the insulin gene in GenBank**

GenBank (b) is an annotated collection of all publicly available DNA sequences. It is is part of the International Nucleotide Sequence Database Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

1. Point your web browser to GenBank: <https://www.ncbi.nlm.nih.gov/genbank/>

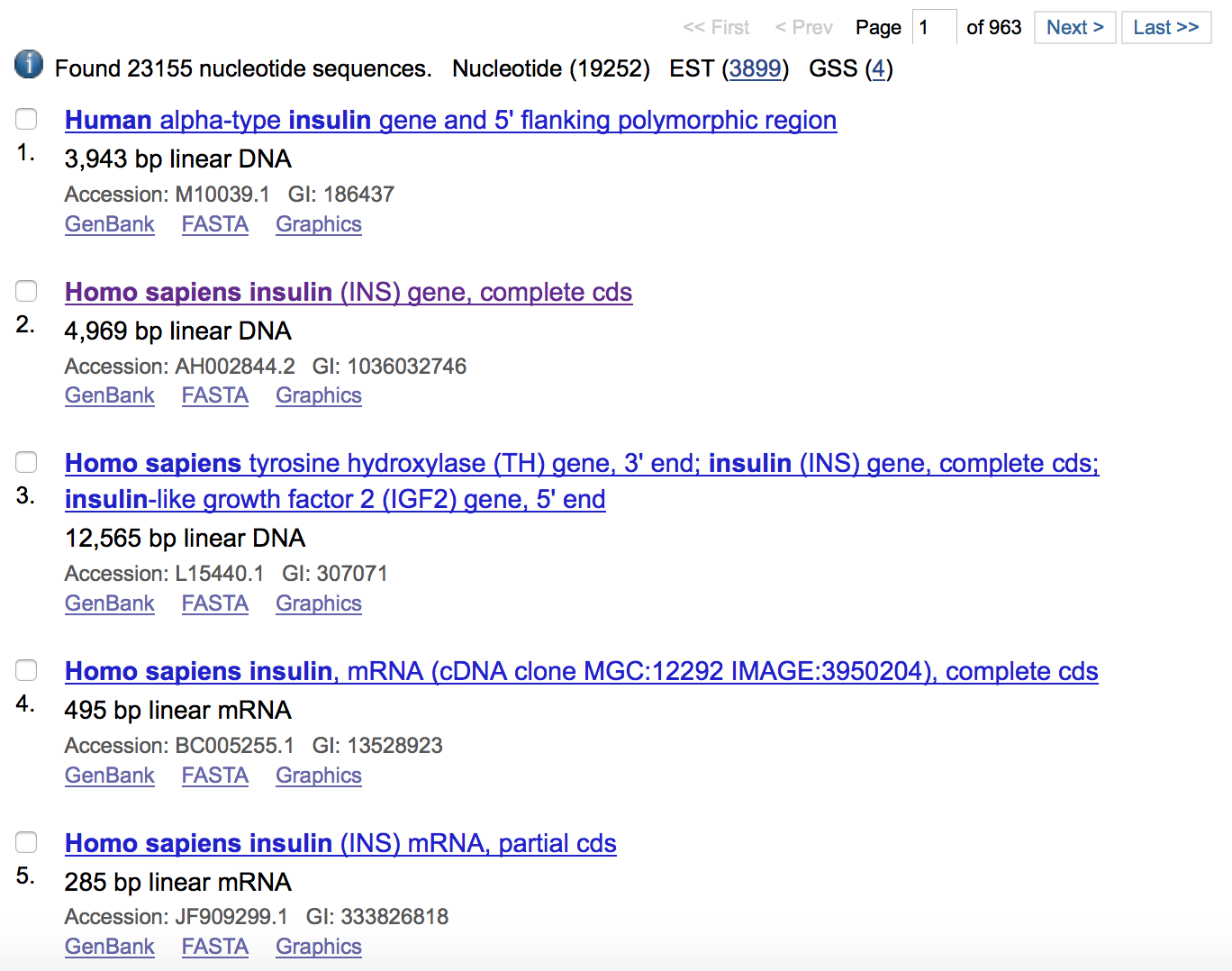
Alternatively, go to the NCBI website: <http://www.ncbi.nlm.nih.gov/> and select the *Nucleotide* database in the *search* pulldown menu to access this same page.

1. In the tops search box type in the words “human insulin”



You may also just search for insulin and then refine the search results to list only results for *Homo sapiens* (human) proteins.

1. The search results show the following (as of September 2016):

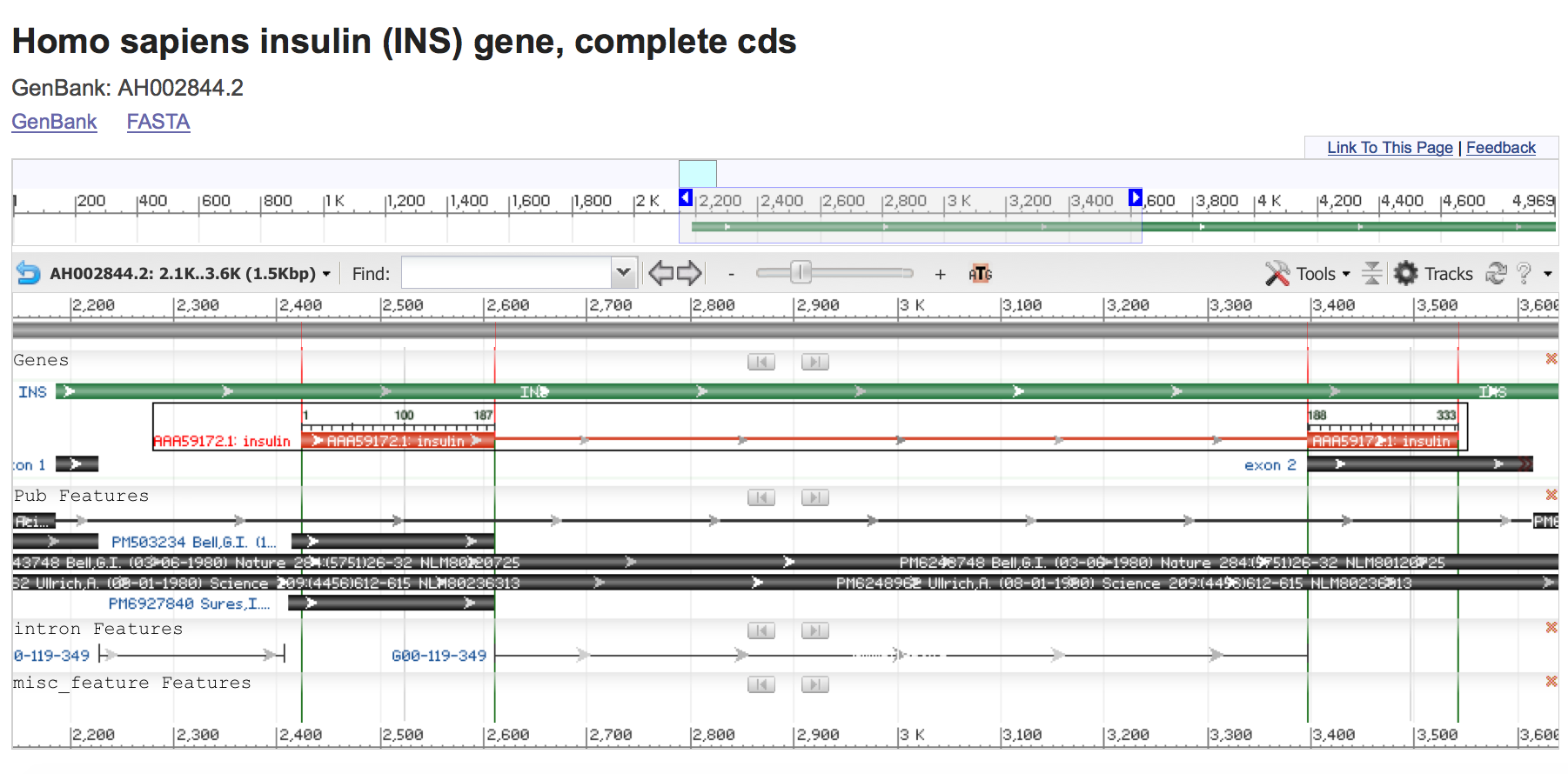


1. Click on the result 2 above since it provides the complete cds (or coding sequence for the protein). Note that the entire gene does not code for the protein – there are introns, untranslated regions (UTRs) etc.
2. Examine the Features section of this entry to find out if and how many introns/exons are there.



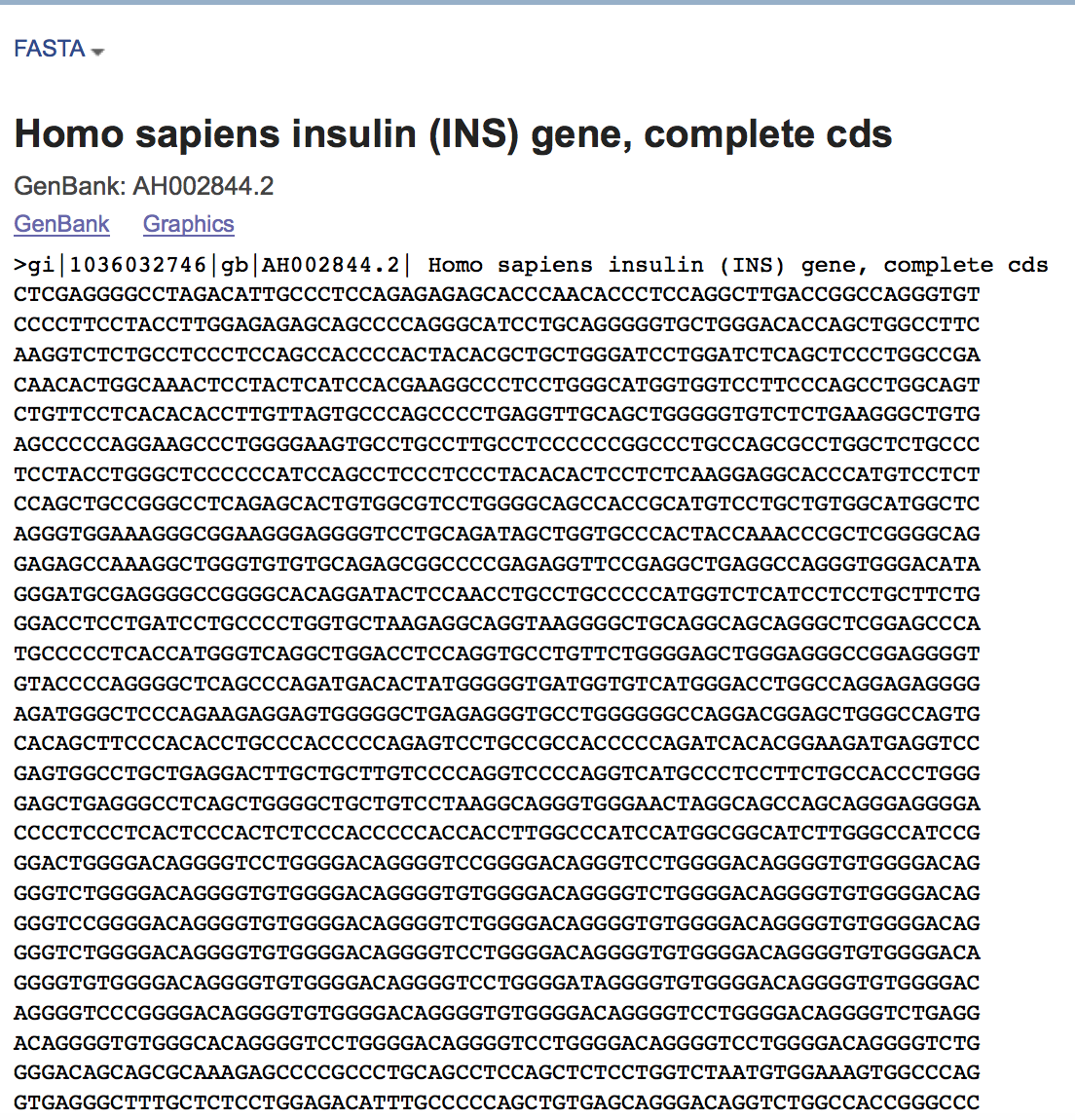
Note that the coding sequence is described as a join between the regions 2424-2610 and 3397-3542.

1. Examine the graphical representation of the gene by clicking on the *Graphics* hyperlink on the top of the page.

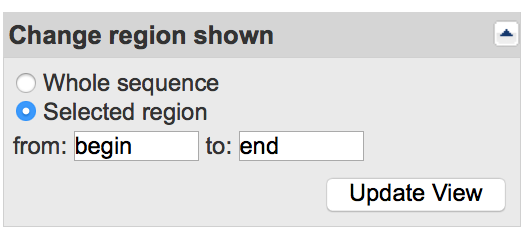


The orange/red track in the middle of the page represents the coding region of the gene. Note that there are 2 exons (denoted by thick lines) and one intron (shown as a thin line) in between.

1. Download the *FASTA* sequence by clicking on the link with that name on the top of the page. You should see the following:



This is only part of the sequence. We need to select from here the regions 2424-2610 and 3397-3542. To get these regions limit the sequence displayed by specifying the begin and end numbers of the region in the top right hand corner of the page:



Type in the numbers 2424 in place of begin and 2610 in place of end. This will show the sequence shown below:

>gb|AH002844.2|:2424-2610 Homo sapiens insulin (INS) gene, complete cds

ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAG

CCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGG

CTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGG

Save this in a file for use later in the exercise. Also find the sequence of the remaining portion of the gene – 3397-3542 and save it.

>gb|AH002844.2|:3397-3542 Homo sapiens insulin (INS) gene, complete cds

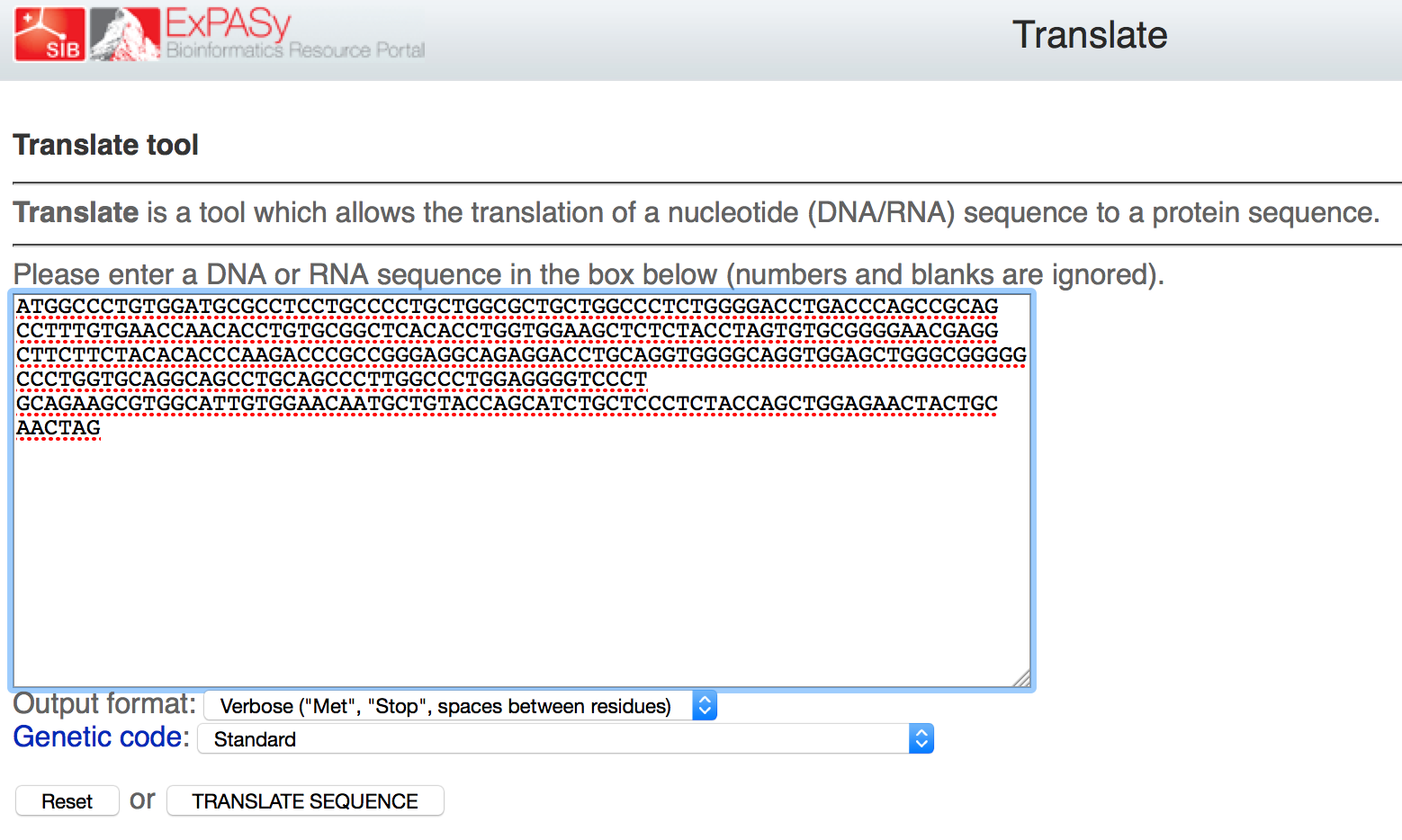
TGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCT

GCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGC

AACTAG

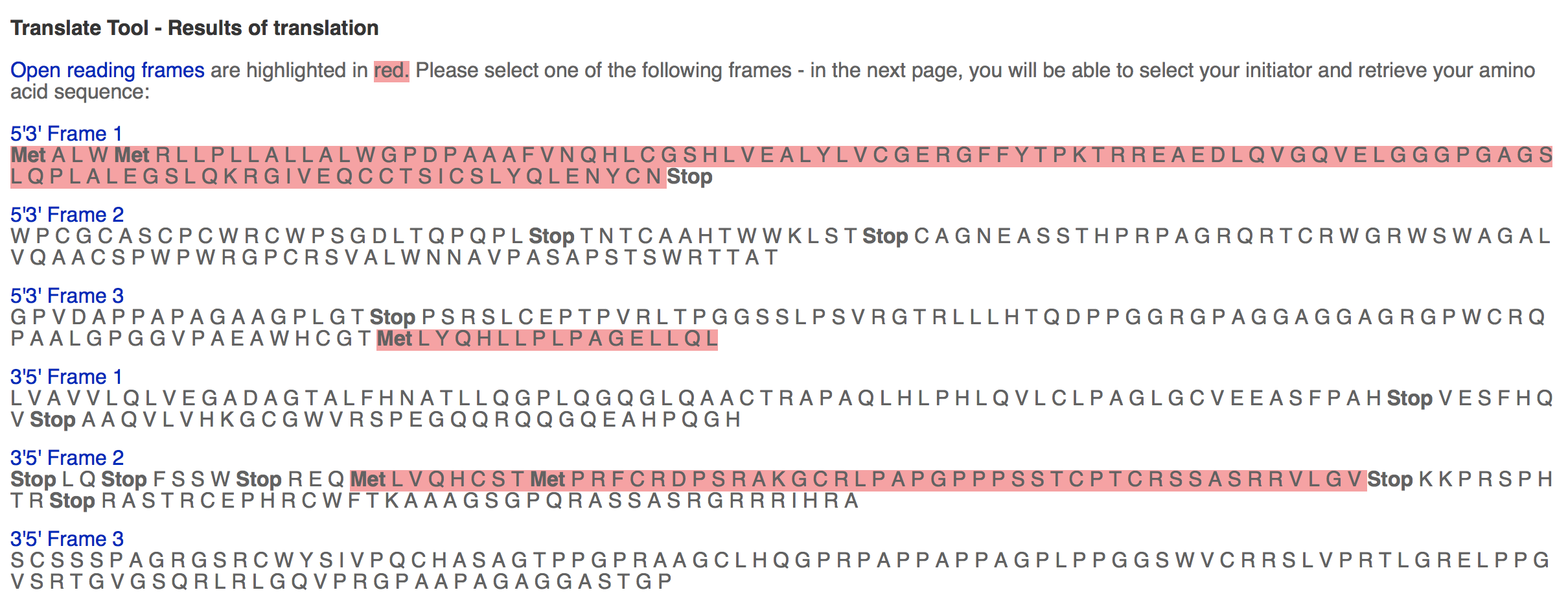
**b. Translate the coding sequence of insulin to figure out its protein sequence**

1. Point your browser to Expasy Bioinformatics Resource Portal - <http://www.expasy.org/tools/>
2. Search for the *Translate* tool on the page and click on it. Paste the sequence that you selected in 7 above in the box as shown below. Click in the text window and paste your sequence using Edit->Paste from the file where you saved the sequences. Be careful that you paste only the FASTA sequence of the CDS.
3. Your browser should look like the following:



Note that the DNA sequence pasted here includes the regions 2424-2610 and 3397-3542 pasted right next to each other.

1. Once you click on the *Translate Sequence* button above, the following screen is presented:



The results show six different sequences representing the different reading frames of DNA (three in one direction and three in the other). Only one of theses frames is the correct and is the one used to translate the protein. Typically, the correct reading frame is the longest, uninterrupted (i.e. no internal stop codons) translation.

Q11. What is the sequence of the insulin gene product? Save this sequence here or in a separate file for later use.

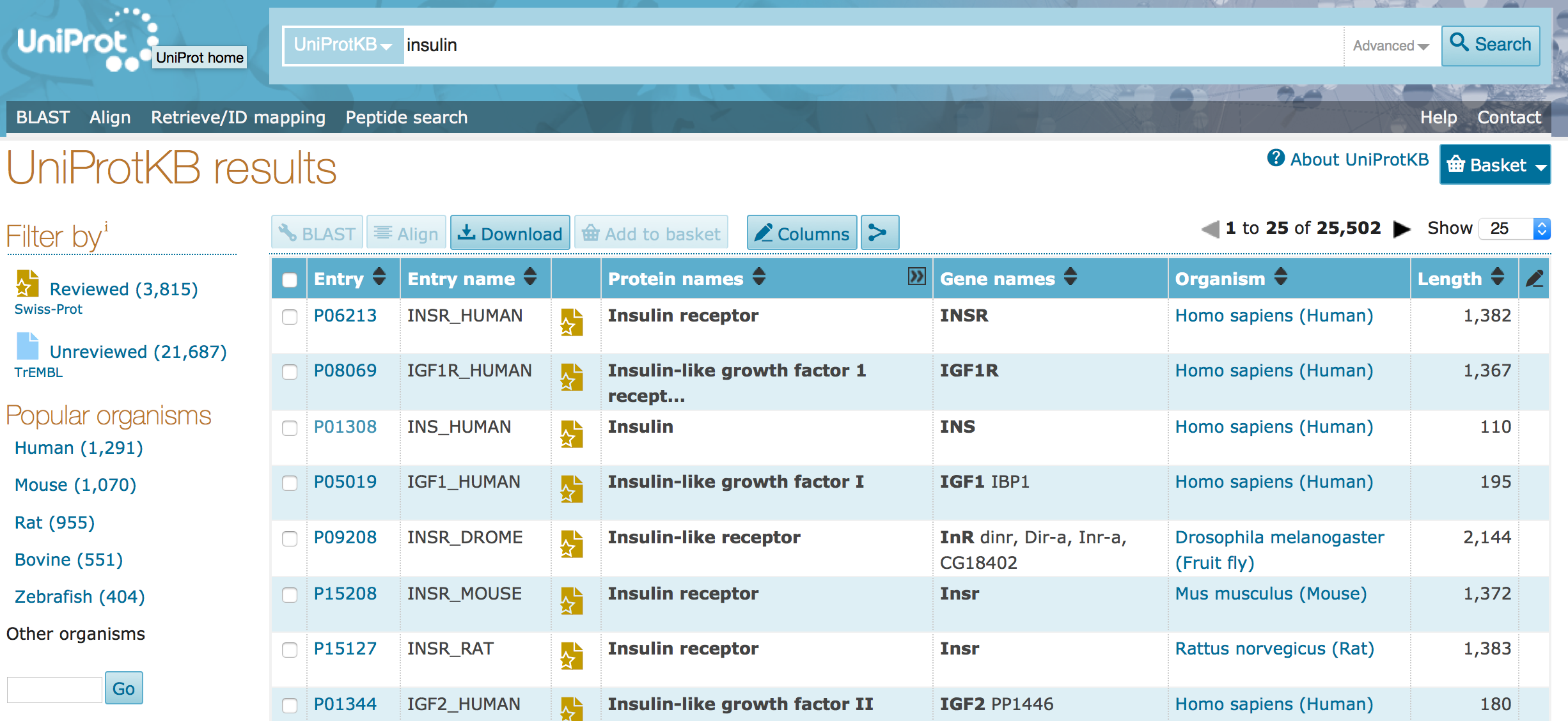
**Extension and Enrichment:**

**c. Verifying the protein sequence of insulin**

The protein sequence of insulin can be accessed from the PDB as well as the sequence database called UniProt (c). This database contains important information for researchers to study the relationship between protein sequence and protein function. There are two kinds of annotations available from this database. One, where staff scientists annotate the sequences manually based on published literature, while the other is an automated process using sophisticated software tools. It is generally accepted that the manually annotated entries have more reliable information, so we will use the human-annotated sequences from a section of the UniProt.

Point your browser to the UniProt at <http://www.uniprot.org/>

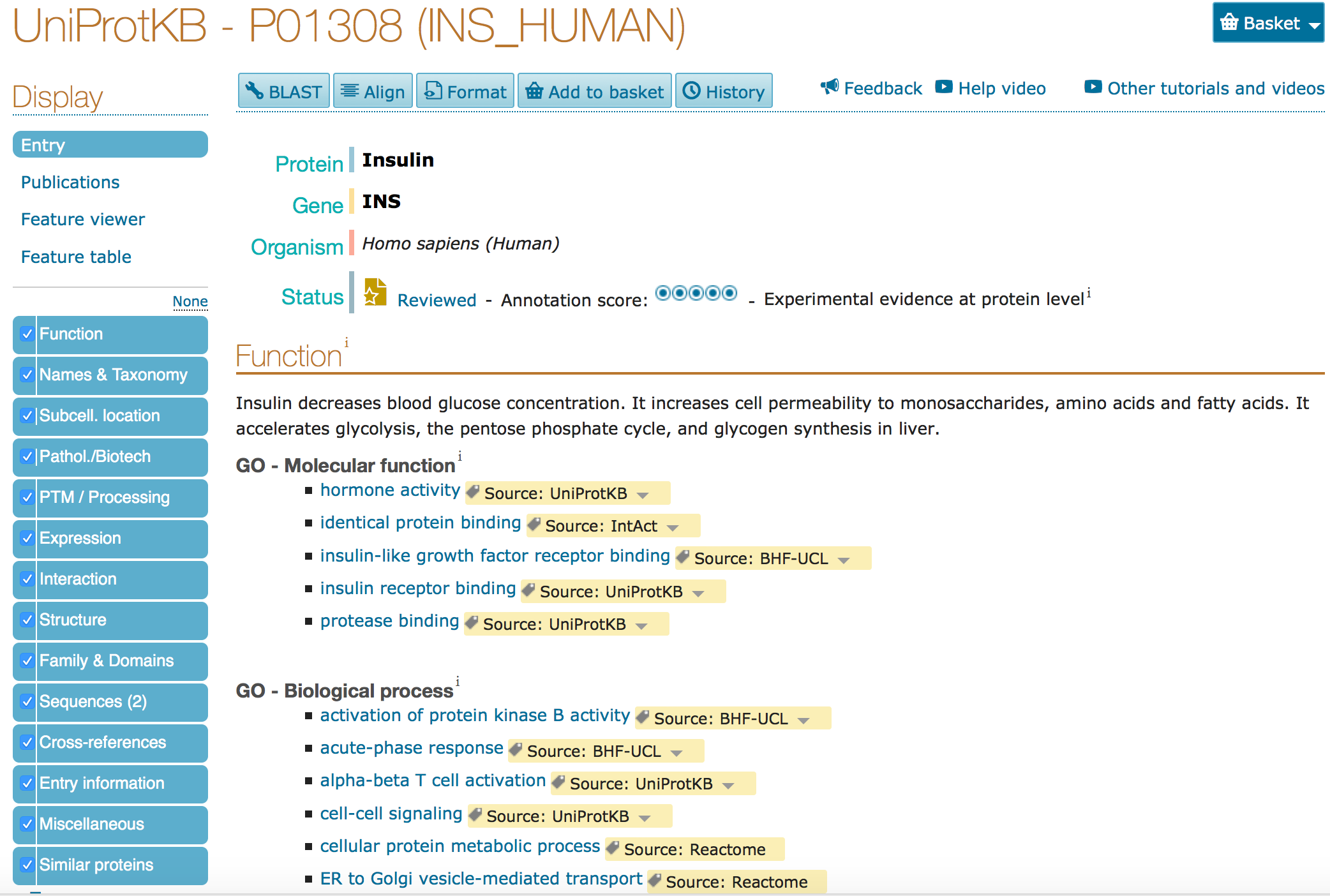
Type the search word Insulin in the top search box, you should see a number of matches to this query, as shown below:



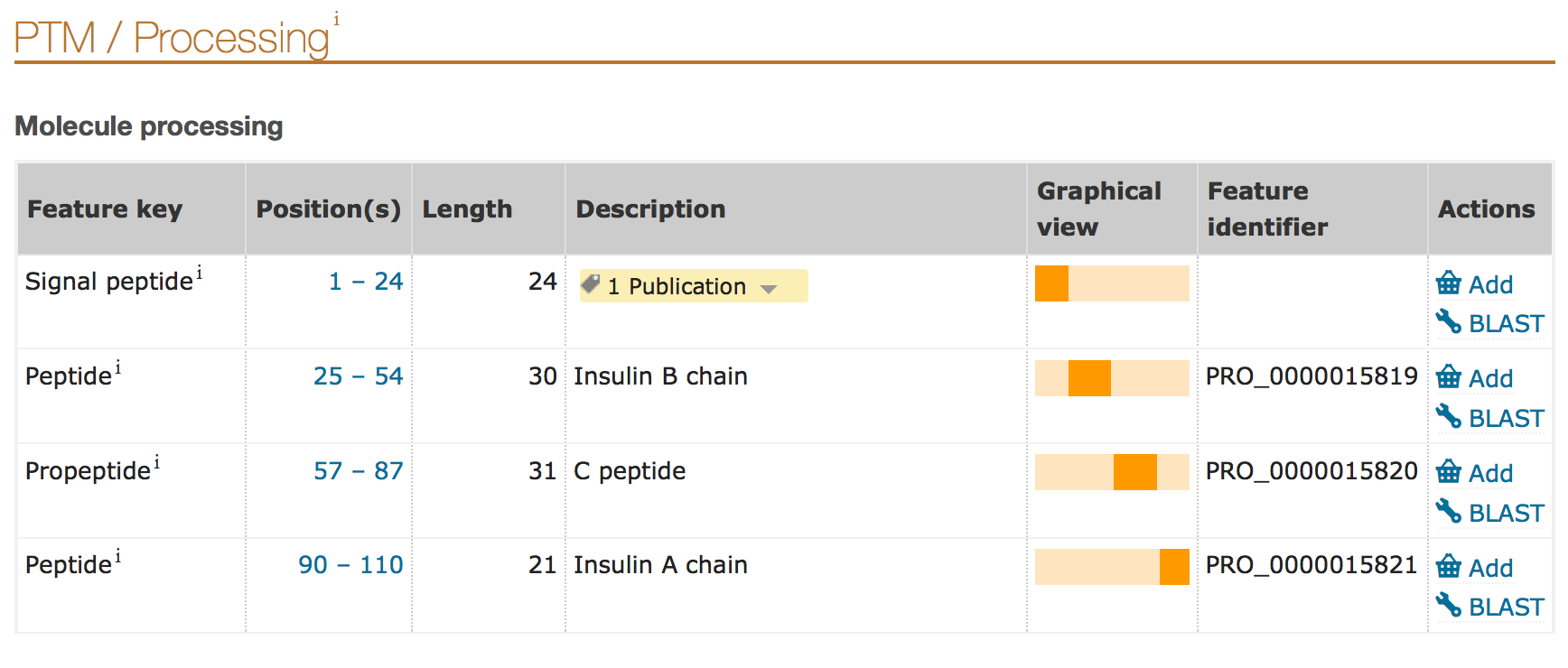
The above screen shot shows the results of running the query in September 2016.

We are interested in finding the sequence of human insulin so open the UniProt entry for INS\_HUMAN (the 3rd result listed in the results shown above).

The human insulin protein page in UniProt provides a wide range of information about the sequence, functions, variations etc. that you can explore to learn more about the protein. Descriptions of functions and mutations/variations are linked to citation literature references. Note that there is a left hand menu that allows you to browse through the various types of information about Insulin. A screen shot of the top portion of the UniProt page for human insulin is shown below

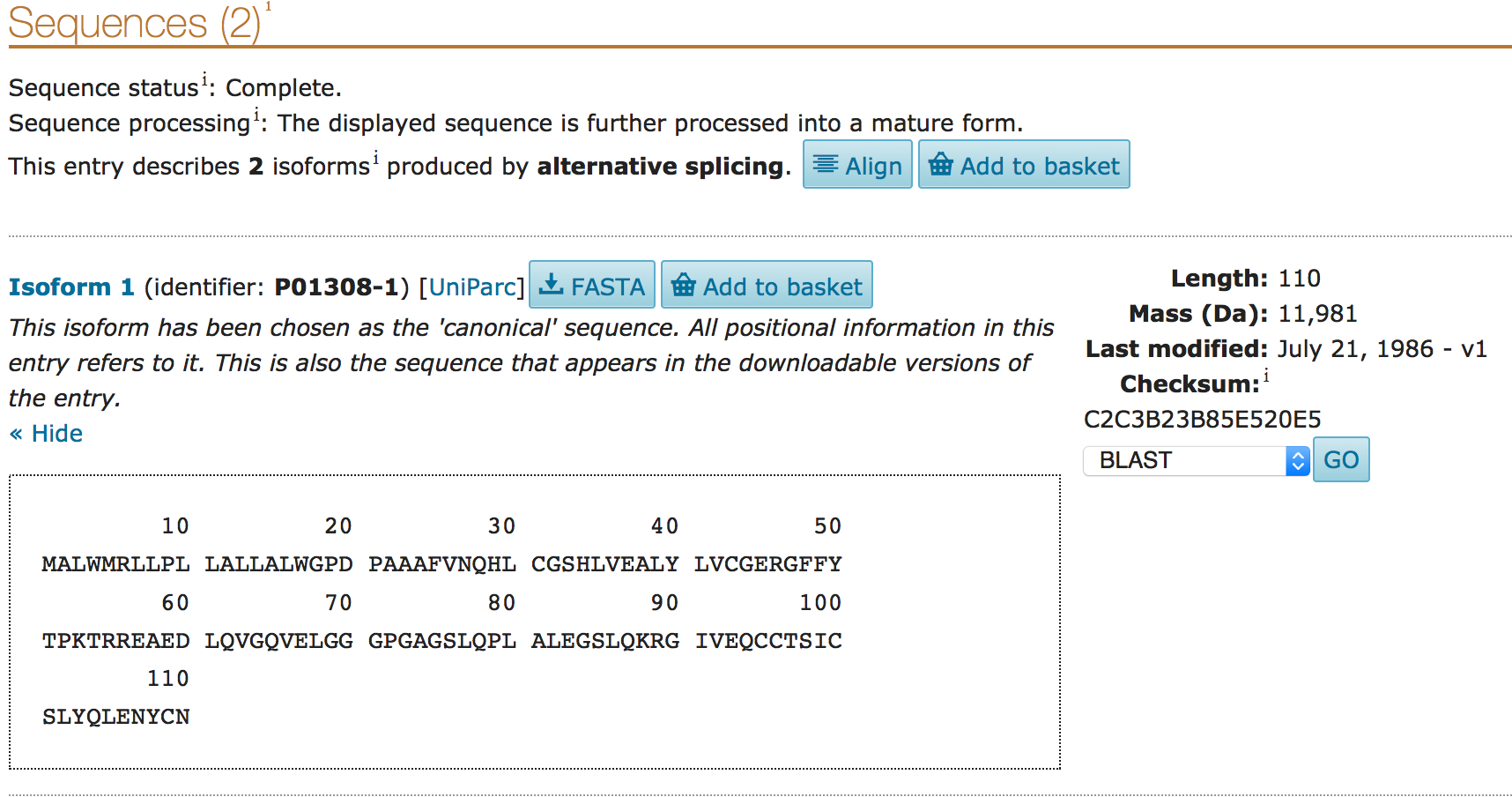


Click on the *PTM/Processing* option in the left hand menu on this page. This shows the Post Translational Modification of the insulin gene product.



Note that the amino acid residue positions for the specific regions/domains of the protein are listed here. Thus the amino acid residues 1-24 form the signal peptide, 25-54 form the B chain of Insulin etc. Keep this in mind for later analysis.

Click on the *Sequences* option in the left hand menu on this page to access the protein sequence of Insulin.



Click on the *FASTA* button to download the sequence of the protein.

>sp|P01308|INS\_HUMAN Insulin OS=Homo sapiens GN=INS PE=1 SV=1 MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

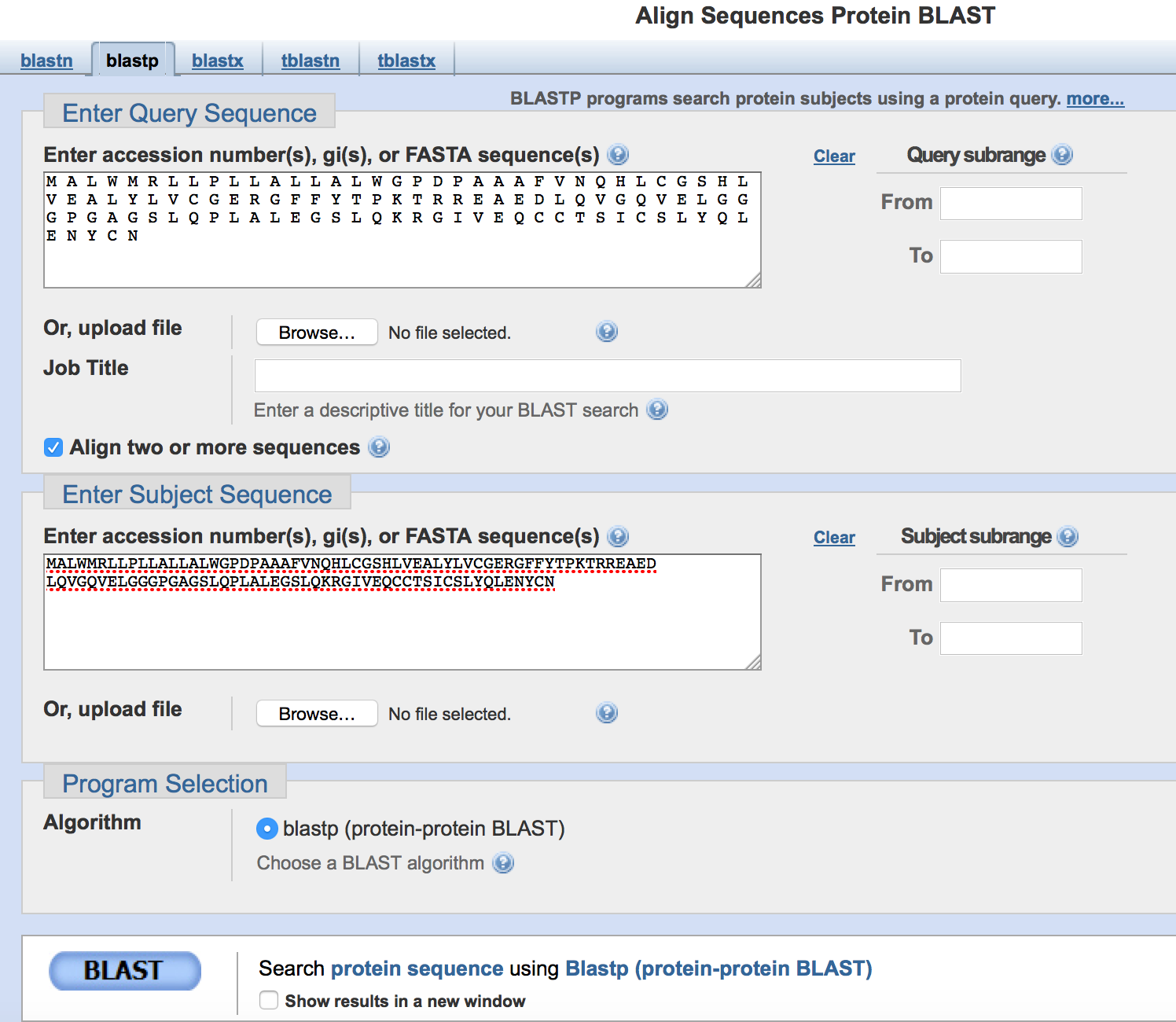
The next few steps help compare the protein sequences as determined by translating the insulin gene and that obtained from the UniProt.

To investigate if this sequence is exactly the same as the one that you derived by translating the gene sequence run a BLAST between the 2 sequences.

Basic Local Alignment Search Tool or BLAST is an algorithm that can be used to find regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Since we are interested in protein sequence comparisons we will use Protein BLAST or blastp (<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>).

While the default page asks you to provide a protein sequence that it can compare with all other known protein sequences, we are specifically interested in comparing the sequences of the insulin gene translated product and the sequence reported in UniProt. In the middle of the page there is a box next to the words *Align two or more sequences.* Click on this box to open a second text box for providing additional protein sequences. Now paste the 2 FASTA sequences of insulin – one from the translation of the insulin gene and the other from UniProt. The screen should now look like the following:



1. Click on the *BLAST* button to see the sequence alignment between the 2 sequences

**Q12.** How well do these sequences align? Explain your answer.

**Q13**. Based on all the different explorations of gene sequence, protein sequence and protein structure, summarize key events in the biosynthesis of insulin? (Hint: consider the modifications that the gene and protein products are likely to undergo to form the final protein)

**Appendix I**

**Letter codes for amino acids in a protein chain:**

A Alanine Ala

C Cysteine Cys

D Aspartic Acid Asp

E Glutamic Acid Glu

F Phenylalanine Phe

G Glycine Gly

H Histidine His

I Isoleucine Ile

K Lysine Lys

L Leucine Leu

M Methionine Met

N Asparagine Asn

P Proline Pro

Q Glutamine Gln

R Arginine Arg

S Serine Ser

T Threonine Thr

V Valine Val

W Tryptophan Trp

Y Tyrosine Tyr

**References:**

(a) The Protein Data Bank. (2000) *Nucleic Acids Research* 28: 235-242.

(b) GenBank. (2013) *Nucleic Acids Research*, Jan; 41: D36-42

(c) UniProt: a hub for protein information. (2015) *Nucleic Acids Research*, 43: D204-D212

**Version:**

This tutorial was last updated in October 2016.

Please send any comments or questions about this tutorial to info@rcsb.org.

1. Crystallographic evidence for dual coordination around zinc in the T3R3 human insulin hexamer. Ciszak, E., Smith, G.D. (1993) Biochemistry 33: 1512-1517 [↑](#footnote-ref-1)