भारतीय प्रौद्योगिकी संस्थान हैदराबाद

Indian Institute of Technology Hyderabad

BM2023 Bioinformatics

Assignment-1

Yashas Chhapola

BM23BTECH11025

Group B

Context

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver. Insulin genes in mouse and rat compose a two-gene system in which *Ins1* was retroposed from the partially processed mRNA of *Ins2*. *Ins1* originated right before the mouse–rat split (approximately 20 million years ago), and both *Ins1* and *Ins2* are under strong functional constraints in these murine species. Because the human and rat II and chicken preproinsulin genes all contain two introns in approximately the same locations, it has been suggested that the two rat insulin genes have evolved by a recent gene duplication followed by the loss of the second intron in the rat insulin I gene.[1]

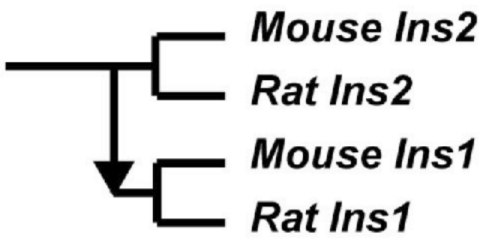

Figure 1: Origin of the duplicated rodent insulin genes. The *Ins1* gene originated by a retroposition event, shown by the arrow, in the common ancestor of the mouse (*Mus musculus*) and rat (*Rattus norvegicus*).

But why do mice and rats have two insulin genes that have identical expression patterns, while other species can survive with a single gene? This might suggest that the two genes differ in function, perhaps specializing function at different sites or times, thus generating a need to retain both genes. While some evidence for differences in the selective constraints acting on the two insulin genes has been detected, no evidence for different functions has been found, thus a convincing explanation has not been reached.[2]

For this reason, in this report, I will analyze both Insulin 1 (Ins1) and Insulin 2 (Ins2).

Problem 1
Find the FASTA sequence of rat insulin protein.

Solution FASTA is a plain text format used to represent nucleotide or peptide sequences, where nucleotides or amino acids are denoted by single-letter codes. In the FASTA format, each sequence starts with a single-line description, followed by the sequence data on subsequent lines.

The FASTA sequences for both *Ins1* and *Ins2* were retrieved from the NCBI database. Below are the details:
Insulin 1 [Rattus norvegicus]:

```
>AAA41439.1 insulin 1 [Rattus norvegicus]
MALWMRFLPLLALLVLWEPKPAQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVEDPQVPQLELGG
GPEAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
```

Insulin 2 [Rattus norvegicus]:

```
>AAA41440.1 insulin 2 [Rattus norvegicus]
MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVAQLELGG
GPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
```

Problem 2
Find the corresponding amino acids.

Solution
Insulin 1 [Rattus norvegicus]:

Table 1: Amino Acid Sequence of Rat Insulin 1 (AAA41439.1)

Position	Amino Acid Sequence
1-10	M A L W M R F L P L
11-20	L A L L V L W E P K
21-30	P A Q A F V K Q H L
31-40	C G P H L V E A L Y
41-50	L V C G E R G F F Y
51-60	T P K S R R E V E D
61-70	P Q V P Q L E L G G
71-80	G P E A G D L Q T L
81-90	A L E V A R Q K R G
91-100	I V D Q C C T S I C
101-110	S L Y Q L E N Y C N

The 110-residue protein (molecular weight: 12,420.51 Da) exhibits the following distribution of amino acids:

Table 2: Amino Acid Composition of Insulin 1

Amino Acid	**Count**	**Percentage (%)**
Alanine (A)	8	7.3
Arginine (R)	6	5.5
Asparagine (N)	2	1.8
Aspartic Acid (D)	3	2.7
Cysteine (C)	6	5.5
Glutamine (Q)	8	7.3
Glutamic Acid (E)	9	8.2
Glycine (G)	8	7.3
Histidine (H)	2	1.8
Isoleucine (I)	2	1.8
Leucine (L)	18	16.4
Lysine (K)	4	3.6
Methionine (M)	2	1.8
Phenylalanine (F)	4	3.6
Proline (P)	8	7.3
Serine (S)	3	2.7
Threonine (T)	3	2.7
Tryptophan (W)	2	1.8
Tyrosine (Y)	4	3.6
Valine (V)	8	7.3

Insulin 2 [Rattus norvegicus]:

Table 3: Amino Acid Sequence of Insulin 2

Position	**Amino Acid Sequence**
1-10	M A L W I R F L P L
11-20	L A L L I L W E P R
21-30	P A Q A F V K Q H L
31-40	C G S H L V E A L Y
41-50	L V C G E R G F F Y
51-60	T P M S R R E V E D
61-70	P Q V A Q L E L G G
71-80	G P G A G D L Q T L
81-90	A L E V A R Q K R G
91-100	I V D Q C C T S I C
101-110	S L Y Q L E N Y C N

The 110-residue protein (molecular weight: 12339.40 Da) exhibits the following distribution of amino acids:

Table 4: Amino Acid Composition of Insulin 2

Amino Acid	**Count**	**Percentage (%)**
Alanine (A)	9	8.2
Arginine (R)	7	6.4
Asparagine (N)	2	1.8
Aspartic Acid (D)	3	2.7
Cysteine (C)	6	5.5
Glutamine (Q)	8	7.3
Glutamic Acid (E)	8	7.3
Glycine (G)	9	8.2
Histidine (H)	2	1.8
Isoleucine (I)	4	3.6
Leucine (L)	18	16.4
Lysine (K)	2	1.8
Methionine (M)	2	1.8
Phenylalanine (F)	4	3.6
Proline (P)	6	5.5
Serine (S)	4	3.6
Threonine (T)	3	2.7
Tryptophan (W)	2	1.8
Tyrosine (Y)	4	3.6
Valine (V)	7	6.4

Problem 3
Collect the information about different amino acids.

Solution The standard 20 amino acids are listed below with their 3-letter and 1-letter symbols:

Table 5: Amino Acids with Symbols

Amino Acid Name	**3-Letter Symbol**	**1-Letter Symbol**
Glycine	Gly	G
Alanine	Ala	A
Phenylalanine	Phe	F
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Asparagine	Asn	N
Glutamine	Gln	Q
Serine	Ser	S
Threonine	Thr	T
Tyrosine	Tyr	Y
Cysteine	Cys	C
Methionine	Met	M
Lysine	Lys	K
Arginine	Arg	R
Proline	Pro	P
Histidine	His	H
Tryptophan	Trp	W

Amino acids can be classified based on various criteria, such as chemical structure, physical and chemical properties, polarity, hydrophobicity, and charge. Below are some common classifications:

Classification Based on Physical and Chemical Properties:

- Acidic: Asp, Glu.
- Basic: Lys, Arg, His.
- Aromatic: Tyr, Trp, Phe.
- Sulfur-containing: Cys, Met.
- Uncharged hydrophilic: Ser, Thr, Asn, Gln.
- Hydrophobic (inactive): Gly, Ala, Val, Leu, Ile.
- Special structure: Pro.

Classification Based on Polarity and Charge:

- Polar/Hydrophilic: Asn, Gln, Ser, Thr, Lys, Arg, His, Asp, Glu (and sometimes Cys and Tyr).
- Nonpolar/Hydrophobic: Gly (sometimes), Ala, Val, Leu, Ile, Pro (and sometimes Tyr), Phe, Trp, Met.

- Negatively charged at neutral pH (acidic): Asp, Glu (and sometimes Cys).
- Positively charged at neutral pH (basic): Lys, Arg (and sometimes His).

The overlapping nature of these classifications can be better visualized using a Venn diagram that groups amino acids based on their nature (aliphatic or aromatic), size (small or tiny), hydrophobicity or polarity (polar or charged), and other properties.

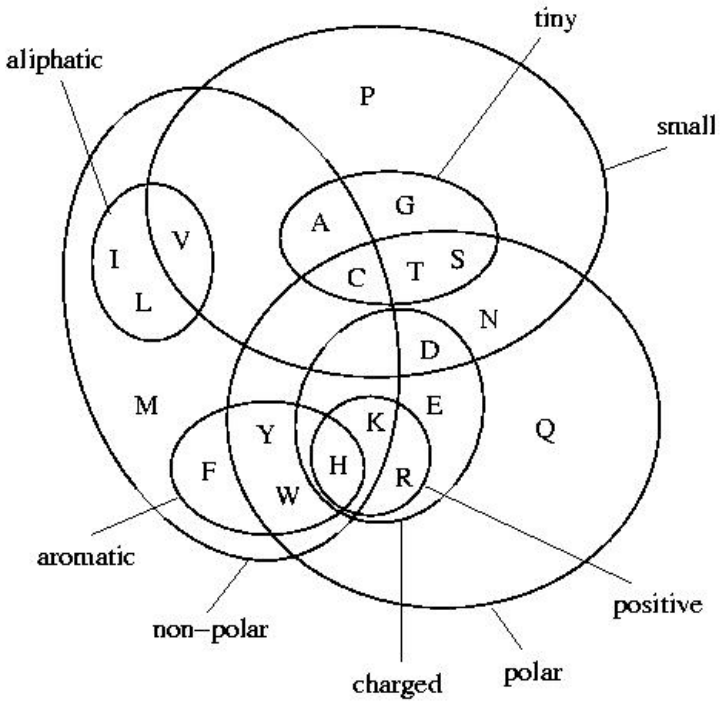

Figure 2: Classification of Amino Acids using Venn Diagram

Problem 4
In which tissues the genes are expressed?

Solution

Expression of Insulin 1

Insulin 1 in Rattus norvegicus (rat) is expressed in pancreatic tissues and six other cell types or tissues. NCBI provides data for some common tissues that express Insulin 1, while expression in pancreatic tissues is confirmed by data from UniProt and Bgee. The absence of pancreatic tissue data in NCBI dataset may explain the discrepancy between NCBI findings and UniProt findings. Moreover, NCBI data also provides insights into expression levels at different ages of the rat.

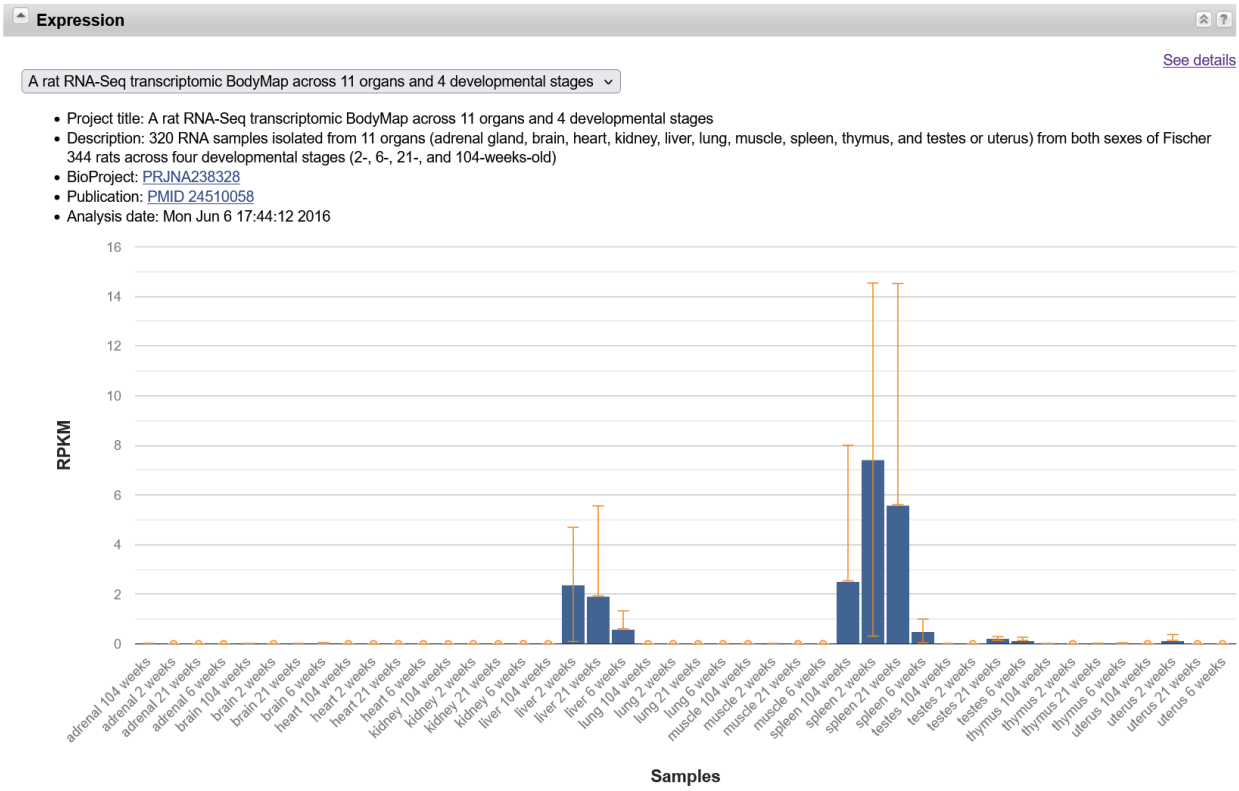

Figure 3: NCBI Dataset for Insulin 1 Expression

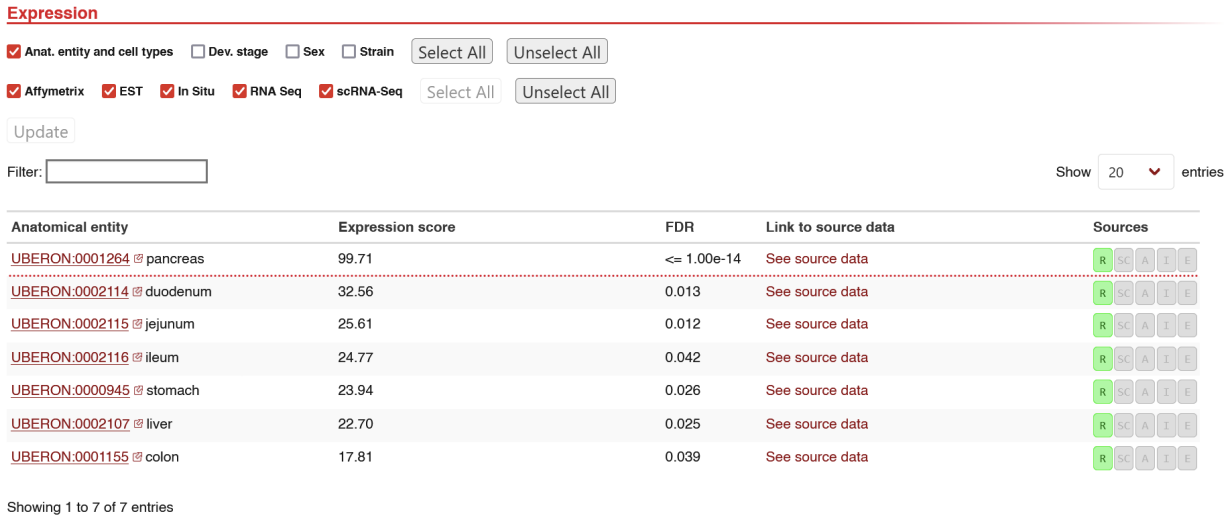

Figure 4: UniProt/Bgee Dataset for Insulin 1 Expression

Expression of Insulin 2

Insulin 2 in Rattus norvegicus (rat) is expressed in pancreatic tissues and 6 other cell types or tissues. A similar discrepancy in datasets between NCBI and UniProt is found in the case of Insulin 2.

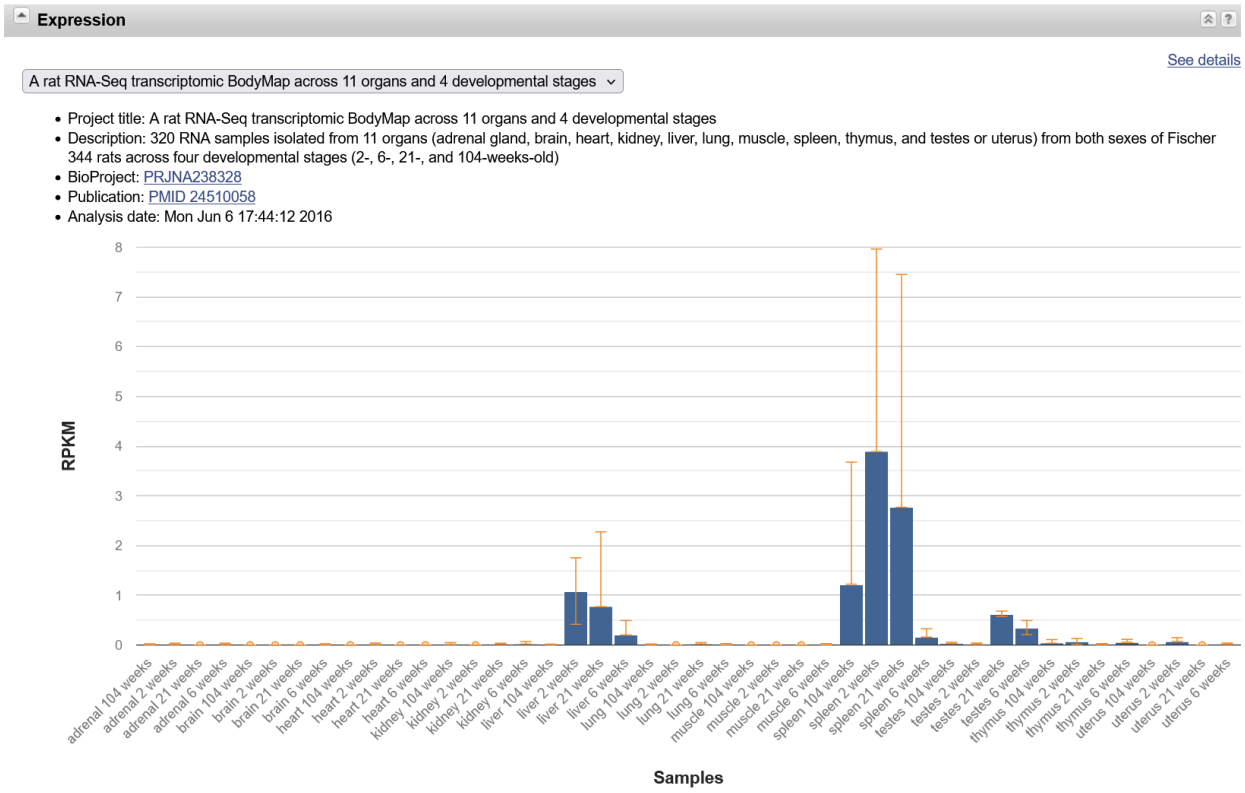

Figure 5: NCBI Dataset for Insulin 2 Expression

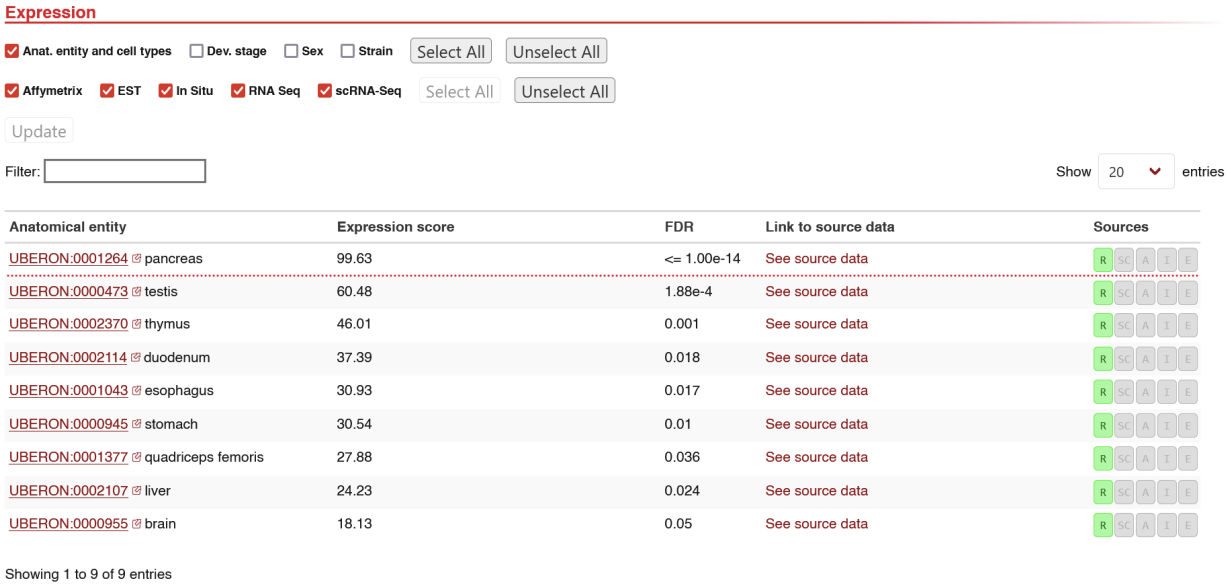

Figure 6: UniProt/Bgee Dataset for Insulin 2 Expression

Problem 5
What secondary structures you can predict from it?

Solution Proteins adopt specific three-dimensional shapes governed by their secondary structures, which are local folded patterns stabilized by hydrogen bonds. The main types of secondary structures are:

- **Beta Sheets and Beta Strands:**
  - Beta-sheets consist of several Beta-strands, stretched segments of the polypeptide chain kept together by a network of hydrogen bonds between adjacent strands.
  - Extended, zigzag segments of a protein chain. β sheets can be **parallel** or **antiparallel**, depending on strand orientation.
  - Provide structural stability and rigidity to proteins.
- **Alpha Helices:**
  - Coiled, rod-like structures stabilized by hydrogen bonds between every fourth amino acid.
  - Play a key role in protein folding and function.
- **Other Secondary Structures:**
  - Regions of the polypeptide chain that do not adopt regular, repeating patterns like alpha-helices (H) or beta-sheets (E). These segments are often referred to as "other" secondary structures due to their unstructured or irregular nature.

For the prediction of secondary protein structure in Insulin 1 and Insulin 2 of rats, I am using JPred. (A Protein Secondary Structure Prediction Server)

Insulin 1

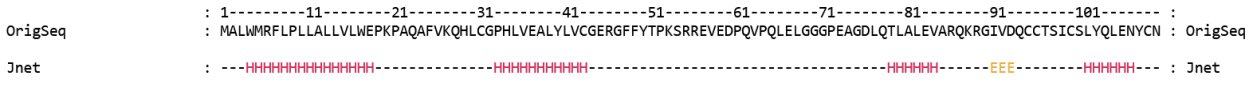

Figure 7: Insulin 1 Secondary Structure Prediction using Jpred

In this sequence, H represents Helical (Alpha-Helices), E represents Extended (Beta-Strands) and Dash (-) represent other types of secondary structure.

The beta-strand (E) predictions in Insulin 1 with **JNETCONF scores of 0/1** (The confidence estimate for the prediction, high values mean high confidence) lack statistical reliability. To confirm whether these regions adopt beta-strand conformations, experimental validation is necessary. Computational predictions alone cannot resolve such low-confidence annotations, underscoring the importance of integrating experimental data for accurate structural characterization.

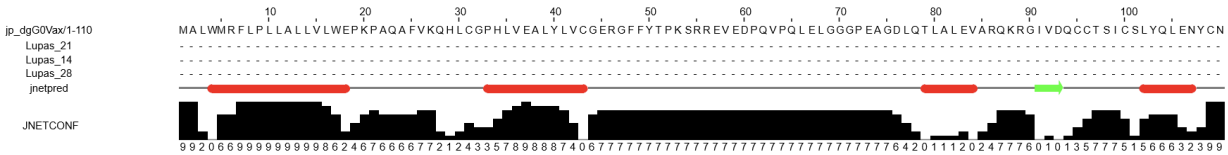

Figure 8: Confidence Factor in Prediction (Insulin 1)

The predicted 3D structure of Insulin 1 largely aligns with the conclusions drawn from the FASTA sequence analysis.

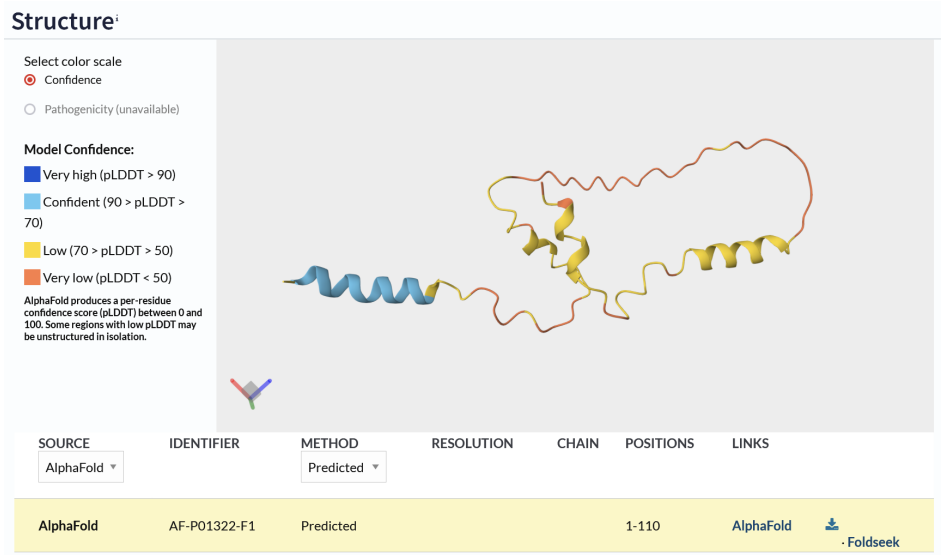

Figure 9: Predicted 3D Structure of Insulin 1

Figure 10 provides a clear picture about the dashes representation.

```
Note on coiled coil predictions  - = less than 50% probability
                                 c = between 50% and 90% probability
                                 C = greater than 90% probability
```

Figure 10: Dashes Representation

Insulin 2

```
                 : 1---------11--------21--------31--------41--------51--------61--------71--------81--------91--------101------- :
OrigSeq          : MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN : OrigSeq

Jnet             : --HHHHHHHHHHHHHHHH----------E--HHHHHHHHHHH---------------------------------HHHHHHH----------------HHHHHH--- : Jnet
```

Figure 11: Insulin 2 Secondary Structure Prediction using Jpred

H represents Helical (Alpha Helix), E represents Extended (Beta Strands) and Dash (-) represent other types of secondary structure.

The predicted beta-strand (E) at residue 30 (Leucine, L) has a **JNETCONF score of 0** (The confidence estimate for the prediction. High values mean high confidence), indicating no statistical confidence in this assignment. This contradiction between the annotation (E) and confidence score (0) suggests the beta-strand prediction at this position is unreliable.

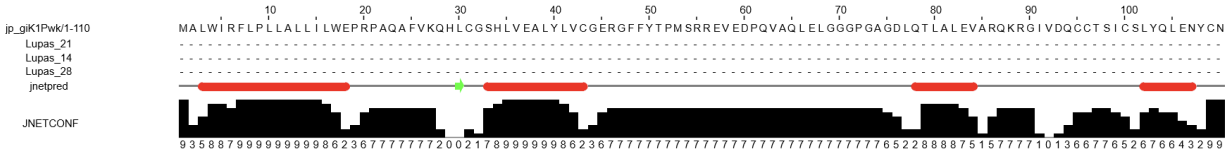

Figure 12: Confidence Factor in Prediction (Insulin 2)

The predicted 3D structure of Insulin 2 largely aligns with the conclusions drawn from the FASTA sequence analysis.

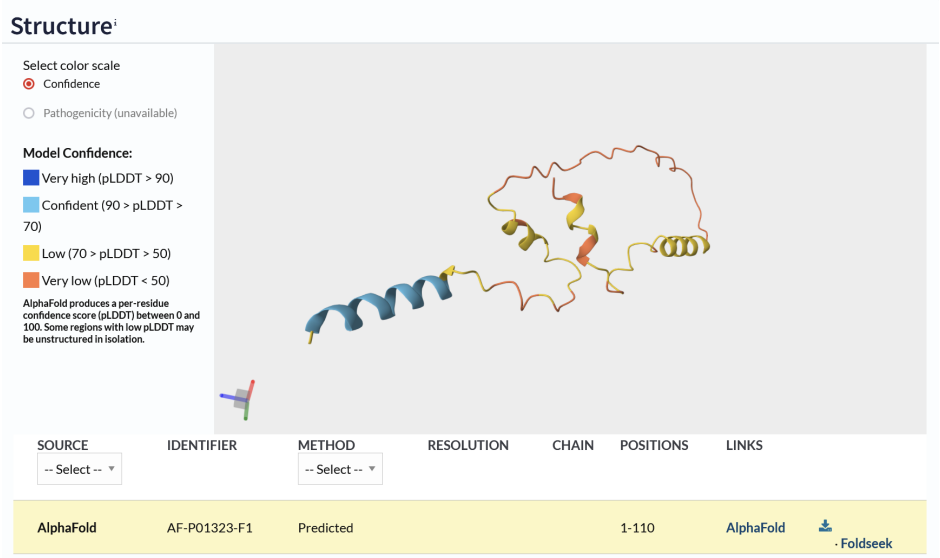

Figure 13: Predicted 3D Structure of Insulin 2

Sources

1. The 3D structures of both proteins were obtained from UniProt.org.

2. Secondary structure predictions from the FASTA sequences were performed using JPred4 (available at www.compbio.dundee.ac.uk/jpred4).

3. Expression data was sourced from UniProt.org, Bgee.org and NCBI.

4. The FASTA sequences were retrieved from the NCBI database.

References

[1] Meng-Shin Shiao, Ben-Yang Liao, Manyuan Long, and Hon-Tsen Yu. Adaptive evolution of the insulin two-gene system in mouse. *Genetics*, 178:1683–91, 04 2008.

[2] David M. Irwin. Evolution of the insulin gene: Changes in gene number, sequence, and processing. *Frontiers in Endocrinology*, 12, 2021.