TASK 2: Multiple Sequence Alignment

Protein Family: Ubiquitin Family

Introduction

Ubiquitin is a small regulatory protein present in almost all eukaryotic cells. It plays a key role in protein degradation, cell cycle control, DNA repair, and signaling. The human genome contains multiple ubiquitin-coding genes (UBB, UBC, UBA52, RPS27A, UBA80) that are highly similar but may differ in their N-terminal extensions or fusion partners.

Because of its small size and high conservation, ubiquitin proteins are ideal candidates for Multiple Sequence Alignment (MSA) to study conserved residues, motifs, and evolutionary relationships.

MSA can highlight:

- Highly conserved residues critical for ubiquitin's structure and function (e.g., Gly76 for conjugation).
- Subtle sequence variations in N-terminal extensions or fusion regions.
- Potential functional motifs shared among the family.

Methodology

1. Sequence Retrieval

1. Open UniProt (https://www.uniprot.org).
2. Search for each protein name (UBB, UBC, UBA52, RPS27A, UBA80).
3. Open the protein entry → click FASTA → copy the sequence.
4. Alternatively, click Download → select FASTA format.
5. Save each sequence for alignment.

2. Multiple Sequence Alignment (MSA)

Using Clustal Omega

1. Open Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/).
2. Paste the 5 FASTA sequences or upload the downloaded files.
3. Click **Submit** to run the alignment.
4. Download alignment in Clustal format, FASTA aligned, or color-coded alignment.

Results

1. Overall Alignment Strategy

- The alignment uses gaps - to line up similar amino acid residues, maximizing the similarity scores.
- The numbering at the end of each line tracks the residue position within the original, full sequence.

2. Conservation and Similarity

The symbols below the alignment blocks indicate the degree of residue conservation across the five sequences:

- *** (Asterisk): Identical Residues.** The amino acid is identical in all sequences. These positions are highly constrained and likely essential for a conserved function (e.g., active site residues or structural integrity).
- **: (Colon): Strong Conservation.** The amino acids are different but possess highly similar physicochemical properties (e.g., both are hydrophobic or both are basic). This is a conservative substitution.
- **. (Period):** Weak Conservation. The amino acids share some, but fewer, properties.

3. The Ubiquitin Core Domain

The most significant finding is the region of high conservation that starts around residue positions 25-34 for the four proteins containing the ubiquitin domain (P0CG47, P62979, P62987) and extending to around position 84 for UBE2C.

- This block represents the ubiquitin fold. The near-perfect identity among P0CG47, P62979, and P62987 confirms that the ubiquitin unit released from the precursors (RS27A and RL40) is structurally and functionally identical to the unit derived from the polyubiquitin precursor (UBB).
- The C-terminal sequence -LRLRGG (Leucine-Arginine-Leucine-Arginine-Glycine-Glycine) is crucial. The terminal double-glycine -GG motif is the site cleaved to produce mature ubiquitin and is the residue used for isopeptide linkage to target proteins.

4. Protein-Specific Domains (Divergence)

The alignment also shows large regions of **dissimilarity** (indicated by extensive gaps and lack of symbols):

- **UBE2C (Ubiquitin-Conjugating Enzyme E2C):** This protein, which is not a ubiquitin precursor, shares very limited similarity with the ubiquitin domain proteins (P0CG47, etc.). Its sequence aligns only loosely in a region where E2 enzymes share similarity, highlighting its distinct role in the ubiquitination cascade.
- **ISG15 (Interferon-Stimulated Gene 15):** This protein is a Ubiquitin-Like Modifier (UBL). It shows poor sequence conservation with canonical ubiquitin (P0CG47), having large gaps and a different overall length. This confirms that ISG15 has a different structure and distinct functional role, primarily in the **immune response**.
- **Ribosomal Proteins (P62979, P62987):** The C-terminal extensions of RPS27A and RPL40 (after the -RGG cleavage site) show complete divergence. These sequences represent the actual ribosomal protein domains (S27a and L40) which are non-homologous to each other or to ubiquitin, reflecting their unique function as structural components of the ribosome.