Task 5: Gene Ontology(GO) Terms Prediction

Instruction

The Gene Ontology project (GO) provides a controlled vocabulary to describe gene and gene product attributes in any organism. There are 3 disjoint categories: cellular component, molecular function and biological process. Predict the GO term for a given protein sequence.

Input

MEFVTNYTLEELKKRFTELGLEPYRAKQVFRWVYKKFVTDFEKMTDLGKKHRELLKEHFAFHPLEKLDRVEAPDAVKYLFKTKDGHILETVLIKERDHYTLCVSSQIGCAVGCTFCATALDGLKRNLS
TAEIIDQYLQVQQDLGEEKIRNVVFMGMGEPLANYENVRKAVEIMVSPEGLDLSKRRITISTSGIVAQIKRMAQDPVMKEVNLAVSLNAVSQKKREELMPLTKTNTLEELMEVLKNYPLPKYRRITLEY
VLIKGVNDSPNDAERLAKLIGRHKKKFKVNLIPFNPDPNLPYERPALTDIMKFQKVLWKYGISNFVRFSKGVEVFGACGQLRTQRLQLQRV

Ground Truth

cytoplasm ; 4 iron, 4 sulfur cluster binding ; metal ion binding ; rRNA (adenine-C2-)-methyltransferase activity ; rRNA binding ; tRNA (adenine-C2-)-methyltransferase activity ; tRNA binding ; rRNA base methylation ; tRNA methylation

OPI-tuned model

cytoplasm ; 4 iron, 4 sulfur cluster binding ; metal ion binding ; rRNA (adenine-C2-)-methyltransferase activity ; rRNA binding ; tRNA (adenine-C2-)-methyltransferase activity ; tRNA binding ; rRNA base methylation

GPT-4o

To predict the Gene Ontology (GO) term for a given protein sequence, one approach is to use machine learning models trained on a dataset of protein sequences with known GO annotations. The process generally involves feature extraction from the protein sequence, model training, and prediction.
Here's a high-level approach:
1. **Feature Extraction**: - Convert the protein sequence into a numerical representation that can be used as input for a machine learning model. Common methods include: - **Amino acid composition**: Count the frequency of each amino acid in the sequence. - **K-mer frequency**: Count the occurrence of subsequences of length k.- **Position-specific scoring matrix (PSSM)**: Represent evolutionary information from multiple sequence alignments.- **Physicochemical properties**: Encode properties like hydrophobicity, charge, and molecular weight.
2. **Model Selection**: - Use a classifier such as Random Forest, Support Vector Machine, or deep learning models like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN). - For deep learning models, techniques like one-hot encoding or embedding layers can be applied directly to the protein sequence.
3. **Training**: - Train the model on a labeled dataset where protein sequences are associated with known GO terms.- The dataset could be sourced from databases like UniProt, where sequences are annotated with GO terms.
4. **Prediction**: - Once the model is trained, input the new protein sequence (like the one provided) into the model to predict the most likely GO term(s).
5. **Post-Processing** - Apply thresholds or confidence scoring to refine the predictions. - Consider hierarchical relationships in GO terms to make more accurate predictions.
For the specific sequence provided:
```
MEFVTNYTLEELKKRFTELGLEPYRAKQVFRWVYKKFVTDFEKMTDLGKKHRELLKEHFAFHPLEKLDRVEAPDAVKYLFKTKDGHILETVLIKERDHYTLCVSSQIGCAVGCTFCATALDGLKRNLSTAEIIDQYLQVQQDLGEEKIRNVVFMGMGEPLA
NYENVRKAVEIMVSPEGLDLSKRRITISTSGIVAQIKRMAQDPVMKEVNLAVSLNAVSQKKREELMPLTKTNTLEELMEVLKNYPLPKYRRITLEYVLIKGVNDSPNDAERLAKLIGRHKKKFKVNLIPFNPDPNLPYERPALTDIMKFQKVLWKYGISNFVR
FSKGVEVFGACGQLRTQRLQLQRV
```
If a model is already in place, you would input this sequence and receive predictions for its GO terms across the three categories (cellular component, molecular function, biological process).
Would you like to see a code example for predicting GO terms, or any further specific details?

Llama-3.1-8B-Instruct

To predict the GO term for a given protein sequence, I'll use a combination of bioinformatics tools and algorithms. Since you didn't provide any specific information about the protein sequence, I'll assume it's a novel protein and try to identify its functional categories using various databases and tools. After analyzing the protein sequence using various tools, including BLAST, Pfam, and InterProScan, I've identified some potential functional domains and motifs. Here are my predictions for the GO terms: 1. Cellular Component: * Based on the presence of a transmembrane domain (TM) and a signal peptide, I predict that this protein is likely to be a membrane protein, possibly a receptor or a transporter. Therefore, I assign the GO term "cell membrane" (GO:0005886) as a potential cellular component. 2. Molecular Function: * The presence of a catalytic domain (Pfam: PF00067) suggests that this protein may have enzymatic activity. I predict that it may be involved in catalyzing a chemical reaction, possibly a hydrolysis reaction. Therefore, I assign the GO term "catalytic activity" (GO:0003824) as a potential molecular function. 3. Biological Process: * The presence of a signal peptide and a transmembrane domain suggests that this protein may be involved in cellular signaling or transport processes. Additionally, the presence of a catalytic domain suggests that it may be involved in metabolic processes. Therefore, I assign the GO term "cellular process" (GO:0009987) as a potential biological process. Please note that these predictions are based on computational analysis and may require experimental validation to confirm the protein's function and GO terms.

Claude 3.5 Sonnet

To predict the Gene Ontology (GO) term for this protein sequence, I would need to use specialized bioinformatics tools and databases, which I don‘t have direct access to. However, I can provide you with a general approach to predict GO terms for a given protein sequence and some insights based on the sequence you provided. General approach to predict GO terms: 1. Sequence similarity search: Use tools like BLAST to find similar proteins with known GO annotations. 2. Protein domain analysis: Identify conserved domains using tools like InterProScan or PFAM. 3. Machine learning methods: Use pre-trained models that can predict GO terms based on sequence features. 4. Literature mining: Search for publications mentioning this or similar proteins. 5. Integrative approaches: Combine multiple methods for more accurate predictions. Observations about the given sequence: 1. Length: The sequence is approximately 420 amino acids long, which is a moderate size for a protein. 2. Composition: It contains a mix of hydrophobic (e.g., L, V, I, F) and hydrophilic (e.g., K, R, E, D) amino acids, suggesting it might have both globular and membrane-interacting regions. 3. Potential functional motifs: - “GEPLANY” might be part of a binding site or active site. - “MSPEG” could be involved in protein-protein interactions. - “KKREELMP” might be a nuclear localization signal due to the basic residues. 4. Recurring patterns: There are several lysine (K) and arginine (R) rich regions, which might indicate DNA/RNA binding or nuclear localization. Based on these observations, some possible GO term predictions could be: 1. Cellular Component: - GO:0005634 (nucleus) - due to potential nuclear localization signals - GO:0005737 (cytoplasm) - as many proteins are present here 2. Molecular Function: - GO:0003677 (DNA binding) - due to the presence of basic amino acid clusters - GO:0005515 (protein binding) - common function for many proteins 3. Biological Process: - GO:0006355 (regulation of transcription, DNA-templated) - if it's a nuclear protein involved in DNA binding - GO:0016032 (viral process) - some viral proteins have similar characteristics Please note that these are speculative predictionsbased on limited information. For accurate GO term prediction, you would need to use specialized bioinformatics tools and databases, and ideally have experimental data to support the predictions.