




Task 6: Function Description Prediction

Instruction	Generate the functional description in free-text form based on the protein sequence.
Input	MAAPSGVHLLVRRGSHRIFSSPLNHIYLHKQSSSQRRNFFFRQRDISHSIVLPAAVSSAHPVPKHIKKPDYVTTGIVPDWGDSEIEVKNEDQIQGLHQACQLARHVLLLAGKSLKVDMTTEEIDALVHREIISHNAYPSPLGYGGFPKSVCTS VNNVLCHGIPDSRPLQDGDIIINIDVTVYYNGYHGDTSETFLVGNVDEC GKKLVEVARRRCRDEAIAACRAGAPFSVIGNTISHITHQNGFQVCPHFVGHGIGSYFHGHPEIWHHANDSDLPMEEGMAFTIEPIITEGSPEFKVLEDAWTVVSLDNQRSAQFEHTVLITSRGAQILTKLPHEA
Ground Truth	Removes the N-terminal methionine from nascent proteins. The N-terminal methionine is often cleaved when the second residue in the primary sequence is small and uncharged (Met-Ala-, Cys, Gly, Pro, Ser, Thr, or Val). Requires deformylation of the N(alpha)-formylated initiator methionine before it can be hydrolyzed (By similarity). May play a role in colon tumorigenesis.
OPI-tuned model	Removes the N-terminal methionine from nascent proteins. The N-terminal methionine is often cleaved when the second residue in the primary sequence is small and uncharged (Met-Ala-, Cys, Gly, Pro, Ser, Thr, or Val). Requires deformylation of the N(alpha)-formylated initiator methionine before it can be hydrolyzed.
 GPT-4o	<p>Based on the provided protein sequence, the following functional description can be inferred:</p> <p>**Functional Description:**</p> <p>This protein appears to have several key functional domains that suggest a role in various cellular processes. The presence of multiple arginine-rich regions (RR) and lysine residues (K) indicates potential roles in nucleic acid binding, possibly involved in transcription regulation or RNA processing. The sequence also features glycine-rich regions (G) which are often associated with flexible linkers or regions that may facilitate interactions with other proteins or nucleic acids.</p> <p>The presence of serine (S) and threonine (T) residues suggests potential sites for phosphorylation, implying regulation by kinases and a role in signal transduction pathways. The conserved regions with histidine (H) and cysteine (C) residues could indicate metal ion binding capabilities, possibly zinc fingers or other metal-binding motifs, which are crucial for DNA-binding and structural stability.</p> <p>The sequence also contains stretches of hydrophobic amino acids (V, L, I), likely forming transmembrane domains or contributing to protein-protein interactions within the cellular membrane or other hydrophobic environments. This could suggest a role in membrane-associated processes or signal transduction across membranes.</p> <p>Given the sequence, the protein might be involved in diverse functions such as gene regulation, signal transduction, and possibly interactions with membranes or other proteins, highlighting its multifaceted role in cellular biology.</p>
 Llama-3.1-8B-Instruct	<p>Here is the functional description in free-text form based on the protein sequence: This protein is a serine protease that plays a crucial role in the blood coagulation cascade. It is a key enzyme in the formation of blood clots and is involved in the conversion of fibrinogen to fibrin. The protein has a serine active site and is activated by thrombin, which is itself an enzyme that is produced during the coagulation process. The protein is composed of multiple domains, including a serine protease domain, a thrombin-binding domain, and a fibrinogen-binding domain. The serine protease domain is responsible for the proteolytic activity of the protein, while the thrombin-binding domain allows the protein to interact with thrombin and be activated. The fibrinogen-binding domain allows the protein to bind to fibrinogen and convert it to fibrin. The protein is also involved in the regulation of platelet aggregation and the formation of blood clots. It is able to bind to platelets and activate them, leading to the formation of platelet aggregates and the release of pro-coagulant factors. Additionally, the protein is able to bind to von Willebrand factor, a protein that plays a key role in platelet adhesion and aggregation. Overall, this protein plays a critical role in the blood coagulation cascade and is essential for the formation of blood clots. Its dysregulation can lead to bleeding disorders or thrombotic events, such as deep vein thrombosis or pulmonary embolism.</p>
 Claude 3.5 Sonnet	<p>Based on the provided protein sequence, I can generate a functional description in free-text form. Please note that this description is based on analysis of the sequence and potential structural motifs, without access to experimental data or specific literature about this exact protein. Here's a possible functional description:</p> <p>This protein appears to be a multi-domain enzyme, likely involved in metabolic processes. The sequence suggests several key features:</p> <ol style="list-style-type: none">N-terminal region: The protein starts with a potential signal peptide or targeting sequence (MAAPSGVHLLVRR...), which might direct it to a specific cellular compartment.Regulatory domain: The presence of serine-rich regions (e.g., SSSQQ) suggests possible phosphorylation sites, indicating that the protein's activity might be regulated by kinases.Catalytic core: The central portion of the protein contains sequences typical of enzymes, particularly oxidoreductases or dehydrogenases. The presence of motifs like GIVPDWGD and VLLLAG suggests potential NAD(P)H binding sites, which are common in these enzyme classes.Substrate binding: Regions rich in hydrophobic amino acids (e.g., ILCHGIPDSRPL) may form binding pockets for organic substrates.Metal binding: The sequence HVLLLAGK could potentially be part of a metal-binding site, suggesting that the protein might require a metal cofactor for its activity.C-terminal domain: The C-terminus (ITSRGAQILTKLPHEA) might be involved in protein-protein interactions or provide additional regulatory functions. <p>Given these features, this protein is likely an oxidoreductase or dehydrogenase involved in cellular metabolism. It may catalyze the oxidation or reduction of specific substrates, possibly playing a role in pathways such as amino acid metabolism, fatty acid oxidation, or carbohydrate processing. The protein's activity is probably regulated through phosphorylation and potentially by cellular localization. It may interact with other proteins as part of a larger metabolic complex. Further experimental characterization would be necessary to confirm its exact function, substrates, and physiological role within the cell.</p>