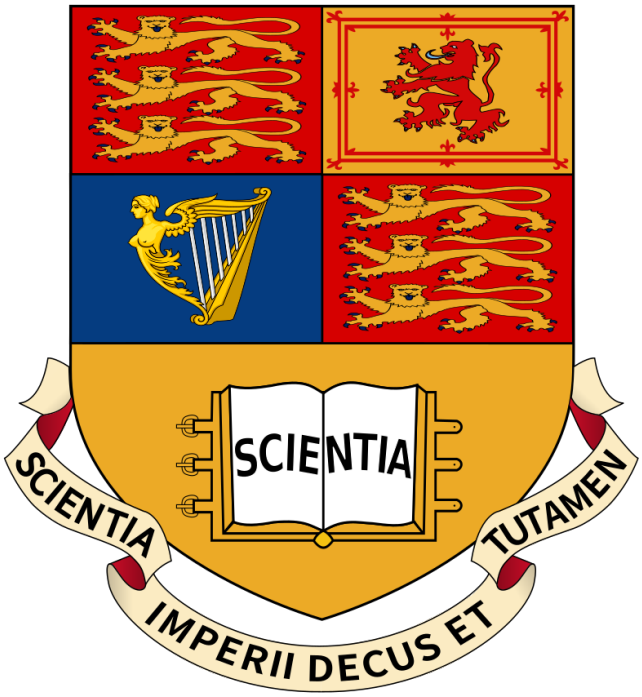
Protein Annotation Coursework

Imperial College London

MRes Bioinformatics and Theoretical Systems Biology

LIFE70004



Student: **01913327**

Words: 800

Question 1.

For the 462-residue protein sequence (AAF94790.1), DeepTMHMM [1] predicted two membrane-spanning regions; posterior probabilities were calculated for each part of the protein sequence to deduce their inside/outside orientation. Figure 1 summarises the residue locations of transmembrane domains and their surrounding intracellular and extracellular regions.

Topology Summary (AA regions)

Domain	Start (AA)	End (AA)
Intracellular	1	21
Transmembrane	22	41
Extracellular	42	171
Transmembrane	172	189
Intracellular	190	462

Figure 1: *DeepTMHMM* results. Amino acid (AA) positions of the topologies calculated by the *DeepTMHMM* algorithm, summarised in a table.

Question 2.

The *BLASTp* algorithm rapidly finds potential homologs for the query, whereas *PSI-BLAST* is an iterative *BLASTp* procedure, building a PSSM to find more distant relationships with potentially greater functional associations [2].

Whilst the top three *BLASTp* hits were from two different sensor kinases (encoded by PhoQ and CarS), all *PSI-BLAST* hits came from a MprB-encoded signal transduction histidine-protein (Figure 2). The combination of far lower average E-values (-117.51 vs -44.74) and a higher average query cover (94.33 vs 68.33) of *PSI-BLAST* hits compared to *BLASTp* respectively reflects its methodology to find more distantly related proteins to the query which share conserved functional domains and 3D structures (Figure 2). *PSI-BLAST* iterations decreased sequence identity to capture evolutionary divergence, reflected by the decrease in average sequence identity of hits compared to *BLASTp*: 21.68% compared to 33.28%.

Upon inspection, HAMP and Histidine Kinase domains were present throughout sequence hits.

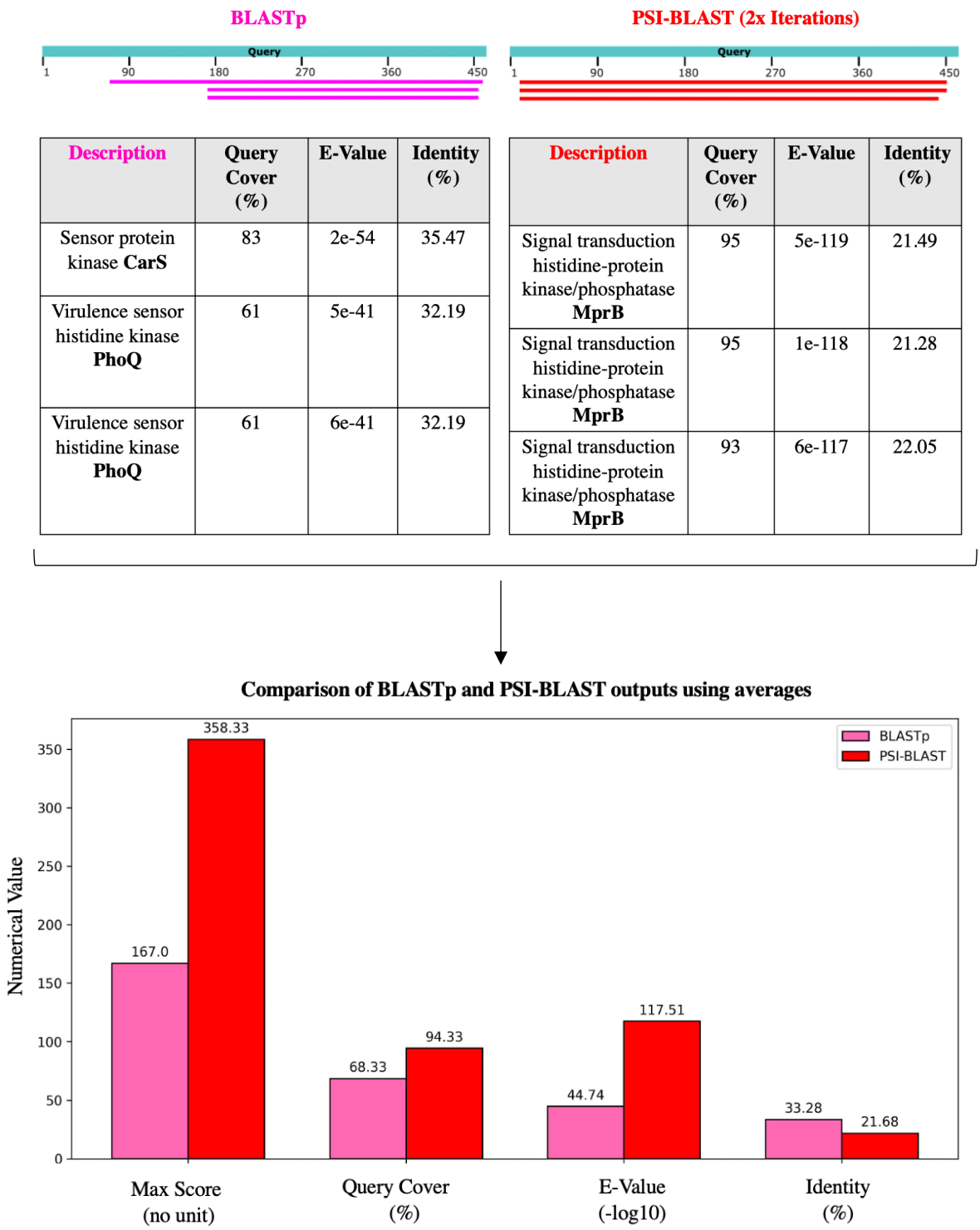


Figure 2: *BLASTp/PSI-BLAST* results. The outputs of *PSI-BLAST* and *BLASTp* in summarised in table format. The results are presented in a bar plot showing averages of key numerical metrics for each hit: a -log10 transformation was applied to present E-values on the same scale.

Question 3.

InterPro collates various sources to provide annotations for a query sequence [3].

Four functional domains were predicted by *InterPro* (Figure 3). A C-terminal signal transduction and HSP90-like ATPase domain were located within a large Histidine Kinase (HK) domain with overlapping residues; the presence of a phosphorylation site and ATP-binding kinase region respectively matches the conserved domain architecture of HK [4]. The HK domain (residues 251-453) was downstream of a HAMP domain (residues 192 – 243).

Regional residue annotations were provided by *PHOBIUS* [5]: cytoplasmic (residues 1-20 and 193-463), transmembrane (residues 21-41 and 172-193) and non-cytoplasmic (residues 251-453). *InterPro* domains were all in the cytoplasm, beginning with the HAMP domain at residue 192. Combining *InterPro* and *PHOBIUS* annotations facilitated a structural visualisation of the protein (Figure 3, bottom). Given slight residue shifts, *PHOBIUS* predictions align closely with DeepTMHMM results (Figure 1), and *InterPro* domain architecture matches BLAST results (HAMP and HK domains).

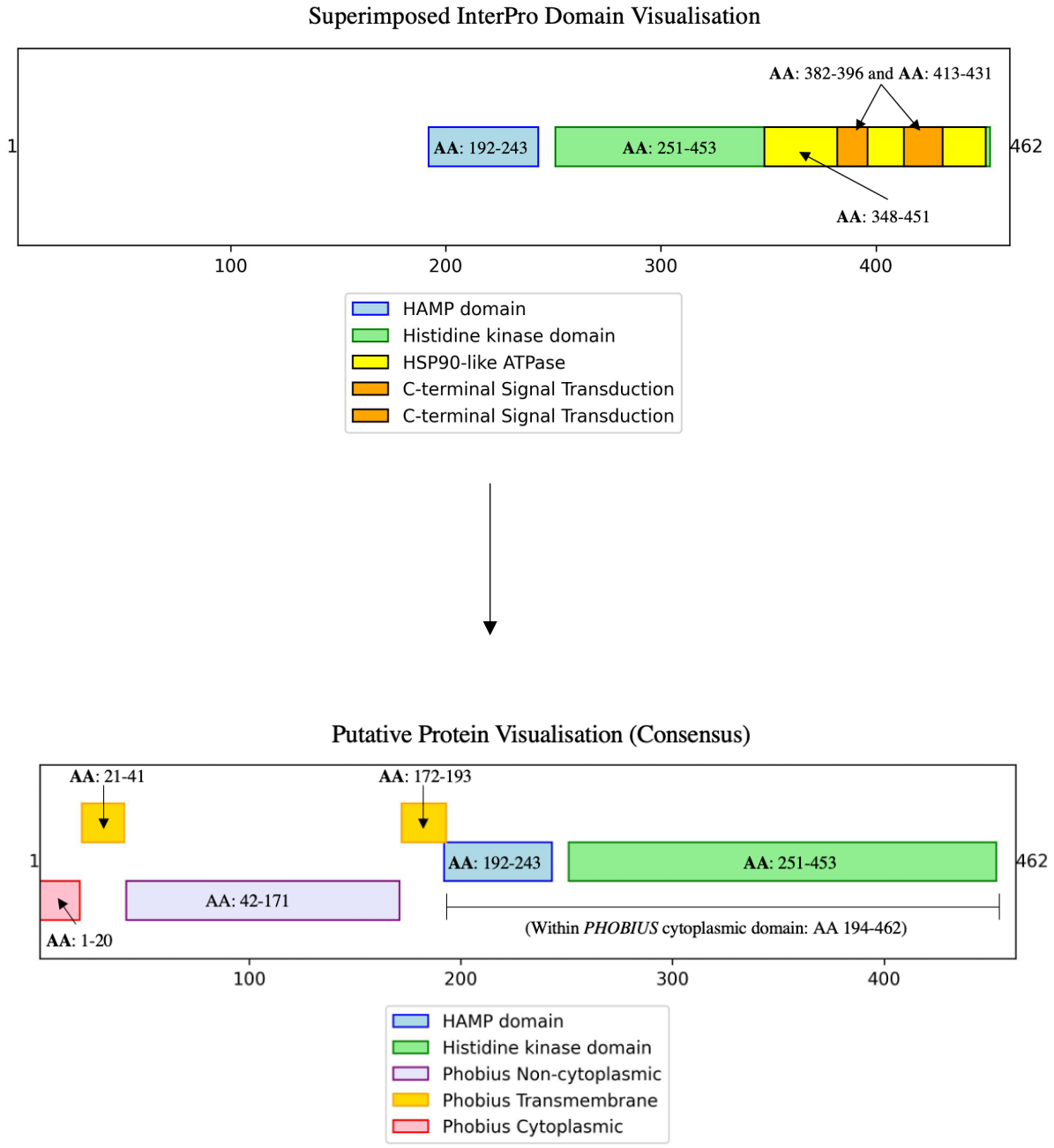


Figure 3: *InterPro* Results. Python graphic visualising the 4 key *InterPro* domains superimposed. This is adapted to another python graphic (via an arrow) aligning the broad *InterPro* domain predictions with *PHOBIUS* topology predictions.

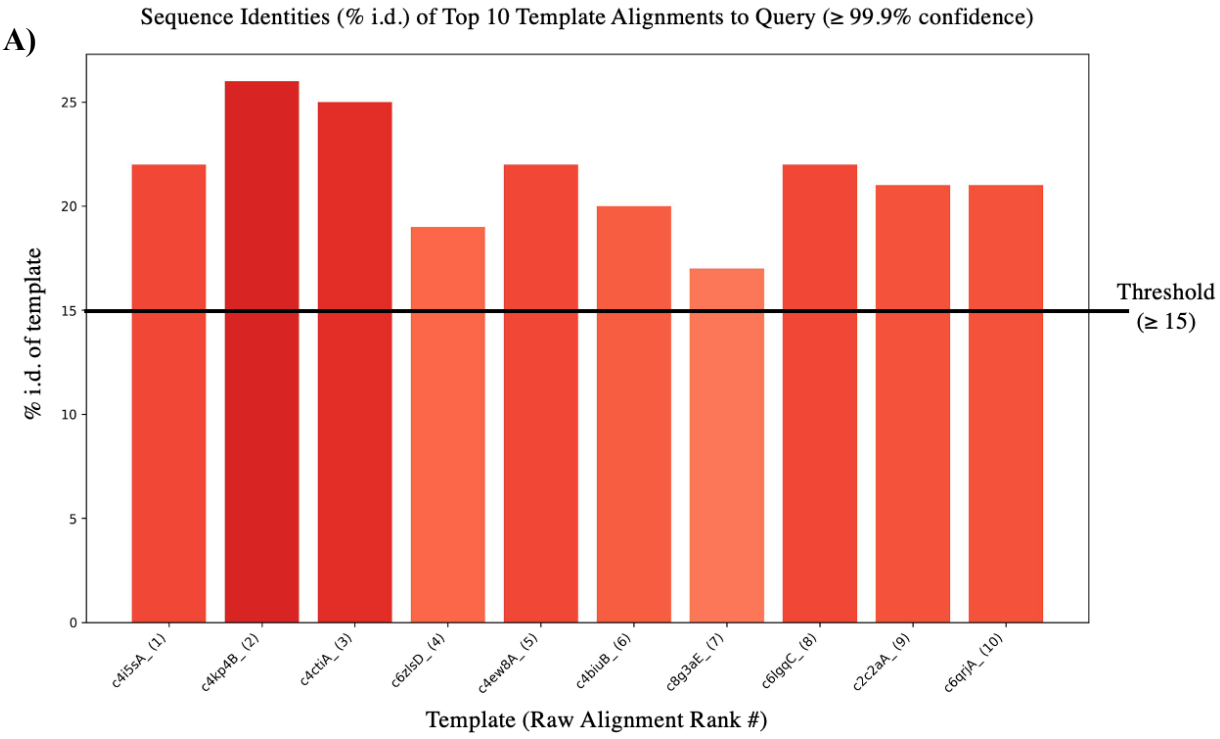
Question 4.

Phyre2 aligned the query to PDB 'templates', predicted unique 3D models and ranked them by raw alignment score. The ten top models had high confidence percentages (>99.9%) and thus a hydrophobic core ~2-4Å RMSD from native. Identity scores between sequences and templates were not high accuracy (≥30%) but adequate to accept 3D model predictions (>15%) (Figure 4A) [6].

Secondary structure and disordered regions can infer tertiary structure for this query (AAF94790.1) with 78-80% accuracy. *Phyre2* confidently predicted two transmembrane regions (residues 21-45 and 172-191) conserved across models, within long alpha-helices extending into the cytoplasm. *Phyre2* also identified intracellular intercalating beta-strands and alpha-helices in all 3D models, typical of a HK catalytic domain [4]. A similar structure was found extracellularly (residues 46-171) but with low confidence, so models were predicted after residue 170 (Figure 4B).

All templates were from kinase/sensor molecules with 'signalling', 'transferase' and 'membrane' annotations - associated with HK sensory autophosphorylation-dependent signal transduction pathways [7].

A)



B)

Template (Rank Order)	Alignment (Coverage)	Protein (PDB)	Molecule (PDB)
c4i5sA_ (1)	195-450	Transferase	Histidine kinase covs
c4kp4B_ (2)	243-452	Transferase/signalling	Osmolarity sensor protein envz
c4ctiA_ (3)	194-449	Signalling	Osmolarity sensor protein envz
c6zlsD_ (4)	190-454	Signalling	Histidine kinase gacs
c4ew8A_ (5)	230-452	Transferase	Sensor protein divl
c4biuB_ (6)	196-451	Transferase	Sensor protein cpxa
c8g3aE_ (7)	170-454	Membrane	Sensor protein bces
c6lgqC_ (8)	239-458	Signalling	Histidine kinase kdpd
c2c2aA_ (9)	232-451	Transferase	Sensor histidine kinase
c6qrjA_ (10)	247-462	Signalling	Hybrid kinase

Figure 4: *Phyre2* Results. A) A bar graph presenting the sequence identities of the top 10 templates. **B)** A table summarising the key information provided by each template PDB. Alignment coverage of the query is also included.

Question 5.

With high average pLDDT across the *AlphaFold* model (Figure 5, left), most of the prediction is high quality, especially the transmembrane portion of the protein and conserved functional domains (pLDDT > 90) [8]. Regions like residues 1-20 have bad pLDDT; *DeepTMHMM* predicted this loop region to be in the cytoplasm, but its 3D orientation appears to be parallel to the transmembrane alpha-helices.

AlphaFold's model spanned the entire protein (residues 1-463) compared to *Phyre2's* top models from residues 190-449, thus *Phyre2* ignored the transmembrane portion of the long central alpha-helices and the extracellular domain (residues 42-171) identified by *AlphaFold*. However, *Phyre2* captured the parallel portion of the long alpha-helices close to the transmembrane region (a characteristic HAMP 'helix-loop-helix' fold) and models agreed on the highly conserved intracellular Histidine Kinase catalytic domain: several alpha-helices packed over one face of large (typically antiparallel) beta-sheet (Figure 5, right) [4].

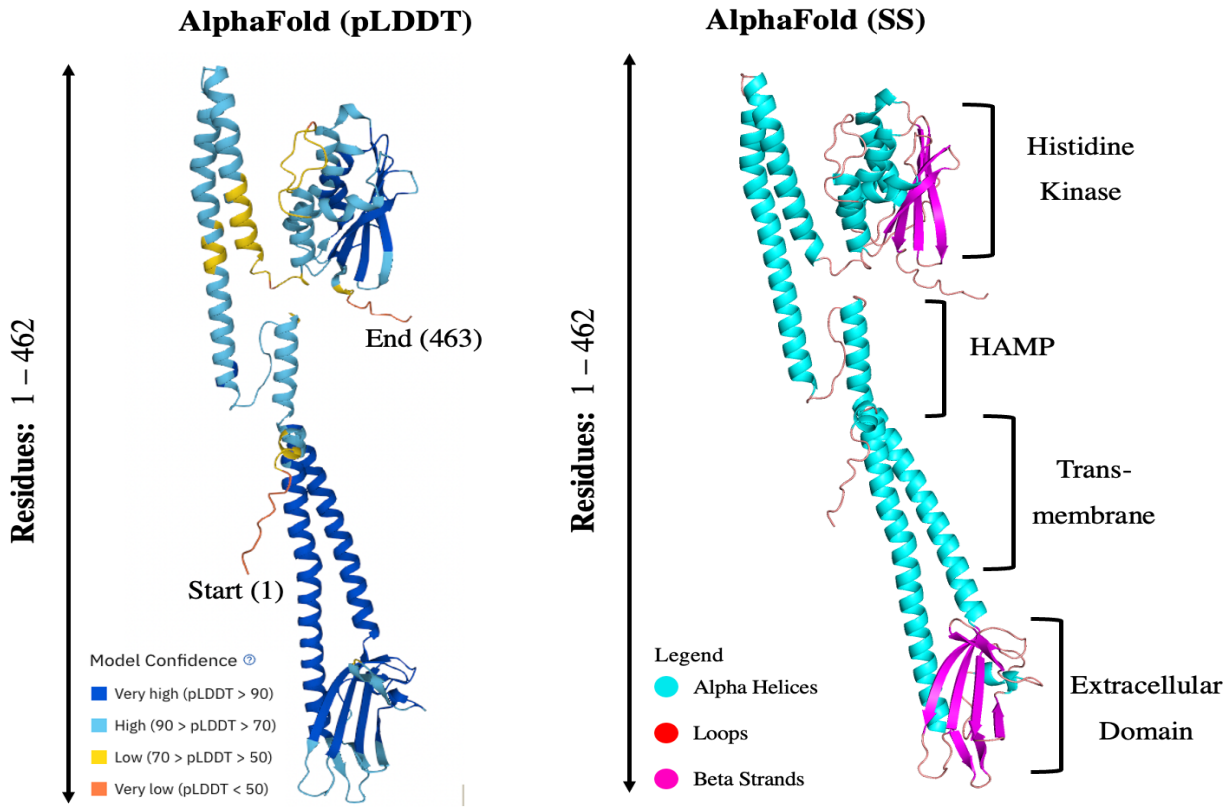


Figure 5: *AlphaFold* Results *AlphaFold* model, coloured by pLDDT on the left and coloured by secondary structure (SS) on the right. Key domains are annotated.

Question 6.

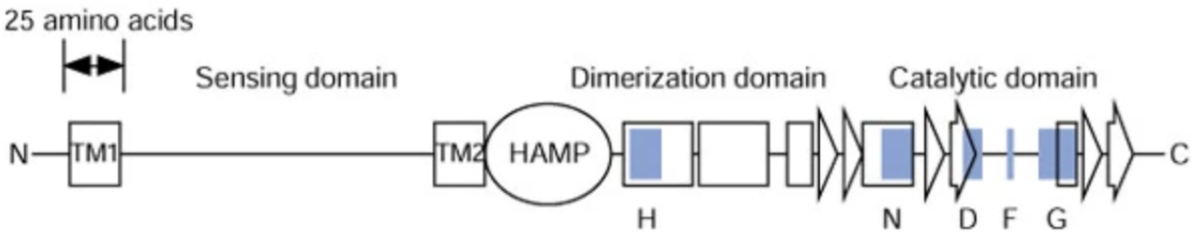
The query protein sequence (AAF94790.1) is a gram-negative bacterium sensor HK protein required for two-component signal transduction systems [9].

Figure 6A shows conserved prokaryotic HK motifs represented by the *E. coli* EnvZ family (identified by *Phyre2* as a close homolog): two transmembrane regions, one extracellular region (sensor) and two intracellular regions [4][6]. The intracellular components include a HAMP linker domain and HK core encompassing a dimerization and catalytic domain.

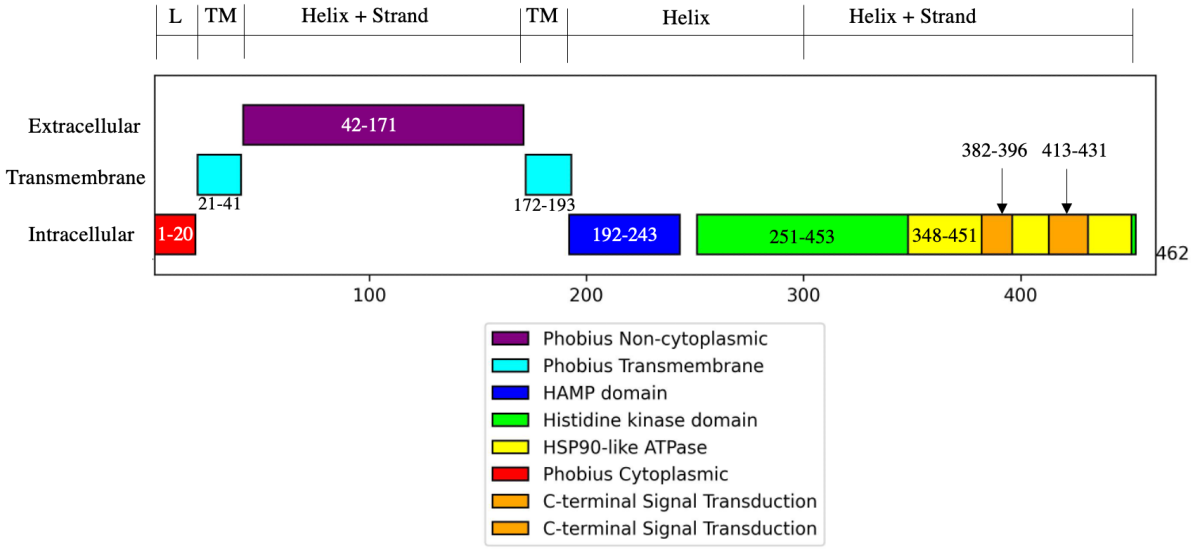
AAF94790.1 has conserved domain architecture and folds with slight residue shifts (Figure 6B). The HAMP domain is annotated whereas the dimerisation domain was missed by software but resembles the initial alpha-helical portion of the HK domain (residues 251-347). The conserved catalytic domain corresponds to the HSP90-like ATPase and signal transduction annotations.

AlphaFold models the key functional domains, folds and their proximities in 3D with high confidence (Figure 6B) [8]. In its active state, the structure is homodimerized via the alpha-helical portion of the HK domain [6].

A) **Conserved motifs in a representative Histidine Protein Kinase (*E. coli* EnvZ) and their relative positions**



B) **Overall Annotation of the Query (AAF94790.1)**



Query (AAF94790.1) 3D Structure

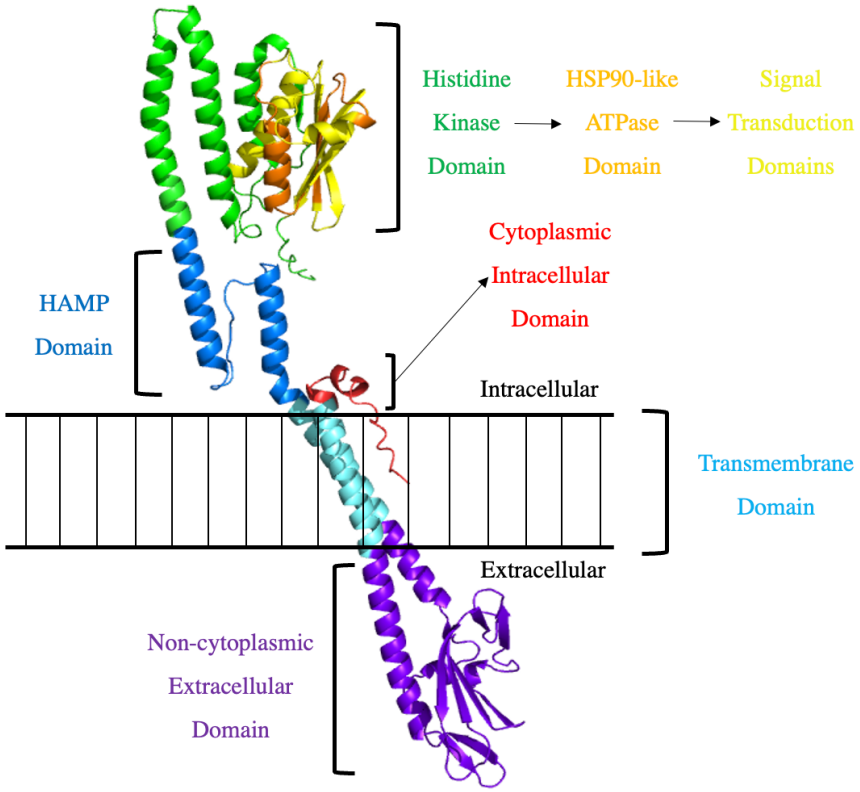


Figure 6: Final Annotation of protein query AAF94790.1/Q9KRK1 A) The relative positions of the conserved motifs in a representative, gram negative HPK (*E. coli* EnvZ), taken from Wolanin PM et al [4]. The 'dimerization domain' and 'catalytic domain' are within a large Histidine Kinase Domain. Rectangles represent alpha-helices and arrows represent beta-sheet. TM1 and TM2 are transmembrane helices. **B)** Python graphic summarising the key motifs throughout the sequence using *PHOBIUS*, *InterPro* and *BLAST*. Regions/domains are colour coded (legend) and their location is annotated on the y-axis. Types of secondary structure are included above the graphic. The corresponding *AlphaFold* predicted structure colour-coded to match the python graphic. Key domains are annotated on the protein.

References

[1] Jeppe Hallgren, Konstantinos D. Tsirigos, Mads D. Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh and Ole Winther (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, Volume 25, Issue 17, 1 September 1997, Pages 3389–3402

[3] Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009 Jan;37(Database issue)

[4] Wolanin PM, Thomason PA, Stock JB. Histidine protein kinases: key signal transducers outside the animal kingdom. Genome Biol. 2002 Sep 25;3(10):REVIEWS3013. doi: 10.1186/gb-2002-3-10-reviews3013. Epub 2002 Sep 25. PMID: 12372152; PMCID: PMC244915.

[5] Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W429-32. doi: 10.1093/nar/gkm256. Epub 2007 May 5. PMID: 17483518; PMCID: PMC1933244.

[6] Kelley, Lawrence A et al. “The Phyre2 web portal for protein modeling, prediction and analysis.” *Nature protocols* vol. 10,6 (2015): 845-58. doi:10.1038/nprot.2015.053

[7] Hirakawa H, Kurushima J, Hashimoto Y, Tomita H. Progress Overview of Bacterial Two-Component Regulatory Systems as Potential Targets for Antimicrobial Chemotherapy. Antibiotics (Basel). 2020 Sep 23;9(10):635.

[8] Jumper, J., Evans, R., Pritzel, A. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[9] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D115-9. doi: 10.1093/nar/gkh131. PMID: 14681372; PMCID: PMC308865.