In Silico Analysis of the Human EGFR Gene: From Genome Annotation to Protein Structure Prediction

BIN 506 : Protein And DNA Sequence Analysis

Term Paper

Taha Ahmad - 2546125

Submission Date: 20.06.2025

Instructor: Prof Dr YESIM AYDIN SON

Table of Contents

Abstract

The epidermal growth factor receptor (EGFR), encoded by the gene with UniProt ID P00533, is a transmembrane tyrosine kinase critical for cellular proliferation and differentiation.. This study presents a comprehensive in silico analysis of EGFR, integrating genomic, proteomic, phylogenetic, and structural bioinformatics approaches. We examined its gene structure and transcript variants, evaluated annotation accuracy using experimental and automated sources, and conducted a comparative analysis of orthologs across six species. Multiple sequence alignments and phylogenetic reconstructions were employed to assess evolutionary conservation, while MEME Suite was utilized to discover conserved sequence motifs within the tyrosine kinase family. Protein domain architecture and functional annotations were mapped onto sequence alignments. Finally, structural comparisons between experimentally determined (PDB: 3POZ) and AlphaFold-predicted EGFR models revealed key secondary features, including a regulatory helix kink at Glycine 749, relevant to kinase activation. This integrative analysis underscores the significance of EGFR as a model system for bioinformatics-driven functional annotation and evolutionary insight.

General information:

The *epidermal growth factor receptor* (**EGFR**) gene in *Homo sapiens* encodes a transmembrane receptor tyrosine kinase which plays a key role in regulating cell proliferation, differentiation, and survival. Upon binding to epidermal growth factor (EGF), EGFR undergoes dimerization and autophosphorylation, initiating downstream signaling cascades such as the MAPK and PI3K/AKT pathways (Yarden & Sliwkowski, 2001). Abnormal EGFR signaling, often due to gene amplification or mutation, is responsible for several cancers, particularly non-small cell lung carcinoma (Zhang et al., 2007).

Structurally, EGFR is a 1210 amino acid protein, and its kinase domain has been resolved by X-ray crystallography (e.g, PDB ID: 3POZ) at high resolution, enabling targeted drug design. The protein is well-curated in UniProt (P00533), where it is annotated with a 5/5 evidence score based on experimental data, confirming its status as a biologically and clinically significant gene.

Genomic Region, Gene Structure, Transcripts, and Protein Function:

According to the GRCh38.p14 genome assembly, the **EGFR** gene is located on **chromosome 7**, specifically at cytogenetic band **7p11.2**, spanning the genomic coordinates **55,019,017–55,211,628** on the forward strand. The gene maps to reference contigs **NC_000007.14** and **AC073324.6**, and consists of **32 exons**, indicating a complex genomic structure that supports extensive transcript diversity.

EGFR produces **13 known transcript variants**, with the **canonical transcript** being **ENST00000275493 (EGFR-201)**. This transcript is **9905 base pairs** long and encodes a **1210-amino-acid** protein.

Functionally, EGFR is expressed broadly across human tissues, with elevated expression observed in the **placenta**, **skin**, and **epithelial tissues**. Upon ligand binding, EGFR activates intracellular pathways critical for **cell growth, wound healing, and development**. Its dysregulation is strongly linked to oncogenic transformation, particularly in epithelial cancers (Arteaga & Engelman, 2014). Moreover, EGFR undergoes **alternative splicing**, which may modulate receptor signaling dynamics or contribute to tissue-specific expression patterns. The presence of over **85,000 genetic variants** in the EGFR locus (NCBI, 2024) underscores its clinical and evolutionary significance.

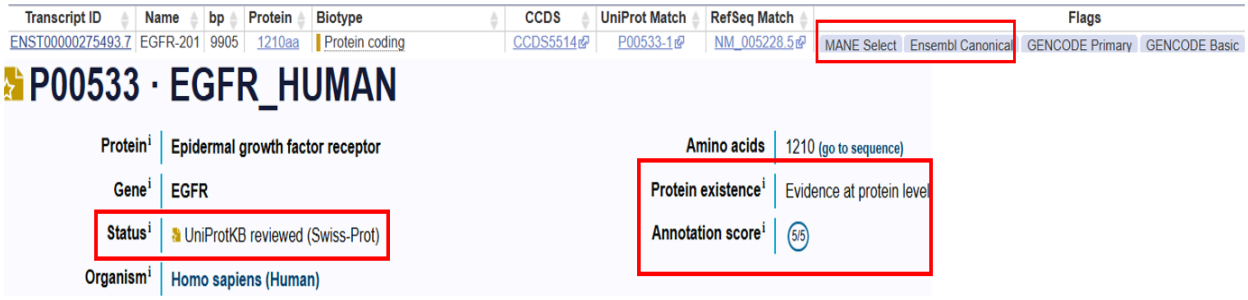
The gene's structure and transcript diversity reflect a high level of regulatory complexity, and its protein product's modular architecture enables precise control of diverse signaling outcomes in response to environmental stimuli.

Supporting Evidence for EGFR Gene Annotation:

The **EGFR gene** is one of the most thoroughly annotated and well-characterized genes in the human genome, supported by robust evidence from both experimental data and automated pipelines. On **UniProt** (P00533), EGFR has an **annotation score of 5/5**, signifying comprehensive validation at the **protein level**, including structural resolution via **X-ray crystallography**, immunochemical expression data, and functional assays (UniProt, 2024). Moreover, the protein is classified as **reviewed** under the **Swiss-Prot** database, indicating manual expert curation based on peer-reviewed literature, rather than solely algorithmic predictions.

Automated annotations from major databases such as **GENCODE** and **RefSeq** converge on the same canonical transcript **ENST00000275493** (EGFR-201 in Ensembl and NM_005228.5 in RefSeq). This transcript is designated as the **MANE Select transcript**, reflecting complete agreement between reference annotation platforms regarding exon-intron structure, coding region, and transcript boundaries. Additionally, EGFR is included in the **Consensus CDS (CCDS)** project (CCDS1146.1), further confirming the high-confidence status of its coding region across annotation systems.

Transcript support level (TSL) data further validate the annotation: the canonical EGFR transcript is labeled as **TSL1**, indicating that its entire exon structure is fully supported by **mRNA or cDNA evidence**. These experimental data are essential to distinguish real biological transcripts from computational artifacts, especially in genes with extensive **alternative splicing**, like EGFR. The alignment of automated and experimental pipelines along with protein-level validation strongly supports the reliability of EGFR gene annotation.



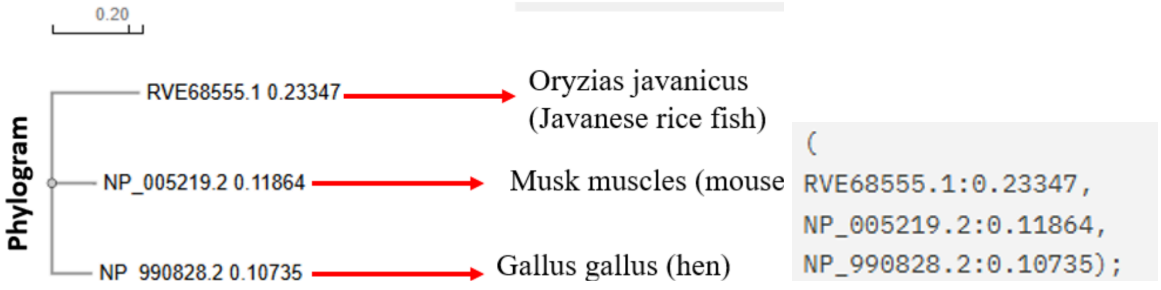
4. (20 points) Find out if any of the following species has proteins homologous to "your protein/gene of interest." M. musculus, G.gallus, O.lapites, A.mellifera, C.elegans, S.cerevisiae.

a. Show the multiple sequence alignment and build a phylogenic tree using only three of the homologous protein sequences.

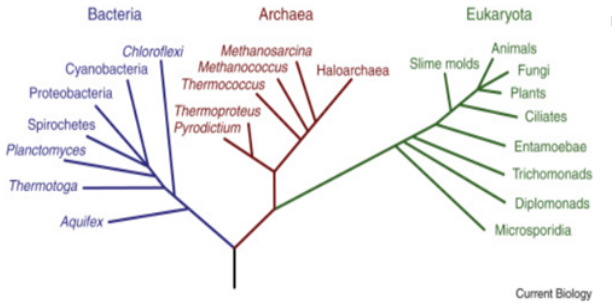
b. Do you see any clusters on the phylogenetic tree which are different from the tree of life?

c. Next, build the phylogenetic tree with all the sequences, describe any differences you have observed and discuss possible effects of adding more sequences to the phylogenetic analysis

```
CLUSTAL O(1.2.4) multiple sequence alignment


RVE68555.1      ------------MAARFLKW--ICVLTSALSCVPAERKVCQGLSNRLNLLGSKDDHYLNM   46
NP_005219.2     ------MRPSGTAGAALLA---LLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSL   51
NP_990828.2     MGVRSPLSASGPRGAAVLVLLLLLLGRVALCSAVEEKKVCQGTNNKLTQLGHVEDHFTSL   60
                            .* .*    :       . .  *:***** .*:*. **  :**: .:

RVE68555.1      VKTYSNCTVVLQNLEITHMEDHHDLSFL------------------RIPLENLRIIRGHS   88
NP_005219.2     QRMFNNCEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNM   111
NP_990828.2     QRMYNNCEVVLSNLEITYVEHNRDLTFLKTIQEVAGYVLIALNMVDVIPLENLQIIRGNV   120
                 : :.** *** *****::: : **:**                  ******:****:

RVE68555.1      LYEGEFALSVLVNADKATGRGSSELLLTNLTEILEGGVKFGTN-QLCNLETIQWFDIVNP   147
NP_005219.2     YYENSYALAVLSNYDANK-TGLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSS   170
NP_990828.2     LYDNSFALAVLSNYHMNKTQGLRELPMKRLSEILNGGVKISNNPKLCNMDTVLWNDIIDT   180
                 *:..:**:** * .  .  *  ** : .* ***.*.*::..*  ***:::: * **:.

RVE68555.1      DKKPKMELPMASNNPRCQKCHSSCFNGSCWAPGPQYCQTFTKLNCAQQCSHRCKGPSPSD   207
NP_005219.2     DFLSNMSMDFQNHLGSCQKCDPSCPNGSCWGAGEENCQKLTKIICAQQCSGRCRGKSPSD   230
NP_990828.2     SRKPLTVLDFASNLSSCPKCHPNCTEDHCWGAGEQNCQTLTKVICAQQCSGRCRGKVPSD   240
                .      : : .:   * **. .* :. **. * : **.:**: ****** **:*  ***

RVE68555.1      CCNEHCAAGCTGPRPTDCLACRDFQDDGVCKDSCPGLMRYDPNLHLLVPNPNGKYSFGAT   267
NP_005219.2     CCHNQCAAGCTGPRESDCLVCRKFRDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGAT   290
NP_990828.2     CCHNQCAAGCTGPRESDCLACRKFRDDATCKDTCPPLVLYNPTTYQMDVNPEGKYSFGAT   300
                **:::********* :***.**.*:*:..***:** *: *:*. : :  **:********

RVE68555.1      CVKNCPHNYVVTDHGACVRTCSGNTYEVEEGGIRKCAKCNGPCPKVCDGLGTGNLTHALS   327
NP_005219.2     CVKKCPRNYVVTDHGSCVRACGADSYEMEEDGVRKCKKCEGPCRKVCNGIGIGEFKDSLS   350
NP_990828.2     CVRECPHNYVVTDHGSCVRSCNTDTYEVEENGVRKCKKCDGLCSKVCNGIGIGELKGILS   360
                **::**:********:***:*. ::**:**.*:*** **:* * ***:*:* *::.  **

RVE68555.1      INATNIDSFKNCTKINGNIAFIHTSIHGDKFTKTPKLDPAKLDVFKTVKEITGYLWIQTW   387
NP_005219.2     INATNIKHFKNCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLIQAW   410
NP_990828.2     INATNIDSFKNCTKINGDVSILPVAFLGDAFTKTLPLDPKKLDVFRTVKEISGFLLIQAW   420
                ******. *****.*.*:: :: .:: ** **:*  *** :**:::*****:*:* **:*
```




```
(
RVE68555.1:0.23347,
NP_005219.2:0.11864,
NP_990828.2:0.10735);
```



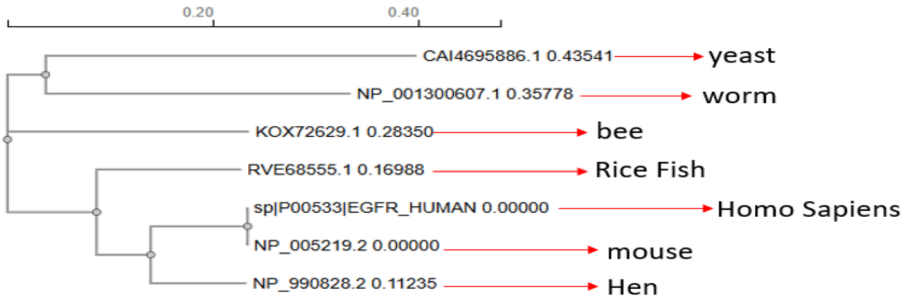
To assess the evolutionary conservation of the **epidermal growth factor receptor (EGFR)** across vertebrate species, homologous protein sequences were retrieved from *Mus musculus* (**NP_005219.2**), *Gallus gallus* (**NP_990828.2**), and *Oryzias javanicus* (**RVE68555.1**). These sequences were aligned using **Clustal Omega**, a widely used multiple sequence alignment tool known for its scalability, high accuracy with divergent sequences, and integration of progressive alignment and Hidden Markov Model refinement steps (Sievers & Higgins, 2014). Clustal Omega's default substitution matrix and guide tree settings were employed to generate both the **MSA** and the **phylogenetic tree**.

The resulting alignment showed strong conservation between the mouse and chicken EGFR homologs, particularly in conserved kinase domains. In contrast, the fish homolog from *Oryzias javanicus* displayed higher sequence divergence. The **phylogenetic tree** supported this observation: *Mus musculus* and *Gallus gallus* clustered closely, with branch lengths of **0.11864** and **0.10735**, respectively, while *Oryzias javanicus* diverged earlier, with a longer branch length of **0.23347**.

These findings are consistent with established vertebrate phylogeny, where mammals and birds share a more recent common ancestor than either does with teleost fish. The branch lengths correspond well with evolutionary distances inferred from whole-genome comparisons, confirming the validity of the alignment approach.

c. Next, build the phylogenetic tree with all the sequences, describe any differences you have observed and discuss possible effects of adding more sequences to the phylogenetic analysis.

```
CLUSTAL O(1.2.4) multiple sequence alignment


CAI4695886.1            ------------------------------------------------------------  0
NP_001300607.1          MKFPFLFLFLIILIDTTIDAASPTSTQGSVTVRR--------------RQRKIVVREWN  45
KOX72629.1              -----------------MEF---------NLSRFFTCFCNNSWKFKRNNLGHSQYVRVQP  34
RVE68555.1              ----------------------------MAARFL------KW------ICVLTSALSCV  19
sp|P00533|EGFR_HUMAN    ----------------------MRPSGTAGAALL------A-------LLAALCPASRA  24
NP_005219.2             ----------------------MRPSGTAGAALL------A-------LLAALCPASRA  24
NP_990828.2             -----------------MGVRSPLSASGPRGAAVL------VLLL----LLLGRVALCSA  33


CAI4695886.1            ------------------------------------------------------------  0
NP_001300607.1          LTDPSVCSGTTNGISRYGTG-NILEDLETMYRGCRRVYGNLEITWIEANEIKKWRESTNS  104
KOX72629.1              LTMKSICIGTNGRLSVPSNKQHHYRNLRDRYTNCTYVDGNLEITWLQNET----------  84
RVE68555.1              PAERKVCQGLSNRLNLLGSKDDHYLNMVKTYSNCTVVLQNLEITHMED-H----------  68
sp|P00533|EGFR_HUMAN    LEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQR-N----------  73
NP_005219.2             LEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQR-N----------  73
NP_990828.2             VEEKKVCQGTNNKLTQLGHVEDHFTSLQRMYNNCEVVLSNLEITYVEH-N----------  82


CAI4695886.1            ------------------------------------------------------------  0
NP_001300607.1          TVDPKNEDSPLKSINFFDNLEEIRGSLIIYRANIQKISFPRLRVIYGDEVFHDNALY---  161
KOX72629.1              -----------FDLSFLQYIREVTGYVLISHVDVKKVVLPRLQIIRGRTLFKLTIHDIEF  133
RVE68555.1              -----------HDLSFL------------------RIPLENLRIIRGHSLYEGEFALSVL  99
sp|P00533|EGFR_HUMAN    -----------YDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVL  122
NP_005219.2             -----------YDLSFLKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVL  122
NP_990828.2             -----------RDLTFLKTIQEVAGYVLIALNMVDVIPLENLQIIRGNVLYDNSFALAVL  131


CAI4695886.1            ------------------------------------------------------------  0
NP_001300607.1          ----IHKNDKVHEVVMRELRVIRNGSVTIQDNPKMCYIGDKIDWKELLYDPDVQKVETTN  217
KOX72629.1              AL--FVTMCQMQNLEMPALRDILNGSVGIYNNYNLCHI-QKINWDEIITGPNATYSYVY-  189
RVE68555.1              VNADKATGRGSSELLLTNLTEILEGGVKFGTN-QLCNL-ETIQWFDIVNPDK-KPKMEL-  155
sp|P00533|EGFR_HUMAN    SNYDANK-TGLKELPMRNLQEILHGAVRFSNNPALCNV-ESIQWRDIVSSDF-LSNMSM-  178
NP_005219.2             SNYDANK-TGLKELPMRNLQEILHGAVRFSNNPALCNV-ESIQWRDIVSSDF-LSNMSM-  178
NP_990828.2             SNYHMNKTQGLRELPMKRLSEILNGGVKISNNPKLCNM-DTVLWNDIIDTSR-KPLTVL-  188
```

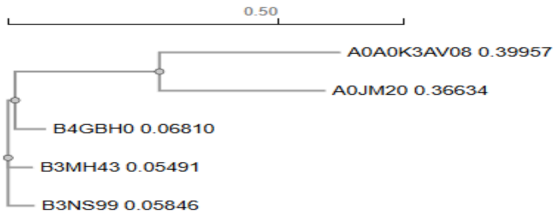



```
(
(
CAI4695886.1:0.43541,
NP_001300607.1:0.35778)
:0.04481,
KOX72629.1:0.28350,
(
RVE68555.1:0.16988,
(
(
sp|P00533|EGFR_HUMAN:0.00000,
NP_005219.2:0.00000)
:0.11364,
NP_990828.2:0.11235)
:0.06359)
:0.10455);
```

To evaluate the evolutionary divergence of the EGFR protein across a broader range of taxa, a phylogenetic tree was constructed using homologous sequences from six species: *Homo sapiens* (sp|P00533), *Mus musculus* (NP_005219.2), *Gallus gallus* (NP_990828.2), *Oryzias javanicus* (RVE68555.1), *Caenorhabditis elegans* (NP_001300607.1), and *Saccharomyces cerevisiae* (CAI4695886.1). These sequences were aligned using Clustal Omega with default parameters, and the resulting tree topology accurately reflected major evolutionary divisions. The human and mouse sequences showed a branch length of 0.00000, indicating 100% identity in the aligned region likely due to a highly conserved domain such as the tyrosine kinase region while the chicken homolog diverged slightly at 0.11235, forming a vertebrate clade. The rice fish (*O. javanicus*) appeared as an early vertebrate branch with a longer distance of 0.16988, consistent with its phylogenetic placement among teleosts. Invertebrate sequences from *C. elegans* and *A. mellifera* showed greater divergence, with branch lengths of 0.35778 and 0.28350, respectively, while the yeast homolog was the most distant (0.43541), reflecting its basal position as a unicellular eukaryote. Adding these evolutionarily distant sequences increased the depth and diversity of the tree but also introduced limitations. First, the presence of identical distances between human and mouse likely reflects localized alignment over conserved motifs rather than full-protein identity, suggesting that deeper divergences may be masked by domain-level similarity. Second, the longer branches associated with yeast, worm, and bee highlight increasing uncertainty in homology, especially when domain architectures differ or annotation confidence is low.

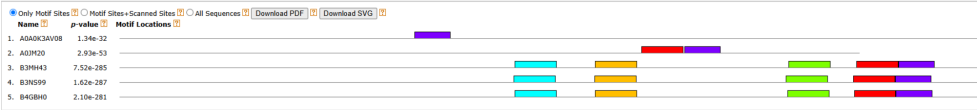
Protein Family, Conserved Domains, and Motif Analysis of EGFR

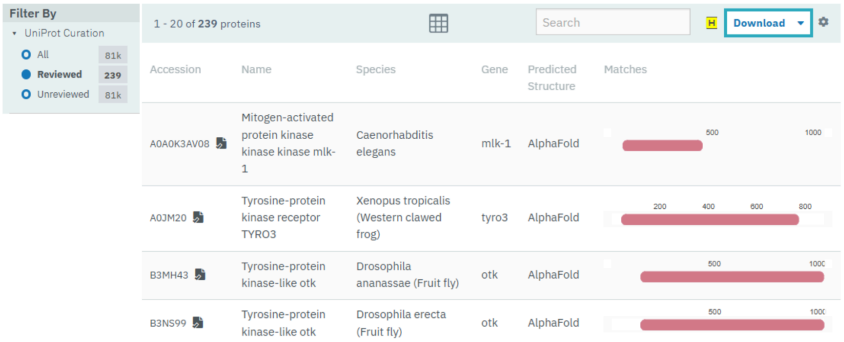
```
CLUSTAL O(1.2.4) multiple sequence alignment

A0A0K3AV08      ------------------------------------------------------------    0
A0JM20          ------------------------------------------------------------    0
B4GBH0          MDMDVMMISMCI-LASTLMAPGWASTSGFLRVPQSQSIVENEAADFGCEA-TDPASYLHY    58
B3MH43          --MAALRISVWILVQALMMALVSSNSSHFLQLPQSQSVVENESVDFECQASTDPSSELHY    58
B3NS99          --MTARMISIYGLVLASMMASVWASSSRFQRLPQSQSVVENESVKFECES-TDSYSELHY    57

A0A0K3AV08      ------------------------------------------------------------    0
A0JM20          ------------------------------------------------------------    0
B4GBH0          EWLHNGREISYDKRVYRIGSHLHIEAVQREEDVGDYVCIATSLASGAREASPPAKLSVIY    118
B3MH43          EWLHNGHRIAYDKRVYQIGSHLHIEAVQRAEDVGDYVCIATSLASGAREASPPAKLSVIY    118
B3NS99          DWLHNGHRIAYDKRVHQIGSNLHIEAARRTEDVGSYVCIATNLASGAREASPPAKLSVIY    117

A0A0K3AV08      ------------------------------------------------------------    0
A0JM20          ------------------------------------------------------------    0
B4GBH0          LESASVQLLGSNRNELLLKCHVEGASGD-EPLQIEWYRDSARLASWGNVHLEEHRLLVRQ    177
B3MH43          IDSASVQLLGSNRNELLLKCHVEAVSGSDDPLQIEWYRNSAKLSSWQNVQLDQHRLIIRQ    178
B3NS99          LESASVQLLGSNRNELLLKCHVEGASGDSEPLEIEWYRNSEKLSTWKNVQLDQHRLIIRQ    177

A0A0K3AV08      ------------------------------------------------------------    0
A0JM20          ------------------------------------------------------------    0
B4GBH0          PSPSDDGLYRCTASNAAGRVMSKQGYVYQANIKCLPRLLK-KNQKLPESWGKQTFLCRGK    236
B3MH43          PSAADDGLYRCTASNAAGRVMSKQGYVYRSSLKCLPRLPRRKNQKLPESWSKEVFLCRGK    238
B3NS99          PGSDDDGLYRCTASNAAGRVMSKQGYAYQSSVKCLPRLARRKNQKMMESWDKQTFLCRGK    237

A0A0K3AV08      ------------------------------------------------------------    0
A0JM20          ------------------------------------------------------------    0
B4GBH0          RGGSGGLDQALSPAPEDLRIVQGPAGQLLIKEGDSAALSCLYELPAELQNQRIQLRWRKD    296
B3MH43          RGGSGGVE-ALPSAPEDLRIVQGPASHAIIKEGDPAALTCLYELPAELQNQRIQLRWRKD    297
B3NS99          RGGAAGLE-ALPAAPEDLRIVQGPVGQSIIKEGEPTALTCLYELPDELKNQRIQLRWRKD    296
```



Discovered Motifs

Motif Locations

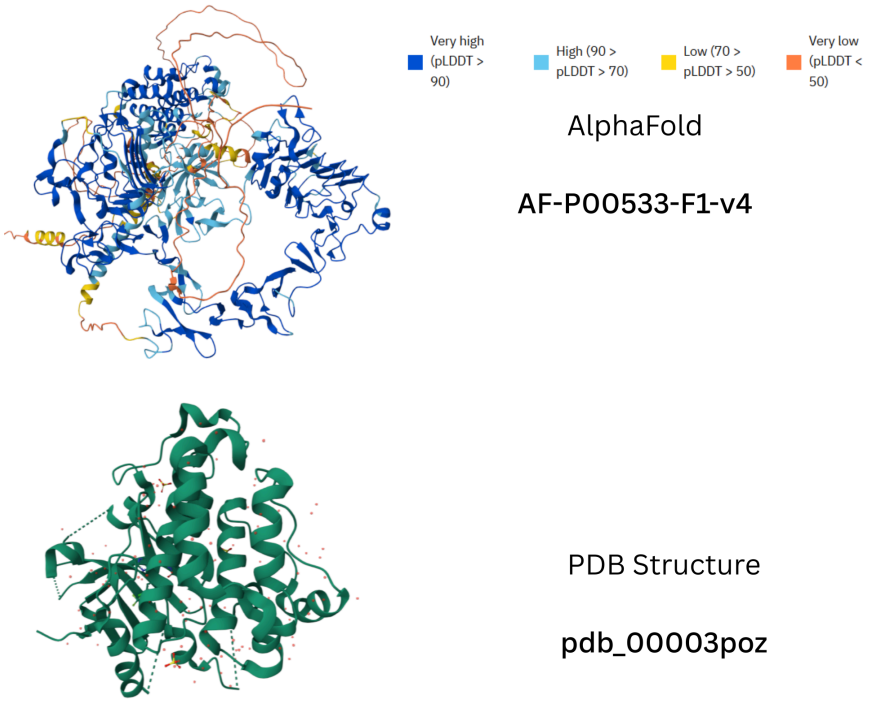




The human EGFR protein (UniProt ID: P00533) is a member of the Receptor Tyrosine Kinase (RTK) family, classified under InterPro entry IPR050122 and further included in the broader tyrosine kinase superfamily IPR016245, which encompasses EGFR/ERB/XmrK-type receptors. This family includes 239 reviewed proteins and over 81,000 predicted entries, underscoring its widespread evolutionary conservation. According to UniProt annotations, EGFR contains several key structural features: a catalytic kinase domain (residues 712–979), a dimerization and phosphorylation region (688–704), and compositionally biased disordered segments toward the C-terminus. To investigate cross-species conservation, five reviewed homologs from *Caenorhabditis elegans*, *Xenopus tropicalis*, and three *Drosophila* species were aligned using Clustal Omega, a widely used multiple sequence alignment tool that offers high sensitivity across divergent sequences by combining progressive alignment with Hidden Markov Models. A phylogram was generated from the alignment to visualize evolutionary distances and detect conserved clustering patterns within the protein family.

To expand on this, motif discovery was performed using MEME Suite (classic mode), yielding five highly significant motifs (E-values ranging from 1.4e–70 to 1.9e–54), each 43–50 amino acids in length. For motif discovery, we used MEME Suite to detect conserved sequence patterns beyond annotated domains. These motifs appeared across all or most sequences and consistently clustered in the mid-to-late protein regions, aligning with the predicted kinase domain. Several motifs particularly Motif 1 and Motif 3 contained key lysine, arginine, and aspartate residues associated with ATP binding and catalytic activity, indicating functional conservation. The motif logos and distribution map further supported this, showing strong positional conservation despite species divergence. This integrated approach combining curated domain data, multiple sequence alignment, and de novo motif discovery demonstrates how bioinformatic tools can be used not only to confirm known functional regions but also to identify novel conserved elements within protein families.

Comparison of EGFR Structure from PDB and AlphaFold Prediction:



The three-dimensional structure of EGFR was analyzed using two independent models: a crystallographic structure obtained from the Protein Data Bank (PDB ID: 3POZ) and the AlphaFold-predicted full-length model from UniProt (P00533). The PDB structure represents the isolated **kinase domain** of EGFR, spanning approximately residues 712–979, and was resolved via **X-ray diffraction at 1.5 Å resolution**. In contrast, the AlphaFold model includes the **entire 1210 amino acid** protein sequence and uses machine learning-based structural inference to predict regions outside experimentally solved domains.

The two structures differ significantly in terms of **coverage** and **confidence resolution**. The PDB entry provides a high-resolution model limited to the core catalytic region, showing a compact fold dominated by well-ordered **α-helices and β-sheets**, particularly in the ATP-binding cleft and activation loop. This structure is suitable for drug-targeting studies and reflects static conformational details of a crystallized domain under controlled conditions. In contrast, the AlphaFold model offers **full-length coverage**, including the extracellular, transmembrane, and C-terminal regions that are absent from crystallographic datasets. AlphaFold's pLDDT confidence scores indicate high reliability (blue, >90) within the kinase domain, but lower confidence (yellow to orange) in the **N-terminal extracellular region** and **disordered C-terminal tail**, which are known to exhibit **conformational flexibility** or lack resolved structural templates.

Functionally, the PDB model enables detailed mechanistic understanding of EGFR's enzymatic activity, while the AlphaFold prediction provides a **broader architectural overview**, useful for mapping inter-domain organization, post-translational modification sites, and potentially disordered regions involved in regulatory interactions. Together, these models are complementary: the crystallographic structure excels in **precision**, while AlphaFold offers **completeness** with probabilistic confidence. This dual-model approach illustrates the value of combining **experimental and computational methods** in structural bioinformatics.

Identification of α-helices and β-sheets

AlphaFold

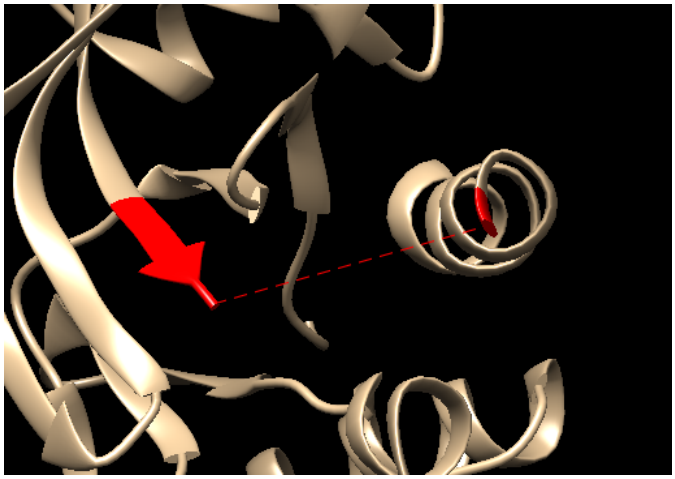
AF-P00533-F1-v4

α-helices → Blue
β-sheets → Green

PDB Structure

pdb_00003poz

Identification of a Helix Break and Responsible Residues:



In the crystallographic structure of EGFR (PDB: 3POZ), a key structural disruption is observed in the **αC-helix** (residues 744–755), where a distinct kink occurs at **Glycine 749 (G749)**. This residue lacks a side chain, allowing enhanced backbone flexibility and enabling the sharp deviation from canonical α-helical geometry (Huse & Kuriyan, 2002). The bend displaces the C-terminal half of the αC-helix outward, stabilizing the **inactive conformation** of the kinase domain and preventing formation of the essential **K745–E762 salt bridge**, which is required for catalytic activity (Zhang et al., 2006). Adjacent residues **E746** and **L748** further contribute by anchoring the N-terminal segment and reinforcing the hydrophobic core, respectively. This localized distortion is critical for regulating conformational dynamics of EGFR and plays a role in **inhibitor binding**

References:

Altenhoff, A. M., Glover, N. M., Train, C. M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., ... & Dessimoz, C. (2016). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life. *Nucleic Acids Research, 46*(D1), D477–D485. https://doi.org/10.1093/nar/gkx1019

Huse, M., & Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell, 109*(3), 275–282. https://doi.org/10.1016/S0092-8674(02)00741-9

Jura, N., Zhang, X., Endres, N. F., Seeliger, M. A., Schindler, T., & Kuriyan, J. (2009). Catalytic control in the EGF receptor and its connection to general kinase regulatory mechanisms. *Molecular Cell, 42*(1), 9–22. https://doi.org/10.1016/j.molcel.2011.03.004

Lemmon, M. A., & Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell, 141*(7), 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology, 1079*, 105–116. https://doi.org/10.1007/978-1-62703-646-7_6

Zhang, X., Gureasko, J., Shen, K., Cole, P. A., & Kuriyan, J. (2006). An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell, 125*(6), 1137–1149. https://doi.org/10.1016/j.cell.2006.05.013