
Análisis bioinformático de sitios conservados de glicoproteínas presentes en eritrocitos humanos, mediante técnicas de modelado por homología.

Maria José Morales Reichenbach

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería



Análisis bioinformático de sitios conservados de glicoproteínas presentes en eritrocitos humanos, mediante técnicas de modelado por homología.

Trabajo de graduación en modalidad de Tesis presentado por
Maria José Morales Reichenbach
para optar al grado académico de Licenciada en Ingeniería
Bioinformática

Guatemala
2023

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería




Análisis bioinformático de sitios conservados de glicoproteínas presentes en eritrocitos humanos, mediante técnicas de modelado por homología.

Trabajo de graduación en modalidad de Tesis presentado por
Maria José Morales Reichenbach
para optar al grado académico de Licenciada en Ingeniería
Bioinformática


Guatemala
2023

Vo.Bo.:


(f) 

Msc. Jorge Chang

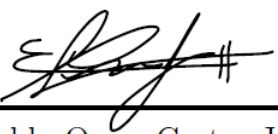
Tribunal Examinador:

(f) 

Msc. Jorge Chang

(f) 

Msc. Luis Augusto Franco López

(f) 

Lic Eddy Omar Castro Jauregui

Fecha de aprobación: Guatemala, 12 de diciembre de 2023.

Agradecimientos

Quiero agradecer a todas las personas que hicieron posible este estudio y que sin su apoyo no hubiera sido posible.

A mi mamá por apoyarme en mis estudios universitarios de manera incondicional.

A mi papá por su acompañamiento.

A mi familia y amigos por ser durante estos cinco años motivación a ser mejor cada día.

A mi asesor Msc. Jorge Chang por su apoyo y ayuda durante la realización de este trabajo de graduación y siempre estar dispuesto a darme consejos.

Agradecimientos	III
Lista de figuras	XII
Lista de cuadros	XIII
Resumen	XIV
1. Introducción	1
2. Objetivos	3
2.1. Objetivo general	3
2.2. Objetivos específicos	3
3. Justificación	4
4. Marco teórico	6
4.1. Aminoácidos	6
4.2. Proteínas	9
4.2.1. Glicoproteínas	10
4.3. Transformación gen a proteína	12
4.4. Árboles filogenéticos para el análisis evolutivo de las proteínas	14
4.5. Alineación de proteínas y las herramientas bioinformáticas	15
4.6. Regiones conservadas	16
4.7. Regiones desordenadas	17
4.7.1. Rol del desorden en proteínas	17
4.8. Técnicas para modelado de proteínas	18
4.9. Factor Rh	19
4.10. Tipos de sangre	20
4.10.1. Conformación de cada grupo	21
4.11. Enfermedades relacionadas a la necesidad del acceso a sangre	21
5. Marco metodológico	23
5.1. Datos iniciales	23
5.2. Procesamiento Expasy	23
5.3. BLAST de las proteínas resultantes	23
5.4. Filtración de proteínas en base de datos Uniprot	24
5.5. Creación de dos grupos de proteínas según su hebra	24

5.6.	Árbol filogenético	24
5.7.	Análisis de regiones desordenadas	24
5.7.1.	IUPred	25
5.7.2.	Algoritmo en R	25
5.8.	Alineación de proteínas en JalView	25
5.9.	Modelado en AlphaFold	25
5.10.	Análisis y alineación de modelados en Chimera	26
6.	Resultados	27
6.1.	Dataset inicial y resultados Expasy	27
6.2.	BLAST de las 30 proteínas mejores rankeadas de cada frame	35
6.3.	Set de Proteínas y árbol filogenético	42
6.4.	Alineación JalView	46
6.5.	Modelados de proteínas	52
6.5.1.	Proteína más desordenada hebra 3'5	53
6.5.2.	Proteína más ordenada hebra 3'5	55
6.5.3.	Proteína más desordenada hebra 5'3	57
6.5.4.	Proteína más ordenada hebra 5'3	59
6.6.	Alineación en Chimera de proteínas seleccionadas	61
6.6.1.	Alineación de cinco proteínas de la hebra 3'5	61
6.6.2.	Alineación de las proteínas presentadas de la hebra 3'5	63
6.6.3.	Alineación de cinco proteínas de la hebra 5'3	65
6.6.4.	Alineación de las proteínas presentadas de la hebra 5'3	67
7.	Discusión de resultados	68
7.1.	Transformación de gen codificante a secuencias proteicas	68
7.2.	Búsqueda BLAST para analizar los grupos de proteínas de cada hebra	68
7.3.	Formación de base de datos del estudio a partir del análisis de ambas hebras	69
7.4.	Ajuste de posiciones en alineación Jalview, detección de regiones conservadas y desordenadas	69
7.5.	Calidad de modelados por técnicas de homología y su relación con el desorden	71
7.6.	Diferencias estructurales entre las proteínas	72
7.7.	Creación y análisis de árboles filogenéticos	73
8.	Antecedentes	75
9.	Alcance	76
10.	Conclusiones	77
11.	Recomendaciones	78
12.	Bibliografía	79
13.	Anexos	84
13.1.	Parámetros para los análisis en las distintas plataformas.	84
13.2.	Análisis proteínas en hebra 3'5	89
13.3.	Análisis proteínas en hebra 5'3	106

Lista de figuras

4.1. Estructura general de los aminoácidos. Extraída de: [NelsonNelson2017]	7
4.2. Códigos de una letra de cada aminoácido según el sistema de Margaret Oakley Dayhoff. Extraída de: [J.J.1984]	8
4.3. Clasificación de los aminoácidos según su grupo R y propiedades químicas. Las estructuras mostradas son los estados en ionización a un pH de 7.0. Extraída de: [NelsonNelson2017]	9
4.4. Proceso de síntesis por deshidratación para la formación de las uniones peptídicas que unen a los aminoácidos. Extraída de: [TeamTeam]	10
4.5. Cuatro niveles estructurales que las proteínas pueden obtener en su estructura. Extraída de: [KnappKnapp]	10
4.6. Tipos principales de glicosilación en humanos. Se muestran los diferentes procesos por lo que los glicanos pueden unirse a lípidos o proteínas. Como resultado de este proceso se obtienen glicoconjugados. Extraída de:[ReilyReily2019]	11
4.7. Proceso de traducción a partir de un gen para obtener una secuencia proteica. Se demuestra el proceso que conlleva la traducción de un gen para obtener las proteínas que este codifica. Extraída de: [TutorialsTutorials]	12
4.8. Tabla que contiene el código para poder leer los diferentes codones y los aminoácidos que representan. Se muestran los códigos utilizados para identificar los diferentes tipos de aminoácidos. Extraída de: [TutorialsTutorials]	13
4.9. Ejemplo de un árbol filogenético generado por el método de unión de vecinos. Las letras muestran los nodos internos, los números grandes las especies de interés y los pequeños la distancia de las ramas. Extraída de: [Saitou NeiSaitou Nei1987]	14
4.10. Diferentes programas del algoritmo BLAST según el tipo de secuencia que se va a trabajar y el resultado que quiere obtenerse. Extraída de: [Donkor, Dayie AdikuDonkor .2014]	16
4.11. Ejemplo de cómo se visualiza un gráfico PAE resultante de AlphaFold. Este gráfico muestra una estructura con varias curvas en sus estructuras. Esto se debe a las áreas verdes que se encuentran alrededor de la diagonal. En este caso el color verde indica una distancia corta entre los aminoácidos. Extraída de: [PhDPhD2022]	19
4.12. Esquema visual de la composición de los grupos sanguíneos según las glicoproteínas presentes. Extraída de: [ByjuByju]	21
6.1. Primer gen de Dataset obtenido de la secuenciación del genoma completo.[JJ2020] Fue procesado en Expasy y del cual se obtienen los diferentes frames de lectura de la hebra 5'3 y 3'5.[60056005]	27
6.2. Resultado Expasy de Gen 1 proteína en Frame 1 Hebra 5'3	28
6.3. Resultado Expasy de Gen 1 proteína en Frame 2 Hebra 5'3	28
6.4. Resultado Expasy de Gen 1 proteína en Frame 3 Hebra 5'3	29

6.5. Resultado Expasy de Gen 1 proteína en Frame 1 Hebra 3'5	29
6.6. Resultado Expasy de Gen 1 proteína en Frame 2 Hebra 3'5	30
6.7. Resultado Expasy de Gen 1 proteína en Frame 3 Hebra 3'5	30
6.8. Segundo gen de Dataset obtenido de la secuenciación del genoma completo.[JJ2020] Fue procesado en Expasy y del cuál se obtiene la hebra 5'3 y 3'5.	31
6.9. Resultado Expasy de Gen 2 proteína en Frame 1 Hebra 5'3	32
6.10. Resultado Expasy de Gen 2 proteína en Frame 2 Hebra 5'3	32
6.11. Resultado Expasy de Gen 2 proteína en Frame 3 Hebra 5'3	33
6.12. Resultado Expasy de Gen 2 proteína en Frame 1 Hebra 3'5	33
6.13. Resultado Expasy de Gen 2 proteína en Frame 2 Hebra 3'5	34
6.14. Resultado Expasy de Gen 2 proteína en Frame 3 Hebra 3'5	34
6.15. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	35
6.16. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	36
6.17. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	36
6.18. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	37
6.19. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	37
6.20. Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	38
6.21. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	38
6.22. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	39
6.23. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	39
6.24. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	40
6.25. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	40
6.26. Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.	41
6.27. Árbol filogenético de las proteínas encontradas en la hebra 3'5	44
6.28. Árbol filogenético de las proteínas encontradas en la hebra 5'3	45
6.29. Parte 1 de la alineación en JalView de las proteínas en la hebra 3'5, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.	47

6.30. Parte 2 de la alineación en JalView de las proteínas en la hebra 3'5, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.	48
6.31. Parte 3 de la alineación en JalView de las proteínas en la hebra 3'5, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.	49
6.32. Parte 1 de la alineación en JalView de las proteínas en la hebra 5'3, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.	50
6.33. Parte 2 de la alineación en JalView de las proteínas en la hebra 5'3, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.	51
6.34. Análisis proteína Q6ZQR8 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.	53
6.35. Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q6ZQR8. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.	54
6.36. Análisis proteína Q8N8C2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.	55
6.37. Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q8N8C2. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.	56

6.38. Análisis proteína Q9UI50 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.	57
6.39. Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q9UI50. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60 %. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.	58
6.40. Análisis proteína M1SZX7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.	59
6.41. Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína M1SZX7. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60 %. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.	60
6.42. Vista cercana a la alineación en Chimera de proteínas Q8WZ27, Q8N8C2, Q6ZQR8, Q1M183, B4E0H0 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.4.	61
6.43. Vista total de la alineación en Chimera de proteínas Q8WZ27, Q8N8C2, Q6ZQR8, Q1M183, B4E0H0 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.4.	62
6.44. Vista completa de la alineación en Chimera de proteínas Q8N8C2 y Q6ZQR8 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.5.	63
6.45. Vista cercana de la alineación en Chimera de proteínas Q8N8C2 y Q6ZQR8 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.5.	64
6.46. Vista completa de la alineación en Chimera de proteínas A0A0A1TSG9, A2NVG1, M1SZX7, Q8WYX4, Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.6.	65
6.47. Vista cercana de la alineación en Chimera de proteínas A0A0A1TSG9, A2NVG1, M1SZX7, Q8WYX4, Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.6.	66

6.48. Alineación en Chimera de proteínas M1SZX7 y Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.7.	67
7.1. Secuencias de proteínas altamente conservadas correspondientes al grupo de estudio de la hebra 5'3.	74
13.1. Parámetros utilizados durante el procesamiento de datos en Expasy.	84
13.2. Parámetros utilizados durante el procesamiento de datos en BLAST.	85
13.3. Parámetros utilizados durante el procesamiento de datos en IUPred2A.	85
13.4. Parámetros de las proteínas alineadas en Chimera de la hebra 3'5 para el análisis de las figuras 6.42 y 6.43.	86
13.5. Parámetros de las proteínas alineadas en Chimera de la hebra 3'5 para el análisis de las figuras 6.44 y 6.45.	86
13.6. Parámetros de las proteínas alineadas en Chimera de la hebra 5'3 para el análisis de las figuras 6.46 y 6.47.	87
13.7. Parámetros de las proteínas alineadas en Chimera de la hebra 5'3 para el análisis de la Figura 6.48.	87
13.8. Código de color de los aminoácidos para analizar la alineación de múltiples proteínas en JalView según coloración Clustal.	88
13.9. Programa en R análisis de desorden.	88
13.10 Análisis proteína Q6ZP21 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	89
13.11 Análisis proteína B4E0H0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	90
13.12 Análisis proteína Q1M183 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	91
13.13 Análisis proteína Q6ZNU7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	92
13.14 Análisis proteína Q6ZP34 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	93
13.15 Análisis proteína Q6ZP99 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	94
13.16 Análisis proteína Q6ZPA0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	95
13.17 Análisis proteína Q6ZPB0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	96
13.18 Análisis proteína Q6ZPB2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	97
13.19 Análisis proteína Q6ZUG4 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	98
13.20 Análisis proteína Q6ZUK0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	99

13.21	Análisis proteína Q8WZ27 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	100
13.22	Análisis proteína Q9H387 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	101
13.23	Análisis proteína Q9H728 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	102
13.24	Análisis proteína Q9H743 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	103
13.25	Análisis proteína Q9NX85 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	104
13.26	Análisis proteína Q9P195 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	105
13.27	Análisis proteína A0A0A1TSG9 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	106
13.28	Análisis proteína A0A0A8IKZ2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	107
13.29	Análisis proteína A0A1A9C9I6 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	108
13.30	Análisis proteína A0A8D5ZBQ7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	109
13.31	Análisis proteína A2NVG1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	110
13.32	Análisis proteína B2RNZ7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	111
13.33	Análisis proteína Q1KSG2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	112
13.34	Análisis proteína Q6ZP50 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	113
13.35	Análisis proteína Q6ZPC1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	114
13.36	Análisis proteína Q8WYX4 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	115
13.37	Análisis proteína Q9UBB8 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	116
13.38	Análisis proteína Q96NR6 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	117

13.39	Análisis proteína Q969H1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	118
13.40	Análisis proteína Q02094 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	119
13.41	Análisis proteína Q16416 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas	120

Lista de cuadros

6.1. Códigos en Uniprot de las proteínas estudiadas de las primeras 15 mejores rankeadas en BLAST separadas por la hebra en que fueron encontradas.	43
6.2. Tabla que muestra la puntuación RMSD obtenida luego de hacer la alineación estructural en Chimera de las proteínas seleccionadas del grupo de la hebra 3'5. Mientras más bajos sean los valores quiere decir que las proteínas tienen un mayor número de similitudes y viceversa.	61
6.3. Tabla que muestra la puntuación RMSD obtenida luego de hacer la alineación estructural en Chimera de las proteínas seleccionadas del grupo de la hebra 5'3. Mientras más bajos sean los valores quiere decir que las proteínas tienen un mayor número de similitudes y viceversa.	65

Este trabajo de graduación se enfoca en el análisis de los sitios conservados y desorden entre las glicoproteínas que forman parte del factor Rh. Este factor es un componente sanguíneo que indica la presencia u ausencia de los antígenos D en los glóbulos rojos de la sangre. Los sitios conservados son las secciones de las proteínas que no cambian a través de la evolución que se sufre. [Stojanovic .Stojanovic .1999] Al realizar este análisis y encontrar los sitios conservados podemos observar la conservación entre familias de proteínas y analizarlas estructuralmente. Estos modelos de las estructuras serán analizados para comprobar la hipótesis de la conservación y variaciones de las glicoproteínas que se encuentran en el mismo Rh. Con estas diferencias se comprueba la conservación estructural y la forma en que los cambios en la estructura podrían llegar a cambiar la función. Toda la investigación de sitios conservados y desorden de la proteína es el inicio para poder dar origen a la transformación experimental de las estructuras para cambiar el tipo de sangre de una muestra. En este trabajo no se realizará nada experimental, se hará todo el análisis bioinformático necesario para poder lograr el trabajo de laboratorio. Este análisis se inició con la búsqueda de proteínas relacionadas a las codificadas del gen presente en el factor Rh. Los datos utilizados son secuencias disponibles en la base de datos Uniprot. Estos datos fueron analizados mediante árboles filogenéticos, modelados por homología, gráficas del desorden, alineación de la secuencia y las estructuras tridimensionales. Los árboles filogenéticos ayudaron a identificar las relaciones evolutivas entre las secuencias. Así como la relación que esto tiene con las regiones conservadas. Al realizar los modelos por homología se generaron gráficas que ayudaron a clasificar los modelados como buenos o malos. A este análisis se le agregan las regiones desordenadas y se concluye cómo estas regiones pueden afectar o no la calidad de la estructura. Mediante el proceso de alineación de secuencias se comparó estructuralmente la conservación de las secuencias así como la calidad de los modelos. Se encontraron varias regiones conservadas en las proteínas estudiadas. También se pudo relacionar esta conservación con la evolución de las proteínas. Como último resultado se logró analizar el desorden de cada modelado para determinar si este era un factor importante en la calidad del modelo.

CAPÍTULO 1

Introducción

Las proteínas forman parte de las funciones esenciales de la vida y son una de las cuatro macromoléculas principales. Están formadas por aminoácidos los cuáles les atribuyen su forma, tamaño y funciones. [NelsonNelson2017] Se tiene un grupo de 20 aminoácidos los cuáles forman todas las proteínas que existen, lo único que cambia entre ellos es el orden en que se unen y propiedades físico-químicas. Dentro del grupo de las proteínas existen las glicoproteínas, estas se encuentran comúnmente en la sangre humana. Son oligosacáridos formados de glicanos unidos a nitrógeno u oxígeno. Dependiendo del tipo de glicoproteínas que se encuentren presentes en la sangre se determina el grupo sanguíneo de la misma.[SpiroSpiro1973]

El sistema utilizado universalmente para clasificar los grupos sanguíneos se basa en los grupos A, B, O, AB y el factor Rhesus positivo o negativo. El grupo obtiene su nombre por los antígenos presentes en el mismo. Estos antígenos hacen que el grupo sanguíneo tenga un papel en las respuestas inmunológicas del cuerpo. Por esta razón existe un sistema de donación de sangre en el cuál un individuo no puede recibir cualquier tipo de sangre.[GoodwinGoodwin2021] El único grupo sanguíneo que puede donar sangre a cualquier individuo es el grupo O-. Este es también el grupo menos común por lo que es importante realizar un estudio de las proteínas que determinan al mismo. Al investigar las glicoproteínas presentes en el grupo Rh se pueden encontrar regiones conservadas que determinen la funcionalidad al determinar el factor Rhesus en la sangre.[ClinicClinic20222]

Las glicoproteínas presentes en el grupo Rh pueden obtenerse a partir de la traducción de un gen que las codifique. Este proceso puede llevarse a cabo con técnicas bioinformáticas como la plataforma Expasy. La plataforma da como resultado las posibles secuencias de glicoproteínas que van a utilizarse. Con estos datos es posible realizar búsquedas de alineación de pares como BLAST. [NIHNIH] Esto hace que se puedan encontrar muestras similares y armar una base de datos concisa para el uso del estudio. Esta base de datos hace referencia a Uniprot esta es una base en la cual las proteínas se ven divididas en dos grupos: las revisadas por lo que son datos confiables para poder estudiarlos según su estructura y características. El otro grupo son proteínas que aún no se han estudiado a profundidad por lo que solo se tiene la secuencia. Tiene aproximadamente 120 millones de entradas. [AokiAoki20171] Las proteínas pueden tener dentro de su secuencia de aminoácidos regiones desordenadas. Esto indica que la proteína no tendrá una estructura tridimensional específica. Estas regiones aunque parecen ser un problema son de hecho la parte esencial de muchas proteínas para que se pueda cumplir su función.

La alineación de múltiples proteínas es una herramienta que permite encontrar regiones conservadas dentro de la secuencia. Este es uno de los objetivos de nuestro estudio ya que a partir

de estas regiones se puede analizar cómo fueron evolucionando las proteínas, junto con los árboles filogenéticos. También se puede ver que tanta diferencia existe entre las regiones conservadas y esas mismas regiones en la estructura tridimensional. Estas regiones conservadas se encuentran utilizando la herramienta bioinformática Jalview.[WaterhouseWaterhouse]

La herramienta de modelado AlphaFold es un sistema de AI desarrollado para predecir la estructura tridimensional de una secuencia de aminoácidos. La base de datos que utiliza tiene aproximadamente 200 millones de entradas, incluye la mayoría de la base de datos de UniProt. Actualmente es la forma más efectiva de predecir una estructura por su alta precisión al modelar. [AlphaFoldAlphaFold] Los modelos van a ser analizados para comprobar que su precisión sea aceptable y tener el análisis completo en la investigación. Para evaluarlos y analizar su estructura se hace uso de las gráficas pLDDT, PAE y sequence coverage, que produce el algoritmo. Con estos modelos se van a buscar los sitios que se alinean para poder describirlos y ver qué papel están cumpliendo.

Chimera es la plataforma en la que se pueden visualizar las estructuras tridimensionales de las proteínas y alinearlas estructuralmente. Se analiza cómo se conforma la secuencia y se pueden realizar análisis como match entre proteínas, seleccionar partes específicas, etc. En Chimera se verá cómo las estructuras cambian entre ellas y también cómo se conservan. Esto hace posible la conclusión entre la diferenciación que existe en las glicoproteínas presentes en el factor Rh. También ayuda a determinar si las secuencias conservadas de las proteínas pueden llegar a alinearse estructuralmente.[Pettersen .Pettersen .2004]

2.1. Objetivo general

Identificar regiones conservadas entre diferentes glicoproteínas encontradas en eritrocitos humanos y analizar las secciones desordenadas presentes en las mismas.

2.2. Objetivos específicos

- Realizar alineamiento de secuencias de glicoproteínas humanas.
- Realizar una búsqueda de proteínas homólogas mediante técnicas de alineación y familias.
- Realizar modelos por homología de glicoproteínas.
- Identificar sitios conservados y diferencias estructurales de glicoproteínas.

Los problemas de falta de sangre O- en hospitales afectan especialmente a un país como Guatemala. Donde el acceso a servicios de salud básicos es limitado y los servicios de emergencia ineficientes, son muy comunes. Agregando a este problema se tiene la rareza de la sangre O-, ya que solo el 7% de la población mundial tiene este tipo de sangre.[ClinicClinic2022] Por lo que aunque se tenga un sistema de salud eficiente sería muy difícil tener un suministro de sangre apto para cualquier paciente para que estos puedan recibir sus tratamientos. La solución que se propone para este problema es el seguimiento de la investigación del análisis de las glicoproteínas. Con esto se podría llegar a tener la forma más eficiente en la que el anticuerpo A y B pueden ser eliminados para tener sangre O-segura para utilizar.

El lograr esto haría que se tuviera una forma de tener a disponibilidad de la población sangre O- en los bancos sin importar el grupo sanguíneo al que pertenecen los donadores. De esta forma se tendrá sangre para cualquier persona en todos los hospitales y servicios de emergencia. La identificación de sitios conservados en las glicoproteínas específicas de grupos sanguíneos permite visualizar las similitudes entre estas. Al saber qué secciones de la secuencia se comparten entre las proteínas se puede empezar a desarrollar una mejor enzima que interactúe con las proteínas y así poder eliminar los anticuerpos necesarios para conseguir un tipo de sangre O-. El análisis de las glicoproteínas es necesario para los avances médicos. Ya que forman parte de funciones esenciales en el cuerpo como el transporte, actividad de anticuerpos y coagulación. Algunos ejemplos de estas glicoproteínas tan importantes son las estimuladoras de tiroides, hidrolasas, oxidasas y transferasas. [SpiroSpiro1973]

Al realizar la investigación se ha observado la falta de documentación que existe de las glicoproteínas de grupos sanguíneos específicos. Según la literatura y documentación se encontró solamente estudios donde secuencian el genoma completo. A partir de eso obtienen genes que codifican específicamente glicoproteínas relacionadas con el grupo Rh. [JJ2020] La importancia del presente estudio se hace más evidente ya que el modelado de las glicoproteínas no se ha realizado. Es importante notar que las bases de datos de datos curados como lo es SwissProt no tienen a las glicoproteínas utilizadas en el estudio. A partir de los datos procesados y analizados se puede llegar a trabajar en conjunto con estas bases de datos para poder expandir el conocimiento que se tienen de estas importantes proteínas.

El análisis bioinformático surgió una década antes de la secuenciación de ADN. Esto quiere decir que las herramientas bioinformáticas no son solamente para este campo, si no vienen de principios fundados desde hace mucho tiempo atrás. Las computadoras empezaron a surgir como herramienta para el análisis biológico desde 1960. La necesidad surgió por la acumulación de datos y su necesi-

dad de analizarlos. Actualmente la bioinformática es crucial para el análisis de conjuntos de datos complejos y grandes. Permite a los investigadores manejar, guardar y analizar datos de genómica, proteómica y biológicos. Esto ha abierto puertas para encontrar tratamientos específicos, la base molecular de enfermedades y medicina personalizada.[HagenHagen2000]

4.1. Aminoácidos

Las proteínas como ya se mencionó son macromoléculas fundamentales en la biología, tienen una amplia gama de funciones esenciales en los seres vivos. La unidad básica de una proteína es el aminoácido, este funciona como un bloque fundamental en la construcción de estas estructuras. Comprender la estructura y la función que los aminoácidos tienen en las proteínas es esencial para la biología y la bioquímica, ya que estos compuestos son los bloques de construcción que determinan la diversidad funcional de las proteínas. El mismo grupo de 20 aminoácidos forman la diversa cantidad de proteínas que existen, lo único que cambia es su conformación. El primer aminoácido descubierto fue la Asparagina en 1806, la última fue la Treonina en 1938. Los nombres de estos aminoácidos surgen generalmente del objeto del que fueron aislados, como es el caso del Glutamato que fue aislado a partir del gluten. [NelsonNelson2017]

Estos aminoácidos que forman a las proteínas son llamados residuos representando al agua que se pierde cuando los aminoácidos se unen. Por esto mismo las proteínas pueden romperse en sus aminoácidos pasando por un proceso en el que se hidrolizan. Los 20 aminoácidos existentes son moléculas orgánicas que tienen un grupo amino (-NH₂) y un grupo carboxilo (-COOH) unidos a un carbono central, llamado carbono alfa (α). Por lo que son llamados α -aminoácidos. Además, cada uno tiene un grupo lateral único, también conocido como cadena lateral o radical (grupo R), que confiere propiedades químicas y físicas específicas de cada uno. Entre estas características se encuentra la carga eléctrica, solubilidad, tamaño.[NelsonNelson2017] Las estructuras de cada aminoácido son conjuntos de un grupo carboxilo, amino y un residuo como se muestra en la Figura 4.1.

Para referirse a los aminoácidos se utilizan códigos de tres letras que en la mayoría de casos son las primeras tres letras del nombre. Cuando se empezaron a realizar investigaciones con las secuencias proteicas en computadoras surgió una nueva forma de referirse a estos aminoácidos. Margaret Oakley Dayhoff fue la inventora de este nuevo sistema el cuál consiste solamente de una letra identificadora a cada aminoácido. Ella es considerada la fundadora de la bioinformática ya que hizo importantes aportes al campo. Este nuevo sistema hace que los archivos de datos que contienen secuencias de proteínas ocupen menos espacio. Esto era de gran interés en aquella época por la poca capacidad que tenían las computadoras. La Figura 4.2 muestra los códigos que se utilizan para cada uno de estos aminoácidos. [NelsonNelson2017]

El grupo R permite clasificar a los aminoácidos en diferentes grupos según su polaridad y tenden-

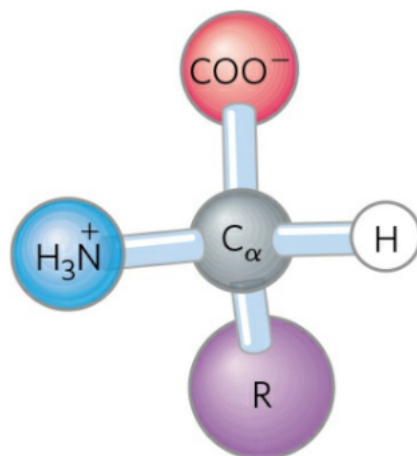


Figura 4.1: Estructura general de los aminoácidos. Extraída de: [NelsonNelson2017]

cia para interactuar con el agua. La asignación del grupo al que pertenecen los aminoácidos Glicina, Histidina y Cisteína se realiza por el juicio de los investigadores más que como una verdad absoluta. Esto ocurre por las cualidades tan específicas de estos aminoácidos que hace que no tengan lugar en uno de los grupos determinados. En la Figura 4.3 se muestra la clasificación de los aminoácidos según su estructura.

Aminoácidos No Polares: La característica de este grupo es que tienen cadenas laterales hidrofóbicas, lo que significa que repelen el agua. En este grupo se incluyen la Glicina, Alanina, Valina, Leucina, Prolina, Isoleucina y Metionina. La Alanina, Valina, Leucina e Isoleucina tienden a agruparse dentro de las proteínas, esto hace que se establezca la estructura proteica a través del efecto hidrofóbico. La Glicina es el aminoácido con la estructura más simple, se clasifica como un aminoácido no polar pero su cadena lateral pequeña no contribuye significativamente a las interacciones hidrofóbicas. La Metionina tiene en su cadena lateral un grupo tioéter ligeramente no polar. La Prolina tiene una cadena lateral alifática con una estructura cíclica distintiva. El grupo amino secundario (imino) de sus residuos se encuentra en una conformación rígida que reduce la flexibilidad estructural en las regiones de polipéptidos. [NelsonNelson2017]

Aminoácidos Polares: Este grupo tiene cadenas laterales hidrofílicas, forman enlaces de hidrógeno con el agua. Lo que significa que interactúan favorablemente con el agua. Se incluyen la Serina, Treonina, Cisteína, Asparagina y Glutamina. La polaridad de la Serina y la Treonina se debe a sus grupos hidroxilos, en la Asparagina y Glutamina se debe a sus grupos amida. Estos aminoácidos son fácilmente hidrolizados por base o ácido. La Cisteína tiene su polaridad contribuida por un grupo sulfhídrico. Es una excepción al grupo ya que es un ácido débil y puede formar enlaces de hidrógeno débiles con oxígeno o nitrógeno. [NelsonNelson2017]

Aminoácidos Aromáticos: En este grupo se incluye la Fenilalanina, Tirosina y Triptófano, tienen sus cadenas laterales aromáticas. Son relativamente no polares lo que los hace ligeramente hidrofóbicos. El grupo hidroxilo de la Tirosina puede formar enlaces de hidrógeno y es un grupo funcional importante en algunas enzimas. La Tirosina y Triptófano son significativamente más polares que la Fenilalanina. Esto es por el grupo hidroxilo de la Tirosina y al nitrógeno del anillo de indol del Triptófano. Estos aminoácidos absorben la luz ultravioleta, los que mejor lo hacen son la Tirosina y Triptófano. Esto les da una absorción grande (280 nm) de luz a las proteínas. [NelsonNelson2017]

Aminoácidos con carga positiva: En este grupo se encuentra la Lisina, Arginina e Histidina. Estos tienen los grupo R más hidrofílicos. Tienen carga significativamente positiva a un pH de 7.0. La Lisina tiene un segundo grupo amino primario en la cadena alifática. La Arginina tiene un grupo

Abbreviation	1 letter abbreviation	Amino acid name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Pyl	O	Pyrrolysine
Ser	S	Serine
Sec	U	Selenocysteine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Aspartic acid or Asparagine
Glx	Z	Glutamic acid or Glutamine
Xaa	X	Any amino acid
Xle	J	Leucine or Isoleucine
TERM		termination codon

Figura 4.2: Códigos de una letra de cada aminoácido según el sistema de Margaret Oakley Dayhoff. Extraída de: [J.J.1984]

guanidinio cargado positivamente. La Histidina tiene un grupo imidazol aromático. Este es el único aminoácido común con una cadena lateral ionizable, puede estar cargada positivamente o sin carga a pH 7.0. Los residuos de este aminoácido facilitan muchas reacciones catalizadas por enzimas ya que pueden actuar como donantes y aceptores de protones.[NelsonNelson2017]

Aminoácidos con carga negativa: Dentro de estos aminoácidos solo se encuentra el Aspartato y Glutamato. Ambos tienen grupos R con carga neta negativa en un pH de 7.0. [NelsonNelson2017]

Los aminoácidos se unen entre ellos mediante enlaces peptídicos para formar las cadenas lineales conocidas como péptidos o proteínas. Pueden unirse mediante un enlace peptídico, este se forma por la deshidratación del agua del grupo α -carboxilo de un aminoácido y se une al α -amino del otro. Este proceso se muestra en la Figura 4.4. Este es un ejemplo de una reacción de condensación, una clase común de reacciones en las células vivas. En condiciones bioquímicas estándar, el equilibrio para la reacción favorece a los aminoácidos en lugar de al dipéptido. Para que la reacción sea termodinámicamente más favorable, el grupo carboxilo debe ser modificado o activado para eliminar el grupo hidroxilo.

Los péptidos pueden identificarse por su comportamiento de ionización. Estos tienen grupos α -amino y α -carboxilo libres en extremos opuestos de la cadena. Los enlaces péptidos que unen los aminoácidos no se ionizan y no afectan el comportamiento ácido-base de los péptidos. Al contrario, los grupos R de algunos aminoácidos pueden ionizarse, lo que contribuye al comportamiento ácido-

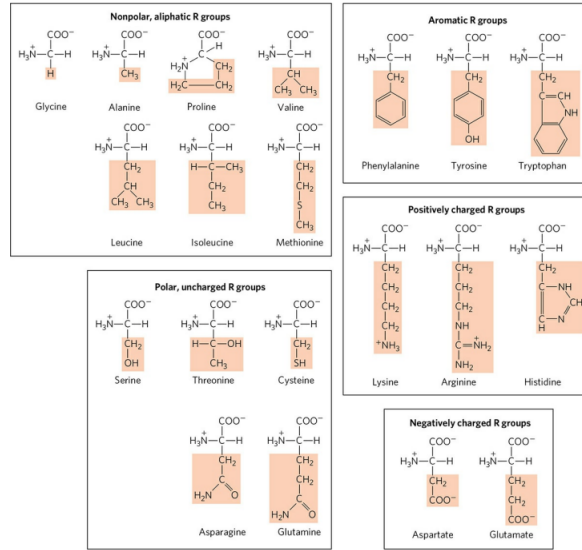


Figura 4.3: Clasificación de los aminoácidos según su grupo R y propiedades químicas. Las estructuras mostradas son los estados en ionización a un pH de 7.0. Extraída de: [NelsonNelson2017]

base general de un péptido. Por lo que se puede decir que la predicción de este comportamiento de un péptido se basa en los grupos α -amino y α -carboxilo libres, junto con la naturaleza y número de grupos R ionizables en el péptido.[Loo, Udseth SmithLoo .1989]

4.2. Proteínas

Existe una gran variedad de proteínas y cientos de estas pueden encontrarse en una sola célula. Esto se debe a la gran cantidad de procesos que dependen de ellas y lo especializadas que son. [NelsonNelson2017] Las diferentes propiedades y funciones de las proteínas se forman a partir de las diferentes combinaciones que las células pueden formar con el mismo set de 20 aminoácidos. [NelsonNelson2017] Las proteínas están formadas de aminoácidos que se unen a polímeros por diferentes tipos de enlaces. Las uniones peptídicas, uno de estos tipos de enlaces, se forman por la síntesis de deshidratación, este proceso se ejemplifica en la Figura 4.4. [TeamTeam] Estas uniones peptídicas hacen posible la formación de proteínas, también pueden confundirse con los péptidos. El término péptido y proteína son generalmente intercambiados y usados para referirse a la misma cosa. La diferencia es que el péptido tiene un peso molecular menor a 10,000. [NelsonNelson2017] Hay seis grupos principales de proteínas, cada grupo cumple una función específica en el organismo en el que se presente. Las funciones por las que se dividen son: transporte, defensa, estructurales, hormonales, contráctiles.

La estructura proteica se divide en cuatro niveles según la complejidad que se está estudiando. A las diferentes estructuras se les llama primaria, secundaria, terciaria y cuaternaria. Su conformación va de más simple a más compleja como puede observarse en la Figura 4.5. La estructura primaria es la más simple y consiste de una sola cadena de aminoácidos.[TeamTeam] Esta es la que codifica el ADN de la proteína, cualquier cambio en los aminoácidos que conforman esta estructura significa una alteración considerable en la función de la proteína. La estructura secundaria se diferencia por la presencia de puentes de hidrógeno, esto hace que la estructura se doble o enrolle. Las formas que puede tomar la proteína son hélices alfa o láminas beta. En algunos casos las proteínas tienen presentes ambas formas formando una proteína más compleja. La estructura terciaria es la que forma

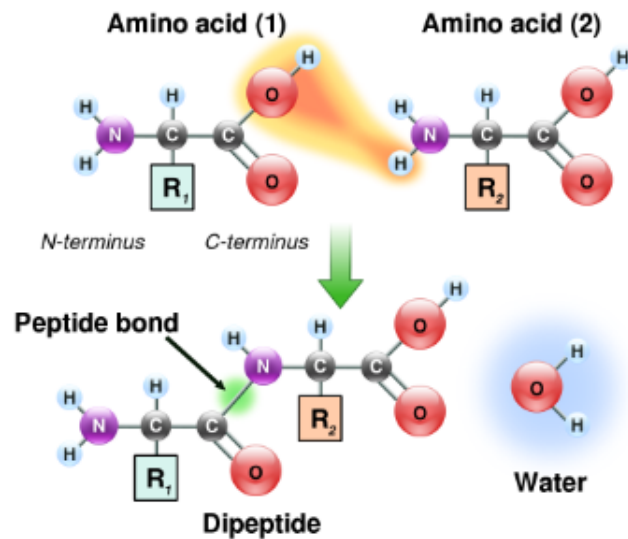


Figura 4.4: Proceso de síntesis por deshidratación para la formación de las uniones peptídicas que unen a los aminoácidos. Extraída de: [TeamTeam]

la conformación 3D de la proteína. Esta depende del tipo de uniones presentes como: Van der Waals, puentes de hidrógeno, enlaces iónicos, etc. La estructura cuaternaria se conforma por subunidades específicas por lo que no todas las proteínas son consideradas cuaternarias. [TeamTeam]

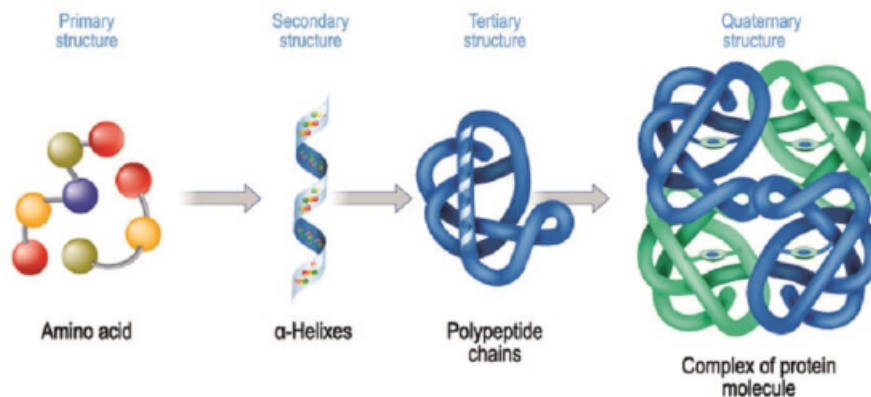


Figura 4.5: Cuatro niveles estructurales que las proteínas pueden obtener en su estructura. Extraída de: [KnappKnapp]

4.2.1. Glicoproteínas

Las glicoproteínas se conforman de cadenas de oligosacáridos, glicanos que se unen a los aminoácidos de forma covalente. Se denominan como glicoconjugados o carbohidratos complejos. [ReilyReily2019] Estos se encuentran en la parte externa de la membrana plasmática, matriz extracelular y la sangre. También tienen funciones dentro de la célula y en este caso se encuentran en el aparato de

Golgi, glándulas secretoras y lisosomas. La parte de los oligosacáridos en las glicoproteínas son bastante heterogéneas, contienen información, reconocen sitios específicos y se unen con lectinas. [NelsonNelson2017]

Estas proteínas se pueden formar por glicanos unidos a nitrógeno u oxígeno. De ahí surgen sus nombres N-glicanos u O-glicanos, estos dos grupos tienen diferentes uniones y subclasificaciones como puede verse en la Figura 4.5. [ReilyReily2019] Las glicoproteínas tienen su unión al carbohidrato en su carbono anomérico mediante un enlace glucosídico al -OH de una Serina o Treonina (O-glicanos). Los enlaces N-glicanos se unen al nitrógeno amino del residuo de Asparagina. [NelsonNelson2017]

Dependiendo a que están unidas las cadenas de carbohidratos se forman glicoproteínas o glicolípidos. Los glicolípidos son componentes de la membrana plasmática en los que la cabeza hidrofílica está formada por oligosacáridos. [NelsonNelson2017] Funcionan igual que las glicoproteínas al unirse con lectinas. En ambos casos la unión de las cadenas de carbohidratos se lleva a cabo por un proceso llamado glicosilación y ocurre en el retículo endoplasmático y aparato de Golgi. Está asociado a enfermedades genéticas. Ahora se puede asociar este proceso a la mejora de respuesta inmune, a la metástasis de las células, etc. [ReilyReily2019]

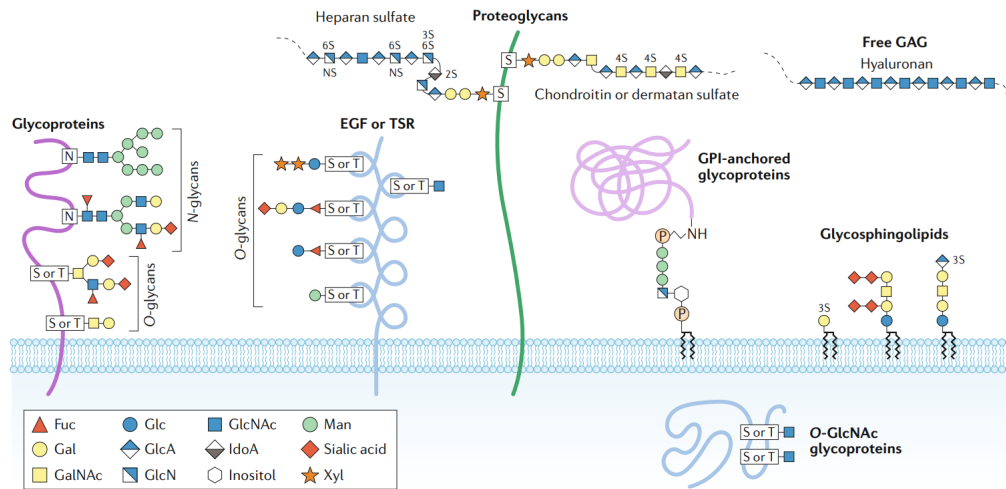


Figura 4.6: Tipos principales de glicosilación en humanos. Se muestran los diferentes procesos por lo que los glicanos pueden unirse a lípidos o proteínas. Como resultado de este proceso se obtienen glicoconjugados.

Extraída de:[ReilyReily2019]

Dentro de la categoría de glicoproteínas se encuentra una extensa cantidad de proteínas y hormonas. Como por ejemplo las proteínas plasmáticas y hormonas peptídicas. También podemos encontrarlas en sustancias del cuerpo sanguíneo. Las alteraciones de las glicoproteínas han sido estudiadas para alterar células cancerosas en el proceso de la metástasis. Esto nos indica que según la estructura de las mismas se puede alterar su función. Esta alteración en la función hace que la determinación del tipo de sangre de una persona se vea ligada al tipo de glicoproteínas que están presentes [BenderBender2019]

Las glicoproteínas que pueden encontrarse en la membrana celular de los eritrocitos, es decir los glóbulos rojos, varían en su estructura por el grupo sanguíneo al que pertenecen. Algunas de las glicoproteínas que encontramos son inmunoglobulina y hormonas como la reguladora de la tiroides. La banda 3, un tipo de glicoproteína, tiene un oligosacárido ligado a N. Las glicoforinas en su mayoría tienen oligosacáridos ligados a O. Este tipo de oligosacáridos están unidos a ácido siálico el cual puede ser NeuAc. Estas no son las únicas glicoproteínas que se encuentran en la sangre por la variedad en la estructura que pueden tomar. [AokiAoki20172] La hipótesis de este trabajo de

graduación es que se puede tener un análisis estructural de estas glicoproteínas. Con esto se busca encontrar los sitios conservados que tienen las glicoproteínas y ver como cambian entre las presentes las pertenecientes al mismo grupo. Esto dará el precedente para la experimentación de cambios de tipo de sangre según sus glicoproteínas.

4.3. Transformación gen a proteína

Los datos utilizados en este estudio se encontraron a partir de dos genes obtenidos de la secuenciación de un genoma completo.[JJ2020] Por esto es importante entender como se realizó este proceso y todo lo que esto implica. La plataforma Expsy es la herramienta bioinformática más utilizada en el área de proteómica. Su nombre significa Expert Protein Analysis System, este surgió cuando era de los primeros servidores para el análisis de las ciencias de la vida. Actualmente es un portal completo para el análisis bioinformático, incluyendo la conversión de un gen a secuencias proteicas.[Artimo .Artimo .2012] Para realizar este proceso se debe tener claro el concepto del Dogma Central. Este es un conjunto de instrucciones que transforma el ADN a un producto funcional, en este caso proteínas. Explica el flujo de información genética que se tiene en el proceso. Para empezar se considera que el ADN tiene la información suficiente para hacer las proteínas que necesitamos. Luego se convierte en el ARN, el cual es el mensajero que pasa esa información a los ribosomas. Acá es donde se genera el producto final. Las secuencias de nucleótidos que codifican un polipéptido contienen un codón de inicio y de fin. Este mismo proceso es el que se lleva a cabo en Expsy solo que bioinformáticamente. En la Figura 4.7 se observa el proceso que se lleva a cabo para traducir el ADN.

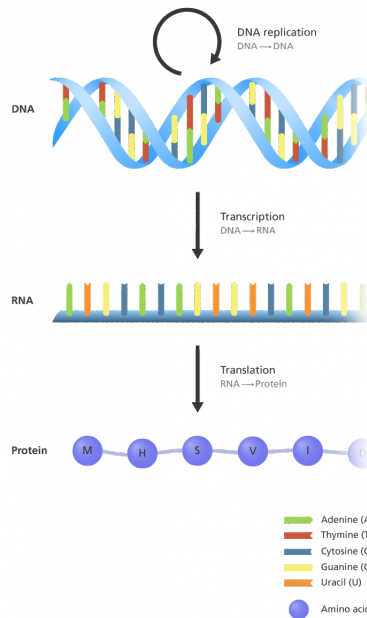


Figura 4.7: Proceso de traducción a partir de un gen para obtener una secuencia proteica. Se demuestra el proceso que conlleva la traducción de un gen para obtener las proteínas que este codifica. Extraída de: [TutorialsTutorials]

La secuencia de ADN consta de cuatro letras A, C, G y T; adenina, citosina, guanina y timina respectivamente. La combinación de estas letras forma las diferentes secuencias de ADN que pueden existir. Dentro de estas secuencias hay divisiones que llamamos codones. Estos codones son combi-

nación de tres letras que pueden ser A, C, G y T, existen 64 codones diferentes. De los 64 solo 61 de ellos representan a uno de los 20 aminoácidos existentes. Los tres codones sobrantes son los encargados de identificar el codón fin de la secuencia. Este sistema se describe como el código genético y es redundante ya que un aminoácido puede ser identificado por varios codones, más un codón solo puede identificar a un aminoácido.[TutorialsTutorials]

Este código se utiliza universalmente, hay algunas excepciones específicas como lo es la mitocondria. Esto significa que una secuencia proteica representa un codón (tres nucleótidos) por cada letra (un aminoácido). La Figura 4.8 muestra todas las combinaciones que puede tener un codón y el aminoácido que va a ser agregado a la secuencia proteica. El codón de inicio generalmente es AUG, el cuál representa una Metionina. El codón final puede ser UAA, UAG o UGA e indican el final de la secuencia, es decir termina la traducción. Para leer la secuencia se utiliza el sistema de lectura de seis cuadros. Como el ADN se divide en codones de tres letras, se tienen tres cuadros distintivos de lectura. La hebra de ADN es doble por lo que se tiene dos hebras anti-paralelas. Como cada hebra tiene su propia secuencia se tienen en total seis cuadros de lectura. Expaty muestra el resultado de la traducción con este sistema[TutorialsTutorials]

RNA codon table

1st position	2nd position				3rd position
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr stop stop	Cys Cys stop Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Amino Acids

Ala: Alanine	Gln: Glutamine	Leu: Leucine	Ser: Serine
Arg: Arginine	Glu: Glutamic acid	Lys: Lysine	Thr: Threonine
Asn: Asparagine	Gly: Glycine	Met: Methionine	Trp: Tryptophane
Asp: Aspartic acid	His: Histidine	Phe: Phenylalanine	Tyr: Tyrosine
Cys: Cysteine	Ile: Isoleucine	Pro: Proline	Val: Valine

Figura 4.8: Tabla que contiene el código para poder leer los diferentes codones y los aminoácidos que representan. Se muestran los códigos utilizados para identificar los diferentes tipos de aminoácidos. Extraída de: [TutorialsTutorials]

4.4. Árboles filogenéticos para el análisis evolutivo de las proteínas

Los árboles filogenéticos son una herramienta utilizada en la biología y la bioinformática, se pueden usar para comprender la evolución de las proteínas y sus relaciones. Los árboles filogenéticos permiten analizar la historia evolutiva de las secuencias proteicas, revelar similitudes y diferencias genéticas. Estos árboles se conforman por nodos y ramas. Los nodos representan los puntos de bifurcación del grafo y las ramas son las líneas de conexión entre los nodos. Existen dos tipos de árboles filogenéticos, enraizado y sin enraizar. El árbol sin enraizar muestra de mejor manera las relaciones evolutivas de las proteínas. Este árbol tiene un nodo raíz, el cual representa el punto del que surgen todas las proteínas. [LageLage2013] Esto muestra la evolución de proteínas, implica cambios en las secuencias de aminoácidos a lo largo del tiempo a partir de un ancestro común. Gracias a esta evolución se forma la diversidad de funciones y estructuras proteicas. Los procesos que generan los cambios entre las secuencias de aminoácidos de las proteínas son mutaciones, selección natural y eventos de duplicación genética. [NelsonNelson2017]

El resultado de un árbol filogenético muestra una hipótesis de los posibles hechos evolutivos, no es un diagrama definitivo. El patrón que se forma a partir de los nodos muestra la forma en la que los organismos evolucionaron. La forma en la que se leen los árboles para entender que tan relacionados están dos puntos es ver su ancestro en común. Mientras menos cercano esté estos estarán menos relacionados y mientras más cercanos estarán más relacionados. Por lo que cada uno de los nodos representan un evento que creó divergencia entre los organismos. Cuando a partir de un nodo surgen varias especies u organismos de interés se genera una Politomía. Esto significa que más de dos organismos surgieron de un evento de divergencia, por lo que no se tiene suficiente información para determinar el orden de las ramas.[AcademyAcademy]

El método que se utiliza para analizar las proteínas es el de unión de vecinos. Este método busca parejas taxonómicas que minimicen la cantidad de ramas del árbol. Se ha comprobado que es el método más efectivo cuando se hacen simulaciones computacionales para obtener la topología correcta. La idea es siempre tener una pareja que surja de cada nodo, de esta forma hay un organismo asignado en la posición 1 y 2. Cuando esto no ocurre es cuando sucede la Politomía. La cantidad de formas que puede tomar una pareja en este método se calcula por $N(N - 1)/2$. Para escoger la forma en la que se colocarán se busca el valor más pequeño de la suma de la longitud de las ramas. Este proceso se repite hasta tener $N-3$ ramas. Este método muestra las especies de interés y la longitud de cada rama como se muestra en la Figura 4.9. Esta figura representa como se miran los nodos a partir de los cuales surgen las especies de interés.[Saitou NeiSaitou Nei1987]

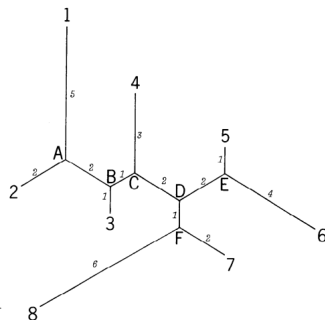


Figura 4.9: Ejemplo de un árbol filogenético generado por el método de unión de vecinos. Las letras muestran los nodos internos, los números grandes las especies de interés y los pequeños la distancia de las ramas.

Extraída de: [Saitou NeiSaitou Nei1987]

4.5. Alineación de proteínas y las herramientas bioinformáticas

La alineación de proteínas es una tarea computacional bastante compleja, esto se aplica para modelados, identificación de grupos funcionales, análisis filogenético y búsqueda de secuencias. Existe la alineación por pares la cuál tiene algoritmos simples y manejables que usan la programación dinámica. La aplicación de estos métodos es costosa computacionalmente y poco factible para números grandes de secuencias. Se han desarrollado algoritmos aproximados para alineaciones de secuencias múltiples como la técnica de alineación progresiva. La forma en la que estos funcionan es que ensamblan una alineación múltiple realizando una serie de alineaciones por pares de secuencias o grupos prealineados. Esto no garantizan una solución óptima y no corrige los errores cometidos en cada paso de alineación por pares. Para corregir o minimizar estos errores, se utilizan dos técnicas: el refinamiento iterativo y la puntuación de consistencia. El refinamiento iterativo se aplica después del ensamblaje progresivo de una alineación. La puntuación de consistencia en la alineación progresiva se ha utilizado para mejorar la calidad de la alineación.[PeiPei2008]

La alineación múltiple de proteínas Clustal es un método progresivo. Este mejora la sensibilidad de los resultados por peso de secuencias, penalizaciones de brechas específicas de posición y elección de matriz de peso. Esto lo hace con una matriz de sustitución, generalmente BLOSUM. Esta matriz califica a las secuencias según su similitud. Inicia el proceso alineando a las secuencias más parecidas entre ellas, luego va agregando secuencias según la alineación. Le da más importancia a las secuencias similares ya que se ponderan. El método es progresivo, esto quiere decir que se dividen las secuencias en clústers y mediante estos se van alineando los aminoácidos. No es tan exacto ya que puede producir algún error que debe corregirse a mano, en caso se desee. Los error se producen por secuencias divergentes, como lo son las proteínas desordenadas.[PeiPei2008]

Jalview es una plataforma de software para la visualización, edición y análisis de alineamientos múltiples de secuencias biológicas. También permite la manipulación de estructuras de proteínas y ácidos nucleicos. La función de Jalview con respecto a el método de alineación Clustal es la visualización y manipulación de alineamientos ClustalW o Clustal Omega. Jalview tiene como una de sus características la integración de este algoritmo por lo que puede correrse directamente desde la plataforma. La alineación que se realiza tiene el mismo enfoque progresivo mencionado anteriormente, a partir de esto se pueden generar árboles filogenéticos.[Clamp, Cuff BartonClamp .1998]

Como ya se sabe este método no es infalible, especialmente cuando se trabaja con proteínas divergentes puede causar problemas. Para este problema Jalview le permite al usuario editar la alineación de los aminoácidos corriendo y eliminando posiciones vacías. Esto permite correcciones que el investigador puede haber observado que le serán útiles para su estudio. Cuando se observa el alineamiento Jalview tiene como una de sus herramientas la coloración Clustal de las secuencias. Esto hace más visible las secciones alineadas así como las regiones conservadas que pueden tener las secuencias.[Clamp, Cuff BartonClamp .1998] Esta coloración se muestra en la Tabla 13.8 como referencia para analizar resultados.

BLAST es una herramienta utilizada en el área de bioinformática para encontrar secuencias proteicas similares. Su nombre significa Basic local alignment search tool y fue inventada en 1990. El objetivo de una búsqueda BLAST es realizar una alineación por pares de la secuencia y así obtener secuencias con similitudes de otras bases de datos. Las aplicaciones de este software incluyen la identificación de especies, mapeo de ADN y búsqueda de secuencias relacionadas. Hay cinco tipos de programas BLAST y cada uno es específico al tipo de secuencia que se quiere estudiar y el resultado que se quiere. Existe un programa para proteínas, nucleótidos, nucleótidos traducidos con sus diferentes resultados, estos se muestran en la Figura 4.10. El BLAST que se usa para analizar proteínas y buscar secuencias proteicas parecidas es BLASTP. Este programa necesita como entrada una secuencia de proteína. Esta es la secuencia que queremos alinear y buscar similares dentro de las bases de datos que se ofrecen. Al realizar la búsqueda cada secuencia se califica según la similitud

que tienen con la secuencia original. Mientras más alta sea la puntuación más similar es la secuencia, estos resultados pueden descargarse en archivo de texto y también puede accederse a las secuencias desde la página web.[Donkor, Dayie AdikuDonkor .2014]

Program	Query sequence type	Target sequence type
BLASTP	Protein	Protein
BLASTN	Nucleotide	Nucleotide
BLASTX	Nucleotide (translated)	Protein
TBLASTN	Protein	Nucleotide (translated)
TBLASTX	Nucleotide (translated)	Nucleotide (translated)

Source: <http://www.ncbi.nlm.nih.gov/blast>

Figura 4.10: Diferentes programas del algoritmo BLAST según el tipo de secuencia que se va a trabajar y el resultado que quiere obtenerse.

Extraída de: [Donkor, Dayie AdikuDonkor .2014]

Chimera es una plataforma programada en Python para el análisis, visualización y selección de secuencias proteicas y de nucleótidos. Está conformado por un centro y extensiones, estas extensiones permiten a los investigadores tener todo lo necesario y extender su estudio hasta donde lo requieran. Las principales utilidades de esta plataforma son la visualización 3D, alineación de estructuras, análisis estructural, manipulación y edición de estructuras, generación de imágenes y secuenciación. Las herramientas que serán útiles para el estudio de las glicoproteínas es la visualización 3D y alineación estructural.[Pettersen .Pettersen .2004]

La parte de alineación estructural se basa en la herramienta de MatchMaker que incluye la plataforma. Esta herramienta crea alineaciones de pares y a partir de esto se sobreponen las estructuras de las secuencias. Para poder iniciar el proceso se debe de seleccionar que secuencia va a ser usada como referencia, generalmente se utiliza la más grande y/u ordenada para que sea un buen modelo. La calidad de la alineación se califica con el valor RMSD (Desviación Cuadrática Media). Este valor representa la diferencia promedio que hay entre las posiciones de los átomos correspondientes en las dos estructuras superpuestas. Mientras más bajo es el valor significa que las estructurales se parecen estructuralmente. Un RMSD alto demuestra estructuras que difieren significativamente.[Meng, Pettersen, Couch, Huang FerrinMeng .2006]

4.6. Regiones conservadas

Al analizar grupos de proteínas se pueden buscar regiones conservadas entre ellas. Estas regiones pueden encontrarse luego de hacer una alineación de las secuencias. Cuando se analiza una proteína en un algoritmo como BLAST se encuentran proteínas con similitud de más de 25 %, sin embargo para algunos estudios es importante conocer que regiones exactamente son las similares. Esto nos provee de información para entender la historia evolutiva de las proteínas, o cualquier secuencia que se esté analizando. Así como también ayuda a identificar que regiones pueden llegar a ser esenciales para la función que un grupo de proteínas está desempeñando. Cuando no existen regiones conservadas significa que las secuencias evolucionaron a tal punto que no puede identificarse a donde pertenecen. También puede significar que las proteínas no comparten ninguna relación.[PietrokovskiPietrokovski1996]

Las regiones conservadas son cruciales para la detección y estudio de proteínas. Estas se vuelven conservadas por su importancia biológica ya que representan un rol muy importante para la función de la proteína. Cuando tenemos un grupo de proteínas, en este caso glicoproteínas del grupo Rh encontradas en la sangre humana, con regiones conservadas quiere decir que estas regiones son importantes para su función. Por lo que es probable que si esta región se viera afectada por evolución o tratamientos la proteína podría cambiar la función que tiene en el cuerpo. Desde el punto de vista bioquímico se sabe que generalmente estas regiones tienen sitios activos o residuos críticos para la

secuencia. También pueden ser parte importante de la unión con sustratos, cofactores, ligandos, etc. Conocer estas regiones permite a los investigadores en el área de salud poder diseñar fármacos que se enfoquen en la región conservada que está causando daño o bienestar.[PietrokovskiPietrokovski1996]

Para poder identificar estas regiones pueden utilizarse plataformas como Jalview, al terminar la alineación puede utilizarse una utilidad que permite colorear las secuencias. Esta coloración puede hacerse de forma clustal y utiliza diferentes colores según el aminoácido que encuentre, los colores pueden observarse en la Tabla 13.8. Los aminoácidos que son coloreados son los que se encuentran en regiones conservadas. En la tabla se puede leer cuál es el porcentaje mínimo de presencia requerido de cada aminoácido para ser considerado conservado. Cuando este porcentaje se cumple el aminoácido se colorea y de esta forma pueden irse alineando las secuencias para identificar que regiones conservadas existen.[Clamp, Cuff BartonClamp .1998] A parte de esta conservación también deben compararse los resultados con los árboles filogenéticos para ver si la evolución de las proteínas influyó de alguna manera con las regiones conservadas. Por ejemplo una proteína que evolucionó en una rama diferente que tanta conservación mantuvo.[PietrokovskiPietrokovski1996]

4.7. Regiones desordenadas

Algunas proteínas contienen dentro de su secuencia regiones desordenadas. Estas proteínas se llaman PID, proteínas intrínsecamente desordenadas. Christian Anfinsen en 1960 dijo que la estructura 3D de una secuencia era única y determinada completamente por sus aminoácidos. Los polipéptidos que no cumplían con este postulado son considerados PID. El estudio de estas regiones ha sido cada vez más común, el nuevo paradigma de las PID dice que aunque las regiones no tengan una estructura 3D definida aún cumplen con su función. Casi un 40% de los proteomas de eucariotas tienen en sus secuencias regiones desordenadas. Esto ha abierto puertas a estudios que han dado como conclusión que estas regiones pueden estar ligadas con enfermedades humanas como el cáncer y problemas neurológicos.[M.M.2006]

Estas regiones son partes de la secuencia que no tienen suficientes aminoácidos hidrofóbicos para mediar el modelado de la estructura. Los aminoácidos comúnmente encontrados en desorden son entonces la Glicina, Alanina, Valina, Leucina, Prolina, Isoleucina y Metionina. También tienen una carga más fuerte de aminoácidos polares como la Serina, Treonina, Cisteína, Asparagina, Glutamina, Lisina, Arginina e Histidina. Esto no quiere decir que la estructura puede tomar todas las conformaciones posibles. Las proteínas con alta carga y poco hidrofóbicas generalmente tienen una conformación completamente extendida. Una conformación compacta se forma cuando hay un balance entre la carga y la parte hidrofóbica.[M.M.2006]

Iupred Anchor2 es una herramienta bioinformática utilizada para la predicción de desorden. La predicción la realiza realizando cálculos aproximados de la energía de la secuencia. Utiliza modelos de aprendizaje y estadísticos para calcular estos niveles. El nivel de energía se basa en las propiedades que tienen los aminoácidos presentes. Hay un límite predefinido del número en el que debería estar esta puntuación para ser considerado ordenado o desordenado. El resultado de esta plataforma es una gráfica que incluye todas las posiciones y en qué umbral se encuentra cada aminoácido. Las regiones que se encuentren por encima de 0.5 son consideradas desordenadas.[DosztányiDosztányi2018]

4.7.1. Rol del desorden en proteínas

Entre las ventajas que tienen estas regiones se encuentra la mediación con los péptidos lo cual da lugar a la unión con múltiples secuencias. Su flexibilidad facilita la regulación de la función de las proteínas por modificaciones post traduccionales que ocurren dentro de la región desordenada. Regula la vida media de las proteínas que fueron marcadas para degradación por el proteasoma. La

flexibilidad les permite unirse a diferentes secuencias, esto hace que puedan señalar y regular más fácilmente.[M.M.2006]

Una de las proteínas más estudiadas pero menos conocidas es la p53. Esta es un ejemplo perfecto cuando queremos hablar del rol de las regiones desordenadas en una proteína. El papel que tiene en el cuerpo es crucial, actúa en el ciclo celular y apoptosis. Se considera uno de los supresores de tumores por estas funciones, así como también reparar el ADN. Las regiones desordenadas encontradas en esta proteínas fueron determinadas como la clave para que esta pueda realizar sus importantes funciones. Si estas regiones se vieran afectadas la proteína perdería su capacidad de interacción y no tendría sus capacidades reguladoras que la hacen capaz de reparar ADN. La presencia de esta proteína mutada ha sido ligada a la presencia de tumores en los individuos de estudio. Esto inició la investigación que descubrió que al tener presencia de desregulaciones en las proteínas desordenadas pueden presentarse enfermedades neurodegenerativas y metabólicas.[UverskyUversky2016]

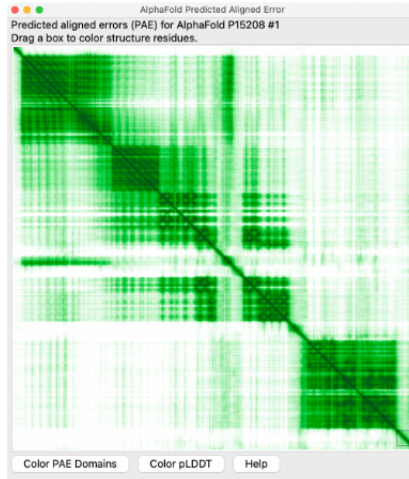
4.8. Técnicas para modelado de proteínas

Los modelados de las estructuras 3D de las proteínas proveen de bastante información al investigador, se analiza su función en un nivel molecular. Desde el 2018 la cantidad de modelados de proteínas en plataformas como PDB ha ido en aumento por su extensa cantidad de aplicaciones y beneficios que trae a esta área de estudio. El modelaje se realiza con técnicas por homología, esto quiere decir que se construye a partir de estructuras conocidas así como aproximaciones para llegar a la mejor forma. El primer algoritmo para generar estos modelos fue SWISS-MODEL, este utiliza la secuencia de aminoácidos como punto de inicio. Desde su creación este ha sido mejorado y también se han generado mejores plataformas para realizar estos modelos.[Waterhouse .Waterhouse .2018]

AlphaFold2 es una de las plataformas que utiliza el modelaje por homología para presentar sus estructuras. Este fue creado por DeepMind, tuvo su lanzamiento oficial a la comunidad científica cuando liberó los modelos del proteoma humano. La plataforma utiliza redes neuronales y técnicas de aprendizaje para predecir la estructura 3D de proteínas con una precisión alta. Se utiliza en biotecnología, farmacia, biomedicina, bioinformática, etc. Aunque no todos los modelos generados son aceptados al 100 %, esta es la mejor herramienta que se tiene en la actualidad para realizar estos modelos. Hasta hace unos años el uso de una plataforma como esta era solo un sueño, por lo que las investigaciones han avanzado más rápidamente.[Jones ThorntonJones Thornton2022]

Cuando se realiza un modelado en AlphaFold2 solo se necesita, en el caso de las proteínas, su secuencia de aminoácidos. El proceso suele ser algo tardado por la cantidad de análisis y lo complejo que es el proceso. Cuando termina se descarga automáticamente un archivo comprimido con las gráficas y archivo utilizado para la visualización de la secuencia. Una de las gráficas generadas sirve para medir la distancia entre los aminoácidos de la secuencia, este es el gráfico de PAE. Esta distancia se obtiene a partir de la posición del primer residuo modelado por AlphaFold2 según la posición real alineada del segundo residuo mediante la prueba de diferencia de distancia local. Los valores pueden estar entre 0-35 Angstroms y se muestra como un mapa de calor. En la Figura 4.11 se muestra un ejemplo de cómo debería verse un diagrama PAE. Los gráficos de AlphaFold2 muestran color azul la distancia es más cercana a cero y rojo cuando es mayor. En los gráficos se espera una línea azul en la diagonal ya que esta es la interacción de cada aminoácido consigo mismo.[PhDPhD2022]

AlphaFold2 genera varios modelados para compararlos y poder determinar cuál es el mejor. Para esto se genera una gráfica pLDDT. Esta gráfica es la que se utiliza para medir el nivel de calidad de la estructura. Muestra cada posición de los diferentes modelos evaluados según el criterio pLDDT. Este indica que si una posición tiene <90 es un modelado con alto grado de confianza. Entre 70 y 90 son buenos modelos, 50 y 70 tienen un grado de confianza bajo y se utilizan con precaución. Los modelos más bajos, con <50 generalmente tienen apariencia de cinta y no deben ser interpretados. La última gráfica que se utiliza es la de la cobertura de secuencia. Esta muestra la cantidad de



Residue-residue alignment confidence (PAE) plot [shown in ChimeraX](#).

Figura 4.11: Ejemplo de cómo se visualiza un gráfico PAE resultante de AlphaFold. Este gráfico muestra una estructura con varias curvas en sus estructuras. Esto se debe a las áreas verdes que se encuentran alrededor de la diagonal. En este caso el color verde indica una distancia corta entre los aminoácidos. Extraída de: [PhDPhD2022]

homólogos identificados y se colorean según su identidad. Este sirve para visualizar la cobertura de la secuencia que se tiene en la estructura.[PhDPhD2022]

Se genera un archivo PDB, este representa la secuencia de la proteína. Lleva su nombre por el Banco de datos de proteínas y contiene toda la información para visualizar el modelo 3D en cualquier plataforma que lo soporte. Este modelo será el que generamos y el que va a ser evaluado a partir de las gráficas mencionadas anteriormente. Una de las plataformas utilizadas para la visualización del modelo es Chimera. Para poder verlo solo debe de cargarse el archivo, pueden cargarse varias secuencias para visualizarlas y analizarlas de forma grupal. Esto es útil para este estudio para poder evaluar grupos de proteínas correspondientes a la misma hebra.[PeitschPeitsch1995]

4.9. Factor Rh

El factor Rhesus, comúnmente abreviado como Rh, es un componente sanguíneo de gran importancia en la medicina y la hematología. Se le llama así en referencia al macaco Rhesus, el cual fue de suma importancia en el descubrimiento de este factor.[GeographicGeographic] Fue investigado por primera vez en 1940 cuando una mujer dio a luz a un bebé fallecido. Esto ocurrió luego de que la mujer recibiera una donación de sangre por parte de su esposo y tuviera una mala respuesta. El sistema Rh se refiere a la presencia o ausencia del antígeno RhD en la superficie de los glóbulos rojos. Su identificación y estudio han sido fundamentales para comprender la inmunohematología y la genética de la sangre.[Avent ReidAvent Reid2000]

El antígeno RhD es una proteína integral de membrana con múltiples dominios. La presencia o ausencia de esta proteína determina si un individuo es Rh positivo o Rh negativo. La estructura del antígeno RhD es esencial para su interacción con anticuerpos y su función en el sistema inmunológico. Ahora se sabe que este sistema es de los más complejos cuando se habla de grupos sanguíneos, incluye aproximadamente 45 antígenos independientes. Este sistema junto con el ABO es el más

utilizado al momento de catalogar los tipos de sangre en donaciones de sangre, su compatibilidad, etc.[WesthoffWesthoff2007]

Su función principal es actuar como un antígeno, lo que significa que puede desencadenar una respuesta inmunológica si un individuo Rh negativo recibe sangre de un donante Rh positivo. Esto es un factor crítico en las transfusiones de sangre y en el embarazo, donde la incompatibilidad Rh puede tener consecuencias graves para el feto. Cuando una madre es Rh negativa y el feto es Rh positivo, existe un riesgo de incompatibilidad Rh que puede llevar a la enfermedad hemolítica del recién nacido. La medicina moderna ha desarrollado estrategias efectivas para prevenir y tratar la incompatibilidad Rh, como la administración de inmunoglobulina Rh(D) (anti-D) a las madres Rh negativas durante el embarazo y después del parto.[WesthoffWesthoff2007]

Las proteínas Rh tienen siempre el antígeno Rh pero solo se expresan en la superficie de los eritrocitos cuando una Glicoproteína asociada al factor Rh está presente. En estudios anteriores se encontró que el 40 % de las secuencias de las proteínas Rh y glicoproteínas Rh muestran homología. Esto quiere decir que tuvieron una relación ancestral por lo que son llamadas la familia de proteínas Rh. Las proteínas accesorias del Rh hacen referencia a otras glicoproteínas que están asociadas a la familia de proteínas Rh debido a su ausencia o deficiencia en los glóbulos rojos Rh nulo. La asociación de la familia de proteínas Rh y las proteínas accesorias del Rh se denomina el 'complejo Rh'. Su densidad es aproximadamente 170 000 daltons.[Avent ReidAvent Reid2000] El complejo Rh tienen como posible función el cotransporte de amonio con otros cationes. También se ha reportado que las glicoproteínas relacionadas al factor Rh tienen gran homología con MEP2, el cual es un sensor de amonio y transportador de levaduras.

Las complicaciones clínicas que pueden surgir relacionadas al factor Rh resultan de la destrucción de los glóbulos rojos debido a la interacción de un aloanticuerpo con los glóbulos rojos que llevan el antígeno correspondiente. El antígeno D es altamente inmunogénico e induce una respuesta inmunológica en el 80 % de las personas con sangre D-negativa cuando se les transfunde con 200 mL de sangre D-positiva. Por esta razón, en la mayoría de los países se realiza rutinariamente la tipificación del antígeno D en cada donante de sangre y receptor de transfusiones. Los pacientes D-negativos reciben productos de glóbulos rojos D-negativos. Como resultado, las complicaciones clínicas debidas a transfusiones incompatibles son poco frecuentes. [Avent ReidAvent Reid2000]

A pesar del uso de terapia inmunosupresora con profilaxis de inmunoglobulina anti-D, la aloinmunización por D durante el embarazo todavía ocurre. [Avent ReidAvent Reid2000] Los aloanticuerpos son producidos cuando el cuerpo se encuentra expuesto a glóbulos rojos que no pertenecen a su grupo sanguíneo. Es decir que se produce una aloinmunización. La producción de estos aloanticuerpos depende de la edad, sexo, cantidad de transfusiones de sangre, etc.[Schonewille, Van De Watering, Loomans BrandSchoonhoven2000] Así como existen problemas inmunológicos en el embarazo también hay otras condiciones relacionadas al factor Rh. Entre estos está el factor Rh nulo y la Leucemia mieloide.

4.10. Tipos de sangre

Existen glicoproteínas específicas al grupo de sangre presente en los eritrocitos humanos y puede analizarse su estructura. El objetivo de este análisis es encontrar las secuencias conservadas entre las proteínas para estudiar si existe alguna relación entre estas y el tipo de sangre en donde se encontró la secuencia. El tipo de sangre de cada persona se determina por los genes de sus padres. Existen cuatro categorías del grupo ABO para clasificar cada tipo los cuales son: A, B, O , AB. De estos grupos surgen 8 tipos de sangre que se diferencian dentro del mismo grupo por su factor Rh. La sangre se compone de glóbulos rojos y blancos, plasma y plaquetas. Los antígenos que se encuentran en la superficie de los glóbulos rojos son los que determinan el grupo al que pertenece la sangre. Estos antígenos pueden ser proteínas o azúcares, se dividen entre antígenos para el grupo ABO y para el factor Rh. Solo algunos glóbulos rojos tienen los antígenos RhD, si lo tienen la sangre es Rh

positiva y si no será Rh negativa. [GoodwinGoodwin2021] Esta estructura de cada grupo sanguíneo se presenta en la Figura 4.12.

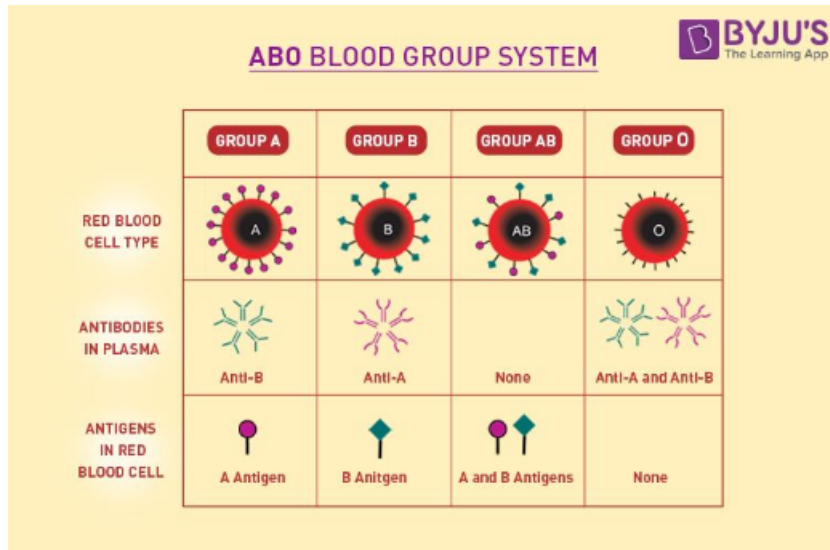


Figura 4.12: Esquema visual de la composición de los grupos sanguíneos según las glicoproteínas presentes. Extraída de: [ByjuByju]

4.10.1. Conformación de cada grupo

La conformación de cada grupo sanguíneo cambia según los anticuerpos presentes. Para el grupo A los glóbulos rojos tienen antígeno A y el plasma anticuerpos anti-B. El grupo B tiene antígeno B y anticuerpos anti-A. La sangre AB tiene antígenos A y B pero el plasma no tiene ningún anticuerpo, por lo que pueden recibir cualquier tipo de sangre del grupo ABO. El tipo de sangre del grupo O no tiene ningún antígeno en su superficie, más tiene anticuerpos A y B. Por esta razón cualquier tipo de sangre del grupo ABO puede recibir sangre O. [GoodwinGoodwin2021] El factor Rhesus, nuestro interés en el estudio, es una proteína que se encuentra en la superficie de los glóbulos rojos. Es importante conocerlo al momento de hacer donaciones de sangre ya que las personas que tienen un factor Rh positivo no producen los anticuerpos anti Rh. Por esto una persona Rh negativa solo puede recibir sangre de otro factor Rhesus negativo. Este factor también puede generar complicaciones en los embarazos de las madres Rh negativo que están esperando un bebé con factor Rh positivo. El cuerpo de la madre puede generar anticuerpos D que dificultan el desarrollo del bebé. [DonationDonation]

4.11. Enfermedades relacionadas a la necesidad del acceso a sangre

La anemia es una enfermedad asociada con la mala composición de la sangre en las personas. Se da cuando hay un conteo bajo de glóbulos rojos aptos para transportar oxígeno al cuerpo. Es una enfermedad que puede estar presente por un período corto o largo de tiempo dependiendo el tipo. Los tipos principales de anemia son aplásica, de células falciformes, por deficiencia de hierro, por deficiencia de vitaminas y talasemia. La anemia aplásica se presenta cuando el cuerpo es incapaz de producir la cantidad de células sanguíneas necesarias. Sus tratamientos incluyen transfusiones sanguíneas o de células madre. [ClinicClinic2021]

La anemia falciforme afecta la estructura de los glóbulos rojos afectando su forma para llevar oxígeno. Este tipo de anemia aún no tiene un tratamiento efectivo por lo que solo se tienen procedimientos para manejar sus síntomas. Cuando su causa es la deficiencia de hierro no se produce la cantidad necesaria de hemoglobina. Su tratamiento principal es tomar suplementos de hierro. En la anemia por deficiencia de vitaminas ocurre lo mismo, solo que por niveles bajos de B12 y folato. Esto hace que los glóbulos rojos sean muy grandes. La talasemia es un tipo de anemia hereditaria, el cuerpo va a tener menos hemoglobina de lo necesario. Para poder tener un transporte normal de oxígeno en la sangre se deben realizar transfusiones sanguíneas de forma regular [ClinicClinic20221]

Según la OMS la salud mundial es un problema grave en niños y mujeres embarazadas. El estimado mundial que se tiene acerca del porcentaje de la población que sufre de algún tipo de esta enfermedad es el siguiente. El 42 % de niños menores a 5 años y el 40 % de mujeres embarazadas son víctimas de esta enfermedad. Esto representa una cantidad grande de personas que necesitaran algún tipo de tratamiento en el transcurso de su vida. Como se mencionó anteriormente el tratamiento viable para los tipos de anemia más graves son las transfusiones sanguíneas. Es por eso que es vital que se tengan centros con acceso constante a sangre para que cualquier paciente pueda llevar a cabo su tratamiento. [de la Saludde la Salud]

Para muchas enfermedades asociadas con la sangre como la anemia el mejor tratamiento que se tiene actualmente es la transfusión de sangre. Así como esta enfermedad hay muchas más condiciones en las que la opción más viable es tener un donador de sangre. Este proceso es largo y tedioso ya que para que se puedan realizar las donaciones el donador debe ser compatible con el paciente. La transfusión de sangre conlleva ciertas reglas que indican qué tipo de sangre pueden recibir los pacientes. Esto se debe a los glóbulos blancos que producen anticuerpos los cuales identifican cualquier objeto o sustancia extraña. Hay un tipo de sangre, O-, la cual es aceptada por todos los grupos sanguíneos. Las personas que tienen este tipo de sangre son la minoría en la población, conformando un 7 % de la misma. Esta pequeña cantidad de posibles donadores hace difícil que los hospitales puedan mantener en disponibilidad este tipo de sangre. [ClinicClinic20222]

5.1. Datos iniciales

Los datos que fueron utilizados para el análisis de glicoproteínas se produjeron a partir de la publicación *Blood group typing from whole-genome sequencing data*. [JJ2020] Para conseguir los datos de esta publicación secuenciaron el genoma completo de 79 individuos de Asia Central para estudiar el gen HLA-DRB1 y 9 antígenos de grupos sanguíneos. [JJ2020] A partir del estudio de estos genomas se obtuvo uno de los genes asociados al factor Rh de la sangre humana. [60056005] La glicoproteína que codifica este gen está asociada al transporte de amonio y dióxido de carbono en la membrana. Es específica a los eritrocitos e interactúa con los antígenos del grupo Rh. Se realizó una conversión del gen 6005 [60056005] a proteína para poder encontrar proteínas similares y así estudiar los sitios conservados. Los datos utilizados fueron las secuencias de ADN de este gen, estos se muestran en las figuras 6.1 y 6.8.

5.2. Procesamiento Expasy

Para poder convertir las secuencias del gen específico a proteínas correspondientes al grupo sanguíneo Rh se utilizó la plataforma de Expasy. Los parámetros que se utilizaron para el análisis se muestran en la Figura 13.1. El proceso se realizó dos veces utilizando cada una de las secuencias mencionadas en la sección anterior. Para el análisis de los resultados se utiliza el sistema de lectura de 6 frames de ADN que maneja el sistema de conversión. El resultado del proceso fueron 12 proteínas en total, estas estaban divididas según la hebra de ADN en la que se encuentran, su frame y gen. Las figuras 6.2 a la 6.7 y 6.9 a la 6.14 muestran el formato en el cual se leen las secuencias de las proteínas que se estarán utilizando en el estudio.

5.3. BLAST de las proteínas resultantes

Para encontrar y formar la base de datos de las glicoproteínas que se analizan en el estudio se realizó una búsqueda en la plataforma BLAST, específicamente un BLASTP. Los parámetros utilizados se muestran en la Figura 13.2, la búsqueda se realizó para encontrar proteínas específicas

de *Homo sapiens*. Este cambio en los parámetros se realizó ya que el estudio busca analizar las secuencias conservadas de las glicoproteínas en humanos. Cada BLAST de los 12 que se realizaron mostraron las proteínas que se relacionaban a cada uno de los frames de interés. Las figuras 6.15 a 6.20 y 6.21 a 6.26 muestran el formato que contiene todas las proteínas que se relacionan a cada uno de nuestros segmentos de interés.[NIHNIH] De todas las proteínas se seleccionaron las primeras 15 de cada documento las cuales son las que más parecidas son a la secuencia ingresada. Esto permite tener un primer filtro de datos para poder armar una base de datos de glicoproteínas que sean adecuadas para el estudio.

5.4. Filtración de proteínas en base de datos Uniprot

Al realizar el primer análisis de los datos resultantes se pudo hacer evidente la repetición de datos que ocurría. Por la forma de lectura del sistema de seis cuadros los datos que resultaron de los frames de cada hebra de ADN tenían duplicados entre ellos. Para poder analizar de la mejor manera los datos se hizo una búsqueda en Uniprot de las primeras 15 proteínas de cada uno de los BLAST para encontrar sus respectivos códigos. Se utilizaron solamente las proteínas que se encontraban en esta base de datos para asegurar la calidad y procedencia de los mismos.[ConsortiumConsortium2018]

5.5. Creación de dos grupos de proteínas según su hebra

Como existían repeticiones se crearon dos listados de códigos Uniprot, se separaron en dos grupos: hebra 3'5 y 5'3. Se realizó una tabla de Excel y se filtró para eliminar a los duplicados. Con esto se logró tener una base de datos compacta de todas las proteínas asociadas a las glicoproteínas del grupo sanguíneo Rh. A partir de este momento las proteínas no se separan por los distintos frames y se refieren solamente a la hebra en la que se encuentran. La Tabla 6.1 hace referencia a los códigos de Uniprot de las glicoproteínas que se utilizaron para el posterior análisis.

5.6. Árbol filogenético

La herramienta Jalview se utilizó para generar los dos árboles filogenéticos que van a ser utilizados en el estudio. Dentro de la aplicación con las secuencias cargadas se siguen los siguientes pasos para obtener el árbol: Calcular -> calcular árbol o ACP -> unir vecinos -> Ver -> Mostrar distancias. Los arboles se encuentran en la Figura 6.27 para la hebra 3'5 y en 6.28 para la hebra 5'3.

5.7. Análisis de regiones desordenadas

Las glicoproteínas que se utilizaron no se encontraban en la base de datos SwissProt por lo que son proteínas no curadas. Al momento de empezar a procesarlas se notó que eran proteínas intrínsecamente desordenadas por la alta cantidad de aparentes regiones desordenadas presentes. Estas regiones deben ser analizadas para saber qué implica esto en las glicoproteínas. Para poder determinar si las proteínas tienen regiones desordenadas se analiza el puntaje que cada posición tiene en la gráfica de análisis. Un puntaje arriba de 0.5 se considera desorden, mientras que todas las posiciones con menos de 0.5 son ordenadas.

5.7.1. IUPred

El primer análisis que se realiza es en la plataforma de IUPred2 Anchor siguiendo los parámetros que se muestran en la Figura 13.3. Se realizó el análisis de las proteínas que fueron modeladas para poder completar el estudio de las mismas. Cada gráfica resultante fue adjunta como una imagen al análisis individual de cada proteína, el primer ejemplo que se presenta es la gráfica D de la Figura 6.34.

5.7.2. Algoritmo en R

Se realizó un algoritmo en R que predice las secciones desordenadas más a profundidad que el gráfico IUPred y las muestra en forma de cuatro distintas gráficas. En la primera gráfica se usan los datos de IUPred para realizar una gráfica que muestra las secciones desordenadas de color rojo y las ordenadas de color azul. La segunda gráfica es un análisis más profundo de las secciones desordenadas, se analiza que porcentaje de los residuos según los aminoácidos presentes en las secuencias se encuentran en una sección desordenada. Estos resultados se muestran en forma de gráfico de barras para facilitar su visualización. En la tercera gráfica se comparan los resultados con los datos pLDDT graficando ambos sets de datos. Al comparar ambos datos se puede observar si hay correlación entre la calidad del modelado y las secciones desordenadas. Esta correlación se muestra en la cuarta gráfica, el programa completo se encuentra en la Figura 13.9. El primer resultado de las gráficas resultantes del programa de una de las proteínas se observa en la Figura 6.35.

5.8. Alineación de proteínas en JalView

El análisis y detección de los sitios conservados de las proteínas se hizo en JalView. Para realizar este análisis se abrió el archivo con todas las secuencias Fasta de ambos grupos a analizar. Con las secuencias cargadas se siguieron los siguientes pasos para realizar la alineación inicial: Servicio Web -> Alignment -> Clustal -> por defecto. Luego de completar estos pasos se trabajó sobre la nueva ventana emergente. Para hacer evidente las secciones conservadas entre las distintas proteínas se colorearon con los siguientes pasos: Color -> Clustal. La alineación al ser de proteínas desordenadas se tenía algunos problemas que tuvieron que ajustarse manualmente. Para poder eliminar o agregar espacios vacíos a las secuencias se utiliza el comando Fn + F2 para editar posición por posición. De esta forma fueron ajustándose las regiones conservadas así como encontrando nuevas.

5.9. Modelado en AlphaFold

La plataforma AlphaFold fue utilizada para realizar los modelados de las proteínas de ambos grupos del estudio. Esta plataforma se encuentra en un Jupyter Notebook y como input se coloca la secuencia Fasta de la proteína a modelar. El resultado es un archivo en formato zip. Los archivos de esta carpeta que nos interesan son un archivo pdb, el cuál se utilizó para visualizar la estructura en Chimera, y tres gráficas. El programa genera cinco archivos pdb y se ordenan según la calidad del modelado. Para el estudio solo se utilizó el primero, es decir el mejor modelado. La primera gráfica lleva por nombre PAE. La segunda gráfica es la pLDDT que muestra el score que califica la aceptación del modelo, mientras más alto mejor score. La tercera gráfica que se analiza de AlphaFold es la cobertura de la secuencia en el modelado. Todas estas gráficas se adjuntaron al modelo de las proteínas de cada grupo de estudio. El primer ejemplo del formato de este análisis puede observarse en la Figura 6.34.

5.10. Análisis y alineación de modelados en Chimera

Chimera es una herramienta para la visualización y análisis de estructuras en 3D. Las proteínas de cada grupo fueron cargadas en la plataforma, se realizaron dos sesiones: para la hebra 3'5 y 5'3. El archivo utilizado fue el pdb rank 1 del archivo zip resultado de AlphaFold. Esto se realizó para tener acceso a todas las proteínas del mismo grupo al momento de querer alinearlas. La imagen de cada modelado individual se adjuntó como la figura A de cada proteína, como puede observarse en la Figura 6.34. Se realizaron dos alineaciones por grupo, la primera tomando en cuenta 5 de las proteínas de la hebra a estudiar. La segunda se realizó con la proteína más desordenada y las más ordenada de cada grupo. La alineación se realizó siguiendo los siguientes pasos: Tools -> Structure Comparison -> MatchMaker. Dentro de esta nueva ventana se selecciona la estructura de referencia, esta será la más ordenada de cada grupo. Las estructuras a alinear serán todas las demás del grupo. Cuando todas las proteínas estuvieron seleccionadas se usa el botón Apply y OK para obtener las estructuras alineadas. Estas imágenes se encuentran en la sección 6.6 de resultados.

6.2. BLAST de las 30 proteínas mejores rankeadas de cada frame

En las siguientes figuras se presentan los resultados de la alineación BLAST de cada uno de los frames de lectura de ambas hebras. Se tienen en total 12 figuras, 6 por cada gen y 3 por cada hebra. Se pueden observar los códigos que mostró el algoritmo, estos identifican a las proteínas que más se asemejan a la secuencia que se está estudiando. Las proteínas están ordenadas según la similitud, es decir que la primera es la que más coincide con la búsqueda. Esto puede verse en la columna de similitud. El formato resultante de BLAST nos indica el nombre científico del organismo en el que se encuentra la proteína. En todos los casos será *Homo sapiens* ya que nos enfocamos en proteínas en la sangre humana. Estos resultados fueron utilizados como la base inicial de datos que se van a trabajar. Esta base de datos fue curada para obtener el set final de datos, este set se encuentra en la Tabla 6.1 con sus códigos de identificación de Uniprot.

```

RID: F1988TW013
Job Title:Gene135Frame1
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from MGS projects
Query #1: Query ID: 1c1|Query_38543 Length: 9984

Sequences producing significant alignments:

Description                               Scientific Name      Common Name      Taxid      Max Score      Total Query cover      E Value      Ident Len      Per. Acc.      Accession
unknown [Homo sapiens]                    Homo sapiens       human           9606       148 148 1%      1e-38 61.48 180      AAL55749.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       144 431 3%      3e-38 70.87 121      BAC85289.1
HCG2038446 [Homo sapiens]                 Homo sapiens       human           9606       143 450 3%      2e-37 81.11 135      EAX11460.1
HCG2039009 [Homo sapiens]                 Homo sapiens       human           9606       140 396 3%      8e-37 70.09 120      EAW64637.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       139 460 4%      6e-36 74.00 137      BAC85304.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       136 396 3%      3e-35 70.10 127      BA015056.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       134 419 4%      1e-34 76.40 128      BAC85285.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       134 427 4%      3e-34 61.02 150      BAC85219.1
PRO1722 [Homo sapiens]                    Homo sapiens       human           9606       132 410 3%      6e-34 66.33 118      AAF69605.1
HCG1744006 [Homo sapiens]                 Homo sapiens       human           9606       132 498 5%      4e-33 72.63 171      EAW62471.1
HCG2039110 [Homo sapiens]                 Homo sapiens       human           9606       129 373 3%      5e-33 71.88 103      EAX06591.1
ubiquitin-conjugating enzyme E2D 4 (putative), isoform CRA_b... Homo sapiens       human           9606       129 444 3%      7e-33 69.57 128      EAW94179.1
HCG1995581 [Homo sapiens]                 Homo sapiens       human           9606       132 431 3%      2e-32 77.91 230      EAM57092.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       129 438 4%      3e-32 73.91 140      BAC85220.1
HCG1814203 [Homo sapiens]                 Homo sapiens       human           9606       129 665 5%      3e-32 63.72 162      EAW81162.1
similar to zinc finger protein 569, isoform CRA_a [Homo sapiens] Homo sapiens       human           9606       127 425 3%      3e-32 65.62 115      EAW56733.1
RefName: Full=Protein G[VQW]; AltName: Full=GVQW motif-containl... Homo sapiens       human           9606       129 380 3%      6e-32 70.21 195      QBN710.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       127 386 3%      7e-32 67.01 129      BAC04333.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       127 416 3%      9e-32 68.89 152      BAA91131.1
PRO2550 [Homo sapiens]                    Homo sapiens       human           9606       125 420 3%      2e-31 65.56 118      AAG35515.1
HCG1991981 [Homo sapiens]                 Homo sapiens       human           9606       125 423 3%      3e-31 72.83 136      EAX02144.1
ubiquitously transcribed tetratricopeptide repeat Y-linked... Homo sapiens       human           9606       122 397 3%      1e-30 71.08 84      ACP43292.1
hypothetical protein FLJ22709, isoform CRA_b [Homo sapiens] Homo sapiens       human           9606       122 397 3%      1e-30 68.82 119      EAW84578.1
PRO2852 [Homo sapiens]                    Homo sapiens       human           9606       124 355 2%      2e-30 70.00 169      AAG35505.1
HCG2033684 [Homo sapiens]                 Homo sapiens       human           9606       124 354 2%      2e-30 70.00 169      EAW58471.1
Solute carrier family 48 (heme transporter), member 1 [Homo... Homo sapiens       human           9606       127 376 2%      2e-30 70.41 239      AAH26344.2
zinc finger protein ENSP00000375192-like [Homo sapiens] Homo sapiens       human           9606       124 379 2%      2e-30 70.97 164      XP_047290968.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       123 358 3%      3e-30 52.94 156      BAC06323.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       124 377 2%      3e-30 68.75 164      BAC86222.1
unnamed protein product [Homo sapiens]    Homo sapiens       human           9606       123 419 5%      5e-30 67.01 179      BAC05191.1

```

Figura 6.15: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.
[NIHNIH]


```

RID: F19GDHJ013
Job Title:Gene135Frame2
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_25016 Length: 9988

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query cover	E Value	Per. Ident	Len	Acc.	Accession
unknown [Homo sapiens]	Homo sapiens	human	9606	135	135	0%	5e-34	87.84	180		AAL55749.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	131	464	3%	3e-33	67.96	140		BAC85220.1
hCG1818479 [Homo sapiens]	Homo sapiens	human	9606	124	344	2%	2e-31	73.40	98		EAW95069.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	123	518	5%	1e-30	67.59	125		BAC04924.1
hCG1744086 [Homo sapiens]	Homo sapiens	human	9606	125	441	4%	1e-30	51.90	171		EAW62471.1
zinc finger protein ENSP00000375192-like [Homo sapiens]	Homo sapiens	human	9606	124	407	2%	2e-30	74.44	164		XP_047290968.1
HCG2030582 [Homo sapiens]	Homo sapiens	human	9606	122	443	3%	2e-30	73.86	102		EAW48014.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	124	403	2%	2e-30	74.44	164		BAC86222.1
mirror-image polydactyly gene 1 protein isoform 2 [Homo sapiens]	Homo sapiens	human	9606	134	374	2%	3e-30	71.28	519		NP_001374998.1
mirror-image polydactyly gene 1 protein isoform 7 [Homo sapiens]	Homo sapiens	human	9606	133	374	2%	3e-30	72.83	522		NP_001399369.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	120	434	3%	1e-29	63.73	127		BAB15056.1
HCG2038961 [Homo sapiens]	Homo sapiens	human	9606	119	374	2%	3e-29	62.86	120		EAW48306.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	120	422	4%	4e-29	59.02	150		BAC85219.1
hCG2039110 [Homo sapiens]	Homo sapiens	human	9606	116	335	2%	2e-28	65.66	103		EA006591.1
hCG1814203 [Homo sapiens]	Homo sapiens	human	9606	118	466	3%	3e-28	62.04	162		EAW81162.1
hCG1995581 [Homo sapiens]	Homo sapiens	human	9606	120	338	2%	4e-28	73.91	230		EAWS7092.1
lysosomal amino acid transporter 1 homolog isoform XI [Homo...]	Homo sapiens	human	9606	123	406	5%	6e-28	70.10	349		XP_047293350.1
HCG2038847 [Homo sapiens]	Homo sapiens	human	9606	116	350	3%	1e-27	53.57	159		EA006302.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	117	338	3%	1e-27	63.89	179		BAC05191.1
hCG1608224, isoform CRA_b [Homo sapiens]	Homo sapiens	human	9606	115	299	2%	1e-27	69.66	131		EAW80679.1
chromosome 14 open reading frame 93, isoform CRA_b [Homo sapiens]	Homo sapiens	human	9606	118	405	4%	2e-27	51.72	228		EAWS6197.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	331	2%	3e-27	70.97	148		BAC85291.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	434	3%	3e-27	70.33	128		BAC85205.1
hCG2033684 [Homo sapiens]	Homo sapiens	human	9606	115	348	3%	4e-27	72.09	169		EAWS8471.1
PRO2852 [Homo sapiens]	Homo sapiens	human	9606	115	348	3%	4e-27	72.09	169		AAG35505.1
hCG2038446 [Homo sapiens]	Homo sapiens	human	9606	113	406	3%	4e-27	71.59	135		EAAX11460.1
PRO0764 [Homo sapiens]	Homo sapiens	human	9606	113	335	2%	7e-27	72.62	133		AAG35521.1
uncharacterized protein LOC124906545 [Homo sapiens]	Homo sapiens	human	9606	114	114	1%	4e-26	48.12	219		XP_047300707.1
hCG1991981 [Homo sapiens]	Homo sapiens	human	9606	110	395	2%	4e-26	69.89	136		EA002144.1
HCG2039054 [Homo sapiens]	Homo sapiens	human	9606	108	355	3%	5e-26	66.67	96		EAW89122.1

Figura 6.16: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.
[NIHNIH]

```

RID: F19MATAU016
Job Title:Gene135Frame3
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_140906 Length: 9981

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query cover	E Value	Per. Ident	Len	Acc.	Accession
hCG1820395 [Homo sapiens]	Homo sapiens	human	9606	136	447	3%	1e-35	79.27	91		EAW91517.1
PRO2550 [Homo sapiens]	Homo sapiens	human	9606	130	519	4%	2e-33	69.79	118		AAG35515.1
similar to zinc finger protein 569, isoform CRA_a [Homo sapiens]	Homo sapiens	human	9606	128	523	4%	1e-32	77.11	115		EAWS6733.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	129	544	4%	2e-32	77.27	152		BAA91131.1
hypothetical protein MGCT2075, isoform CRA_c [Homo sapiens]	Homo sapiens	human	9606	126	467	4%	1e-31	72.73	132		EAW93800.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	125	458	3%	4e-31	71.91	152		BAC86261.1
PDZ and LIM domain protein 5 isoform i [Homo sapiens]	Homo sapiens	human	9606	125	447	4%	5e-31	75.29	136		NP_001243358.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	124	477	3%	8e-31	72.41	132		BAC85397.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	127	365	3%	3e-30	60.40	260		BAC87615.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	121	429	3%	6e-30	70.93	122		BAC85329.1
DISC1 scaffold protein [Homo sapiens]	Homo sapiens	human	9606	134	489	4%	2e-29	69.47	755		KAI2521821.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	121	468	5%	2e-29	63.21	165		BAC86633.1
intraflagellar transport protein 25 homolog isoform f [Homo...]	Homo sapiens	human	9606	122	459	4%	3e-29	73.26	214		NP_001369190.1
death-associated protein kinase 2 isoform 14 [Homo sapiens]	Homo sapiens	human	9606	128	485	3%	4e-29	75.61	434		NP_001382220.1
hCG1742852 [Homo sapiens]	Homo sapiens	human	9606	123	469	4%	5e-29	62.50	247		EAWS5807.1
DNA polymerase-transactivated protein 3 [Homo sapiens]	Homo sapiens	human	9606	117	436	4%	5e-29	68.75	101		AAR21084.1
heat shock protein family B (small) member 11 [Homo sapiens]	Homo sapiens	human	9606	122	464	4%	6e-29	73.26	214		KAI2480768.1
ubiquitin-conjugating enzyme E2D 4 (putative), isoform CRA_b...	Homo sapiens	human	9606	117	499	4%	1e-28	61.11	128		EAW94179.1
v-myb myeloblastosis viral oncogene homolog (avian), isoform...	Homo sapiens	human	9606	128	490	4%	1e-28	76.83	478		EAW47968.1
UPF0764 protein (c1orf89 isoform X3) [Homo sapiens]	Homo sapiens	human	9606	120	429	4%	2e-28	76.62	216		XP_011520694.1
disrupted in schizophrenia 1 protein isoform c [Homo sapiens]	Homo sapiens	human	9606	130	483	4%	3e-28	68.42	755		NP_001158011.1
HCG2039055 [Homo sapiens]	Homo sapiens	human	9606	119	609	6%	3e-28	60.00	184		EAW89455.1
DISC1 scaffold protein [Homo sapiens]	Homo sapiens	human	9606	130	483	4%	3e-28	68.42	755		KAI24085338.1
hCG2042307 [Homo sapiens]	Homo sapiens	human	9606	115	321	2%	3e-28	72.22	94		EAW98491.1
PRO1722 [Homo sapiens]	Homo sapiens	human	9606	115	474	4%	5e-28	74.67	118		AAB49085.1
hypothetical protein [Homo sapiens]	Homo sapiens	human	9606	118	405	3%	6e-28	62.24	190		QR280143.1
hypothetical protein [Homo sapiens]	Homo sapiens	human	9606	118	405	3%	7e-28	62.24	190		ASF87458.1
c-MYB [Homo sapiens]	Homo sapiens	human	9606	127	490	4%	9e-28	76.83	666		AAB49034.1
MYB proto-oncogene, transcription factor [Homo sapiens]	Homo sapiens	human	9606	128	494	4%	1e-27	76.83	726		KAI2543948.1
v-myb myeloblastosis viral oncogene homologue (avian) [Homo...]	Homo sapiens	human	9606	127	492	4%	1e-27	76.83	659		CAE55174.1

Figura 6.17: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 3'5 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.
[NIHNIH]

RID: F18K9NC701R
 Job Title:Gene153Frame1
 Program: BLASTP
 Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
 Query #1: Query ID: lcl|Query_380966 Length: 9971

Sequences producing significant alignments:

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query cover	E Value	Per. Ident	Acc. Len	Accession
unknown [Homo sapiens]	Homo sapiens	human	9606	176	176	1%	5e-48	50.94	282	AAL55829.1
Rh blood group null antigen complex glycoprotein subunit [Homo sapiens]	Homo sapiens	human	9606	140	140	0%	2e-37	100.00	59	AAB36027.1
uncharacterized protein LOC124905955 [Homo sapiens]	Homo sapiens	human	9606	123	123	1%	2e-29	50.32	221	XP_047299889.1
uncharacterized protein LOC124905955 [Homo sapiens]	Homo sapiens	human	9606	123	123	1%	3e-29	50.32	221	XP_047302847.1
hCG2007893 [Homo sapiens]	Homo sapiens	human	9606	118	323	1%	5e-29	75.32	112	EA089784.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	115	115	0%	1e-28	80.88	80	AGG39759.1
erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	113	113	0%	5e-28	100.00	52	AAC04250.1
Rh 50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-27	64.84	190	SBUS7553.1
uncharacterized protein LOC124909461 [Homo sapiens]	Homo sapiens	human	9606	121	121	0%	4e-27	57.76	358	XP_047305382.1
Rh blood group antigen complex glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	4e-27	64.84	206	CEJ95650.1
ammonium transporter Rh type A isoform X1 [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	6e-27	64.84	217	XP_011513090.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	109	109	0%	1e-26	98.08	52	UYW66190.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	2e-26	64.84	263	BAP94457.1
truncated Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	7e-26	64.84	321	BCY16066.1
mutant erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	116	179	1%	1e-25	64.84	351	AAC04248.1
lysine demethylase 5B [Homo sapiens]	Homo sapiens	human	9606	112	294	1%	1e-25	69.62	229	KAI2520984.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	2e-25	64.84	381	KAI2542629.1
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	2e-25	64.84	381	AAF23100.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	401	CD067716.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	AH12605.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	QPD99008.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	QPD99009.1
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	NP_000315.2
Rh-null regulator protein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	AAD54392.1
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	BCI65376.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	UUA08075.1
RHAG [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	AHY04440.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	QPD99007.1
mutant Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	QB867486.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	116	0%	3e-25	64.84	409	UUA08074.1

Figura 6.18: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.

[NIHNIH]

RID: F191GX3F016
 Job Title:Gene153Frame2
 Program: BLASTP
 Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
 Query #1: Query ID: lcl|Query_132332 Length: 9943

Sequences producing significant alignments:

Description	Scientific Name	Common Name	Taxid	Max Score	Total Score	Query cover	E Value	Per. Ident	Acc. Len	Accession
uncharacterized protein LOC124909461 [Homo sapiens]	Homo sapiens	human	9606	211	211	3%	6e-58	40.23	358	XP_047305382.1
uncharacterized protein LOC124901689 [Homo sapiens]	Homo sapiens	human	9606	175	175	2%	4e-46	39.74	319	XP_047277114.1
hCG2038675 [Homo sapiens]	Homo sapiens	human	9606	167	167	1%	8e-45	51.79	220	EAA53957.1
hCG2038372 [Homo sapiens]	Homo sapiens	human	9606	157	157	0%	2e-42	81.11	138	EAA09876.1
hCG2038360 [Homo sapiens]	Homo sapiens	human	9606	146	146	1%	2e-37	49.01	227	EAA05501.1
Rh 50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	5e-34	92.54	190	SBUS7553.1
Rh blood group antigen complex glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	8e-34	92.54	206	CEJ95650.1
ammonium transporter Rh type A isoform X1 [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	1e-33	92.54	217	XP_011513090.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	354	1%	4e-33	92.54	263	BAP94457.1
truncated Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	358	1%	2e-32	92.54	321	BCY16066.1
mutant erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	135	368	1%	4e-32	92.54	351	AAC04248.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	419	2%	7e-32	92.54	381	KAI2542629.1
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	419	2%	7e-32	92.54	381	AAF23100.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	481	2%	1e-31	92.54	401	CD067716.1
truncated Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92.54	402	AAF04566.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92.54	409	QPD99008.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	489	2%	1e-31	92.54	409	QPD99009.1
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92.54	409	NP_000315.2
Rh-null regulator protein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92.54	409	AAD54392.1
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92.54	409	BCI65376.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	483	2%	1e-31	92.54	409	UUA08075.1
RHAG [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92.54	409	AHY04440.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92.54	409	QPD99007.1
mutant Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	484	2%	1e-31	92.54	409	QB867486.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92.54	409	UUA08074.1
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	483	2%	1e-31	92.54	409	UUA09137.1
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92.54	409	AQM55598.1
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92.54	409	AAF23097.1
50 kDa erythrocyte plasma membrane glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92.54	409	CAA45883.1
unnamed prbtein product [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92.54	409	BAG36285.1

Figura 6.19: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.

[NIHNIH]

```

RID: F1972TN4013
Job Title:Gene153Frame3
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_48719 Length: 9943

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Len	Accession
NCGI781135 [Homo sapiens]	Homo sapiens	human	9606	194	194	1%	3e-55	66.22	155
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	181	181	1%	3e-50	61.25	164
hCG2038441 [Homo sapiens]	Homo sapiens	human	9606	144	144	1%	1e-37	61.54	148
PRO0657 [Homo sapiens]	Homo sapiens	human	9606	134	319	2%	7e-35	78.48	81
hCG2036967 [Homo sapiens]	Homo sapiens	human	9606	124	124	1%	4e-31	63.41	123
hCG2038530 [Homo sapiens]	Homo sapiens	human	9606	110	110	0%	2e-26	63.95	107
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	113	221	2%	3e-26	71.79	196
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	110	401	3%	1e-25	69.62	161
Rh50 [Homo sapiens]	Homo sapiens	human	9606	103	103	0%	5e-24	97.92	102
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	108	0%	6e-24	98.04	263
ammonium transporter Rh type A isoform XI [Homo sapiens]	Homo sapiens	human	9606	106	106	0%	1e-23	98.00	217
truncated Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	211	1%	2e-23	98.04	321
hCG1813079 [Homo sapiens]	Homo sapiens	human	9606	102	145	1%	2e-23	81.82	108
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	110	305	1%	3e-23	100.00	409
Rhesus blood group-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	99.4	99.4	0%	4e-23	100.00	46
mutant erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	108	212	1%	4e-23	98.04	351
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	7e-23	98.04	381
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	7e-23	98.04	381
unknown [Homo sapiens]	Homo sapiens	human	9606	103	150	2%	7e-23	41.01	202
truncated Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	9e-23	98.04	402
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
Rh-null regulator protein [Homo sapiens]	Homo sapiens	human	9606	108	300	1%	1e-22	98.04	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
RHAG [Homo sapiens]	Homo sapiens	human	9606	108	301	1%	1e-22	98.04	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	300	1%	1e-22	98.04	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	108	303	1%	1e-22	98.04	409

Figura 6.20: Resultado de la búsqueda BLAST para el gen 1 en la Hebra 5'3 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.
[NIHNIH]

```

RID: F1936EVX016
Job Title:Gene235Frame1
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_39263 Length: 10021

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Len	Accession
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	144	386	2%	3e-38	70.87	121
hCG2038446 [Homo sapiens]	Homo sapiens	human	9606	143	336	2%	2e-37	81.11	135
hCG2039009 [Homo sapiens]	Homo sapiens	human	9606	140	331	2%	7e-37	70.09	120
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	139	328	2%	5e-36	74.00	137
hCG1820395 [Homo sapiens]	Homo sapiens	human	9606	136	471	3%	1e-35	79.27	91
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	137	329	2%	2e-35	70.10	127
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	135	330	2%	1e-34	76.40	128
PRO1722 [Homo sapiens]	Homo sapiens	human	9606	134	488	3%	2e-34	67.35	118
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	135	325	2%	2e-34	61.02	150
unknown [Homo sapiens]	Homo sapiens	human	9606	135	335	0%	5e-34	87.84	100
PRO2550 [Homo sapiens]	Homo sapiens	human	9606	131	529	4%	2e-33	69.79	113
ubiquitin-conjugating enzyme E2D 4 (putative), isoform CRA_b...	Homo sapiens	human	9606	131	494	3%	2e-33	70.65	128
hCG1744006 [Homo sapiens]	Homo sapiens	human	9606	132	383	3%	4e-33	72.63	171
hCG2039110 [Homo sapiens]	Homo sapiens	human	9606	129	300	2%	5e-33	71.88	103
similar to zinc finger protein 569, isoform CRA_a [Homo sapiens]	Homo sapiens	human	9606	128	522	4%	1e-32	77.11	115
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	129	527	4%	2e-32	77.27	152
hCG1995581 [Homo sapiens]	Homo sapiens	human	9606	132	326	2%	2e-32	77.91	230
RecName: Full=Protein GVQM1; AltName: Full=GVQM motif-containi...	Homo sapiens	human	9606	131	484	4%	2e-32	71.28	195
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	129	315	2%	2e-32	73.91	140
hCG1814203 [Homo sapiens]	Homo sapiens	human	9606	129	559	3%	3e-32	63.72	162
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	127	311	2%	7e-32	67.01	129
hypothetical protein MCG72075, isoform CRA_c [Homo sapiens]	Homo sapiens	human	9606	126	479	4%	1e-31	72.73	132
hCG1991981 [Homo sapiens]	Homo sapiens	human	9606	125	297	2%	2e-31	72.83	136
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	126	494	3%	3e-31	71.91	152
PDE and LIM domain protein 5 isoform 1 [Homo sapiens]	Homo sapiens	human	9606	125	467	3%	4e-31	75.29	136
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	124	497	4%	6e-31	72.41	132
ubiquitously transcribed tetrapeptide repeat Y-linked...	Homo sapiens	human	9606	121	444	3%	1e-30	71.88	84
hCG2033684 [Homo sapiens]	Homo sapiens	human	9606	124	303	2%	2e-30	74.70	169
PRO2852 [Homo sapiens]	Homo sapiens	human	9606	124	303	2%	2e-30	70.00	169
Solute carrier family 48 (heme transporter), member 1 [Homo...	Homo sapiens	human	9606	127	309	2%	2e-30	70.41	239

Figura 6.21: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.
[NIHNIH]

```

RID: F18806ED013
Job Title:Gene235Frame2
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lc1|Query_92513 Length: 9972

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Acc. Len	Accession
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	131	482	4%	4e-33	67.96	BAC85220.1
hCG1818479 [Homo sapiens]	Homo sapiens	human	9606	124	417	2%	2e-31	73.48	EAWS5089.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	123	537	5%	1e-30	67.59	BAC84924.1
hCG1744086 [Homo sapiens]	Homo sapiens	human	9606	124	488	4%	2e-30	51.98	EAWS2471.1
zinc finger protein ENSP00000375192-like [Homo sapiens]	Homo sapiens	human	9606	124	464	3%	2e-30	74.44	XP_047290968.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	124	461	3%	2e-30	74.44	BAC86222.1
mirror-image polydactyly gene 1 protein isoform 2 [Homo sapiens]	Homo sapiens	human	9606	134	427	3%	3e-30	71.28	NP_001374998.1
mirror-image polydactyly gene 1 protein isoform 7 [Homo sapiens]	Homo sapiens	human	9606	133	426	3%	3e-30	72.83	NP_001399369.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	120	462	4%	2e-29	63.73	BAB15056.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	121	440	4%	2e-29	59.84	BAC85219.1
hCG2038961 [Homo sapiens]	Homo sapiens	human	9606	119	435	3%	3e-29	62.86	EAWS48386.1
hCG1814283 [Homo sapiens]	Homo sapiens	human	9606	118	568	4%	3e-28	62.04	EAWS1162.1
hCG2039110 [Homo sapiens]	Homo sapiens	human	9606	115	397	3%	3e-28	65.66	EAWS6591.1
zinc finger protein 30 homolog (mouse) [Homo sapiens]	Homo sapiens	human	9606	115	115	0%	3e-28	64.95	EAWS6745.1
hCG1995581 [Homo sapiens]	Homo sapiens	human	9606	119	482	3%	5e-28	73.91	EAWS7092.1
lysosomal amino acid transporter 1 homolog isoform XI [Homo...]	Homo sapiens	human	9606	123	502	4%	6e-28	70.18	XP_047279350.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	117	489	4%	1e-27	63.89	BAC85191.1
hCG1660224, isoform CRA_b [Homo sapiens]	Homo sapiens	human	9606	114	357	3%	1e-27	69.66	EAWS8679.1
hCG2038847 [Homo sapiens]	Homo sapiens	human	9606	115	414	4%	2e-27	53.57	EAWS6302.1
chromosome 14 open reading frame 93, isoform CRA_b [Homo sapiens]	Homo sapiens	human	9606	117	475	4%	2e-27	51.72	EAWS6197.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	429	3%	3e-27	70.33	BAC85205.1
PR02852 [Homo sapiens]	Homo sapiens	human	9606	115	415	3%	4e-27	72.09	EAWS3505.1
hCG2038684 [Homo sapiens]	Homo sapiens	human	9606	115	414	3%	4e-27	72.09	EAWS8471.1
hCG2038446 [Homo sapiens]	Homo sapiens	human	9606	113	460	3%	5e-27	71.59	EAWS11460.1
PR00764 [Homo sapiens]	Homo sapiens	human	9606	112	431	3%	8e-27	72.62	EAWS5521.1
hCG2031845 [Homo sapiens]	Homo sapiens	human	9606	111	111	0%	1e-26	78.87	EAWS6532.1
neuronal thread protein AD7c-NTP [Homo sapiens]	Homo sapiens	human	9606	119	553	6%	2e-26	38.59	EAWS8737.1
hCG1991981 [Homo sapiens]	Homo sapiens	human	9606	110	455	3%	5e-26	69.89	EAWS2144.1
hCG2039054 [Homo sapiens]	Homo sapiens	human	9606	108	483	3%	6e-26	66.67	EAWS9122.1
KIAA1651 protein [Homo sapiens]	Homo sapiens	human	9606	108	350	2%	7e-26	65.17	BAB33321.1

Figura 6.22: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado. [NIHNIH]

```

RID: F18D068K016
Job Title:Gene235Frame3
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lc1|Query_6185 Length: 9951

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Acc. Len	Accession
unknown [Homo sapiens]	Homo sapiens	human	9606	144	144	1%	2e-37	60.74	AAL55749.1
hCG2038582 [Homo sapiens]	Homo sapiens	human	9606	122	211	1%	2e-30	73.86	EAWS48014.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	127	308	2%	3e-30	68.48	BAC87615.1
hCG2042387 [Homo sapiens]	Homo sapiens	human	9606	115	289	2%	3e-28	72.22	EAWS8091.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	114	0%	3e-27	70.97	BAC85291.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	291	2%	4e-27	66.32	BAG64432.1
hCG2014340 [Homo sapiens]	Homo sapiens	human	9606	113	358	3%	5e-27	77.78	EAWS8386.1
hCG1979495 [Homo sapiens]	Homo sapiens	human	9606	112	269	2%	9e-27	59.32	EAWS5411.1
HMG2a [Homo sapiens]	Homo sapiens	human	9606	112	328	1%	2e-26	79.69	AAL43855.1
hCG2041539 [Homo sapiens]	Homo sapiens	human	9606	108	215	2%	9e-26	73.08	EAWS4334.1
hCG1641844, isoform CRA_a [Homo sapiens]	Homo sapiens	human	9606	111	154	1%	1e-25	72.97	EAWS2036.1
hCG1817437 [Homo sapiens]	Homo sapiens	human	9606	107	258	1%	2e-25	66.67	EAWS47553.1
high mobility group protein HMG1-C isoform d [Homo sapiens]	Homo sapiens	human	9606	108	318	1%	3e-25	78.12	NP_001287848.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	105	252	2%	6e-25	77.61	BAB12795.1
hCG1828679, isoform CRA_b [Homo sapiens]	Homo sapiens	human	9606	107	219	1%	1e-24	77.46	EAWS2980.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	106	338	2%	3e-24	75.36	BAA01131.1
hCG2039055 [Homo sapiens]	Homo sapiens	human	9606	107	345	3%	4e-24	52.46	EAWS8455.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	105	280	2%	6e-24	57.60	BAB71591.1
LAG1 longevity assurance homolog 5 (S. cerevisiae), isoform...	Homo sapiens	human	9606	107	154	1%	1e-23	75.00	EAWS8131.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	105	197	1%	1e-23	65.96	BAC86879.1
leucine rich repeat and sterile alpha motif containing 1 [Homo...]	Homo sapiens	human	9606	104	172	1%	2e-23	75.76	KAL2553983.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	103	211	1%	3e-23	70.13	BAC86448.1
hCG2038848 [Homo sapiens]	Homo sapiens	human	9606	100	233	1%	3e-23	68.18	EAWS9577.1
hCG2028158 [Homo sapiens]	Homo sapiens	human	9606	102	284	2%	3e-23	79.69	EAWS48919.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	103	258	2%	4e-23	64.13	BAC86723.1
tRNA splicing endonuclease subunit 2 [Homo sapiens]	Homo sapiens	human	9606	100	199	1%	4e-23	72.97	KAL2528324.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	103	256	2%	5e-23	64.13	BAC87102.1
abraxas 1, BRCA1 A complex subunit [Homo sapiens]	Homo sapiens	human	9606	99.8	254	1%	7e-23	72.68	KAL4026032.1
PRO1902 [Homo sapiens]	Homo sapiens	human	9606	99.4	286	1%	1e-22	67.47	AAF22026.1
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	103	153	1%	2e-22	48.76	BAA91396.1

Figura 6.23: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 3'5 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado. [NIHNIH]

```

RID: F1A0K3M016
Job Title:Gene253Frame1
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_93476 Length: 9946

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Acc. Len	Accession
hCG2038372 [Homo sapiens]	Homo sapiens	human	9606	155	155	0%	1e-41	80.00	138
PR00657 [Homo sapiens]	Homo sapiens	human	9606	135	232	1%	3e-35	78.48	81
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	115	115	0%	1e-28	80.88	80
erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	113	113	0%	4e-28	100.00	52
Rh 50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	174	1%	3e-27	64.84	190
uncharacterized protein LOC124909461 [Homo sapiens]	Homo sapiens	human	9606	120	120	0%	4e-27	57.76	358
Rh blood group antigen complex glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	209	1%	4e-27	64.84	206
uncharacterized protein LOC124905955 [Homo sapiens]	Homo sapiens	human	9606	116	116	1%	4e-27	45.40	221
ammonium transporter Rh type A isoform XI [Homo sapiens]	Homo sapiens	human	9606	116	223	1%	6e-27	64.84	217
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	110	110	0%	7e-27	98.08	52
uncharacterized protein LOC124905955 [Homo sapiens]	Homo sapiens	human	9606	116	116	1%	8e-27	45.40	221
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	225	1%	2e-26	64.84	263
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	114	114	0%	2e-26	71.79	196
truncated Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	327	2%	7e-26	64.84	321
mutant erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	116	328	2%	1e-25	64.84	351
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	2e-25	64.84	381
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	2e-25	64.84	381
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	405	2%	3e-25	64.84	401
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	416	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	417	2%	3e-25	64.84	409
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	3e-25	64.84	409
Rh-null regulator protein [Homo sapiens]	Homo sapiens	human	9606	116	416	2%	3e-25	64.84	409
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	116	421	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	3e-25	64.84	409
Rh6 [Homo sapiens]	Homo sapiens	human	9606	116	417	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	417	2%	3e-25	64.84	409
mutant Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	417	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	3e-25	64.84	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	116	419	2%	3e-25	64.84	409

Figura 6.24: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 1. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.

[NIHNIH]

```

RID: F1AAW53301R
Job Title:Gene253Frame2
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #1: Query ID: lcl|Query_272507 Length: 9928

Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Acc. Len	Accession
uncharacterized protein LOC124909461 [Homo sapiens]	Homo sapiens	human	9606	212	212	3%	2e-58	40.23	358
uncharacterized protein LOC124901689 [Homo sapiens]	Homo sapiens	human	9606	175	175	2%	4e-46	39.74	319
hCG2038675 [Homo sapiens]	Homo sapiens	human	9606	168	168	1%	5e-45	51.79	220
hCG2038441 [Homo sapiens]	Homo sapiens	human	9606	144	144	1%	1e-37	61.54	148
hCG2038360 [Homo sapiens]	Homo sapiens	human	9606	147	147	1%	1e-37	49.01	227
Rh blood group null antigen complex glycoprotein subunit [Homo...]	Homo sapiens	human	9606	139	139	0%	7e-37	100.00	59
B lymphocyte activation-related protein BC-1514 [Homo sapiens]	Homo sapiens	human	9606	117	117	1%	1e-28	55.05	130
hCG2007893 [Homo sapiens]	Homo sapiens	human	9606	117	338	2%	2e-28	75.32	112
Unknown (protein for MGC1168815) [Homo sapiens]	Homo sapiens	human	9606	112	112	1%	5e-26	55.56	178
lysine demethylase 5B [Homo sapiens]	Homo sapiens	human	9606	112	345	2%	2e-25	69.62	229
hCG2036516 [Homo sapiens]	Homo sapiens	human	9606	108	108	0%	3e-25	61.63	149
BC-1514 protein-like [Homo sapiens]	Homo sapiens	human	9606	104	104	0%	5e-24	63.53	131
hCG2038624 [Homo sapiens]	Homo sapiens	human	9606	102	341	3%	1e-23	53.47	99
hCG1811714 [Homo sapiens]	Homo sapiens	human	9606	100	298	2%	2e-23	70.83	69
hCG1813079 [Homo sapiens]	Homo sapiens	human	9606	101	144	1%	4e-23	81.82	108
unknown [Homo sapiens]	Homo sapiens	human	9606	102	149	2%	2e-22	41.01	202
hCG2000782 [Homo sapiens]	Homo sapiens	human	9606	99.0	242	2%	3e-22	51.30	106
MRE11 meiotic recombination 11 homolog A (S. cerevisiae),...	Homo sapiens	human	9606	98.2	244	1%	7e-22	71.62	125
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	96.6	309	2%	9e-22	66.22	136
hCG2042782 [Homo sapiens]	Homo sapiens	human	9606	97.4	305	2%	1e-21	65.82	114
hCG2038546 [Homo sapiens]	Homo sapiens	human	9606	97.1	97.1	0%	2e-21	79.66	125
hCG2042305 [Homo sapiens]	Homo sapiens	human	9606	95.1	225	1%	3e-21	70.00	84
Intraflagellar transport 74 [Homo sapiens]	Homo sapiens	human	9606	101	251	1%	3e-21	66.25	280
Intraflagellar transport 74 [Homo sapiens]	Homo sapiens	human	9606	101	251	1%	3e-21	66.25	280
hCG2038304 [Homo sapiens]	Homo sapiens	human	9606	97.8	97.8	1%	5e-21	37.36	174
amiloride-sensitive amine oxidase [copper-containing] isoform ...	Homo sapiens	human	9606	105	259	2%	1e-20	71.43	723
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	95.9	322	3%	2e-20	54.55	173
PRO2411 [Homo sapiens]	Homo sapiens	human	9606	92.0	296	2%	2e-20	64.79	68
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	100	306	2%	3e-20	65.82	368
hCG2038486 [Homo sapiens]	Homo sapiens	human	9606	92.4	92.4	1%	8e-20	44.44	123

Figura 6.25: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 2. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por los organismos en donde fue encontrado.

[NIHNIH]

```

RID: F1AY22VR016
Job Title:Gene253Frame3
Program: BLASTP
Database: nr All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Query #: Query ID: lcl|Query_33128 Length: 9972
Sequences producing significant alignments:

```

Description	Scientific Name	Common Name	Taxid	Max Score	Total Query Score	E Value	Per. Ident	Len	Accession
HCG1781136 [Homo sapiens]	Homo sapiens	human	9606	194	194	1%	5e-55	66,22	155
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	181	181	1%	4e-50	61,25	164
unknown [Homo sapiens]	Homo sapiens	human	9606	177	177	1%	3e-48	50,94	282
Rh 50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	5e-34	92,54	190
Rh blood group antigen complex glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	8e-34	92,54	206
ammonium transporter Rh type A isoform XI [Homo sapiens]	Homo sapiens	human	9606	135	239	1%	1e-33	92,54	217
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	354	1%	4e-33	92,54	263
truncated Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	368	1%	2e-32	92,54	321
mutant erythrocyte membrane glycoprotein Rh50 [Homo sapiens]	Homo sapiens	human	9606	135	368	1%	4e-32	92,54	351
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	419	1%	7e-32	92,54	381
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	419	1%	7e-32	92,54	381
lysine demethylase 5B [Homo sapiens]	Homo sapiens	human	9606	130	315	1%	8e-32	76,71	229
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	481	2%	1e-31	92,54	401
truncated Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92,54	402
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92,54	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	489	2%	1e-31	92,54	409
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92,54	409
Rh-null regulator protein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92,54	409
ammonium transporter Rh type A [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92,54	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	483	2%	1e-31	92,54	409
RHAG [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92,54	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	487	2%	1e-31	92,54	409
mutant Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	484	2%	1e-31	92,54	409
Rh associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92,54	409
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	483	2%	1e-31	92,54	409
Rh-associated glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92,54	409
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92,54	409
50 kDa erythrocyte plasma membrane glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	485	2%	1e-31	92,54	409
unnamed protein product [Homo sapiens]	Homo sapiens	human	9606	135	486	2%	1e-31	92,54	409
mutant Rh50 glycoprotein [Homo sapiens]	Homo sapiens	human	9606	135	481	2%	1e-31	92,54	409

Figura 6.26: Resultado de la búsqueda BLAST para el gen 2 en la Hebra 5'3 Frame 3. Se muestran las secuencias similares con su puntaje de similitud. También se clasifican por las organismos en donde fue encontrado.
[NIHNIH]

6.3. Set de Proteínas y árbol filogenético

En esta sección se presentan los árboles filogenéticos realizados con la herramienta JalView. Se muestra la filogenética de los dos grupos de proteínas presentados en la Tabla 6.1. La imagen de cada árbol fue colocada en un ángulo de 90 grados para visualizar mejor los nombres de las proteínas, las cuales se encuentran en sus respectivas ramas. Los grupos de datos presentados en la Tabla 6.1 son los datos seleccionados de Uniprot a partir de los códigos resultantes en el proceso de BLAST. Cada columna representa a una de las hebras donde encontramos las proteínas. Se unieron ambos genes mostrados en las figuras 6.1 y 6.8, así como los frames de lectura de cada una de las hebras. Con esto se obtuvieron solamente los datos utilizados ya que permitió eliminar resultados duplicados. A partir de estos datos se modeló la proteína más desordenada y menos desordenada de cada grupo. Para el grupo perteneciente a la hebra 5'3 se modeló la proteína M1SZX7 en la Figura 6.40 y la Q9UI50 en la Figura 6.38. Las proteínas de la hebra 3'5 fueron Q8N8C2 en la Figura 6.36 y Q6ZQR8 en la Figura 6.34.

En la Figura 6.27 se muestra el árbol de las proteínas encontradas en la hebra 3'5. Podemos observar que la proteína Q6ZTX9 es la menos relacionada con el grupo. Esta se divide a partir del primer nodo. A pesar de no estar relacionada con el grupo de proteínas de la hebra 3'5 se observa que tiene algunas secciones conservadas. Estas secuencias conservadas se muestran en las figuras 6.29, 6.30 y 6.31. Las demás proteínas surgen de las mismas ramas y se encuentran más relacionados filogenéticamente. El árbol de las proteínas en la hebra 5'3 se encuentra en la Figura 6.28. La proteína fuera del grupo de clados relacionados es la B7Z407. A pesar de ser la proteína menos relacionada al resto del grupo se pueden observar algunas secuencias conservadas en las figuras 6.32 y 6.33. Este árbol muestra diferencias respecto al árbol de la hebra 3'5. Pueden observarse clados mucho más grandes, es decir que del mismo nodo surge una cantidad de proteínas más grande. Esto puede sugerir una relación más estrecha entre las proteínas.

Proteínas encontradas en la hebra 5'3	Proteínas encontradas en la hebra 3'5
Q8WYX4	Q6ZPB0
Q16416	Q6ZP21
M1SZX7	Q9H743
Q9UBB8	Q6ZPB2
A0A1A9C9I6	Q6ZPA0
A0A0A1TSG9	Q9P195
A0A0A8IKZ2	Q9H728
A0A8D5ZBQ7	Q6ZP99
Q02094	Q8N9K0
Q9UHG9	Q9NX85
X5MNR9	Q9H387
Q96E98	C3UJR9
A0A7S8RGC4	Q9H397
A0A7S8RFX7	Q6P1K1
Q9UK70	Q6ZUA3
A0A6S6PEI8	Q8WZ27
A0A023UN48	Q8N8C2
A0A7S8MU46	Q6ZUK0
A0A411HBG8	Q6ZP34
Q6ZPC1	Q96JR5
Q9UI50	Q9BYA9
Q969H1	Q8N210
A2NVG1	Q6ZUG4
Q6ZP50	Q6ZNU7
Q9UP76	Q6ZQR8
Q9UK69	Q6ZLN6
B2RNZ7	Q6ZTF6
Q1KSG2	Q6STG2
Q96NR6	A0A895FY86
Q6ZNU8	A0A248J1I3
Q9H389	P10242
B4DHX3	Q70AC3
Q6ZP44	Q708E2
B7Z407	B4XZE4
Q6ZPB6	Q306F7
Q6ZP20	Q6ZUH1
Q6FG63	B4E0H0
Q13629	Q1M183
Q0VFX3	A0A024RC64
	B7Z559
	Q96LS9
	Q6ZSR7
	Q6ZTX9
	Q6ZT71
	Q6ZS53
	Q9UHT1
	Q9NWI4

Tabla 6.1: Códigos en Uniprot de las proteínas estudiadas de las primeras 15 mejores rankeadas en BLAST separadas por la hebra en que fueron encontradas.

[ConsortiumConsortium2018]

6.4. Alineación JalView

Los resultados de la alineación clustal de las proteínas en JalView se muestran en las figuras de esta sección. Las figuras 6.29, 6.30 y 6.31 muestran la alineación de las proteínas encontradas en la hebra 3'5. Las secciones con coloración indican que es una sección conservada. Las imágenes fueron recortadas para ser mostradas por partes y poder visualizar bien cada posición. Las figuras 6.32 y 6.33 muestran la alineación de las proteínas encontradas en la hebra 5'3. Las figuras de los resultados muestran la alineación en diferentes secciones, es decir que la imagen original tuvo que ser recortada en diferentes partes para poder visualizar todas las posiciones de los aminoácidos. Los colores que se observan se leen según la tabla de coloración para alineación clustal en JalView en la Figura 13.8. Estos colores indican que existen secciones conservadas así como las propiedades químicas de las mismas, los porcentajes mostrados en la tabla indican que criterio se utiliza para considerarlos conservados. El porcentaje es la cantidad mínima que debe estar presente del grupo de aminoácidos para considerarlo ordenado. Podemos ver que en ambos grupos de proteínas existen estas secciones por lo que se tienen grupos de proteínas altamente conservadas.

Para las proteínas pertenecientes a la hebra 3'5 podemos observar en la Figura 6.29 que todas las secuencias, menos C3UJR9 y Q9BYA9 inician con una Metionina en su secuencia. Esta es considerada como la primera region conservada del grupo de proteínas. Las proteínas P10242, Q70AC3 y Q708E2 presentan una alta conservación. En las figuras 6.29, 6.30 y 6.31 se observa que las secuencias de estas tres proteínas cambian en lo más mínimo y la diferencia más significativa se encuentra en la terminal. Las otras proteínas aunque no se ven iguales entre ellas se pudieron encontrar varias secuencias conservadas. La segunda sección conservada significativa esta en la posición 104 a la 121. Se observan los aminoácidos Leucina, Valina, Fenilalanina, Alanina, Metioninas e Isoleucina. La última posición de esta secuencia se encuentran aminoácidos Serina. En esta sección también se encuentran Glicinas y Prolinas.

La tercera sección conservada se encuentra en las posiciones 146 a la 168. En esta sección se encuentran basicamente los aminoácidos Leucina, Valina, Fenilalanina, Alanina, Metioninas e Isoleucina. Aunque tambien hay algunas Prolinas y Glicinas como en la sección anterior. Estas regiones conservadas se observan en la Figura 6.29. En esta figura también se observa una región con alta cantidad de Prolinas a partir de la posición 220. En la Figura 6.30 se observan dos de las secciones conservadas más grandes del grupo de proteínas en la hebra 3'5. La sección conservada que se consideraría como la cuarta inicia en la posición 241 y termina en la 283. Los aminoácidos que se encuentran en esta sección son similares a las secciones anteriores. La diferencia es que acá se agrega Ácido Glutámico, Glutamina y Treonina. Estos aminoácidos son también los que están presentes en la quinta sección conservada del grupo, se encuentra en la posición 328 a la 355. La ultima región conservada que podemos considerar es en la posición 388 donde los aminoácidos conservados se muestran con coloración celeste. Las terminales de las proteínas observadas en la Figura 6.31 no tienen ningún tipo de conservación, a excepción de las tres proteínas mencionadas al inicio.

La conservación de las glicoproteínas pertenecientes a la hebra 5'3 se encuentran en las figuras 6.32 y 6.33. La primera sección conservada se observa en la primera posición, todas inician con una Metionina a excepción de Q16416 y Q9UP76. Luego se puede observar que toda la primera parte de las proteínas, es decir de la posición 4 a la 196 está altamente conservada. En la primera franja de esta sección se nota una gran dominación de los aminoácidos Leucina, Valina, Fenilalanina, Alanina, Metioninas e Isoleucina con coloración celeste. También hay franjas con coloración morada con aminoácidos Ácido Glutámico. Y con coloración verde los aminoácidos Glutamina, Serina y Treonina. Hay varias posición con los aminoácidos Prolinas y Glicinas conservados. Esta sección conservada se presenta en la Figura 6.32. La terminal de estas proteínas así como en el grupo de la hebra 3'5 no se encuentran conservadas de una forma que podamos considerar significativa. En la Figura 6.33 se pueden observar estas terminales, solo tenemos a los aminoácidos Prolinas y Glicinas conservados.

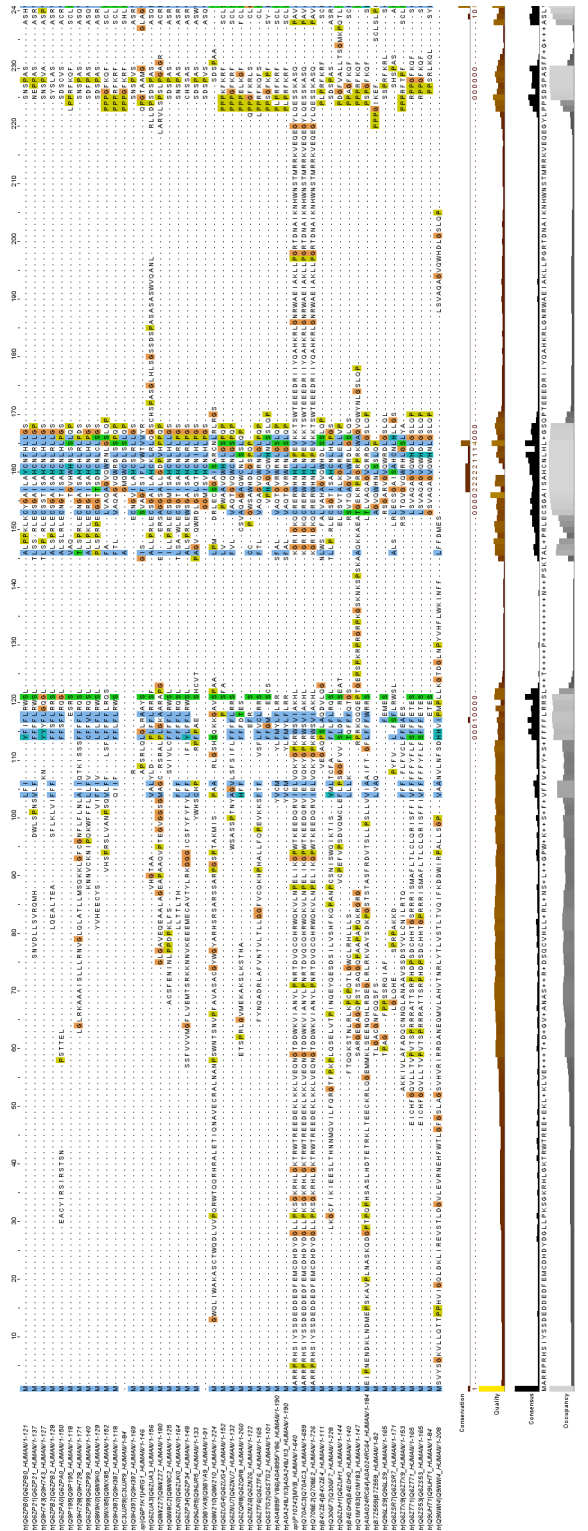


Figura 6.29: Parte 1 de la alineación en JalView de las proteínas en la hebra 3'5, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.

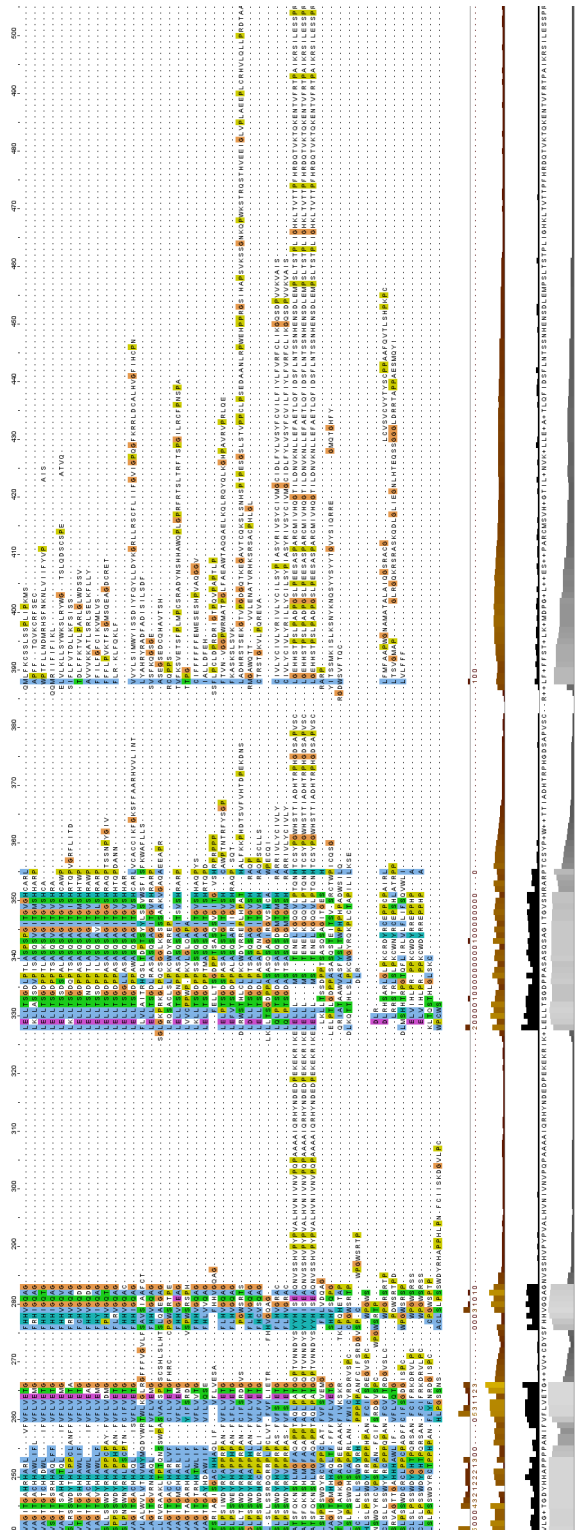


Figura 6.30: Parte 2 de la alineación en JalView de las proteínas en la hebra 3'5, se presentan las secuencias conservadas encontradas. La coloración depende del tipo de aminoácido conservado, el código de colores se encuentra en la Figura 13.8.

6.5. Modelados de proteínas

En esta sección se muestran los cuatro modelados que mejor representan los casos que se quieren estudiar. Se escogió la proteína más ordenada y menos desordenada de cada uno de los grupos (hebra 5'3 y 3'5). El análisis de los modelados incluye cuatro gráficas para analizar cada aspecto. La figura a) es el modelado 3D realizado en Chimera a partir del archivo PDB resultante de AlphaFold. La figura b) muestra la gráfica PAE resultante de AlphaFold que muestra la distancia entre los aminoácidos en el modelado, de esta gráfica solo se toma en cuenta la rank 1, ya que es el mejor modelo. En la figura c) se tiene la cobertura de la secuencia incluida en el modelado, la gráfica es resultado de AlphaFold. La figura d) tiene la gráfica pLDDT resultado de AlphaFold, se observa cuál de los modelos tiene mejor puntaje de calidad de modelado en cada posición de la secuencia. La figura e) es la gráfica resultante de IUPred Anchor2 que muestra el desorden de las diferentes secciones de la proteína, una puntuación arriba de 0.5 es desorden. Este análisis se muestra en las figuras 6.34, 6.36, 6.38 y 6.40. Los demás modelados realizados de las otras proteínas se muestran en la sección de Anexos para futura referencia. También se tienen las gráficas resultantes del programa en R para analizar las secciones desordenadas. Este análisis solo se realizó con las proteínas más desordenadas y ordenadas para visualizar cómo funciona el programa y comprobar los resultados. Para presentar estos resultados se muestran cuatro gráficas, la a) se realiza haciendo uso del archivo de IUPred Anchor2 para recrear la gráfica de desorden. La gráfica b) hace un análisis de los residuos y compara que porcentaje de los aminoácidos presentes se encuentran en una sección ordenada y desordenada. En la gráfica c) se compara el gráfico pLDDT y IUPred para saber como se relacionan. Esta correlación también se muestra en la gráfica d). Este análisis se muestra en las figuras 6.35, 6.37, 6.39 y 6.41.

6.5.1. Proteína más desordenada hebra 3'5

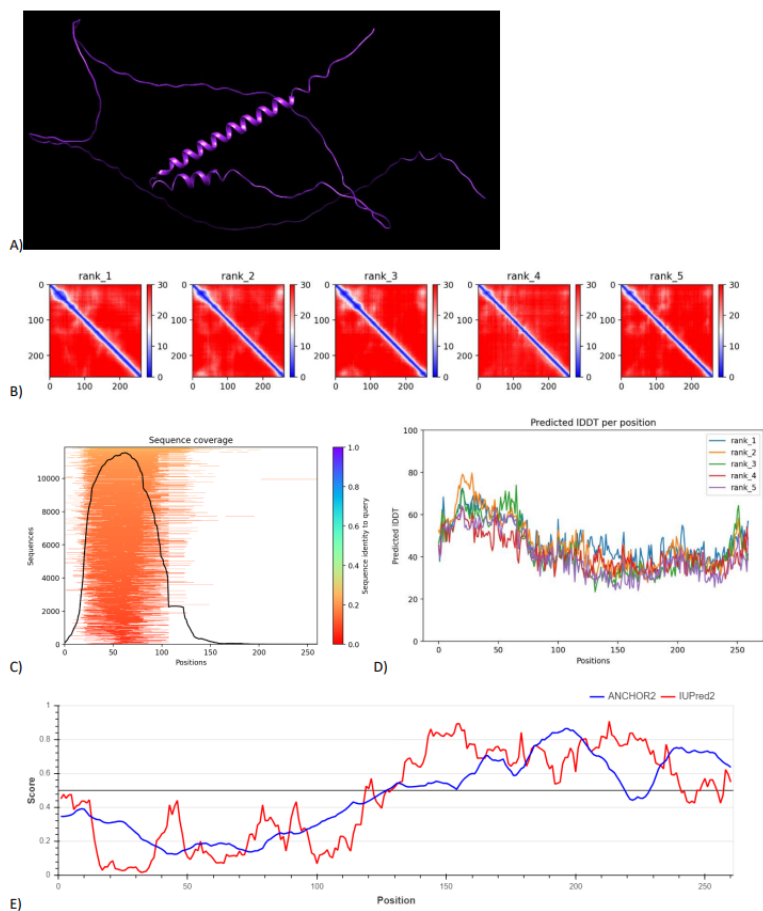


Figura 6.34: Análisis proteína Q6ZQR8 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.

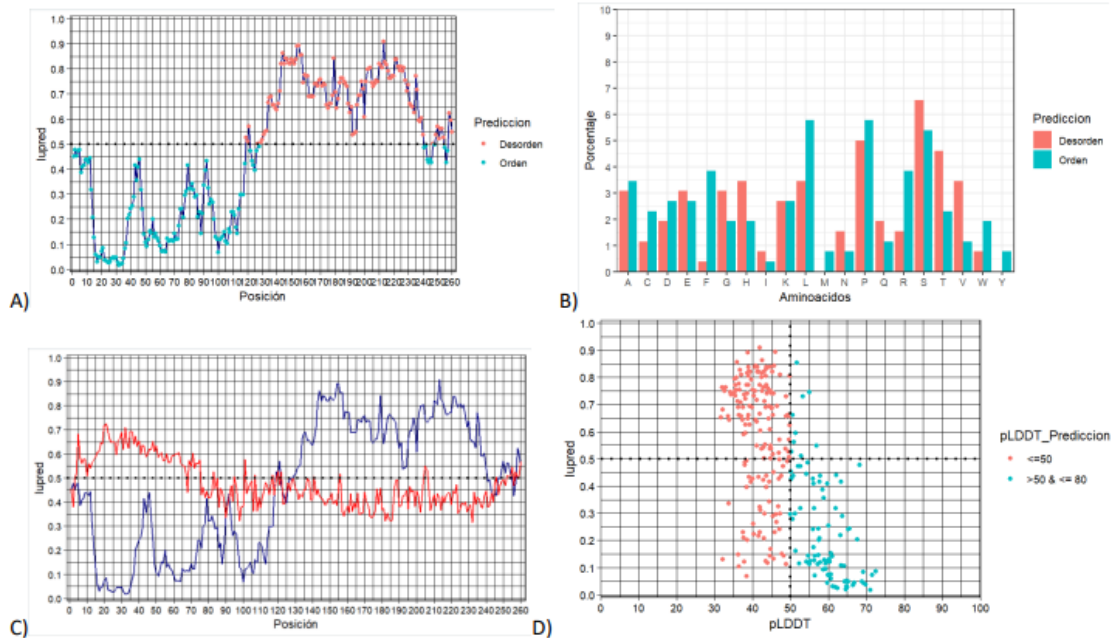


Figura 6.35: Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q6ZQR8. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.

6.5.2. Proteína más ordenada hebra 3'5

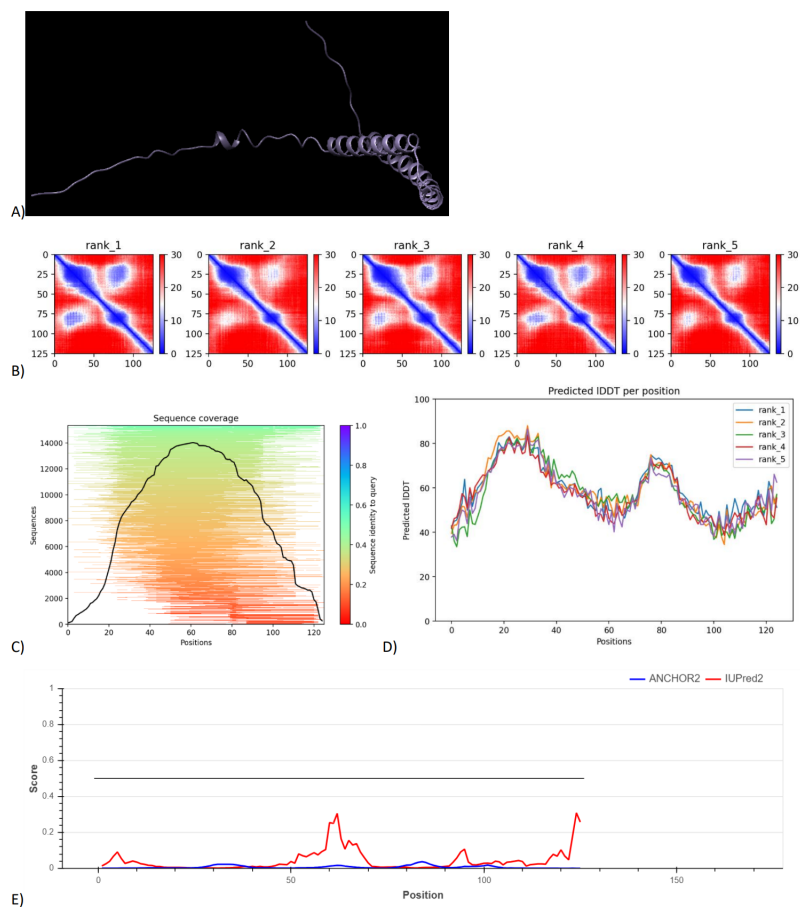


Figura 6.36: Análisis proteína Q8N8C2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.

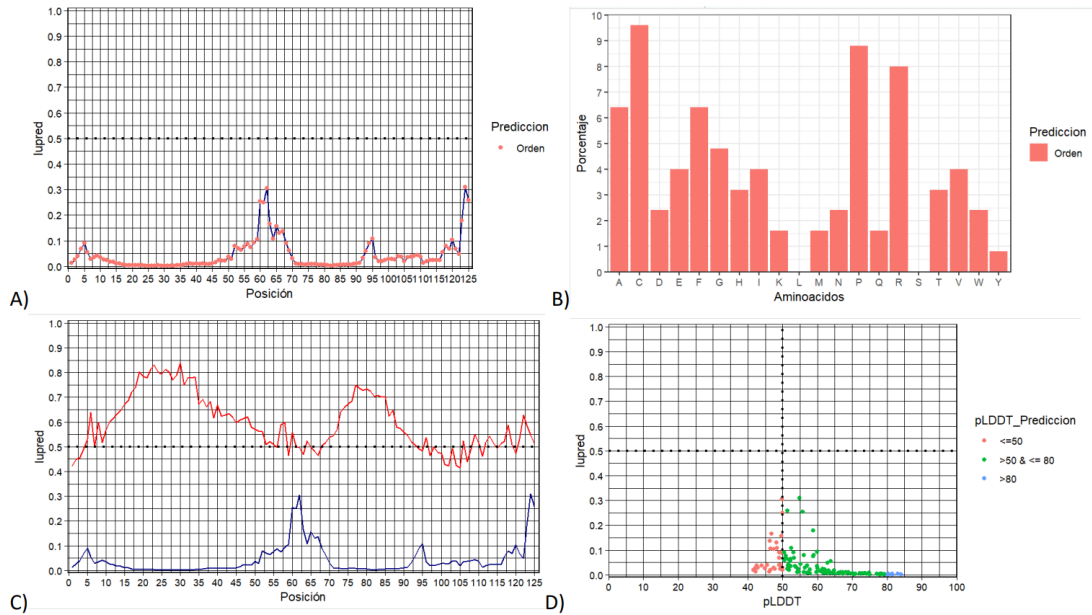


Figura 6.37: Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q8N8C2. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.

6.5.3. Proteína más desordenada hebra 5'3

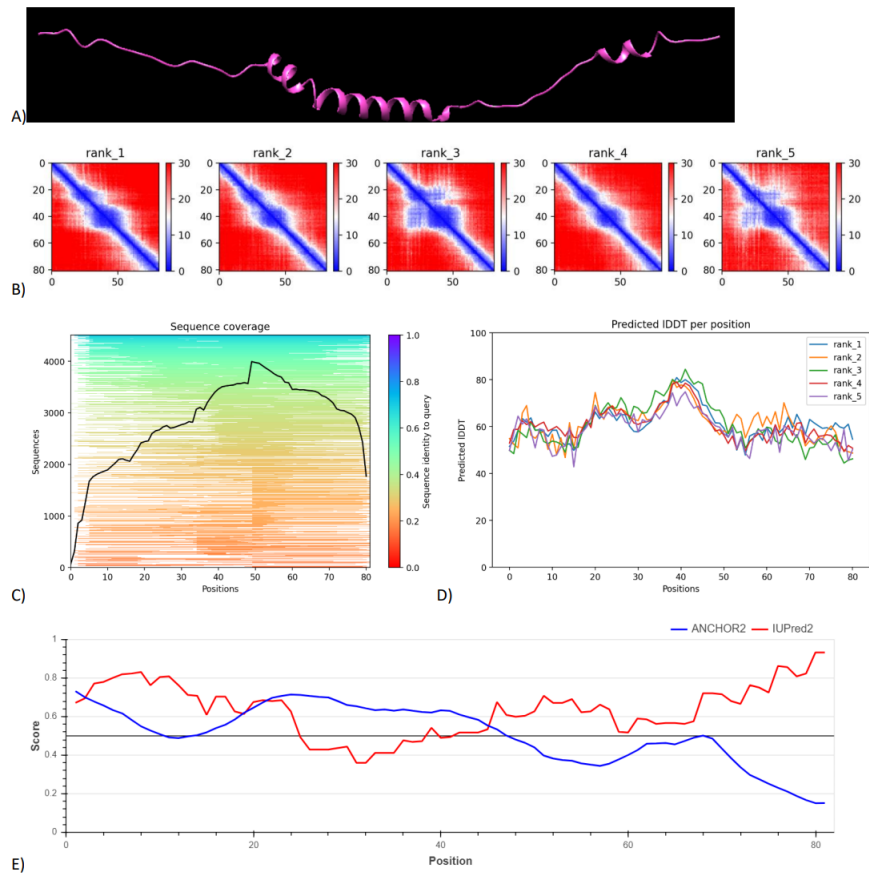


Figura 6.38: Análisis proteína Q9UI50 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.

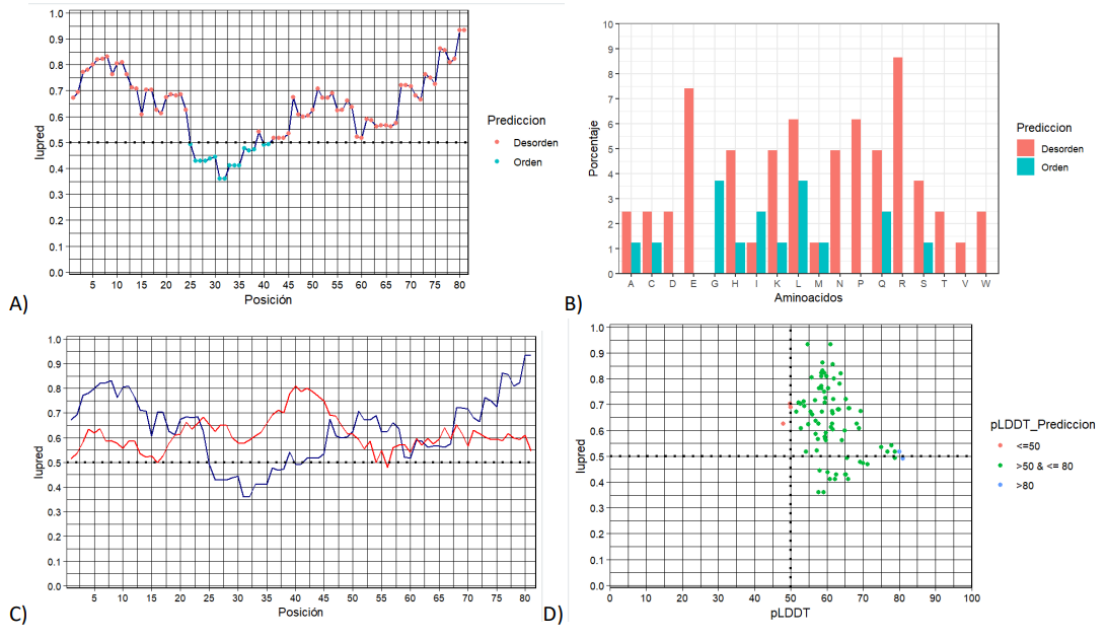


Figura 6.39: Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína Q9UI50. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.

6.5.4. Proteína más ordenada hebra 5'3

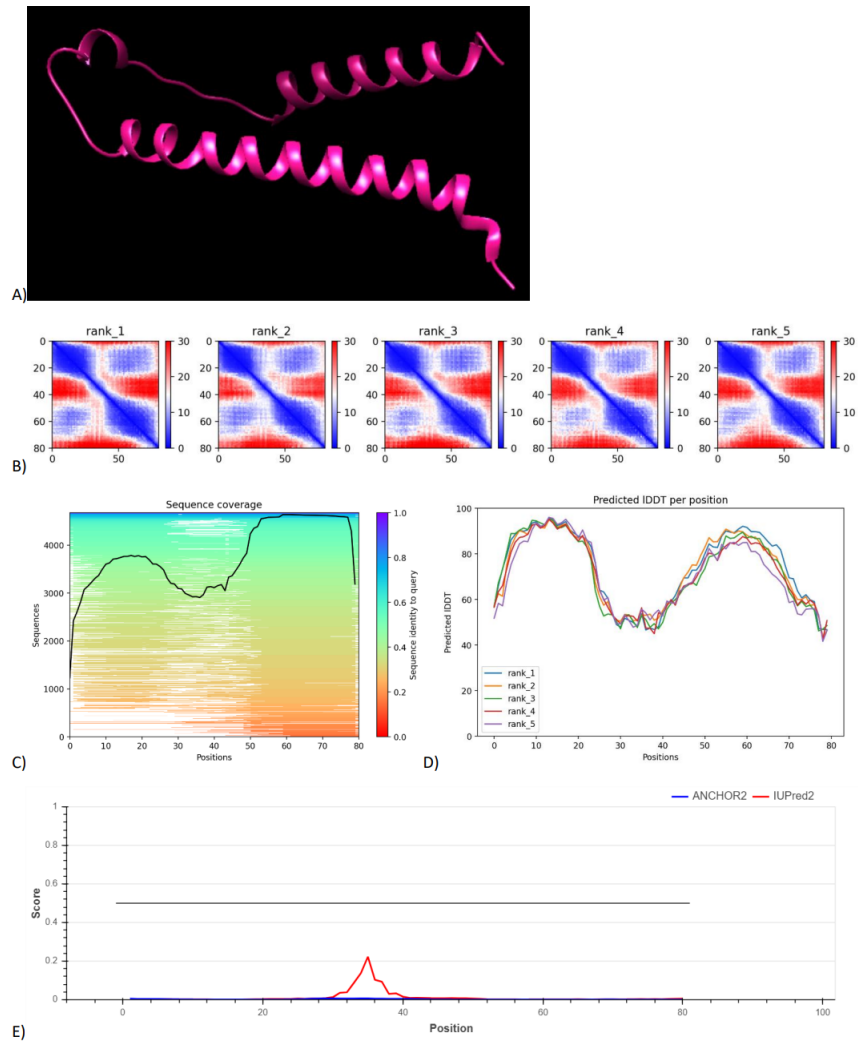


Figura 6.40: Análisis proteína M1SZX7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE de los cinco modelos generados ordenados de mejor a peor. C) Cobertura de la secuencia en el modelo. Muestran todas las secuencias alineadas e indica su similitud con la secuencia, se ordenan de arriba para abajo. La línea negra califica la cobertura relativa de la secuencia respecto al número de secuencias alineadas. D) Gráfico pLDDT de los cinco modelos generados. Muestra el puntaje de calidad de cada uno. E) Análisis IUPred Anchor 2 de secciones desordenadas.

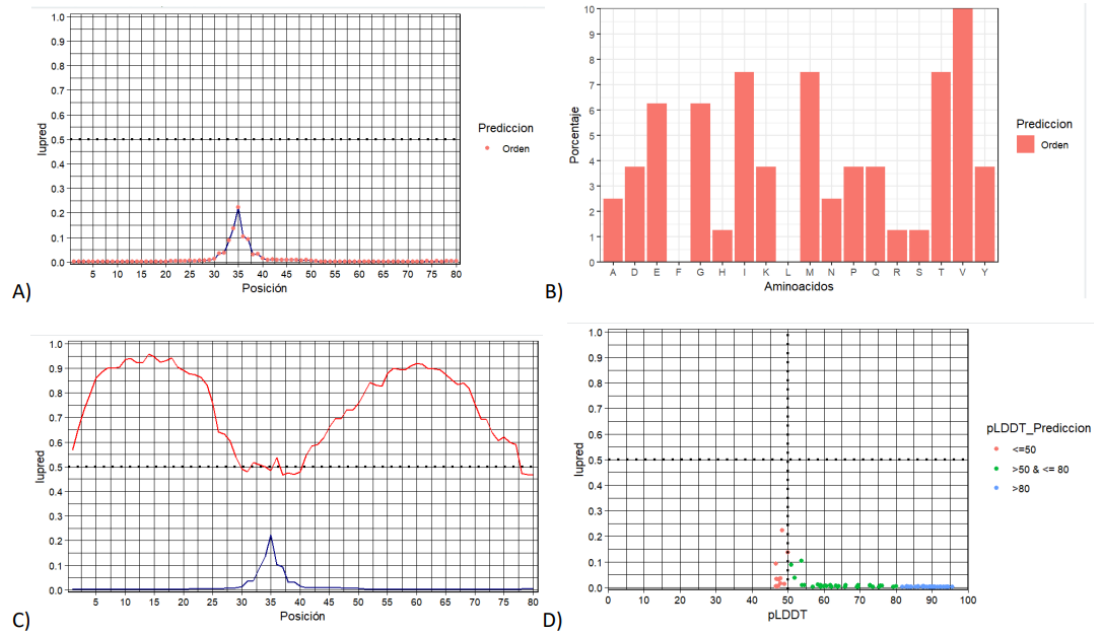


Figura 6.41: Gráficos de programación R para analizar a fondo las secciones desordenadas de la proteína M1SZX7. A) Gráfico de desorden según el análisis IUPred. Los colores rojos muestran posiciones en desorden y los azules en orden. Se realiza un corte en 0.5 para indicar que todo lo que esté por arriba es desorden y por abajo es orden. B) Porcentaje de residuos en orden y desorden de la secuencia. La gráfica se lee viendo el eje y y tomando en cuenta el porcentaje. Por ejemplo si la barra de un aminoácido llega a 6 se toma como el 60%. C) Gráfico IUPred (azul) y pLDDT (rojo). Muestra como el desorden puede verse relacionado con la calidad del modelo que se genera. D) Correlación entre IUPred y pLDDT. Se grafican los puntos de acuerdo a la calidad del modelo y su respectivo orden. Los colores que analizan los diferentes rangos de calidad del modelo que se tienen se muestran del lado derecho.

6.6. Alineación en Chimera de proteínas seleccionadas

Las figuras en esta sección muestran las alineaciones de las proteínas en Chimera, se observan las diferentes secuencias y las secciones que estructuralmente coinciden. Esto sirve para comparar la relación de las regiones conservadas de la sección 6.4 y las secciones vistas estructuralmente. En la primera parte de este análisis se muestran 5 proteínas de cada grupo (hebra 5'3 y 3'5). Las proteínas se seleccionaron buscando las más desordenadas y menos desordenadas. Cada alineación contiene dos proteínas ordenadas, dos desordenadas y una proteína intermedia. Para saber que proteína corresponde a cada color se pueden utilizar las imágenes de referencia citadas en cada imagen. Se indica que proteína se utilizó como referencia en Chimera al momento de hacer la alineación. La proteína seleccionada como referencia es la proteína más ordenada. Al inicio del análisis de cada hebra se muestra una tabla con las puntuaciones RMSD de cada interacción. Se tiene en la fila la proteína que fue utilizada como referencia y en cada fila las proteínas que fueron superpuestas. Para el análisis del grupo de la hebra 3'5 se tomó como referencia la proteína Q8N8C2 y para el grupo 5'3 se toma como referencia la proteína M1SZX7.

6.6.1. Alineación de cinco proteínas de la hebra 3'5

	B4E0H0	Q8WZ27	Q1M183	Q6ZQR8
Q8N8C2	8 pruned atom pairs is 0.902 angstroms; (across all 124 pairs: 26.924)	6 pruned atom pairs is 1.089 angstroms; (across all 19 pairs: 10.086)	6 pruned atom pairs is 1.461 angstroms; (across all 63 pairs: 30.905)	20 pruned atom pairs is 0.718 angstroms; (across all 102 pairs: 28.213)

Tabla 6.2: Tabla que muestra la puntuación RMSD obtenida luego de hacer la alineación estructural en Chimera de las proteínas seleccionadas del grupo de la hebra 3'5. Mientras más bajos sean los valores quiere decir que las proteínas tienen un mayor número de similitudes y viceversa.

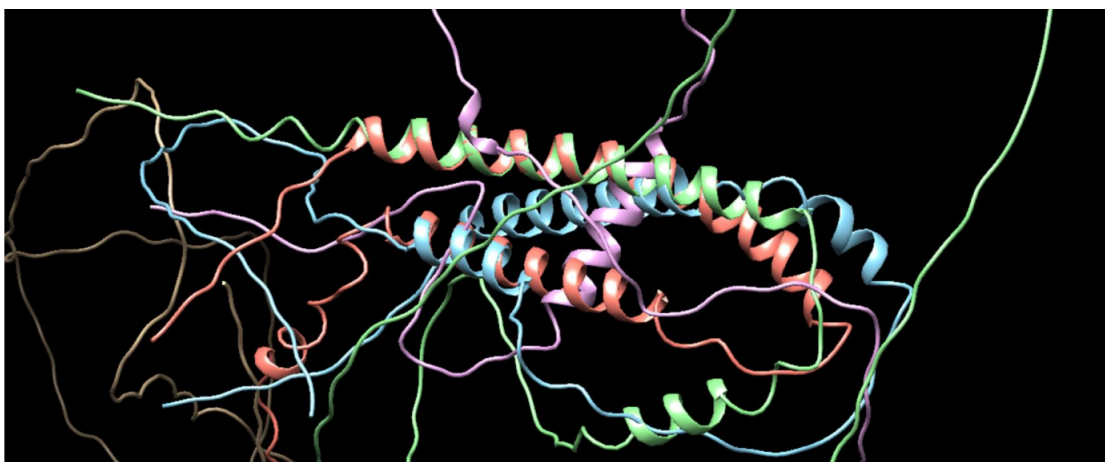


Figura 6.42: Vista cercana a la alineación en Chimera de proteínas Q8WZ27, Q8N8C2, Q6ZQR8, Q1M183, B4E0H0 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.4.



Figura 6.43: Vista total de la alineación en Chimera de proteínas Q8WZ27, Q8N8C2, Q6ZQR8, Q1M183, B4E0H0 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.4.

6.6.2. Alineación de las proteínas presentadas de la hebra 3'5

Luego de analizar cinco proteínas de cada grupo se seleccionaron solo las dos proteínas presentadas en la sección 6.5, es decir la más desordenada y ordenada. En la Figura 6.44 y 6.45 se observan las proteínas correspondientes al grupo de la hebra 3'5. Como referencia se utilizó la proteína más ordenada del grupo.



Figura 6.44: Vista completa de la alineación en Chimera de proteínas Q8N8C2 y Q6ZQR8 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.5.

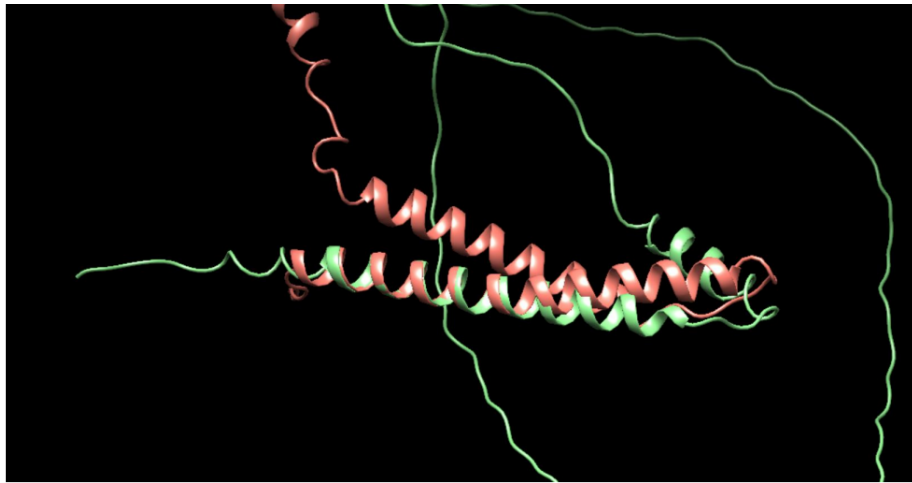


Figura 6.45: Vista cercana de la alineación en Chimera de proteínas Q8N8C2 y Q6ZQR8 haciendo referencia a Q8N8C2. El color correspondiente a cada proteína se muestra en la Figura 13.5.

6.6.3. Alineación de cinco proteínas de la hebra 5'3

	A0A0A1TSG9	Q9UI50	Q8WYX4	A2NVG1
M1SZX7	31 pruned atom pairs is 1.117 angstroms; (across all 80 pairs: 5.647)	11 pruned atom pairs is 1.133 angstroms; (across all 51 pairs: 10.961)	4 pruned atom pairs is 1.202 angstroms; (across all 39 pairs: 17.689)	7 pruned atom pairs is 1.396 angstroms; (across all 70 pairs: 12.931)

Tabla 6.3: Tabla que muestra la puntuación RMSD obtenida luego de hacer la alineación estructural en Chimera de las proteínas seleccionadas del grupo de la hebra 5'3. Mientras más bajos sean los valores quiere decir que las proteínas tienen un mayor número de similitudes y viceversa.

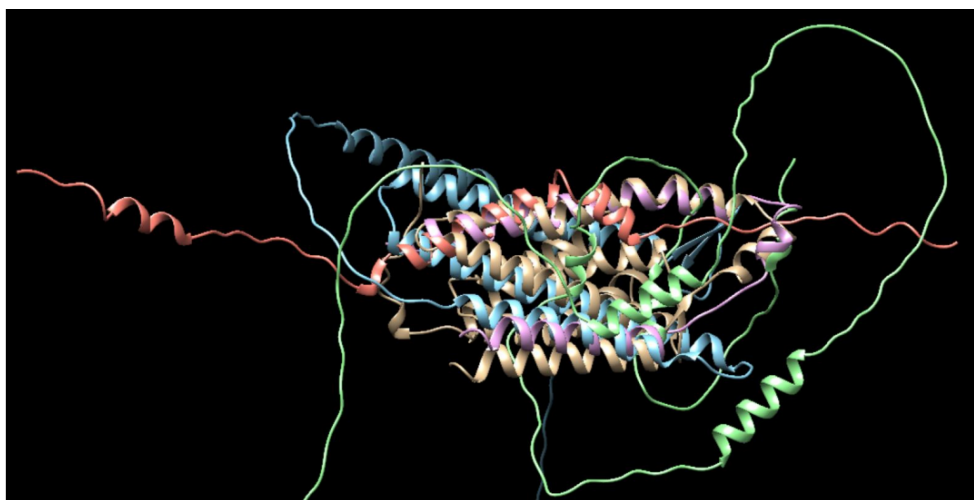


Figura 6.46: Vista completa de la alineación en Chimera de proteínas A0A0A1TSG9, A2NVG1, M1SZX7, Q8WYX4, Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.6.

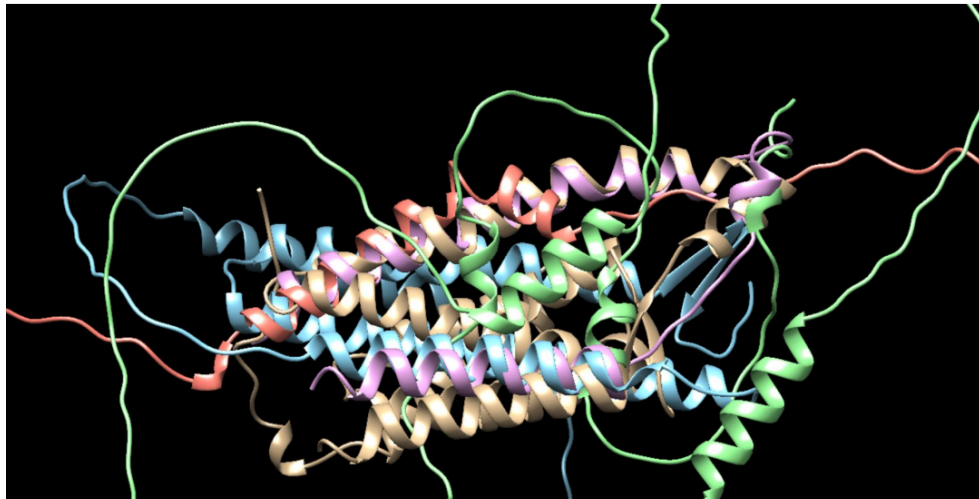


Figura 6.47: Vista cercana de la alineación en Chimera de proteínas A0A0A1TSG9, A2NVG1, M1SZX7, Q8WYX4, Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.6.

6.6.4. Alineación de las proteínas presentadas de la hebra 5'3

En la Figura 6.48 se muestran las proteínas resultantes expuestas en la sección 6.2 correspondientes al grupo de la hebra 5'3. Se presenta la proteína más desordenada y ordenada para compararlas estructuralmente. La proteína más ordenada se utilizó como referencia para realizar la alineación. Con esto se termina de analizar los resultados y se observan los cambios de cómo se mira la secuencia en teoría y estructuralmente.

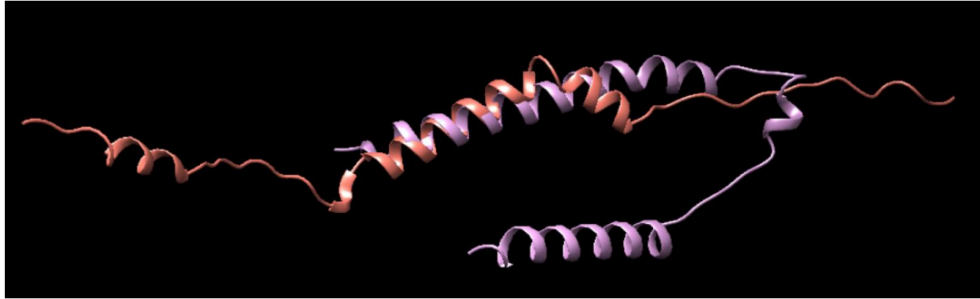


Figura 6.48: Alineación en Chimera de proteínas M1SZX7 y Q9UI50 haciendo referencia a M1SZX7. El color correspondiente a cada proteína se muestra en la Figura 13.7.

7.1. Transformación de gen codificante a secuencias proteicas

La recolección de los datos de glicoproteínas específicas a un grupo sanguíneo disponibles en bases de datos bioinformáticas inició al principio del estudio. Rápidamente se pudo notar que los datos específicos que se querían encontrar no estaban disponibles en ningún recurso. Por lo que se cambió el enfoque de búsqueda para encontrar un gen codificante de glicoproteínas relacionadas al grupo Rh. Este gen se encontró en un estudio de secuenciación del genoma humano.[JJ2020] Este cambio de enfoque permite utilizar herramientas de traducción de ADN así como tener dos grupos de proteínas, provenientes de la hebra 5'3 y 3'5 del ADN. Al analizar glicoproteínas relacionadas al factor Rh se determina el proceso que se debe de llevar con las secuencias pertenecientes a otros grupos sanguíneos y comprobar que pueden existir regiones conservadas. Este análisis bioinformático es una predicción de las regiones conservadas y desorden más tiene que ser comprobado experimentalmente.

La transformación gen a proteína fue realizada por la plataforma ExPasy. Podemos observar en las figuras 6.2 - 6.7, 6.9 - 6.14 las secuencias proteicas resultantes. Estas secuencias son de un tamaño bastante grande y están divididas por secciones. De cada uno de los dos genes utilizados para la transformación se obtuvieron las lecturas basadas en el sistema de seis cuadros. Hay tres cuadros por cada hebra del gen codificante, las secuencias de aminoácidos de cada uno son las diferentes lecturas que se puede tener de la secuencia de ADN del gen.[TutorialsTutorials] Estas secuencias contienen la información de las glicoproteínas que se codifican, por lo que son los datos de entrada para la búsqueda por pares.

7.2. Búsqueda BLAST para analizar los grupos de proteínas de cada hebra

Las figuras 6.15 - 6.20, 6.21 - 6.23 muestran las proteínas que fueron encontradas en cada uno de los análisis. La principal observación que se puede realizar es la diferencia en el tipo de secuencia que se encontró en cada hebra. En las búsquedas pertenecientes a la hebra 3'5 de ambos genes se encuentran proteínas como dedos de Zinc, los cuales son estabilizadores de pliegues. También hay proteínas transportadoras de aminoácidos y proteínas hipotéticas, es decir que no hay evidencia que existan pero se predice que son expresadas. Hace sentido que no se hayan encontrado glicoproteínas

en esta hebra ya que es la hebra molde del gen, esto también puede darse por no haber encontrado otras glicoproteínas con secuencias similares.[TutorialsTutorials]

Ahora en cambio cuando observamos los resultados de los cuadros en la hebra 5'3 hay bastante diferencia en el tipo de proteínas encontradas. Entre estas podemos encontrar subunidades de glicoproteínas del complejo de antígeno nulo del grupo sanguíneo R, glicoproteínas asociadas a Rh, glicoproteínas Rh50 de la membrana de los eritrocitos, glicoproteínas Rh 50, transportadores de amonio Rh tipo A isoforma X1 y glicoproteínas del complejo antígeno del grupo sanguíneo Rh. Este grupo de datos muestra un mejor enfoque al tipo de datos que se quieren analizar. El hecho que esta hebra 5'3 codificó este tipo de secuencias se debe a que es la hebra con ARN mensajero el cuál es el que participa en el proceso de transcripción. Esto se debe a que es la hebra informativa y la que muestra la funcionalidad. Por esto se encuentran otras glicoproteínas que probablemente tengan un ARN parecido que los lleva a tener secuencias similares. [TutorialsTutorials]

7.3. Formación de base de datos del estudio a partir del análisis de ambas hebras

Cuando se comparan los datos de las secuencias encontradas por BLAST en los diferentes *frames* es notorio que las secuencias se repiten. Lo único cambiante es el porcentaje de similitud por lo que se presentan en orden diferente. Como los *frames* son posibles lecturas de la misma secuencia de nucleótidos es normal que se identifiquen las mismas proteínas. Por esta razón se decide unir todos los *frames* pertenecientes a la misma hebra e iniciar la creación de la base de datos a estudiar mostrada en la Tabla 6.1. Estos códigos de identificación se buscaron en la plataforma Uniprot, los datos no pertenecían al conjunto de datos curados. Esto quiere decir que no se ha hecho una revisión de estas secuencias por lo que se decide también analizar sus regiones desordenadas.

7.4. Ajuste de posiciones en alineación Jalview, detección de regiones conservadas y desordenadas

Las secuencias cargadas en Jalview al ser alineadas mediante Clustal mostraron secuencias que se veían conservadas más no estaban alineadas. Como esta herramienta permite añadir y eliminar los espacios vacíos, que la plataforma crea para poder alinear las diferentes regiones, se alineó manualmente. La primera secuencia que aparece es la base de la alineación. Esto permitió alinear la secuencia de forma que la mayor cantidad de aminoácidos mostrara su región conservada. Uno de los primeros ajustes y el más importante fue alinear la primera posición de las secuencias. Las figuras 6.29 y 6.32 muestra que en ambos grupos de estudio se observa que la primera posición se conserva como una Metionina. Esto hace sentido ya que se sabe que el codón de inicio de las secuencias es representado por nucleótidos que identifican a este aminoácido.[TutorialsTutorials]

Mediante la coloración Clustal es evidente la alta cantidad que se tiene de Glicinas y Prolinas. Estos aminoácidos están en la agrupación de no polares. Las Glicinas en las proteínas tienen la característica de actuar como los aminoácidos flexibles.[NelsonNelson2017] Es decir que crean las bisagras necesarias para permitir el movimiento de las proteínas. Al ser este un aminoácido tan pequeño permite cambios en la conformación de la cadena polipéptida.[NelsonNelson2017] La Prolina en cambio se ha asociado con la rigidez de una proteína. La presencia de estos dos aminoácidos (Glicina y Prolina) no polares pueden asociarse al desorden que se observa en nuestras secuencias.[Francois-Xavier TheilletFrancois-Xavier Theillet2013] Estos aminoácidos se asocian a las regiones desordenadas por sus grupos R.[Francois-Xavier TheilletFrancois-Xavier Theillet2013] Vamos a analizar primero la alineación de la agrupación 3'5.

La proteína intrínsecamente desordenada analizada del grupo 3'5 fue Q6ZQR8, en la Figura 6.34 en la gráfica E se observa la región desordenada de la secuencia. Se puede observar que a partir de la posición 130 a la 240 se predice desorden. Según la alineación en la Figura 6.30 en estas últimas posiciones hay una alta concentración de Glicinas y Prolinas. Para comprobar si estos aminoácidos se relacionan con el desorden se puede ver la Figura 6.35 en la gráfica B. Este gráfico de barras indica que del porcentaje de Glicinas presentes aproximadamente un 30% está en desorden a comparación de un 20% de orden. Esto quiere decir que la Glicina si tiene un papel importante en el desorden. Las Prolinas no muestran un porcentaje mayor que este presente en el desorden. Según esta gráfica el aminoácido con más presencia en desorden son las Serinas. Este aminoácido polar según la teoría también está presente en desorden. Esto se debe a su baja hidrofobicidad, bajas restricciones en las conformaciones de la columna y la capacidad de interactuar fácilmente con el solvente [Francois-Xavier TheilletFrancois-Xavier Theillet2013]

La proteína menos desordenada de este grupo es la Q8N8C2, en la Figura 6.36 en la gráfica E se observa que la línea que predice el desorden está completamente bajo el nivel de 0.5. En la alineación mostrada en las figuras 6.29 y 6.30 se muestra que hay menos presencia de Glicina y Prolina. La terminal que fue la que mostró el desorden en la proteína anterior no presenta concentración de aminoácidos asociados al desorden. Por lo que se puede llegar a determinar que las proteínas con terminal con Glicinas, Prolinas y Serinas pueden mostrar altos grados de desorden.[Francois-Xavier TheilletFrancois-Xavier Theillet2013]

La primera de las regiones conservadas observadas en resultados se conforma de aminoácidos con coloración celeste (ver Tabla 13.8). Estos aminoácidos son polares y aromáticos. Aunque la conservación es del grupo en general de estos aminoácidos se observa una concentración de Fenilalanina y Leucina. Esto indica que esta región puede ser capaz de hacer interacciones hidrofóbicas con otras moléculas. Esta región termina con coloración verde, específicamente con Serinas las cuáles son polares. La segunda región conservada se muestra en la Figura 6.29. Hay más Glicinas y Prolinas presentes, específicamente al centro y final de la región. Los aminoácidos no polares y aromáticos indican estabilidad en la estructura terciaria. En la posición 223 también hay una alta concentración de Prolinas en la mayoría de las secuencias. Esto puede indicar rigidez en esta sección de las proteínas.[Francois-Xavier TheilletFrancois-Xavier Theillet2013]

En la Figura 6.30 se muestra la región conservada más grande de la hebra 3'5. Esta región también es predominada por la coloración celeste. Es decir que tiene aminoácidos hidrofóbicos. Por lo que cuenta con las mismas características. También hay presencia de Prolinas y una alta concentración de Glicinas en la posición 261, 280 y 283. En las regiones con coloración verde hay Serinas, Treoninas y Glutaminas.[NelsonNelson2017] También hay en coloración azul regiones con Histidinas, aminoácido con carga positiva. Estas colaboran a la síntesis de las proteínas, producción de glóbulos rojos y blancos así como relación con el sistema inmunológico. La región hidrofóbica es importante para las proteínas ya que es un punto de acceso para las mismas, por ejemplo pueden crear puentes de hidrógeno.[Eisenhaber1 ArgosEisenhaber1 Argos1996]

Con coloración azul también hay presencia de Tirosinas, aminoácido aromático. Esto puede indicar presencia de sitios activos y potencial para ligarse con otras moléculas.[Ramnath AnjanaRamnath Anjana2012] Otra de las características de esta región conservada, que también veremos en la siguiente es la coloración morada. Esta coloración indica presencia de ácido glutámico. La presencia de esto ayuda a la interacción con iones metálicos.[María D. López-LeónMaría D. López-León2006] Esto hace sentido ya que se habían encontrado varios dedos de Zinc en este grupo de proteínas. La cuarta región conservada de este grupo tiene los mismos tipos de aminoácidos conservados. Es una región polar e hidrofóbica. En las terminales mostradas en la Figura 6.31 se observa una alta concentración de Glicinas y Prolinas. Con esto se podría concluir que las proteínas tendrán regiones desordenadas al final de sus secuencias.[Francois-Xavier TheilletFrancois-Xavier Theillet2013]

Ahora se analizará el grupo de proteínas en la hebra 5'3. La proteína más desordenada de este grupo es la Q9UI50 en la Figura 6.38 en la gráfica E se observa el análisis del desorden. En la Figura

6.32 se observa que es una secuencia relativamente corta cuando la comparamos con las demás. El orden se presenta en toda la secuencia menos de la posición aproximadamente 25 a 40. Los aminoácidos que presentan un mayor porcentaje en regiones de desorden se presentan en la Figura 6.39 gráfico B. La Arginina y ácido glutámico son los aminoácidos más desordenados en esta proteína. El ácido glutámico está en desorden por ser un aminoácido ácido.[CEDILLOCEDILLO2021] En la alineación podemos ver que ambos están dispersos por toda la secuencia, lo cual se comprueba en el desorden de toda la proteína. Según la gráfica de aminoácidos en desorden se puede ver que el porcentaje mayor en todos los casos está en región desordenada. Lo cual se asocia con el análisis anterior de los aminoácidos presentes en desorden.

La proteína más ordenada correspondiente a la hebra 5'3 es M1SZX7, en la Figura 6.40 gráfica E se observa que la línea del desorden es prácticamente cero en toda la secuencia. Esto quiere decir que todos los aminoácidos deberían encontrarse en orden. En la Figura 6.41 gráfica B comprobamos que todas las barras están de color rojo y el código de color indica que eso es orden. El programa de R que genera estos gráficos coloca el color rojo como orden en este caso así como en la proteína ordenada Q8N8C2 ya que no hay regiones desordenadas. En el análisis de la proteína ordenada de la hebra 3'5 se vio que era la secuencia más corta del grupo. En la Figura 6.32 vemos que este fenómeno también se cumple para M1SZX7. Aunque hay aminoácidos asociados al desorden, es decir Prolinas y Glicinas, estas regiones desordenadas no existen. Esto significa que la presencia de estos aminoácidos no quiere decir que habrá desorden si no que en las regiones desordenadas estarán estos aminoácidos.[M.M.2006]

En las secuencias de este grupo vemos una alta conservación de Leucinas, Metioninas, Isoleucinas, Valinas y Fenilalaninas. Estos son aminoácidos no polares y la Fenilalanina aromática, esto hace esta región conservada hidrofóbica. Se tiene un comportamiento bastante parecido al análisis de la hebra 3'5. Podemos ver la coloración verde más presente, especialmente de la Treonina y Serina. Estos aminoácidos se relacionan con la capacidad de la proteína de tener sitios de anclaje de carbohidratos. Esto ayuda a la glicosilación que en las glicoproteínas presentes es bastante importante. Otro aminoácido muy presente es el ácido glutámico que a comparación con la hebra 3'5 está más frecuente. Vemos también Glicinas y Prolinas bastante alineadas, esto puede indicar regiones desordenadas pero se deben analizar las gráficas de desorden IUPred ya que no es el único factor. Hay proteínas en este grupo que son prácticamente la misma secuencia desde el inicio hasta el final. Esto es buen indicio del estudio ya que por la forma en la que las proteínas fueron seleccionadas se nota que estas regiones seguirán conservadas. En la Figura 6.33 se muestran las terminales de las secuencias, hay alta conservación de Prolinas y Glicinas. Como se mencionó anteriormente esto puede indicar desorden de la terminal.

7.5. Calidad de modelados por técnicas de homología y su relación con el desorden

La proteína más desordenada de la hebra 3'5 muestra su estructura y gráficas de AlphaFold en la Figura 6.34. Podemos ver en la gráfica A que la estructura tiene hélices alfa y secciones aparentemente desordenadas en sus extremos. Cuando vemos la gráfica de PAE esto hace sentido, ya que no muestra distancias cortas entre los aminoácidos más que un leve cambio en donde se asume están las hélices. Podemos observar en la gráfica C que solamente 100 posiciones aproximadamente se ven cubiertas en la estructura. La calidad del modelo se muestra en la gráfica D, ninguna sección muestra un puntaje por arriba de 90 por lo que no se considera una estructura con grado alto de confianza. El modelo mejor rankeado es el celeste y se observa que el puntaje más alto apenas llega a 70. Luego se tienen secciones con un grado de confianza bajo y la mayoría está con menor a 50. Esto quiere decir que es un modelado con nivel de confianza bajo. En la gráfica C de la Figura 6.35 podemos ver la relación que tiene este análisis con el desorden. En general podemos ver que la gráfica pLDDT sube cuando el desorden baja y viceversa. Por lo que el desorden si afecta la calidad del modelado.

Esto se debe a la conformación no fija que tienen las regiones desordenadas, por lo que AlphaFold no puede predecirlas.[M.M.2006]

Al analizar esta misma gráfica C en la Figura 6.37 de la proteína más ordenada de la hebra 3'5 vemos que se tiene un comportamiento similar. Es decir que mientras más baja la gráfica del desorden más sube la de la calidad del modelado. En la Figura 6.36 gráfica D vemos que la calidad del modelo celeste bastante buena en la mayoría de la región. Ambas cúspides se encuentran en una puntuación de 50-70 confianza baja (pequeña región) y 70-90 confianza buena (la mayoría de la región) lo que indica que es un modelo apto para usarse, siempre analizándolo adecuadamente.[PhDPhD2022] Podemos ver en la gráfica A que la estructura tridimensional tiene dos hélices alfa paralelas y en los extremos estructuras aparentemente desordenadas. Aunque estos extremos parecen no tener estructura definida la gráfica E del desorden confirma que son ordenadas. La gráfica B muestra la distancia corta que se tiene entre las dos hélices alfa como dos pequeñas áreas azules. La gráfica C indica una buena y alta cobertura de la secuencia en la estructura.

La proteína más desordenada de la hebra 5'3 se encuentra en la Figura 6.38, la gráfica A muestra que tiene una estructura con hélices alfa de forma bastante lineal, aunque hay dos secciones de las hélices cercanas. El gráfico B confirma esta cercanía de las hélices ya que hay áreas azules presentes. La gráfica C también indica una buena cobertura de la secuencia en la estructura. Por último calificaremos la calidad del modelado según la gráfica D, prácticamente toda la secuencia se encuentra arriba de 50 por lo que si puede usarse. Solo una pequeña cúspide de la gráfica indica un puntaje mayor a 70.[PhDPhD2022] Cuando comparamos esta gráfica con la del desorden en la gráfica C de la Figura 6.39 tiene el comportamiento esperado. La cúspide con mejor puntaje de calidad es la región de la proteína que más ordenada está. En cambio las regiones desordenadas muestran un mal modelado.

La gráfica C de la Figura 6.41 muestra el análisis de la proteína más ordenada de la hebra 5'3 y comprueba el comportamiento esperado, con esto podemos confirmar que mientras más desorden exista en una región de más baja calidad será su modelado. Esto se comprueba con la teoría del desorden la cual dice que estas regiones tienen dificultad encontrando su estructura fija.[M.M.2006] En la Figura 6.40 gráfica D podemos ver que el modelado se encuentra mayoritariamente en una puntuación de 70 a >90. Es un modelo con un alto grado de confianza, justamente es el modelo que analizamos con una puntuación de desorden de prácticamente cero.[PhDPhD2022] La gráfica A nos muestra una estructura con forma de ù formada por dos hélices alfa. La gráfica B muestra que en efecto estas hélices se encuentran bastante cercanas y son prácticamente toda la estructura. La cobertura que se muestra en la gráfica C es excelente, incluye prácticamente todo.[PhDPhD2022]

7.6. Diferencias estructurales entre las proteínas

A parte de la alineación de secuencias también se realizó la alineación de las estructuras tridimensionales de un grupo seleccionado de cada hebra. En las figuras 6.42 y 6.43 se tienen cinco proteínas alineadas. La Figura 6.2 nos muestra la puntuación a base de la similitud entre proteínas, todas se comparan con Q8N8C2 por ser la más ordenada de las cinco. Podemos ver que la puntuación RMSD está entre 6 y 8 a excepción de la alineación con Q6ZQR8. Estos números no son tan buenos indicadores ya que un RMSD con valor menor a 4 es señal de proteínas similares.[LiebeschuetzLiebeschuetz2015] A pesar de tener un valor RMSD alto se observan regiones alineadas y conservadas. Esto quiere decir que la conservación no siempre se ve reflejada estructuralmente.[PietrokovskiPietrokovski1996] En las imágenes podemos ver que hay algunas secciones "sobrantes". Esto es algo esperado ya que no todas las secuencias son del mismo tamaño como se vio en las imágenes de su alineación en Jalview. La proteína que más difiere en su puntaje RMSD es la proteína más desordenada de este grupo. La alineación de solo estas dos puede verse en las figuras 6.44 y 6.45. Aunque el RMSD fue de 20, un poco más alto a las demás, podemos ver que estructuralmente si se tiene una hélice alfa prácticamente igual.

En las figuras 6.46 y 6.47 vemos la alineación de las proteínas correspondientes a la hebra 5'3. Los puntajes RMSD en la Figura 6.3 muestran valores más diversos entre 4 y 31. Las tres proteínas que tienen más similitud a la de referencia (M1SZX7) tienen valores de 4, 7 y 11. Solo la proteína Q8WYX4 tiene una similitud estructural aceptable con un valor RMSD de 4.[LiebeschuetzLiebeschuetz2015] Esta alineación tiene más hélices presentes por lo que hace la vista un poco más obstruida que en el análisis anterior. La proteína que muestra mayor diferencia es la A0A0A1TSG9, su modelado se encuentra en la Figura 13.27. Este modelado muestra una proteína altamente ordenada y con una calidad muy buena de modelado. En la gráfica B (PAE) de esta figura podemos ver que prácticamente toda el área de un color azul.[PhDPhD2022] Esto nos indica que hay mucha cercanía entre aminoácidos y esto es comprobable al ver la gráfica A. Esto dificulta la alineación con otras proteínas, en la alineación de Jalview (6.32) se observa que la secuencia si tiene regiones conservadas. Se demuestra que la conservación no siempre se observa de forma estructural. La Figura 6.48 se alinea la proteína más desordenada y ordenada del grupo. Podemos ver que la puntuación RMSD es de 11, lo cual no indica una similitud significativa. En la alineación (Figura 6.32) vemos que las secuencias no tienen las mismas regiones conservadas. Por lo que hace sentido que no tengan similitudes significativas.

7.7. Creación y análisis de árboles filogenéticos

Los dos árboles filogenéticos, uno de la hebra 5'3 y otro de la 3'5 del ADN se muestran en las figuras 6.27 y 6.28 respectivamente. La cercanía de las secuencias entre ellas indica una alta relación evolutiva. Se analizaron las secuencias tomando en cuenta las regiones conservadas encontradas y su cercanía evolutiva en el árbol. Primero se discutirán las relaciones evolutivas de las proteínas en la hebra 3'5. En la alineación de Jalview mostrada en la Figura 6.29 hay tres proteínas con prácticamente la misma secuencia. Estas son P10242, Q70AC3 y Q708E2. En el árbol se puede ver que estas proteínas están en la rama más alejada y su ancestro en común más cercano está a poca distancia. Esto hace que se pueda relacionar la similitud de sus secuencias con su evolución. Ya que las regiones conservadas no cambian por su importancia en la funcionalidad.[PietrokovskiPietrokovski1996]

Dos de las proteínas que no tienen la primera región conservada son Q9P195 y Q6ZNX6. Según la alineación de Jalview estas secuencias no tenían mucho en común, al buscarlas en el árbol filogenético se pudo determinar que son dos proteínas que surgen del mismo nodo. Cuando se volvió a ver sus secuencias se notaron las similitudes que estas tienen en esa región. Las otras regiones conservadas se mantuvieron, por lo que se puede determinar que la primera región conservada no es tan esencial como las otras.[PietrokovskiPietrokovski1996] La proteína A0A024RC64 es una proteína cuya secuencia es más corta que las demás por lo que no llega a la última región conservada. Sin embargo puede analizarse su posición en el árbol y ver que del nodo más cercano surgen proteínas que si tienen esta región conservada. Estos análisis de las secuencias pueden realizarse con todas las proteínas con comportamientos inusuales y que se quieran comprender de donde vienen.[LageLage2013]

En la Figura 6.28 se muestra el árbol filogenético de las proteínas correspondientes a la hebra 5'3. La primera observación que es interesante ver en el árbol son las secuencias seleccionadas en la Figura 7.1. Podemos ver que las secuencias son prácticamente iguales. En el árbol tenemos un nodo que tiene a 9 proteínas, estas proteínas son las más similares en sus secuencias. Estas secuencias no son las únicas seleccionadas lo que significa que esta región se conservó por otras generaciones como pueden verse en las proteínas de ramas vecinas. Las proteínas que tienen menos similitud son la M1SZX7 y Q9UBB8. Estas proteínas aunque vienen del mismo ancestro que las demás seleccionadas se separan desde el inicio y se quedaron más alejadas. Esto da como indicio que mientras más se alejan las proteínas las regiones conservadas pueden perderse y van quedando solo las más esenciales. A partir del árbol filogenético construido se observa que la mayoría de las proteínas, menos 1 surgen del mismo nodo. Esto explica la alta cantidad de regiones conservadas que encontramos e indica que son regiones esenciales para el funcionamiento de estas proteínas.[LageLage2013]

Se puede encontrar un trabajo similar en la investigación *Intrinsic disorder in PRAME and its role in uveal melanoma*. En este trabajo se realiza un análisis del desorden similar al de nuestro estudio, se utilizan proteínas de la base de datos Uniprot y se analiza el desorden en IU-Pred. También se analizan las características y calidad de la estructura por el algoritmo de Alpha-Fold. Concluyen acerca de las regiones desordenadas y posibles funciones que pueden tener en la estructura.[Antonietti .Antonietti .2023]

Con este estudio se puede llegar a generar un análisis de las consecuencias que pueden tener los cambios en las regiones conservadas de las proteínas. También se puede analizar la forma en que el desorden contribuye a la función de las proteínas. Es interesante analizar si la edición de este desorden y conservación tiene consecuencias graves como lo es en la proteína p53. Ya que se comprobó que las proteínas correspondientes al grupo Rh están conservadas puede hacerse este mismo análisis con proteínas presentes en los eritrocitos de cualquier grupo sanguíneo.

- El desorden en las secuencias de las proteínas influye en la calidad de modelado que se realiza por técnicas de homología.
- Se encontraron regiones conservadas entre las proteínas, estas se conservan a través de varios nodos de árboles filogenéticos por lo que se concluye que estas regiones son esenciales para el funcionamiento.
- No existen similitudes estructurales significativas según los valores RMSD generados por Chimera. Por lo que la conservación no indica similitud estructural.
- Las regiones conservadas de los grupos de proteínas son hidrofóbicas, polares y aromáticas.
- Se encontraron funciones importantes en estas regiones como sitios activos y anclaje de carbohidratos que contribuyen a la glicosilación de las glicoproteínas.
- Las regiones desordenadas se asocian con la presencia de aminoácidos Prolinas, Glicinas y Serinas.

- Documentar cada paso del proceso que se va realizando, especialmente de los códigos de proteínas. Los códigos llegan a ser difíciles de memorizar por lo que es útil tener varios archivos para ir escribiendo que se realiza.
- Tener los archivos y programas a utilizar ordenados. Se recomienda tener una carpeta por cada parte del análisis. Las carpetas recomendadas son AlphaFold, árboles, BLAST, FASTA, IUPred, Jalview. También es útil tener un archivo de texto donde se apunten las observaciones preliminares de los resultados.
- Se recomienda tener la teoría a la par de los resultados para asociar teóricamente lo que se observa y comprobar si hace sentido el resultado.
- Para estudios futuros se recomienda analizar las regiones conservadas en una alineación estructural según las proteínas relacionadas en los árboles filogenéticos.
- Analizar la proteína más ancestral con la más lejana en el árbol para analizar que tanto cambian las regiones conservadas y el desorden por la evolución.

CAPÍTULO 12

Bibliografía

Bibliografía

[60056005] GEN_NIH6005, *GI. Bethesda (MD) : National Library of Medicine (US), National Center for Biotechnology*

arboles2Academy, K. . Árboles filogenéticos. Árboles filogenéticos. <https://es.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees>

AlphaFoldAlphaFold. . AlphaFold Protein Structure Database. Alphafold protein structure database. <https://alphafold.ebi.ac.uk>

antecedenteAntonietti, M., Taylor, DJ., Djulbegovic, M., Dayhoff, GW., Uversky, VN., Shields, CL. Karp, CL. 2023. Intrinsic disorder in PRAME and its role in uveal melanoma Intrinsic disorder in prame and its role in uveal melanoma. *Cell Communication and Signaling*211222.

BatemanAoki, T. 20171. A Comprehensive Review of Our Current Understanding of Red Blood Cell (RBC) Glycoproteins A comprehensive review of our current understanding of red blood cell (rbc) glycoproteins. *PMB*. <https://doi.org/10.3390>

AokiAoki, T. 20172. A Comprehensive Review of Our Current Understanding of Red Blood Cell (RBC) Glycoproteins A comprehensive review of our current understanding of red blood cell (rbc) glycoproteins. *PMB*. <https://doi.org/10.3390>

expasyteoriaArtimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E.others 2012. ExpASy: SIB bioinformatics resource portal Expasy: Sib bioinformatics resource portal. *Nucleic acids research*40W1W597–W603.

RhReviewAvent, ND. Reid, ME. 2000. The Rh blood group system: a review The rh blood group system: a review. *Blood, The Journal of the American Society of Hematology*952375–387.

BenderBender, DA. 2019. Glucoproteínas Glucoproteínas. *Harper Bioquímica ilustrada, 31e*. Harper bioquímica ilustrada, 31e. New York, NYMcGraw-Hill Education. [accessmedicina.mhmedical.com/content.aspx?aid=1166872015](https://www.accessmedicina.mhmedical.com/content.aspx?aid=1166872015)

ByjuByju. . Blood Groups-ABO Blood Group and Rh Group System. Blood groups-abo blood group and rh group system. <https://byjus.com/biology/blood-groups/>

acidoglu2CEDILLO, ASS. 2021. Análisis in silico de la interacción entre un péptido de la sialoproteína ósea y el cristal hidroxapatita Análisis in silico de la interacción entre un péptido de la sialoproteína ósea y el cristal hidroxapatita. *CENTRO DE INVESTIGACION Y ESTUDIOS AVANZADOS DEL INSTITUTO POLITECNICO NACIONAL*.

jalviewclustalClamp, M., Cuff, J. Barton, G. 1998. JalView–analysis and manipulation of multiple sequence alignments Jalview–analysis and manipulation of multiple sequence alignments. *EMBnet News*5416–21.

AnemiaSíntomasycausasClinic, M. 20221. Anemia Síntomas y causas. Anemia síntomas y causas. <https://www.mayoclinic.org/es-es/diseases-conditions/anemia/symptoms-causes/syc-20351360>

TransfusióndesangreClinic, M. 20222. Transfusión de sangre. Transfusión de sangre. <https://www.mayoclinic.org/es-es/tests-procedures/blood-transfusion/about/pac-20385168>

UniprotConsortium, TU. 201811. UniProt: a worldwide hub of protein knowledge UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research47D1D506-D515. <https://doi.org/10.1093/nar/gky1049> 10.1093/nar/gky1049

OMSAnemiade la Salud, OM. . Anemia. Anemia. https://www.who.int/es/health-topics/anaemia#tab=tab_1

RhfactorexplainedDonation, RCB. . Rh Factor Explained. Rh factor explained. <https://www.redcrossblood.org/local-homepage/news/article/what-is-the-rh-factor--why-is-it-important-.html>

blastteoriaDonkor, ES., Dayie, N. Adiku, TK. 2014. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA) Bioinformatics with basic local alignment search tool (blast) and fast alignment (fasta). Journal of Bioinformatics and sequence analysis611–6.

iupredteoriaDosztányi, Z. 2018. Prediction of protein disorder based on IUPred Prediction of protein disorder based on iupred. Protein Science271331–340.

ExpasyE., G. 2003. ExpASy: the proteomics server for in-depth protein knowledge and analysis Expasy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res.. 31:3784-3788

hydrofEisenhaber1, F. Argos, P. 1996. Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. Protein Engineering912.

PID2Francois-Xavier Theillet, LK. 2013. The alphabet of intrinsic disorder The alphabet of intrinsic disorder. Intrinsically Disord Proteins11.

MacacoRhesusGeographic, RN. . Macaco Rhesus. Macaco rhesus. <https://www.nationalgeographic.es/animales/macaco-rhesus>

GoodwinGoodwin, M. 2021. Tipos de sangre: ¿cuáles son y qué significan? Tipos de sangre: ¿cuáles son y qué significan? <https://www.medicalnewstoday.com/articles/es/tipos-de-sangre#abo-y-tipos-comunes>

bioinformaticaHagen, JB. 2000. The origins of bioinformatics The origins of bioinformatics. Nature Reviews Genetics13231–236.

codigoetraJ., E. 1984. IUPAC-IUB Joint Commission on Biochemical Nomenclature.Nomenclature and Symbolism for Amino Acids and Peptides Iupac-iub joint commission on biochemical nomenclature.nomenclature and symbolism for amino acids and peptides. Biochem13839-37.

PaperDatosJ, P. 2020. Blood group typing from whole-genome sequencing data Blood group typing from whole-genome sequencing data. PLoS ONE.

alphafoldteoJones, DT. Thornton, JM. 2022. The impact of AlphaFold2 one year on The impact of alphafold2 one year on. Nature methods19115–20.

SarahKnappKnapp, S. . Glycoprotein. Glycoprotein. <https://biologydictionary.net/glycoprotein/>

arboles1Lage, AM. 2013. APLICACIONES DE LA BIOINFORMÁTICA EN LA ELABORACIÓN DE FILOGENIAS MOLECULARES Aplicaciones de la bioinformática en la elaboración de filogenias moleculares. CINVESTAV.

rmsdpostLiebeschuetz, J. 201505. What is the importance of the RMSD value in molecular docking? What is the importance of the rmsd value in molecular docking? <https://www.researchgate.net/post/What-is-the-importance-of-the-RMSD-value-in-molecular-docking/55677ac85f7f7126fa8b4591/citation/download>

ionizacionLoo, JA., Udseth, HR. Smith, RD. 1989. Peptide and protein analysis by electrospray ionization-mass spectrometry and capillary electrophoresis-mass spectrometry Peptide and protein analysis by electrospray ionization-mass spectrometry and capillary electrophoresis-mass spectrometry. *Analytical Biochemistry*1792404-412. <https://www.sciencedirect.com/science/article/pii/000326978990153X> [https://doi.org/10.1016/0003-2697\(89\)90153-X](https://doi.org/10.1016/0003-2697(89)90153-X)

PIDM., BM. 2006. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society transactions*4451185–1200. <https://doi.org/10.1042/BST20160172>

acidoglutamicoMaría D. López-León, MDG., Paloma Arranz. 2006. Síntesis, Caracterización y Estudio en Disolución del Ácido N-2-(4-amino-1,6-dihidro-1-metil-5-nitroso-6-oxopirimidinil)- L-Glutámico Síntesis, caracterización y estudio en disolución del Ácido n-2-(4-amino-1,6-dihidro-1-metil-5-nitroso-6-oxopirimidinil)- l-glutámico. Universidad de Jaén. <https://doi.org/10.1371/journal.pone.0242168>

rmsdMeng, EC., Pettersen, EF., Couch, GS., Huang, CC. Ferrin, TE. 2006. Tools for integrated sequence-structure analysis with UCSF Chimera Tools for integrated sequence-structure analysis with ucsf chimera. *BMC bioinformatics*71–10.

NelsonNelson, D. 2017. Lehninger principles of biochemistry Lehninger principles of biochemistry. WH Freeman.

BLASTNIH. . BLAST: Basic Local Alignment Search Tool. Blast: Basic local alignment search tool. <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

clustalPei, J. 2008. Multiple protein sequence alignment Multiple protein sequence alignment. *Current Opinion in Structural Biology*183382-386. <https://www.sciencedirect.com/science/article/pii/S0959440X08000407> *Nucleic acids / Sequences and topology* <https://doi.org/10.1016/j.sbi.2008.03.007>

pdbchimeraPeitsch, MC. 1995. Protein modeling by E-mail Protein modeling by e-mail. *Bio/technology*137658–660.

chimerateoriaPettersen, EF., Goddard, TD., Huang, CC., Couch, GS., Greenblatt, DM., Meng, EC. Ferrin, TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*25131605–1612.

paeplddtPhD, SL. 2022. Explained: how to plot the prediction quality metrics with AlphaFold2. Explained: how to plot the prediction quality metrics with alphafold2. <https://blog.biostrand.be/explained-how-to-plot-the-prediction-quality-metrics-with-alphafold2>

conservedPetrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic acids research*24193836–3845.

aromaticRamnath Anjana, MKV. 2012. Aromatic-aromatic interactions in structures of proteins and protein-DNA complexes: a study based on orientation and distance Aromatic-aromatic interactions in structures of proteins and protein-dna complexes: a study based on orientation and distance. *Bioinformation*824.

ColinReily, C. 2019. Glycosylation in health and disease Glycosylation in health and disease. *Nat Rev Nephrol*. <https://doi.org/10.1038/s41581-019-0129-4>

arboles3Saitou, N. Nei, M. 198707. The neighbor-joining method: a new method for reconstructing phylogenetic trees. The neighbor-joining method: a new method for reconstructing phylogenetic

trees. *Molecular Biology and Evolution* 44:406-425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>

aloantibodies after transfusion: factors influencing incidence and specificity Red blood cell alloantibodies after transfusion: factors influencing incidence and specificity. *Transfusion* 46:250-256.

glycoproteins Spiro, RG. 1973. Glycoproteins. *Glycoproteins*. C. Anfinsen, JT. Edsall FM. Richards (), (27, 349-467). Academic Press. <https://www.sciencedirect.com/science/article/pii/S0065323308604519> [https://doi.org/10.1016/S0065-3233\(08\)60451-9](https://doi.org/10.1016/S0065-3233(08)60451-9)

Stojanovic Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M. Hardison, R. 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research* 27:193899-3910. <https://doi.org/10.1093/nar/27.19.3899>

Albert Team, A. . Proteins: AP[®]. Proteins: Ap[®]. <https://www.albert.io/blog/proteins-ap-biology-crash-c>

traduccion Tutorials, O. . Converting DNA to Protein Sequence. Converting dna to protein sequence. <https://omicstutorials.com/convert-dna-to-protein-sequence/#:~:text=Transcription%20and%20Translation&text=In%20transcription%2C%20the%20information%20in,read%27%20to%20make%20specific%20proteins>.

p53 Uversky, VN. 2016. p53 proteoforms and intrinsic disorder: An illustration of the protein structure-function continuum concept. *International journal of molecular sciences* 17:111874.

Waterhouse Waterhouse, A. . Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Jalview Home Page - Jalview. Jalview version 2 - a multiple sequence alignment editor and analysis workbench. jalview home page - jalview. <https://www.jalview.org>

homologia Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R. Schwede, T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* 46:W1W296-W303. <https://doi.org/10.1093/nar/gky427>

Rh Structure Westhoff, CM. 2007. The structure and function of the Rh antigen complex. *Seminars in hematology* 44: 42-50.

13.1. Parámetros para los análisis en las distintas plataformas.

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

Please enter a DNA or RNA sequence - numbers and blanks are ignored

Output format

- Verbose: Met, Stop, spaces between residues
- Compact: M, -, no spaces
- Includes nucleotide sequence
- Includes nucleotide sequence, no spaces

DNA strands

forward reverse

Genetic codes - [See NCBI's genetic codes](#)

Standard

reset TRANSLATE!

Figura 13.1: Parámetros utilizados durante el procesamiento de datos en Expasy.
[E.E.2003]

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [?](#)

MFPGCFQQNFRNEFGKAGM-SGWVR-MFYSFIFLQKQFLFPVLSKLE--
 LPKIFLFQY-FLTRTVK-TIKCPQTCFFKGP-DEIT-QSVPWTW-PQAGT-
 SLNTIPALQLPFLSRKSK-TKRE-PKRVWTRVNRNRS-
 VPNGQCKNVLMSDAVSCSSMIN-GDVTHTKTDNQIDLLQDSQKAFGSE-

From To

Or, upload file Ninguno archivo selec. [?](#)

Job Title
 Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Databases Standard databases (nr etc.): New Experimental databases [Try experimental clustered nr database](#) [?](#)
 For more info see [What is clustered nr?](#)

Compare Select to compare standard and experimental database [?](#)

Standard

Database [?](#)

Organism exclude
 Optional Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences
 Optional

Program Selection

Algorithm Quick BLASTP (Accelerated protein-protein BLAST)
 blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
 Choose a BLAST algorithm [?](#)

Figura 13.2: Parámetros utilizados durante el procesamiento de datos en BLAST. [NIHNIH]

Protein Sequence

Enter SWISS-PROT/TrEMBL identifier or accession number:

or paste the amino acid sequence:

or provide your email address and upload a (multi)FASTA file (max 1MB):
 Email: Ninguno archivo selec.

Prediction type:

IUPred2 long disorder (default)
 IUPred2 short disorder
 IUPred2 structured domains
 Context-dependent predictions:
 ANCHOR2
 Redox state (experimental)

Figura 13.3: Parámetros utilizados durante el procesamiento de datos en IUPred2A. [NIHNIH]



Figura 13.4: Parámetros de las proteínas alineadas en Chimera de la hebra 3'5 para el análisis de las figuras 6.42 y 6.43.

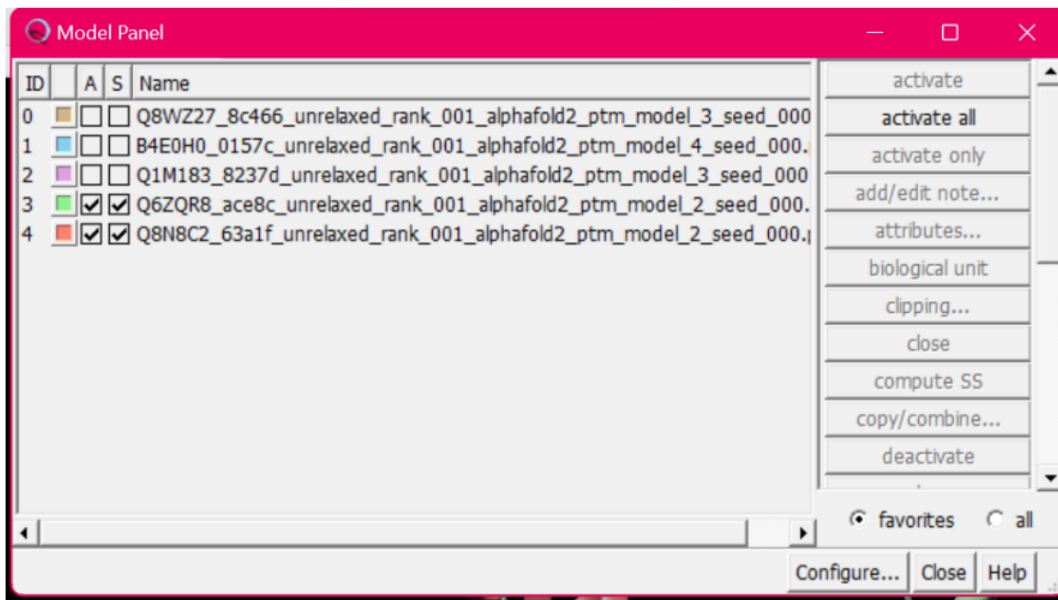


Figura 13.5: Parámetros de las proteínas alineadas en Chimera de la hebra 3'5 para el análisis de las figuras 6.44 y 6.45.



Figura 13.6: Parámetros de las proteínas alineadas en Chimera de la hebra 5'3 para el análisis de las figuras 6.46 y 6.47.

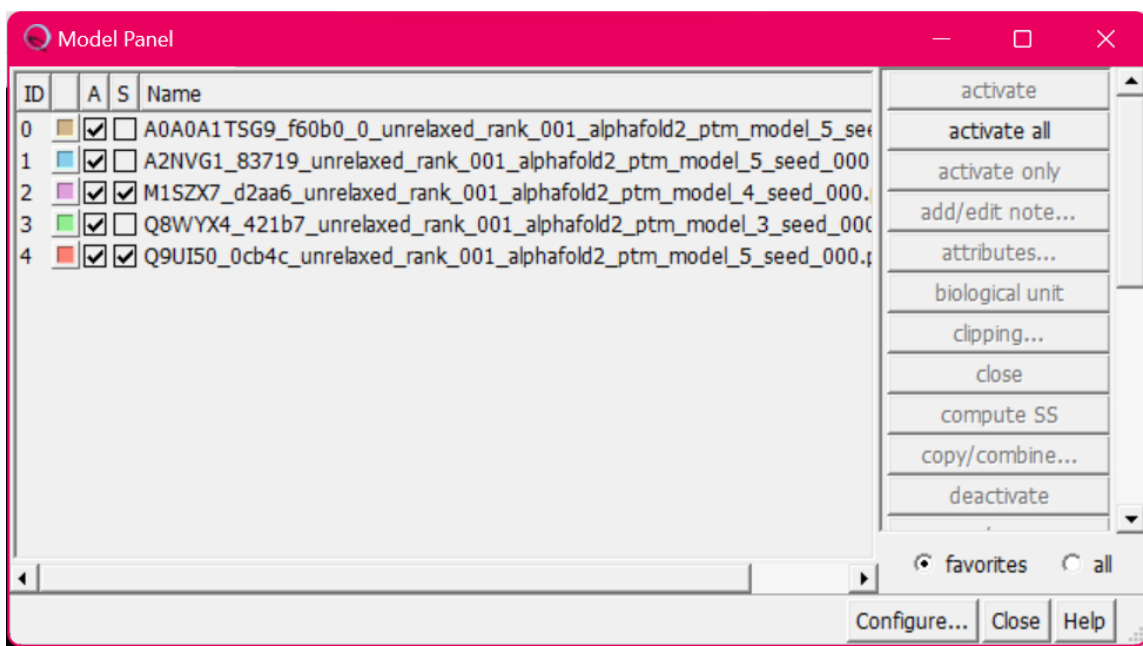


Figura 13.7: Parámetros de las proteínas alineadas en Chimera de la hebra 5'3 para el análisis de la Figura 6.48.

Clustal X Default Colouring		
Residue at position	Applied Colour	{ Threshold, Residue group }
A,I,L,M,F,W,V	BLUE	{+60%. WLVIAMFCHP}
R,K	RED	{+60%.KR},{+80%. K.R.Q}
N	GREEN	{+50%. N}. {+85%. N.Y}
C	BLUE	{+60%. WLVIAMFCHP}
C	PINK	{100%. C}
Q	GREEN	{+60%.KR},{+50%.QE},{+85%.Q.E.K.R}
E	MAGENTA	{+60%.KR},{+50%.QE},{+85%.E.Q.D}
D	MAGENTA	{+60%.KR}, {+85%. K.R.Q}, {+50%.ED}
G	ORANGE	{+0%. G}
H,Y	CYAN	{+60%. WLVIAMFCHP}, {+85%. W,Y.A.C.P.Q.F.H.I.L.M.V}
P	YELLOW	{+0%. P}
S,T	GREEN	{+60%. WLVIAMFCHP}, {+50%. TS}, {+85%.S.T}

Figura 13.8: Código de color de los aminoácidos para analizar la alineación de múltiples proteínas en JalView según coloración Clustal.

```

1 library(ggplot2)
2 library(bio3d)
3
4 #Q6Z0R8
5 setwd("/Users/mjnre/Documents/")
6
7 p53 <- read.csv(file="/Users/mjnre/Documents/ejemplo.iupred", header=F, sep="\t", col.names=c("Posición","Aminoácido","Iupred","Anchor"), comment.char="#")
8
9 umbral <- 0.5
10 p53Prediccion <- NA
11 p53Prediccion[p53Iupred<umbral] <- "Desorden"
12 p53Prediccion[p53Iupred>umbral] <- "Orden"
13
14 plot_p53 <- ggplot(p53, aes(x=Posición,y=Iupred)) +
15   scale_x_continuous(n.breaks = 20, expand = c(0.01, 0.01)) +
16   scale_y_continuous(n.breaks = 10, limits = c(0,1), expand = c(0,0.01)) +
17   geom_line(color="navyblue") +
18   geom_point(aes(color=p53Prediccion)) +
19   geom_hline(yintercept = 0.5, lty="dotted", size=1) +
20   theme_linedraw()
21 print(plot_p53)
22
23 cuentaTotal <- table(p53Prediccion)
24 porcentaje <- 100*cuentaTotal/length(p53Posición)
25
26 aminoacidos <- table(p53Aminoácido, p53Prediccion)
27 aminoacidos_porcentaje <- 100*aminoacidos/length(p53Posición)
28
29 aminoacidos_df<-as.data.frame(aminoacidos_porcentaje)
30 colnames(aminoacidos_df) <- c("Aminoacidos","Prediccion","Porcentaje")
31
32 plot_aa <- ggplot(aminoacidos_df, aes(x=Aminoacidos,y=Porcentaje, fill=Prediccion)) +
33   geom_col(position = "dodge") +
34   scale_y_continuous(n.breaks = 10, limits = c(0,100), expand = c(0,0.01)) +
35   theme_bw()
36 print(plot_aa)
37
38 pdbP53 <- read.pdb(file = "/Users/mjnre/Documents/ejemplo.pdb")
39
40 p53pLDDT <- pdbP53$atom[pdbP53$alpha,"b"]
41
42 p53pLDDT_Prediccion <- ">80"
43 p53pLDDT_Prediccion[which(p53pLDDT<=50)]<-"<=50"
44 p53pLDDT_Prediccion[which(p53pLDDT>50 & p53pLDDT<=80)]<-">50 & <= 80"
45
46 pLDDT_con_Iupred <- ggplot(p53, aes(x=Posición,y=Iupred)) +
47   scale_x_continuous(n.breaks = 20, expand = c(0.01, 0.01)) +
48   scale_y_continuous(n.breaks = 10, limits = c(0,1), expand = c(0,0.01)) +
49   geom_line(color="navyblue") +
50   geom_line(data = p53, mapping = aes(x=Posición,y=pLDDT/100), col="red") +
51   geom_hline(yintercept = 0.5, lty="dotted", size=1) +
52   theme_linedraw()
53 pLDDT_con_Iupred
54
55 pLDDT_vs_IUPred<-ggplot(p53, aes(x=pLDDT,y=Iupred)) +
56   scale_x_continuous(n.breaks = 10, expand = c(0,0.01), limits=c(0,100)) +
57   scale_y_continuous(n.breaks = 10, limits = c(0,1), expand = c(0,0.01)) +
58   geom_point(aes(col=pLDDT_Prediccion)) +
59   geom_hline(yintercept = 0.5, lty="dotted", size=1) +
60   geom_vline(xintercept = 50, lty="dotted", size=1) +
61   theme_linedraw()
62 pLDDT_vs_IUPred

```

Figura 13.9: Programa en R análisis de desorden.

13.2. Análisis proteínas en hebra 3'5

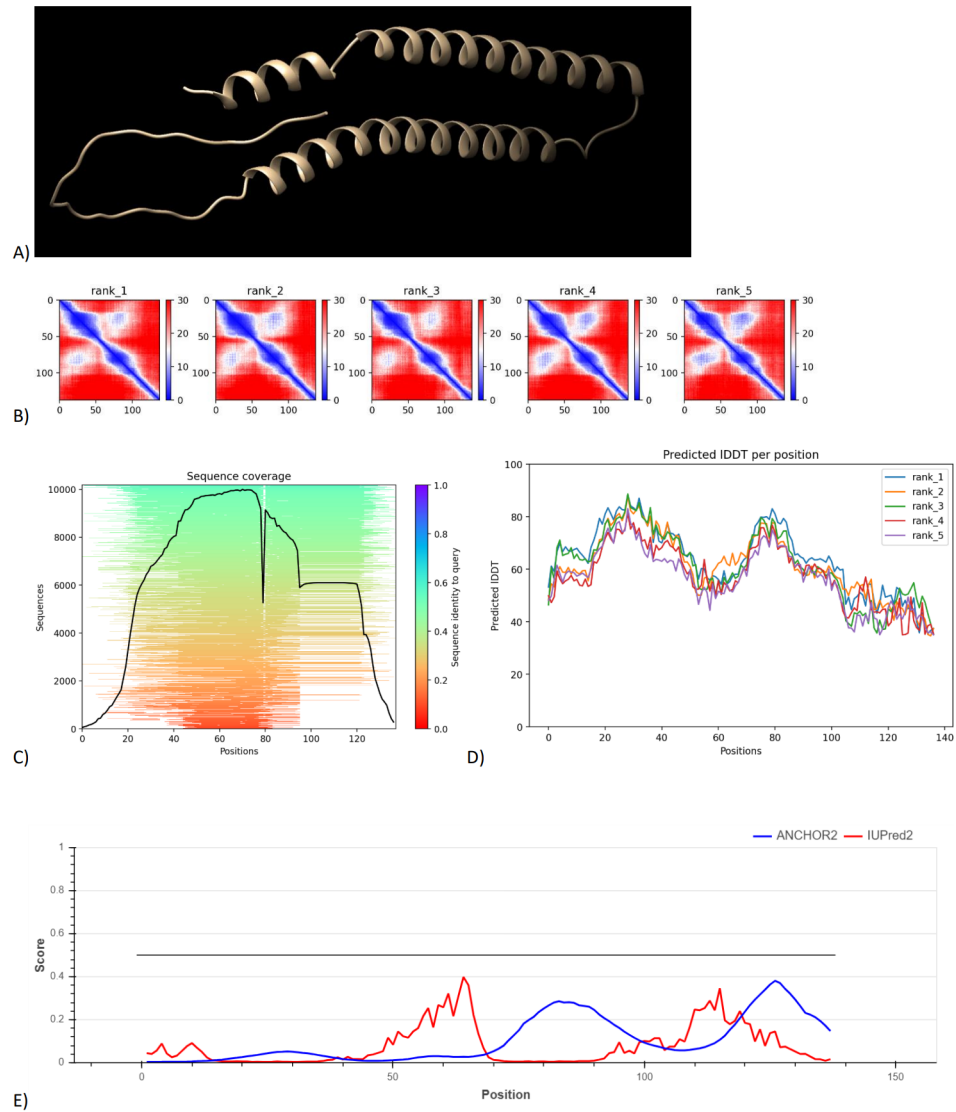


Figura 13.10: Análisis proteína Q6ZP21 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

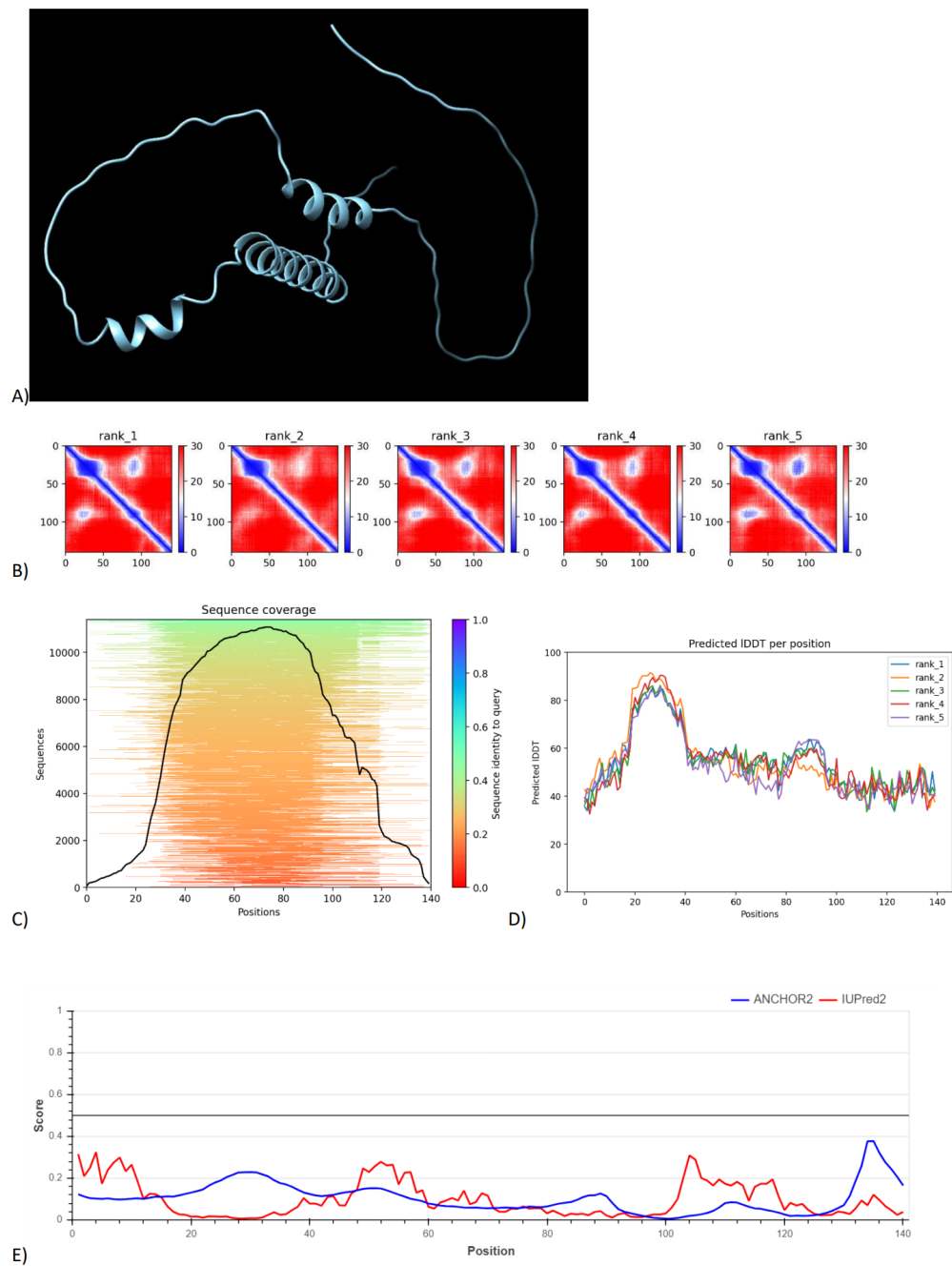


Figura 13.11: Análisis proteína B4E0H0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

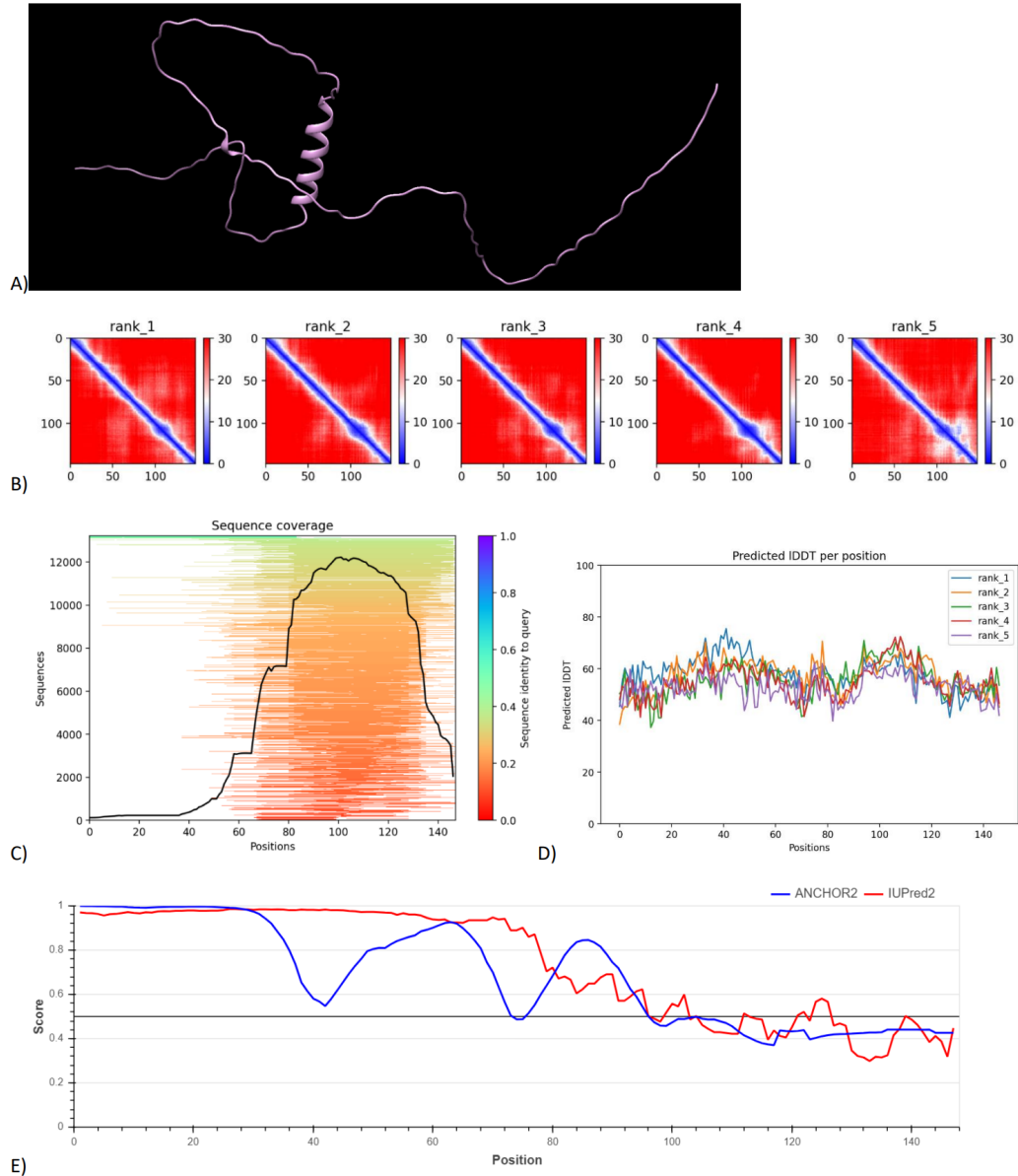


Figura 13.12: Análisis proteína Q1M183 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

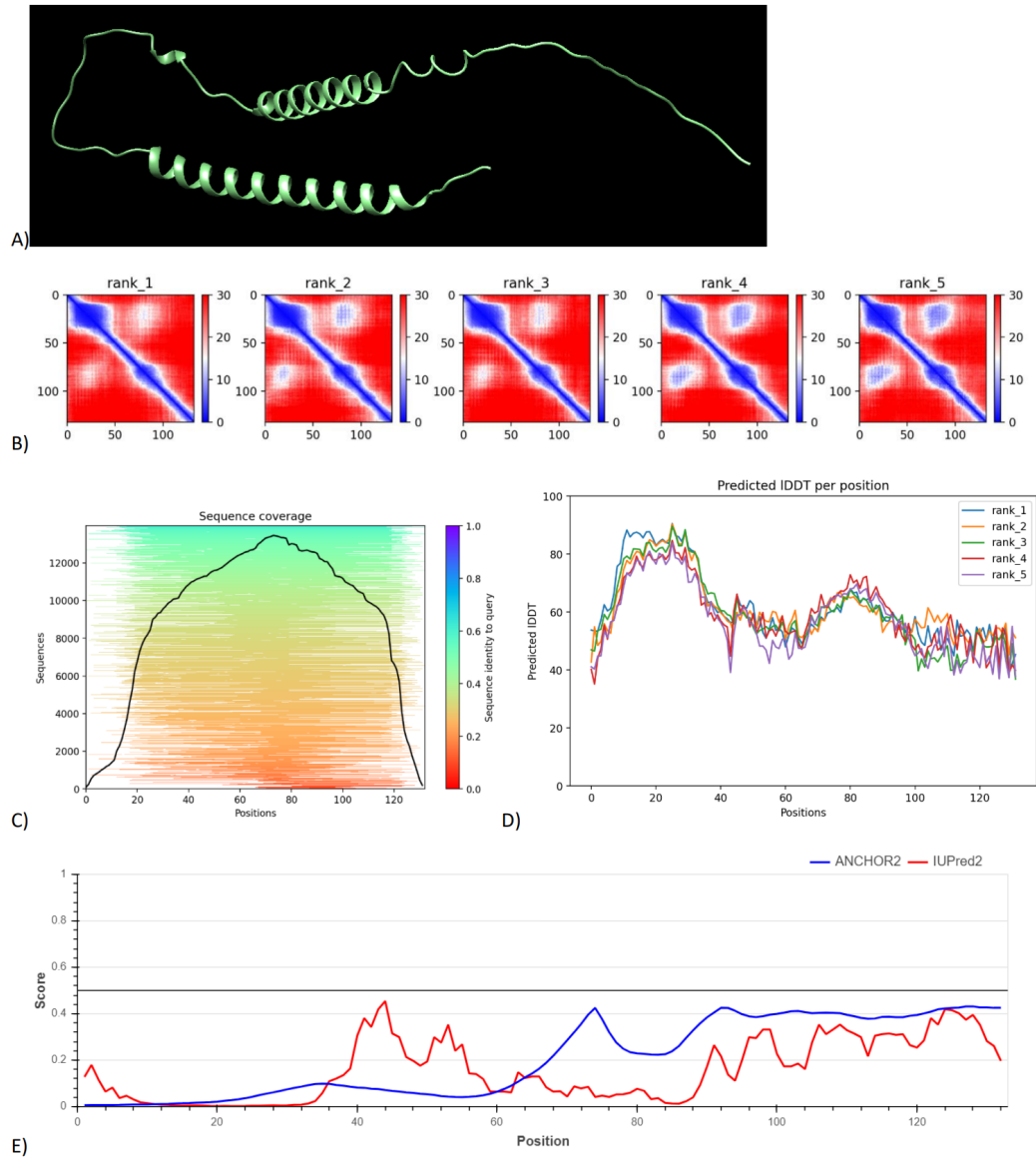


Figura 13.13: Análisis proteína Q6ZNU7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

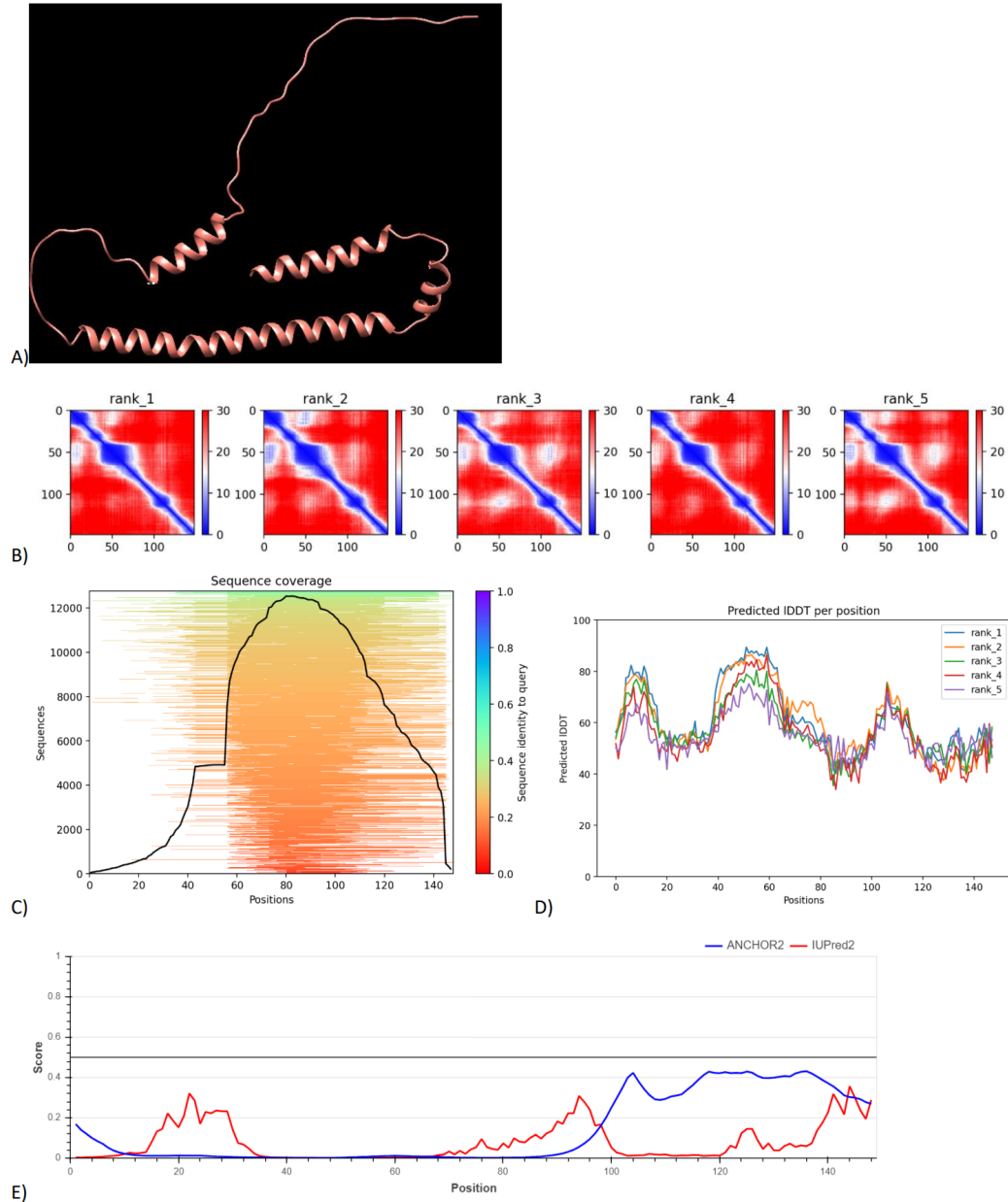


Figura 13.14: Análisis proteína Q6ZP34 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

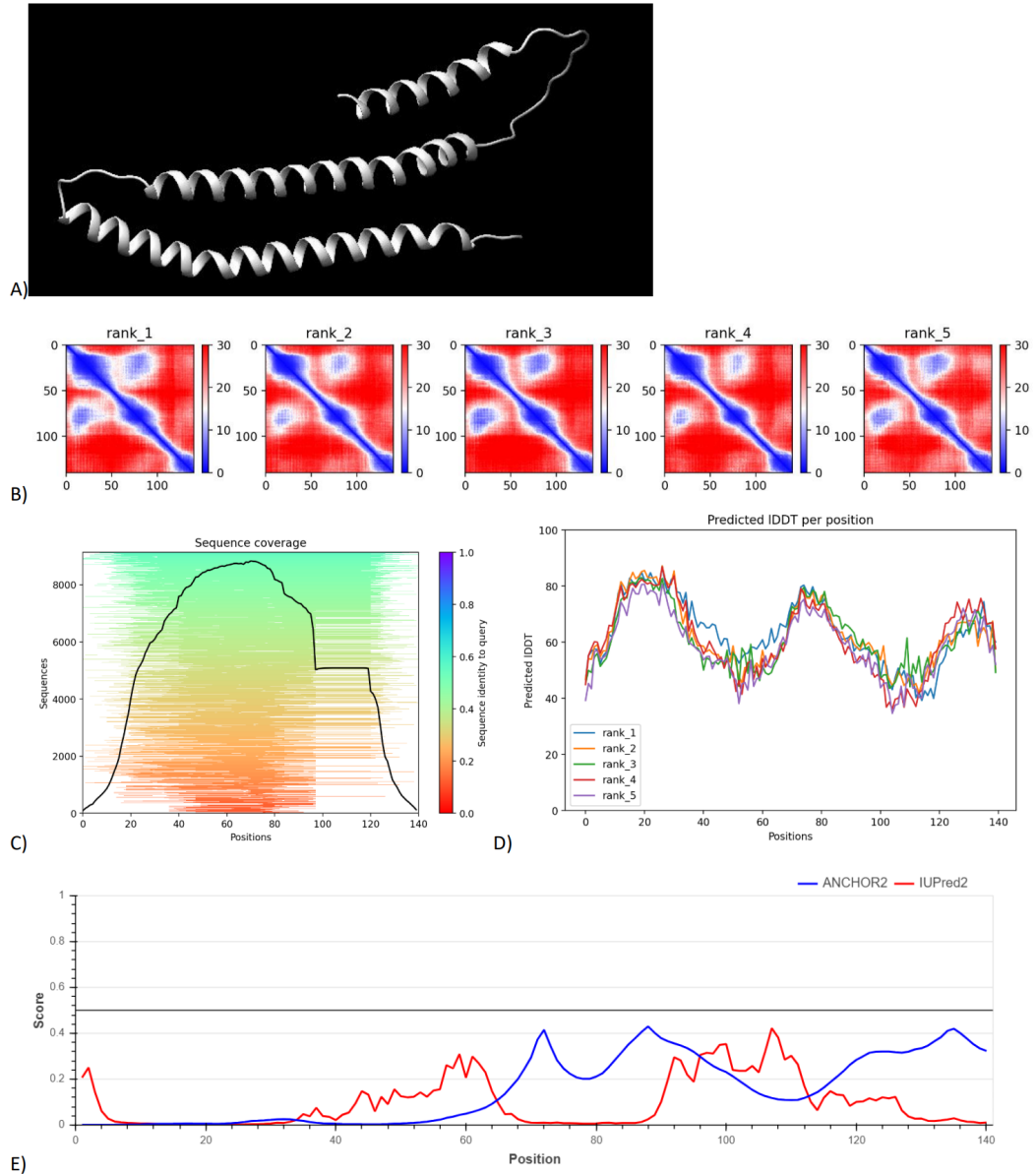


Figura 13.15: Análisis proteína Q6ZP99 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

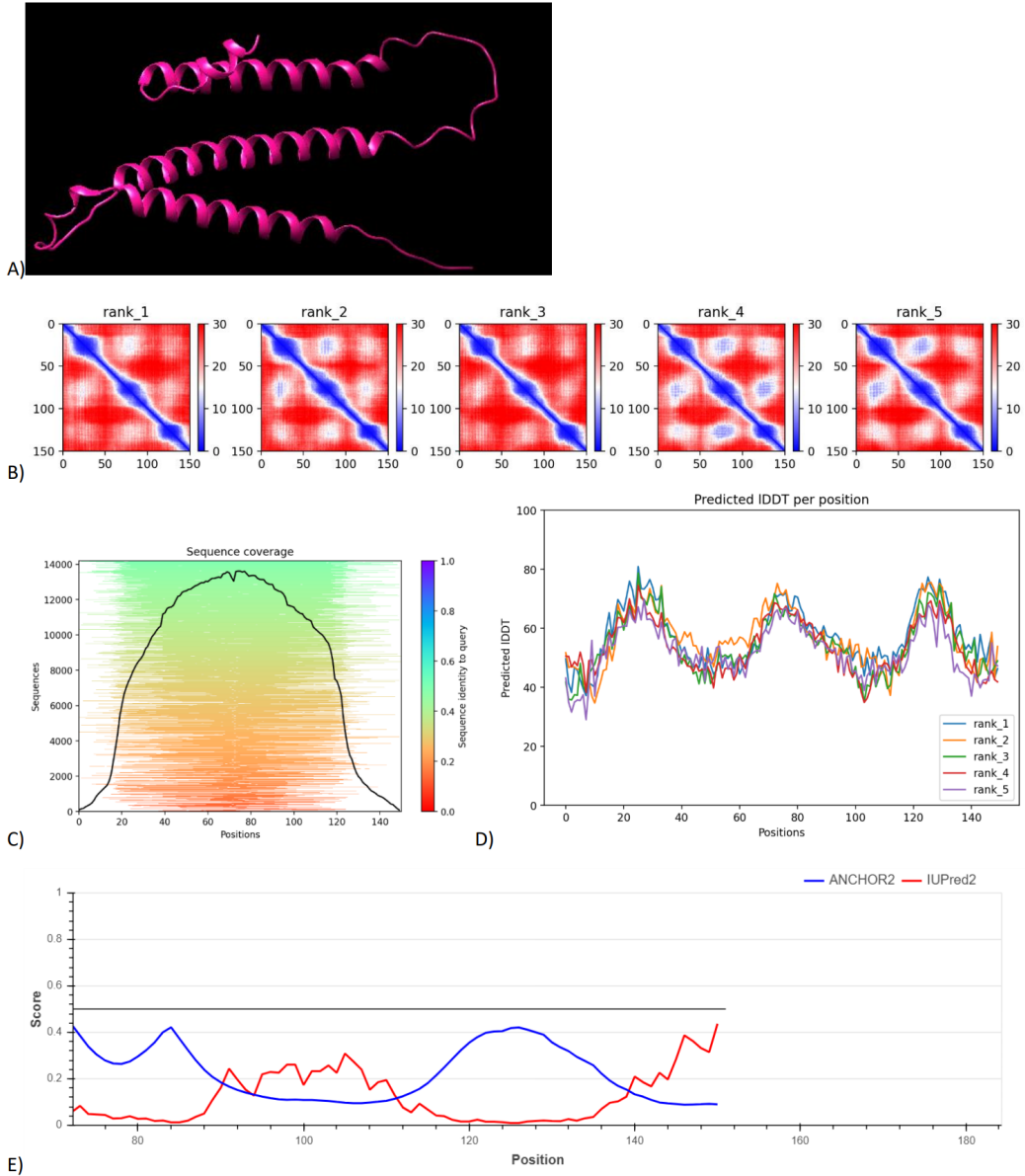


Figura 13.16: Análisis proteína Q6ZPA0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

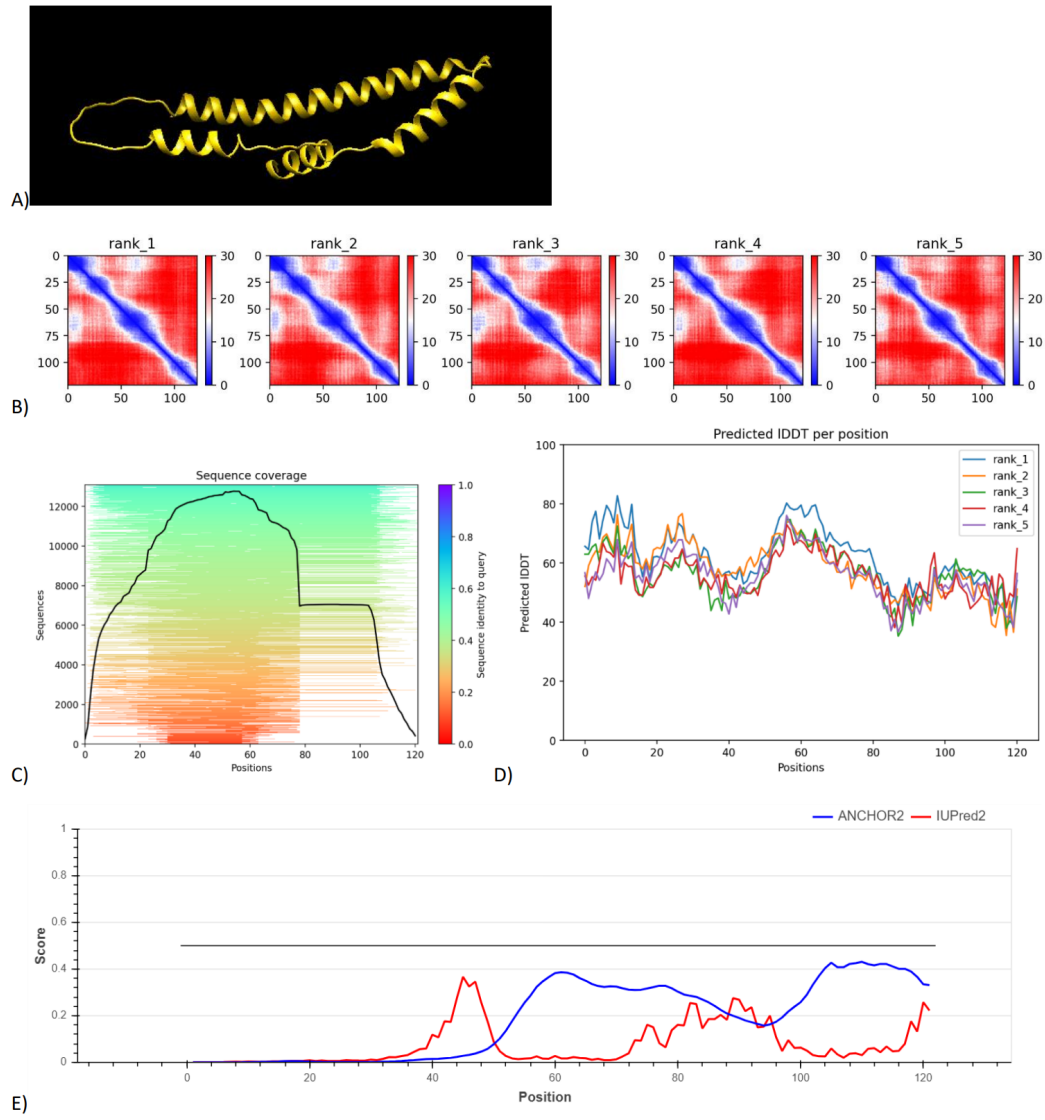


Figura 13.17: Análisis proteína Q6ZPB0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

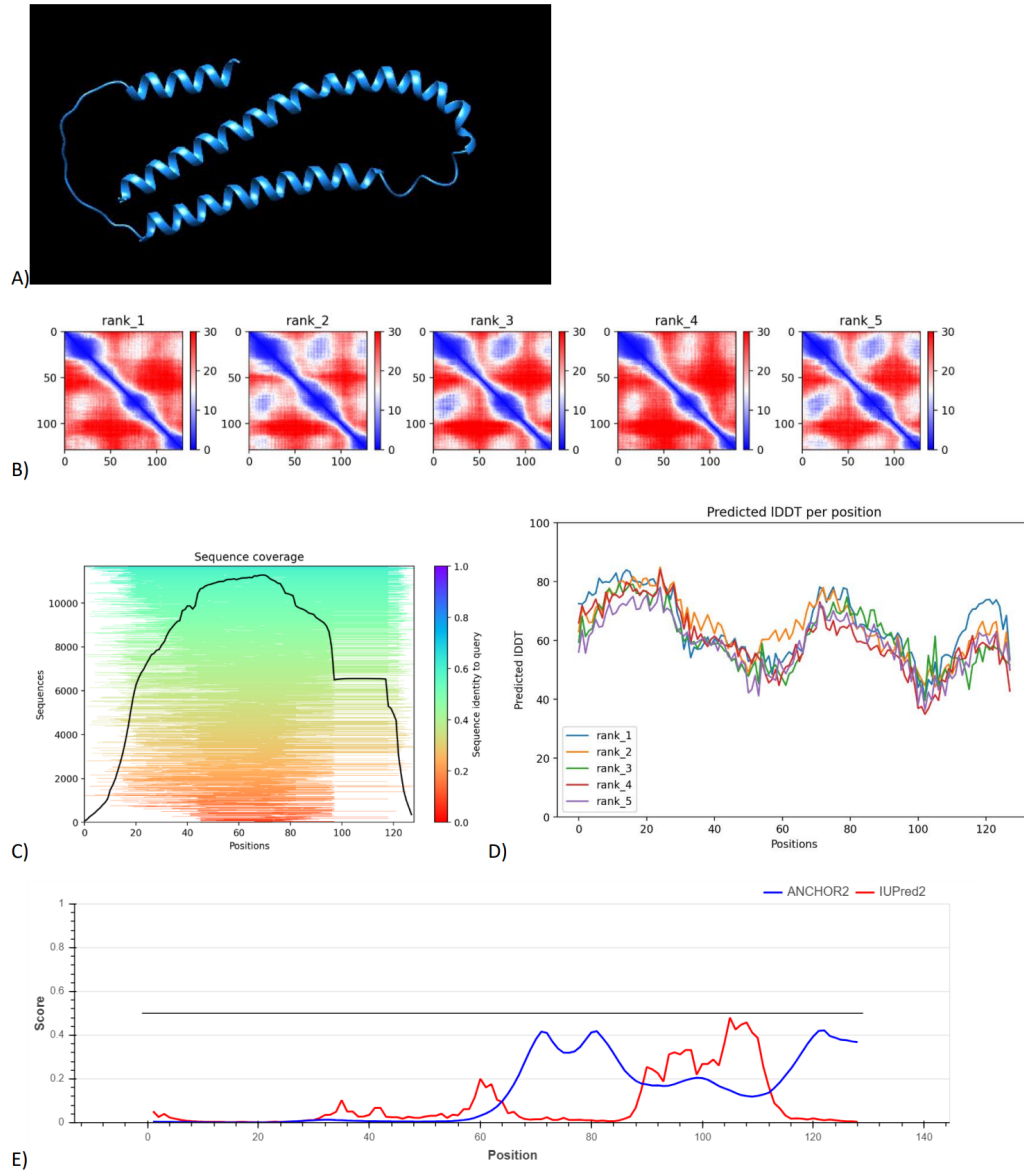


Figura 13.18: Análisis proteína Q6ZPB2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

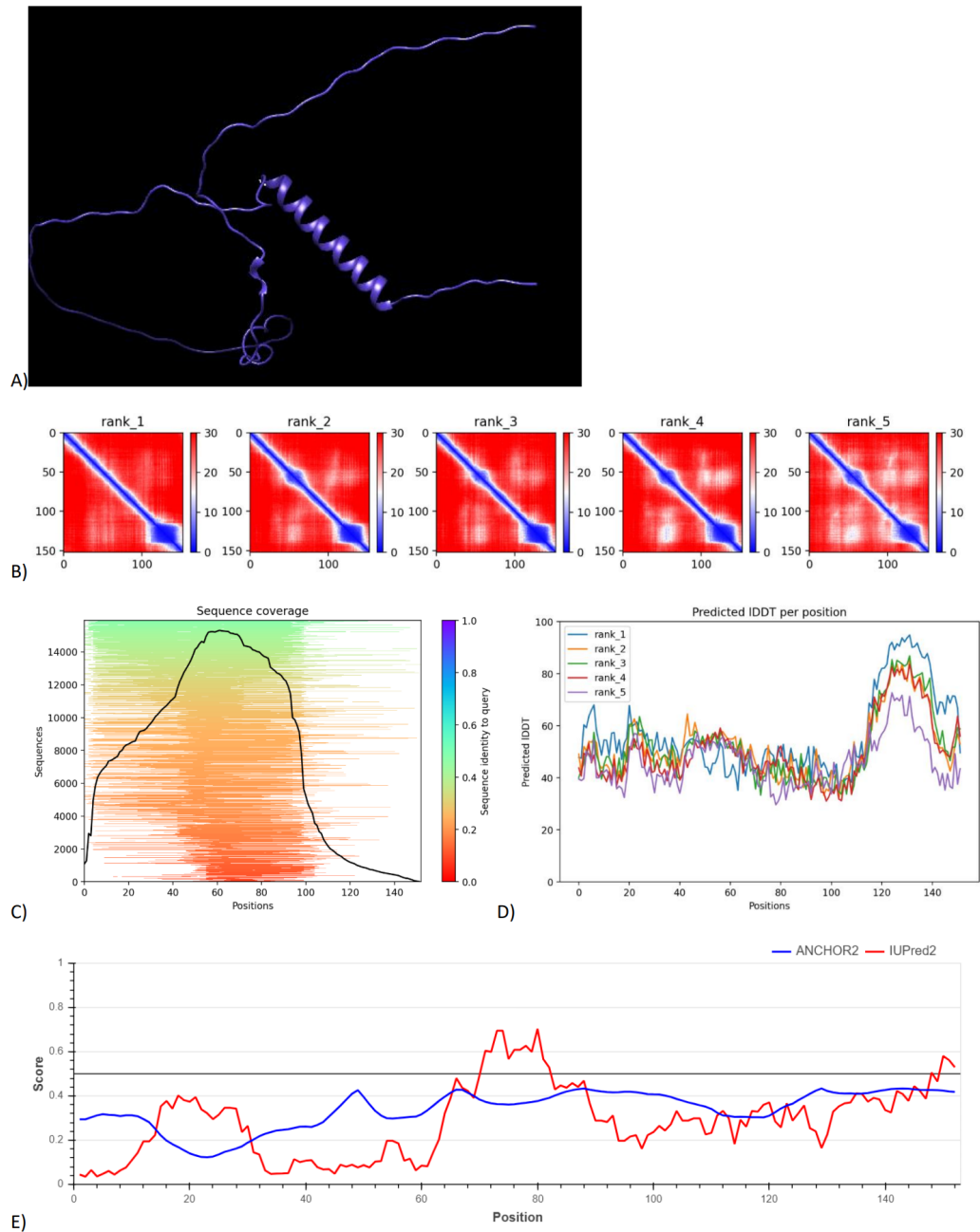


Figura 13.19: Análisis proteína Q6ZUG4 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

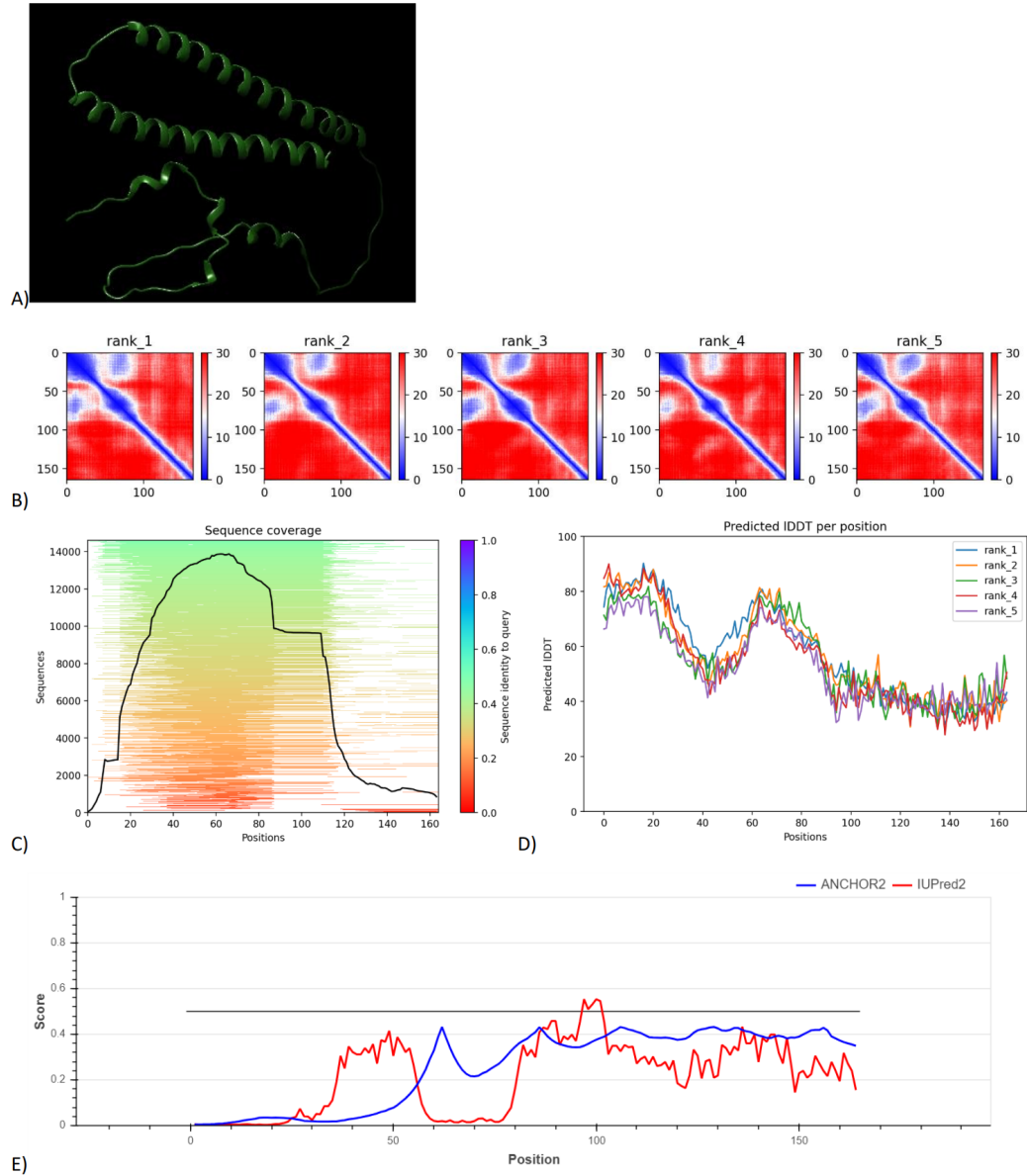


Figura 13.20: Análisis proteína Q6ZUK0 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

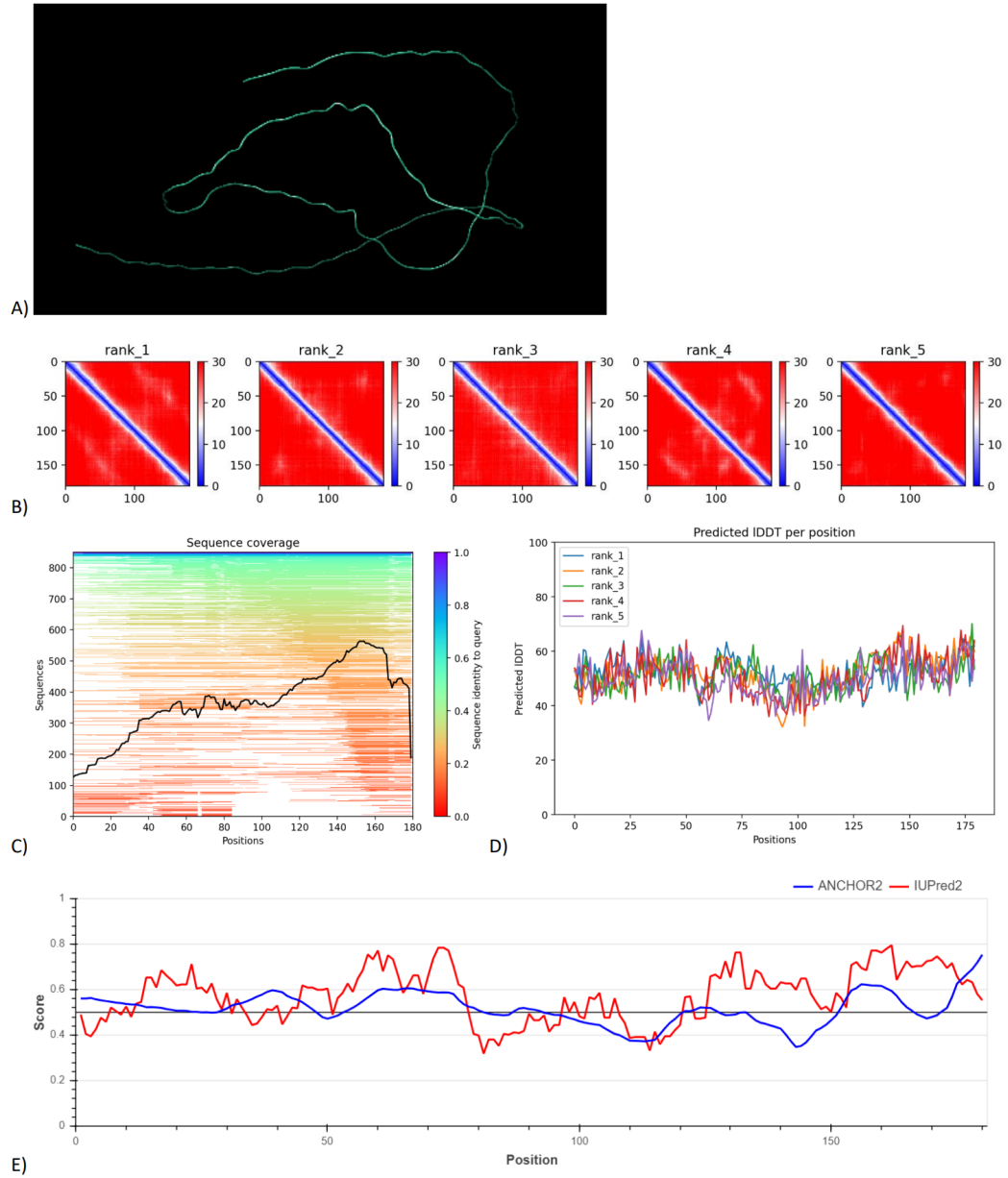


Figura 13.21: Análisis proteína Q8WZ27 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

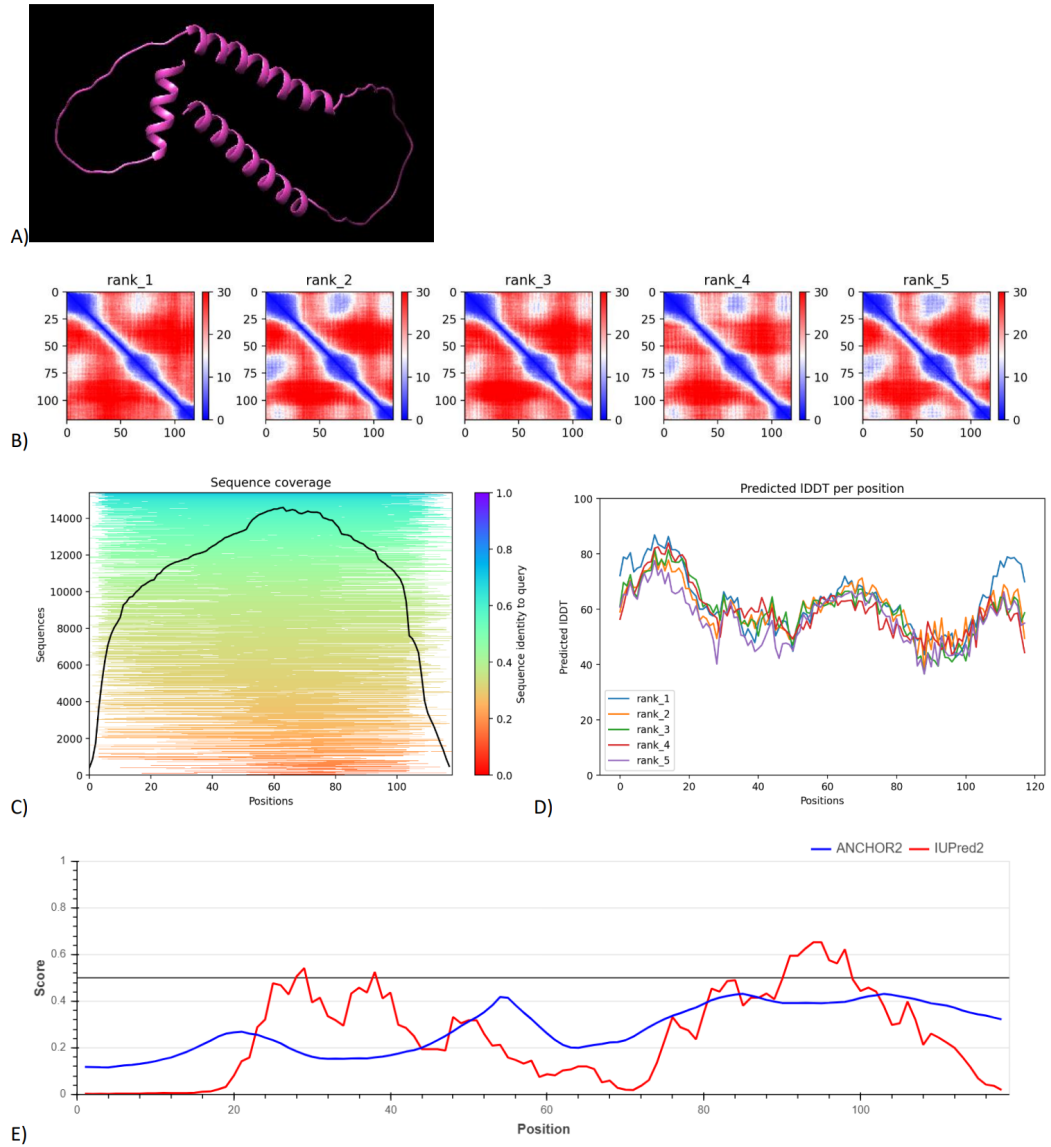


Figura 13.22: Análisis proteína Q9H387 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

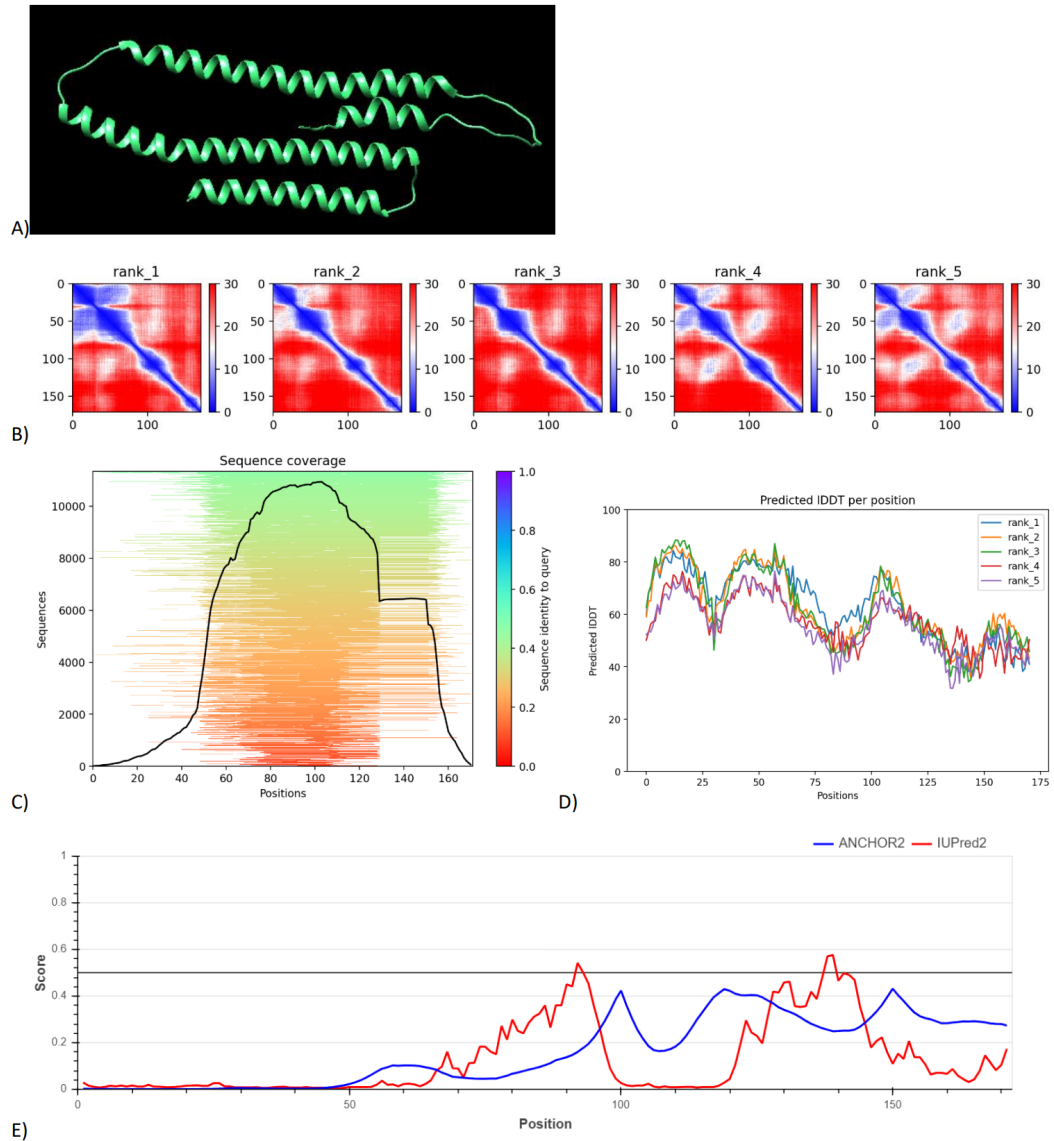


Figura 13.23: Análisis proteína Q9H728 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

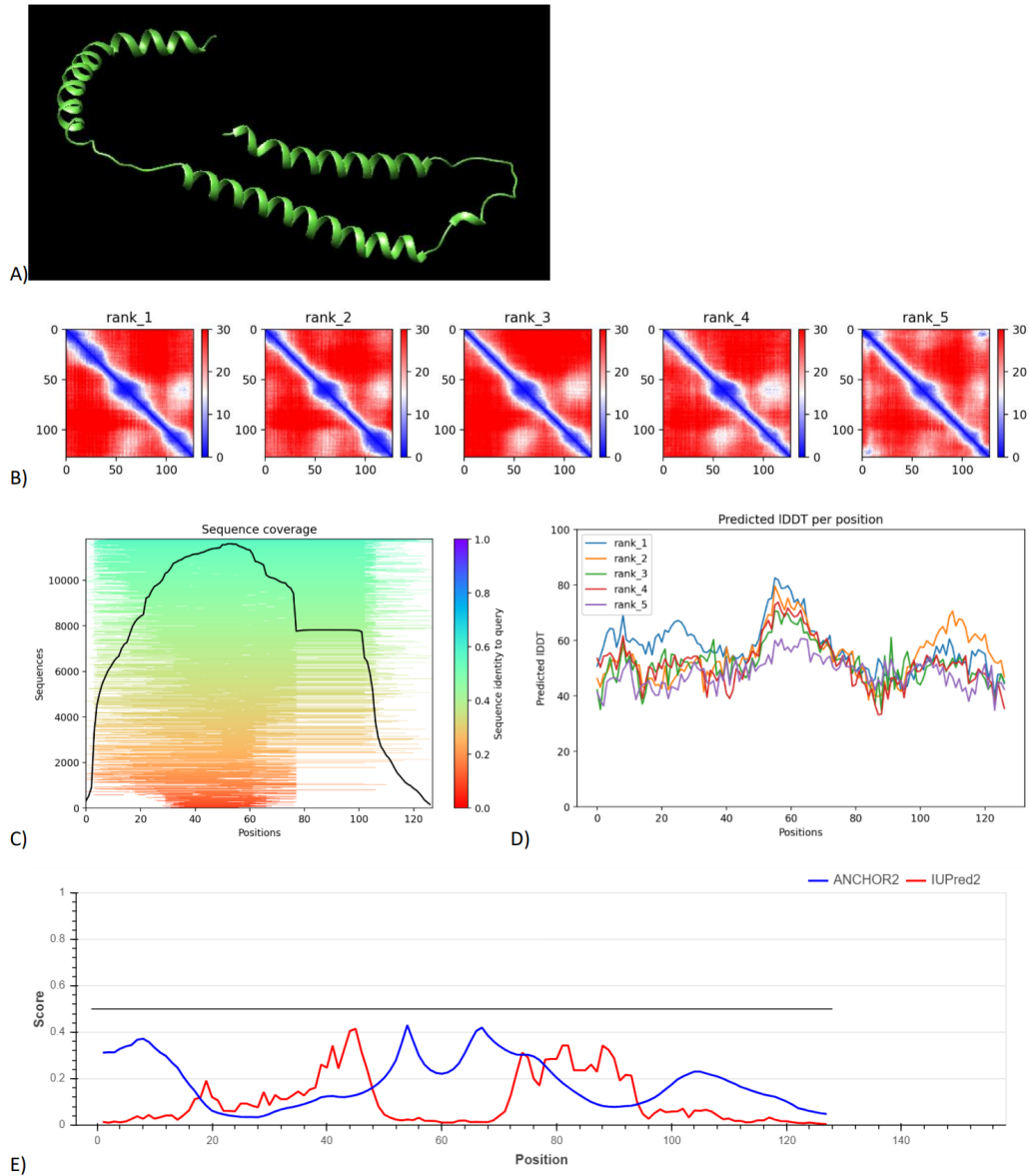


Figura 13.24: Análisis proteína Q9H743 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

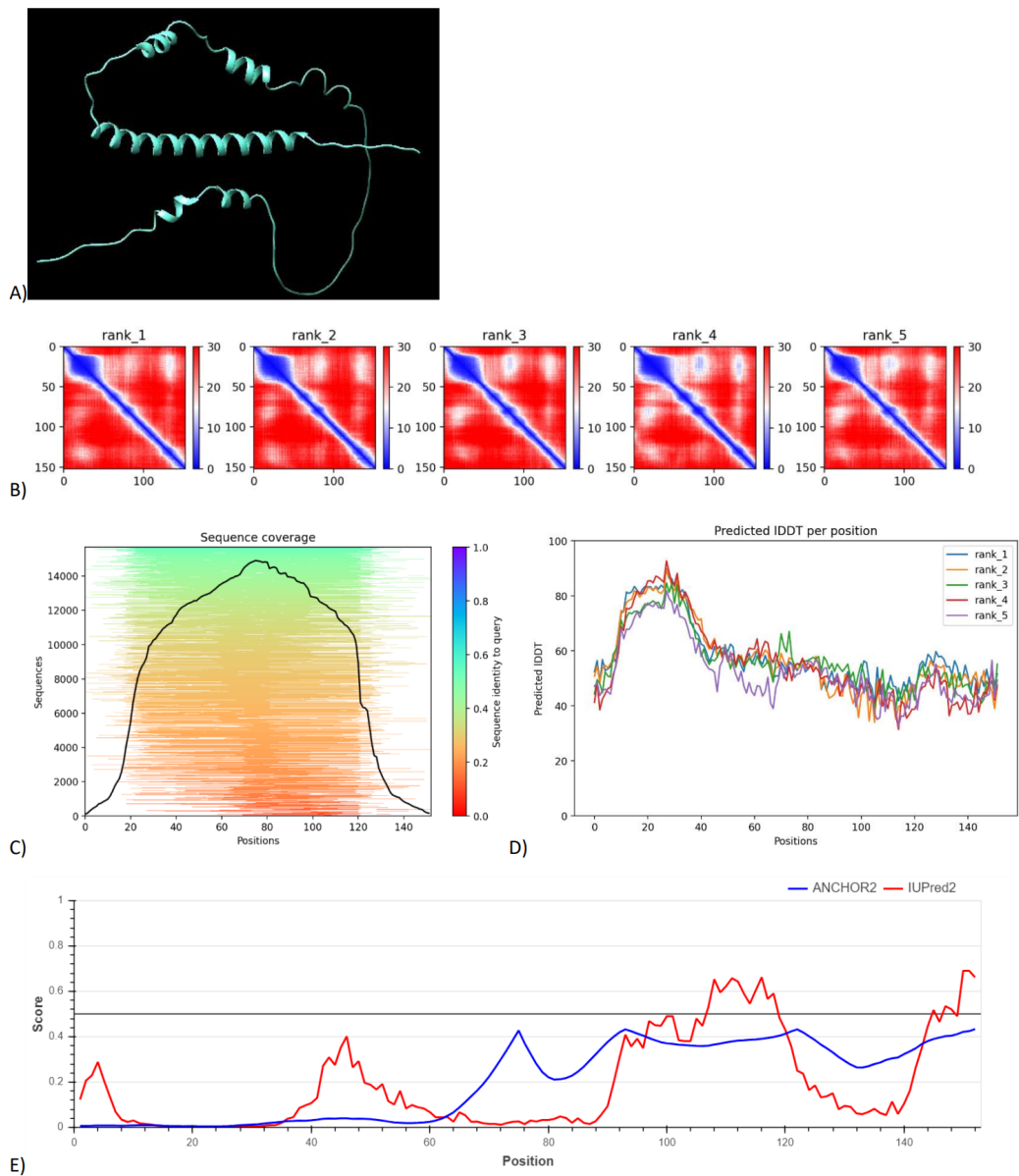


Figura 13.25: Análisis proteína Q9NX85 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

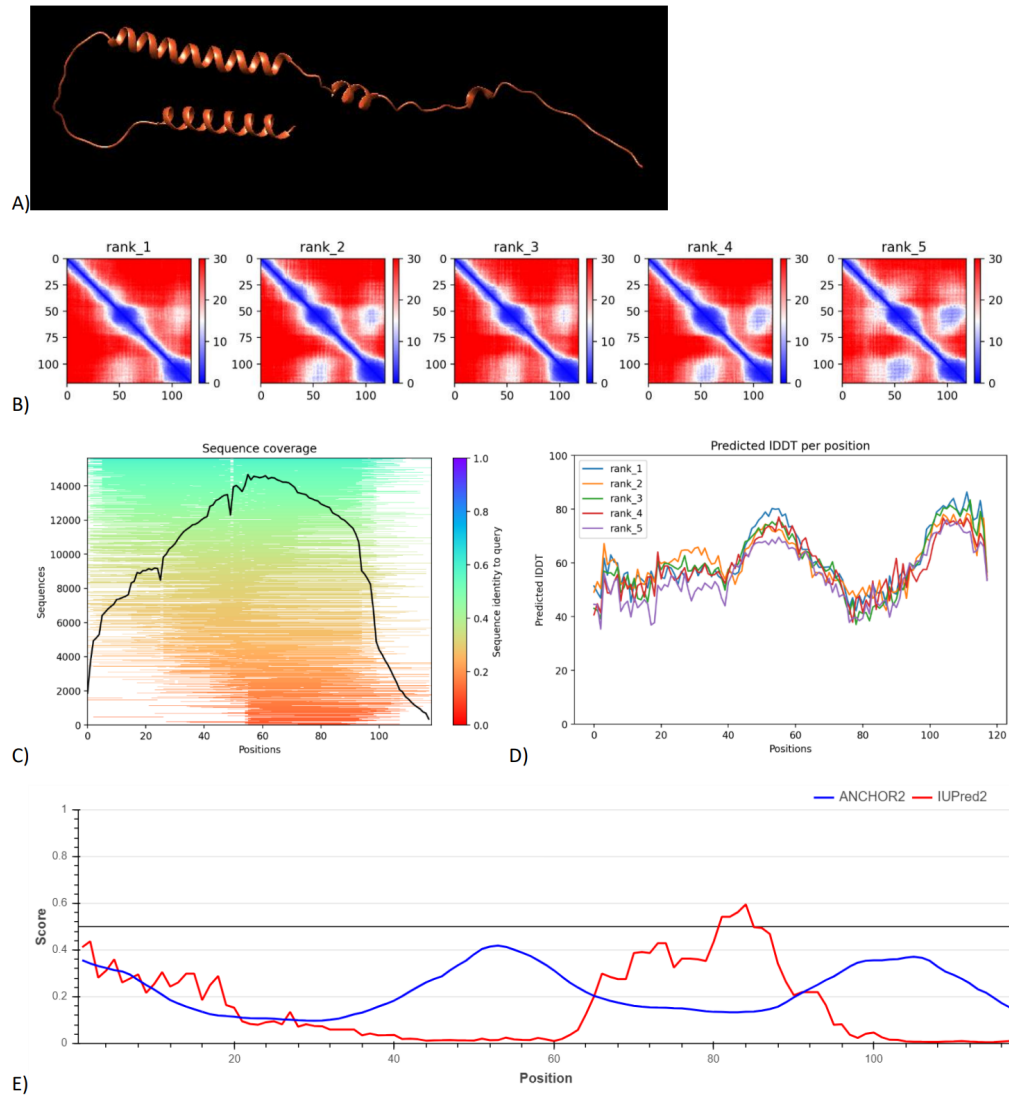


Figura 13.26: Análisis proteína Q9P195 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

13.3. Análisis proteínas en hebra 5'3

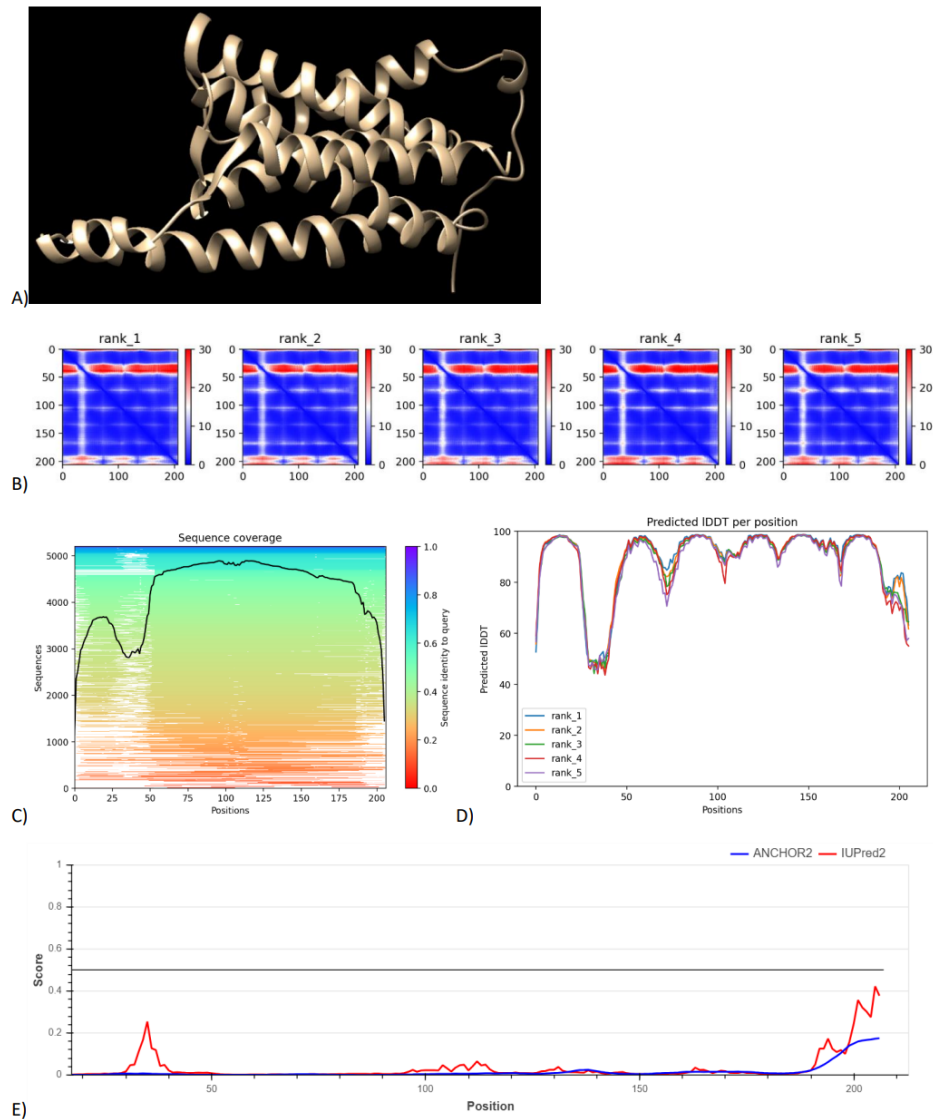


Figura 13.27: Análisis proteína A0A0A1TSG9 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

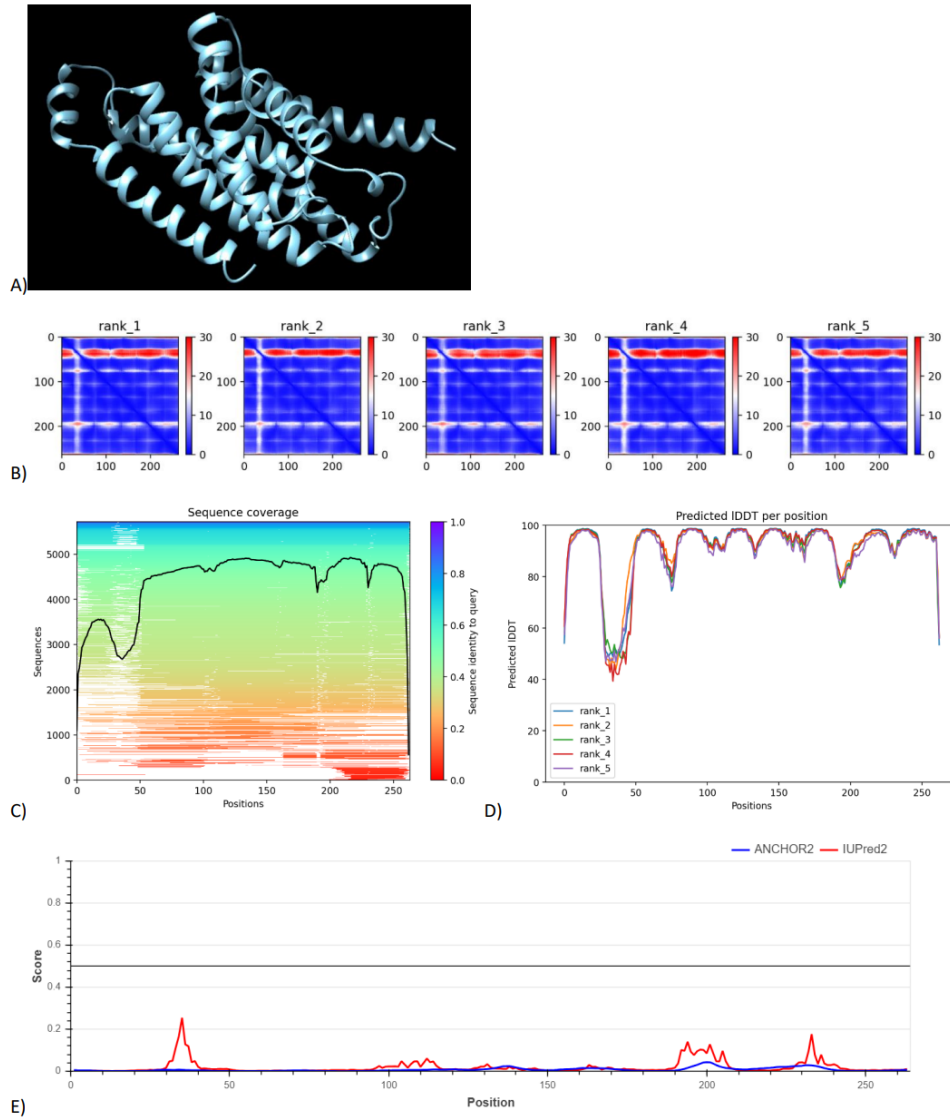


Figura 13.28: Análisis proteína A0A0A8IKZ2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

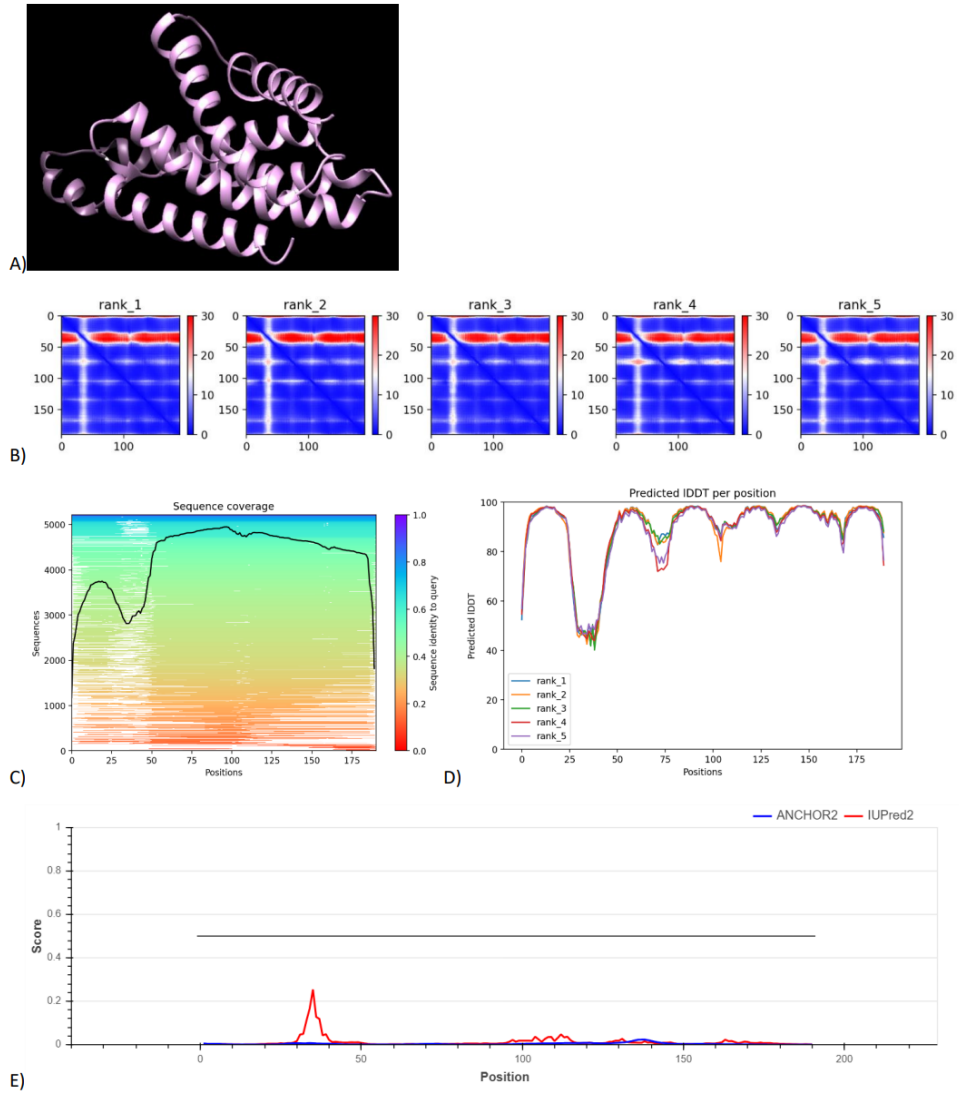


Figura 13.29: Análisis proteína A0A1A9C9I6 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

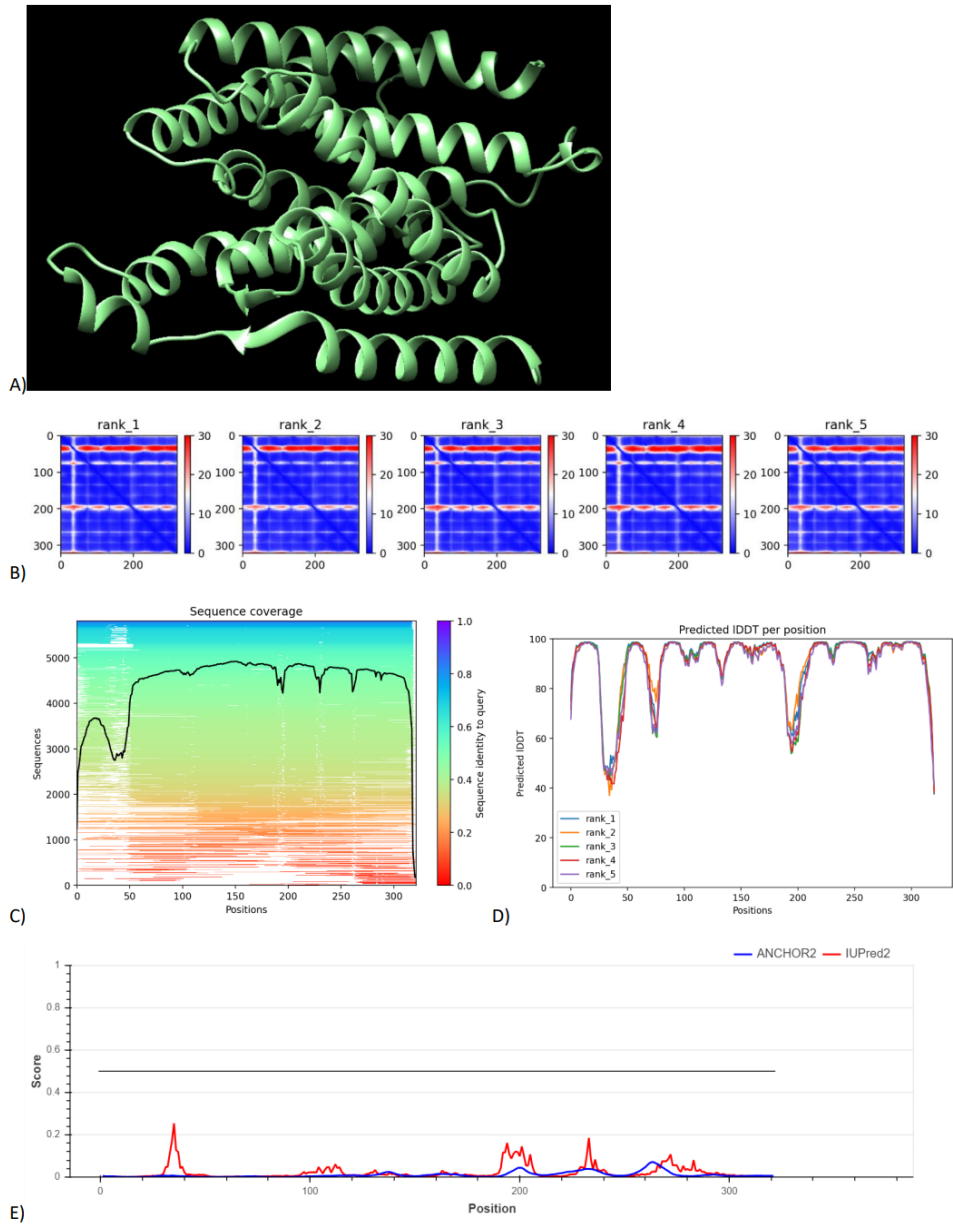


Figura 13.30: Análisis proteína A0A8D5ZBQ7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

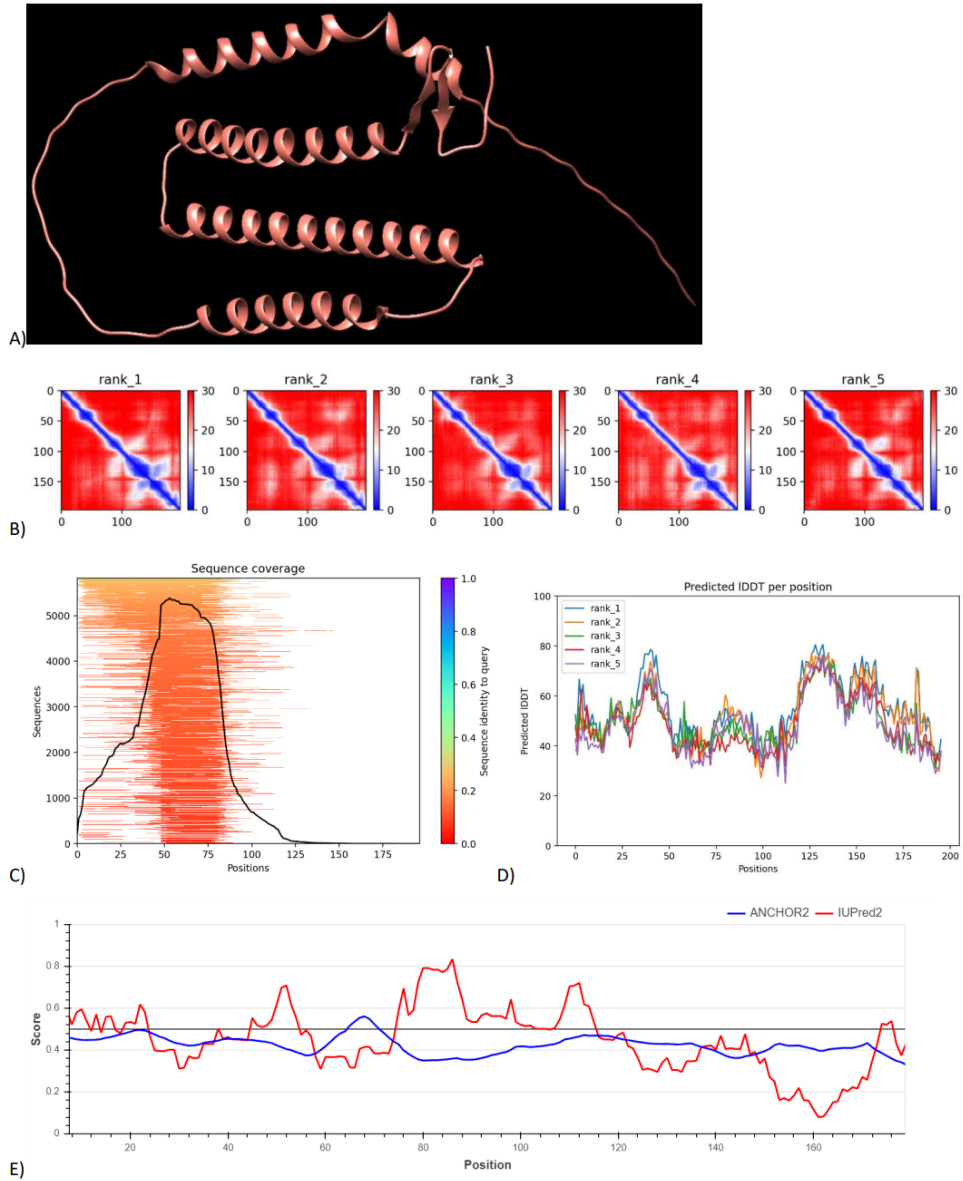


Figura 13.31: Análisis proteína A2NVG1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

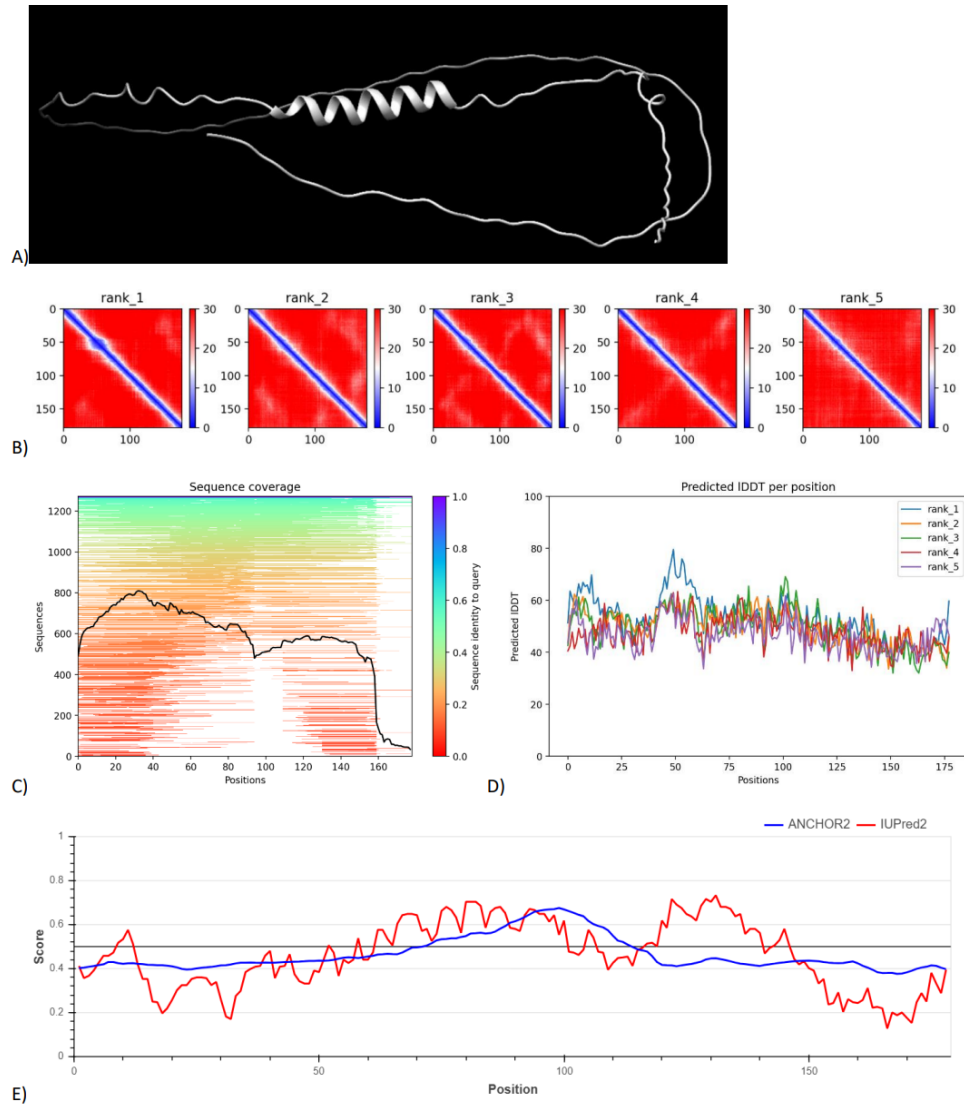


Figura 13.32: Análisis proteína B2RNZ7 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

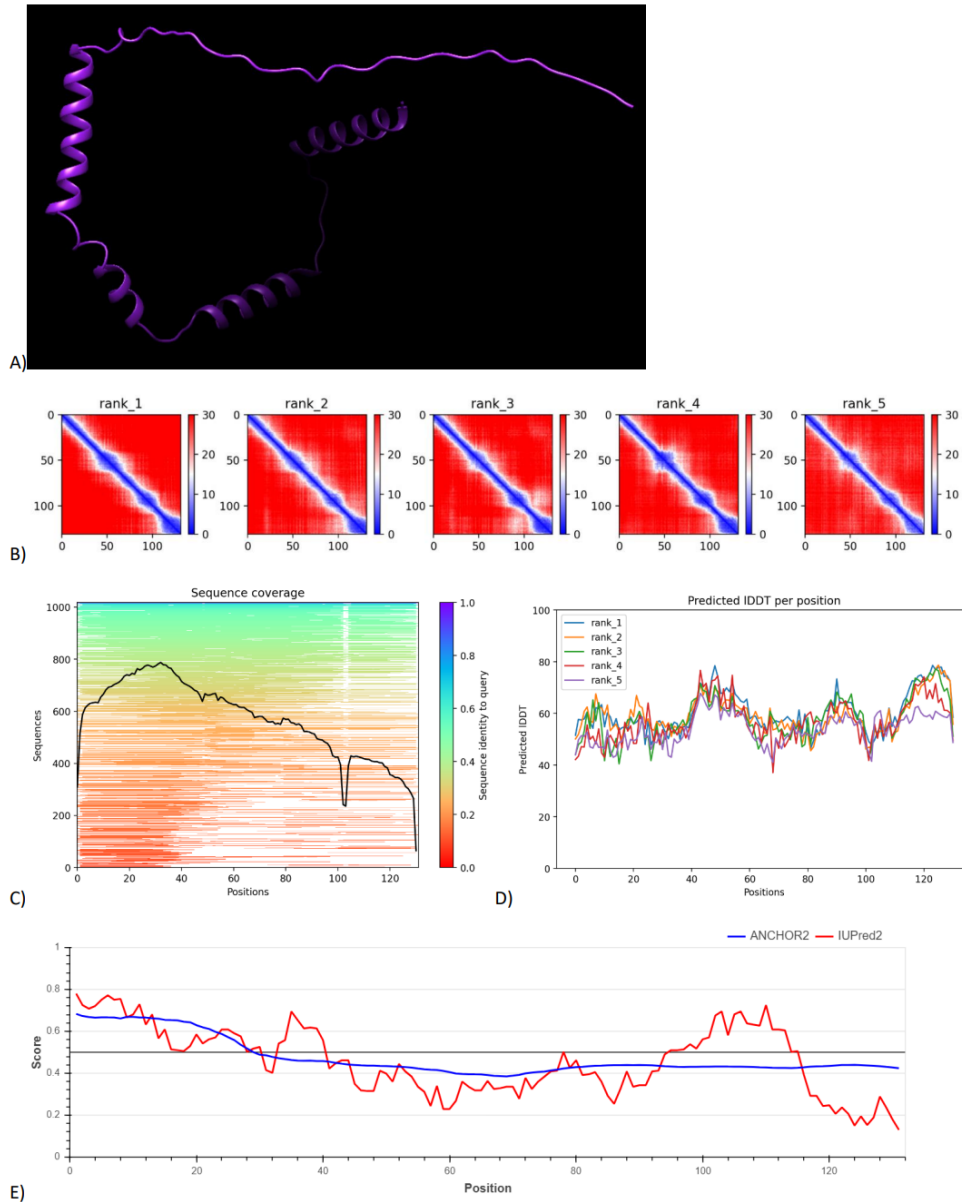


Figura 13.33: Análisis proteína Q1KSG2 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

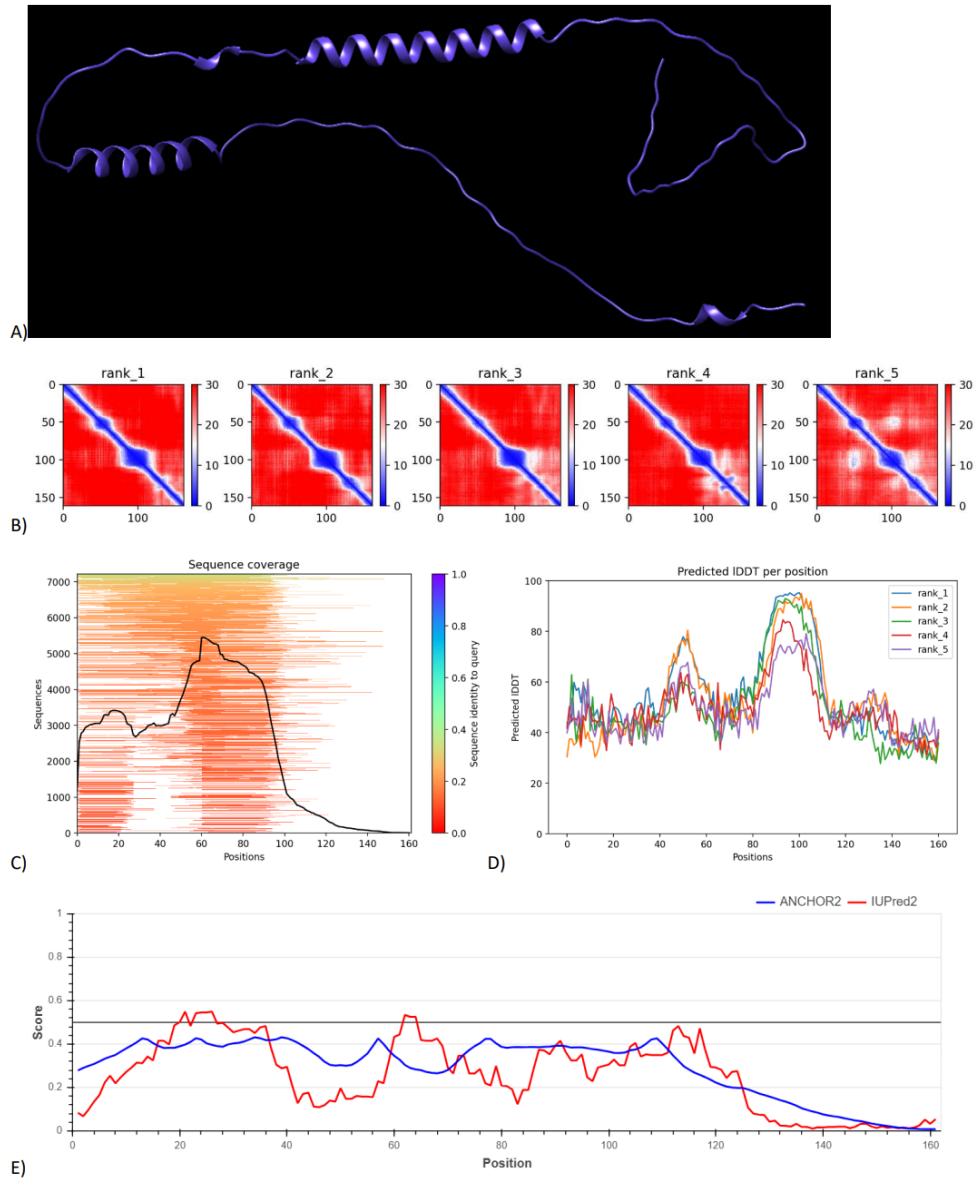


Figura 13.34: Análisis proteína Q6ZP50 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

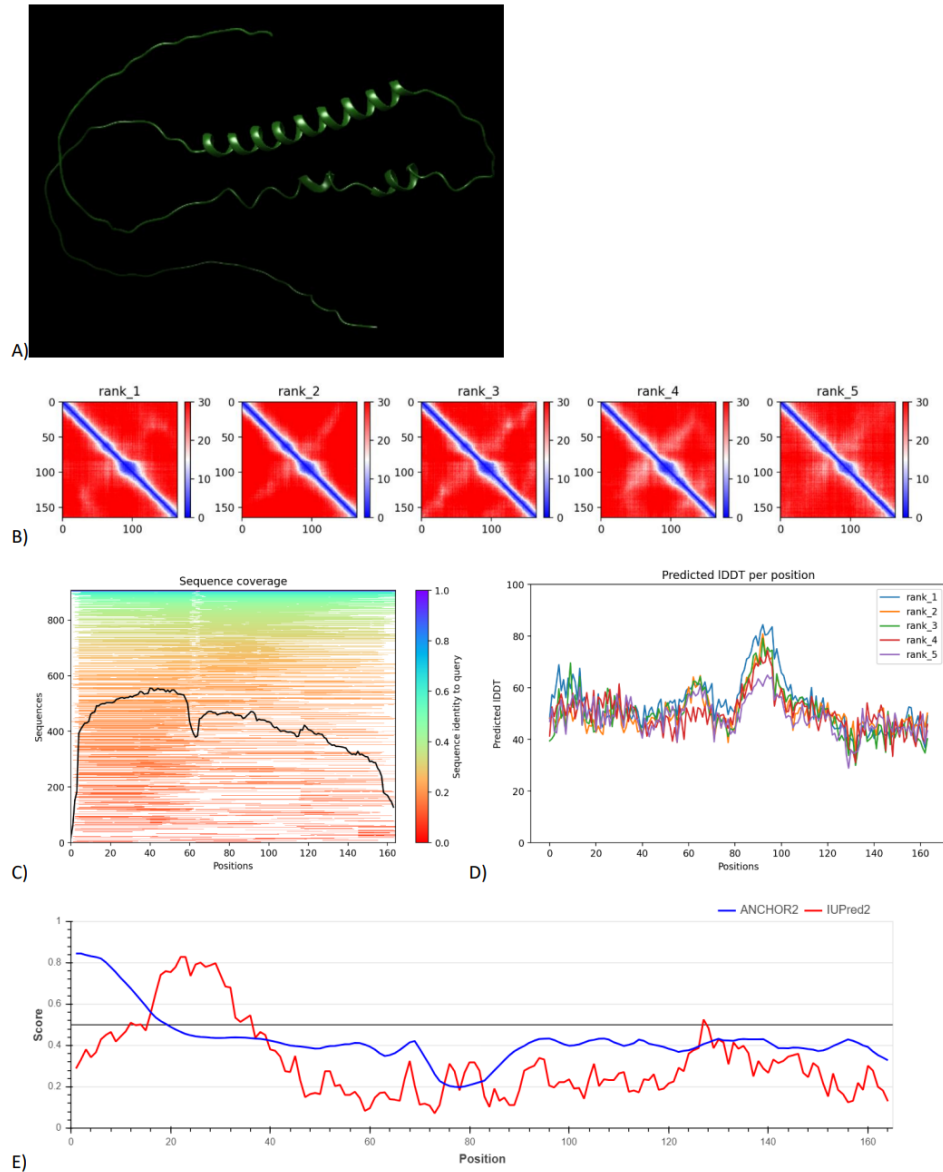


Figura 13.35: Análisis proteína Q6ZPC1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

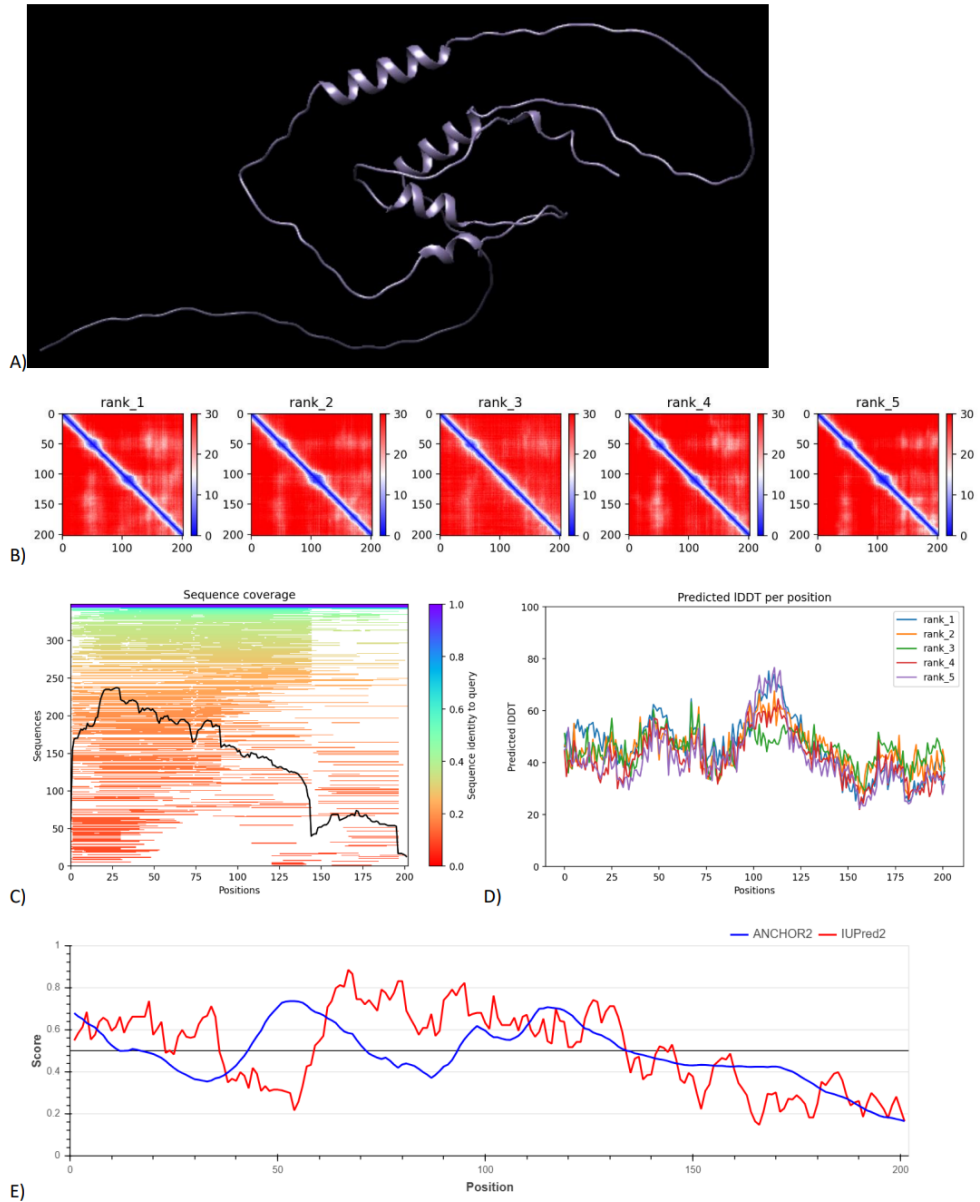


Figura 13.36: Análisis proteína Q8WYX4 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

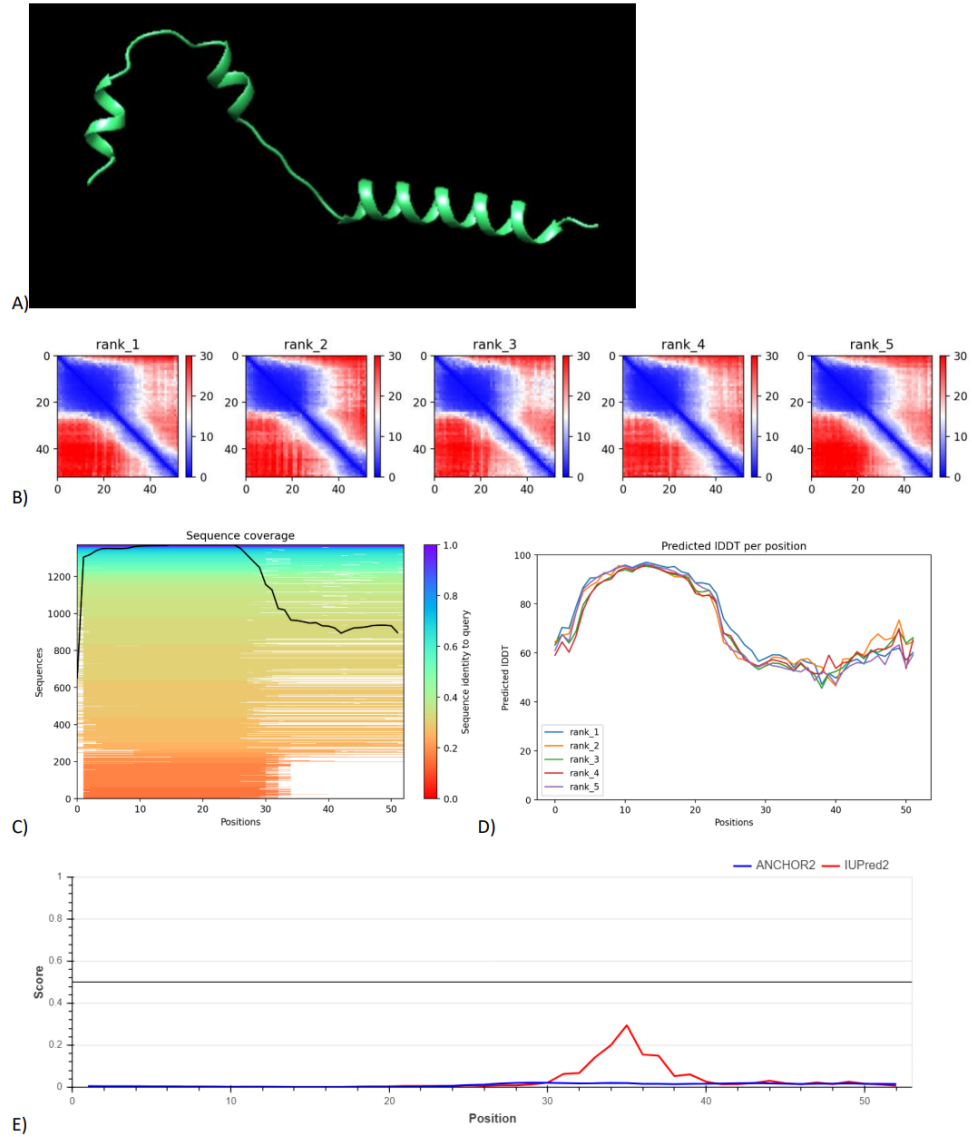


Figura 13.37: Análisis proteína Q9UBB8 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

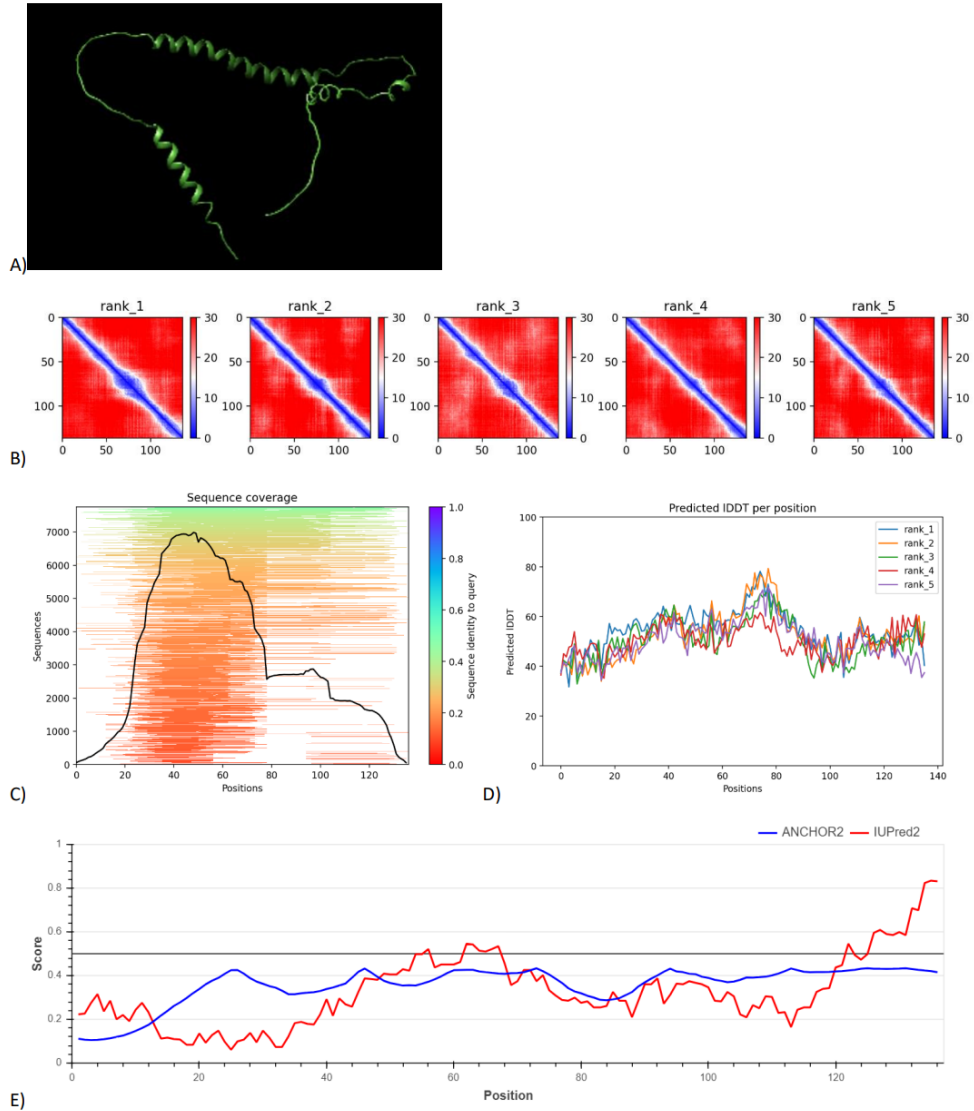


Figura 13.38: Análisis proteína Q96NR6 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

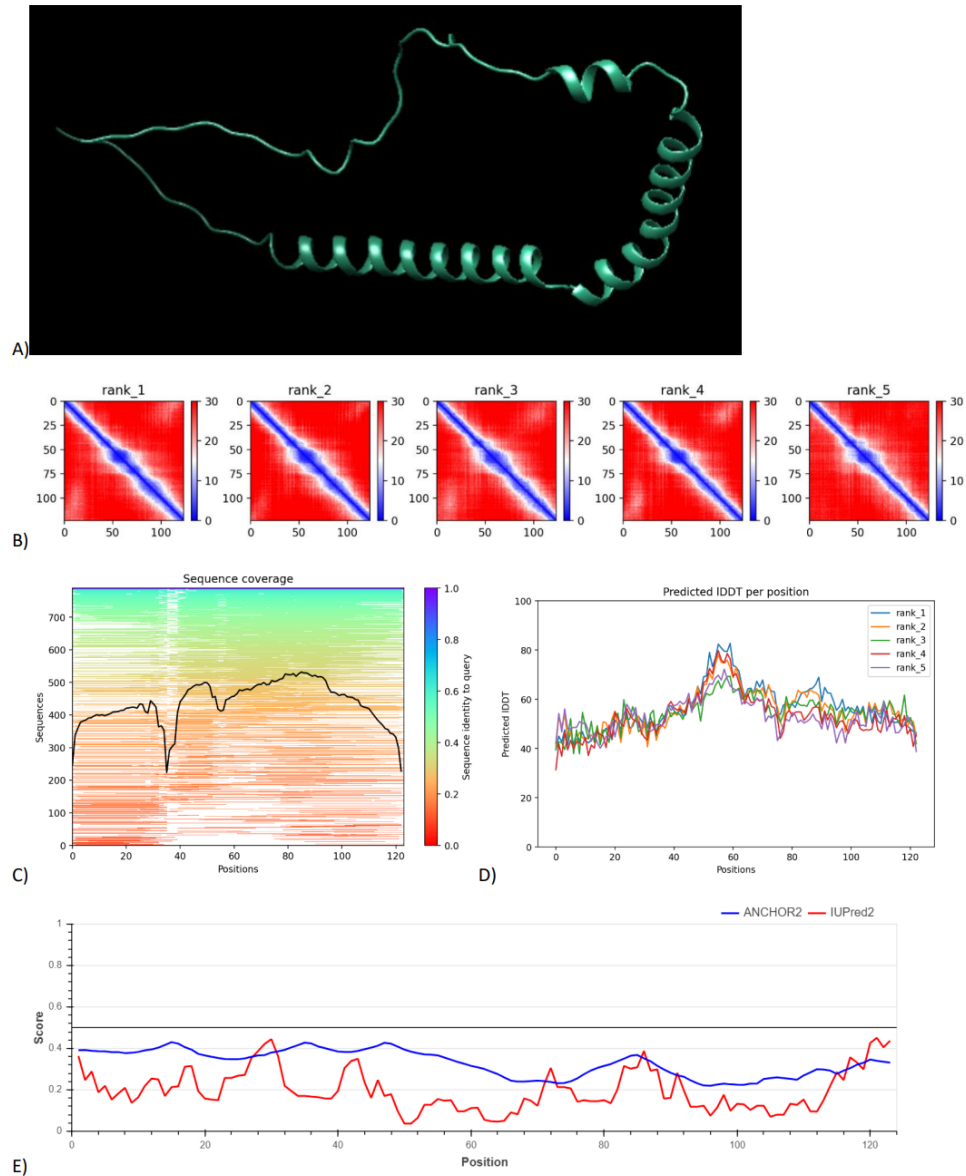


Figura 13.39: Análisis proteína Q969H1 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

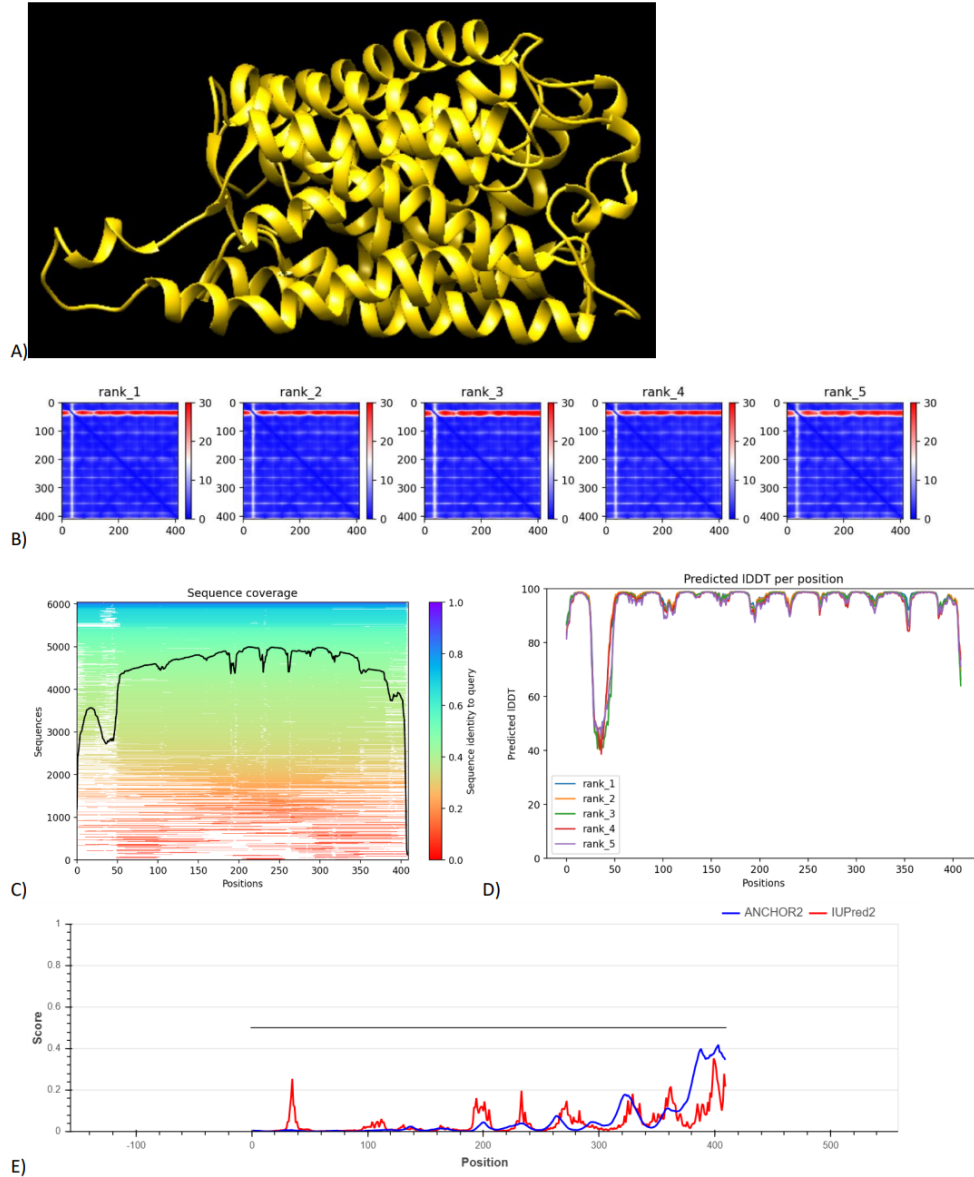


Figura 13.40: Análisis proteína Q02094 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas

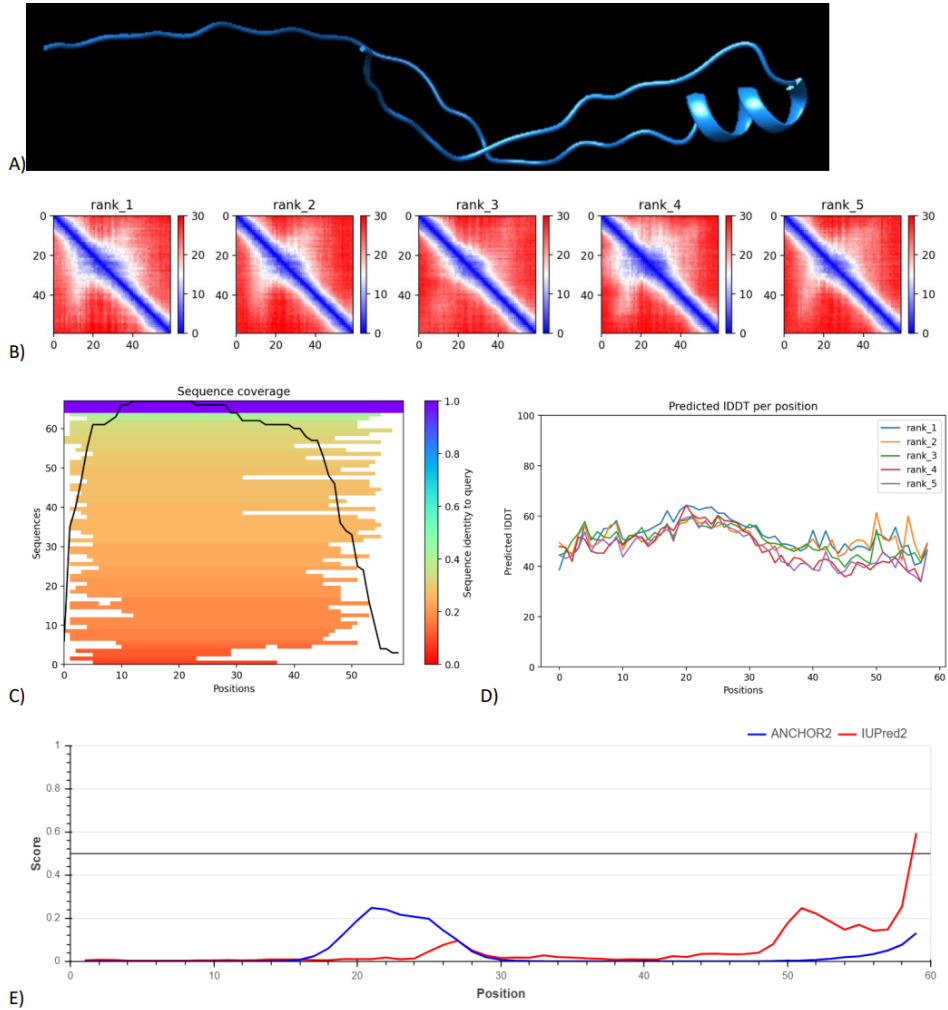


Figura 13.41: Análisis proteína Q16416 con gráficos generados por AlphaFold y IUPred. A) Modelado 3D en Chimera. B) Gráficos PAE. C) Cobertura de la secuencia en el modelo D) Gráfico pLDDT E) Análisis IUPred de secciones desordenadas