**Análisis de Herramientas y Métodos: Predicción de Estructura Proteica Terciaria**

*Daniel F. Rengifo*

**Marco Teórico:**

La predicción de estructuras proteicas especialmente su estructura terciaria, a partir de una secuencia de aminoácidos, es un problema abierto para la biología computacional. Este problema está íntimamente relacionado con un amplio abanico de disciplinas, como lo son la medicina y la farmacología, entre otras. El problema consiste en poder determinar, con cierto grado de seguridad, la estructura que una cierta proteína adoptara en distintos escenarios.

Este problema, por ende, es derivado de una serie de sub-problemas, los cuales deben ser resueltos de manera secuencial. Estos sub-problemas hacen referencia a diferentes facetas de la estructura de una proteína[c]:

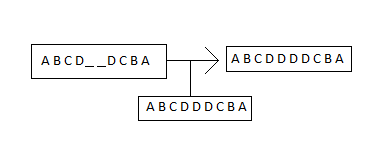
* En primer lugar, está la estructura primaria, la cual describe la secuencia neta de aminoácidos que generará la proteína.
* La estructura secundaria describe la estructura bidimensional de la proteína. Para poder predecir esta estructura es necesario tener en cuenta las interacciones moleculares registradas en el campo de la química orgánica.
* La estructura terciaria describe su estructura tridimensional a partir de una estructura bidimensional. Si bien este paso es similar al anterior, su complejidad es sustancialmente mayor, dado su estado espacial. Sin embargo, es en este paso donde se puede dilucidar, con claridad aceptable, la funcionalidad de una proteína especifica.
* También se puede hacer referencia a la estructura cuaternaria, la cual describe la forma en la cual una proteína puede relacionarse con otras a nivel estructural.

Como se ha descrito, es necesario tener una estructura secundaria como referencia para determinar una estructura terciaria coherente. Por esta razón, en este proyecto se llevarán a cabo ambos problemas, de forma conjunta. Sin embargo, es importante anotar que el enfoque del proyecto se hará sobre la predicción de estructuras terciarias.

Para llevar a cabo esta predicción, existen una serie de diferentes estratagemas, protocolos y métodos, que van desde *aprendizaje de maquina* hasta modelos probabilísticos y algoritmos del tipo “divide y vencerás”.

Entre los métodos a analizar se encuentran dos principales vertientes:

* Modelado comparativo[b]: El cual usa estructuras previamente conocidas para inferir posibles patrones sobre las secuencias a predecir.

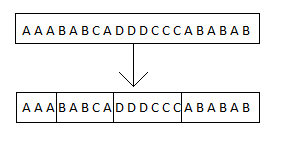
*ej. Modelo Comparativo*

Esta aproximación toma un enfoque sobre las simetrías existentes en la naturaleza, para poder acercarse a un modelo fidedigno de la estructura de una proteína. Específicamente, estos modelos toman similitudes presentes tanto en estructuras primarias como secundarias para poder dar una respuesta al problema de predicción terciaria.

La efectividad de estos algoritmos recae en la forma como las moléculas se comportan, es decir de forma predecible, para poder recrear eventos similares al propuesto. Sin embargo, no intentan hacer un análisis sobre estos comportamientos molécula a molécula, más bien operan asumiendo la repetición de estos patrones.

Estos algoritmos, por consiguiente, pueden obtener buenos resultados, dependiendo de las bases de datos sobre las cuales se estén soportando, además de la flexibilidad y confiabilidad de los estándares bajo los cuales opera, sin embargo, usualmente fallan en configuraciones de aminoácidos ambiguas, con historiales no tan claramente unilaterales hacia una u otra configuración terciaria.

* Predicción *de-novo*[e]: Una aproximación de fuerza bruta sobre la predicción, la cual ocupa grandes cantidades de recursos de procesamiento, para poder revisar un colosal número de posibilidades, frente a patrones de termodinámica e interacciones moleculares determinados por la física y la química orgánica.
* Procesamiento *Ab-Initio*[h]: Es usual utilizar otros modelos de predicción, para dividir una secuencia de aminoácidos en segmentos de alta concentración de interacciones moleculares. Esta división, facilita la labor de algoritmos predictores a nivel secundario y terciario.

*ej. Procesamiento Ab-Initio*

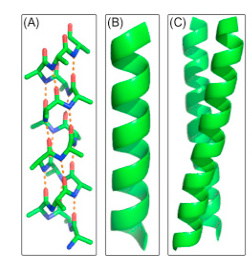
Esta técnica puede ser extremadamente útil, sobre todo a la hora de reducir considerablemente el abanico de análisis sobre el cual el algoritmo debe correr, sin embargo, puede dar una cierta parcialización a los resultados, puesto que da un mayor peso a interacciones entre los segmentos, que a interacciones entre los mismos. Por esta razón es una herramienta que no debe ser tomada a la ligera pues puede interferir drásticamente con la confiabilidad del software.

* *Aprendizaje de Maquina*: Esta herramienta ha sido utilizada en casi todas las soluciones modernas propuestas para este problema, dada la gran cantidad de posibilidades o de estructuras homologas sobre las cuales se debe encontrar patrones pertinentes.

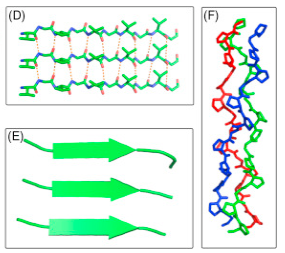
Debe tenerse en cuenta, sin embargo, que los métodos utilizados para alimentar y entrenar estos algoritmos difieren sustancialmente entre diferentes propuestas de solución. Su correcto funcionamiento depende en gran medida del pre-procesamiento de los datos con los cuales trabaja, asi como en su entrenamiento por lo que no se consideran una solución para la predicción de estructuras proteicas, por su propia cuenta.

A la hora de utilizar estas herramientas sin embrago, es necesario tomar en cuenta los patrones y tendencias encontradas en la estructura proteica general. Se han identificado un gran número de patrones, que como se ha explicado anteriormente, surgen gracias al comportamiento termodinámico de los átomos y las moléculas, en este caso se han identificado un buen número de variantes, entre las cuales, las más comunes son[c].:

* Hélice *alpha*: Estructura con forma de hélice, con propiedades hidrofóbicas, usualmente encontrado en el núcleo de las proteínas.

*Ejemplos de hélices alpha*

* Lamina *beta*: Conjunto continúo ininterrumpido de enlaces de hidrogeno.

*Ejemplo lamina beta*

* Loop: Estructuras que se encuentran entre hélices *alpha* y láminas *beta*, y se encuentran en la superficie de la estructura. No tienen un componente estructural definido.

Para el caso de la predicción de estructura secundaria la herramienta a utilizar será EVA[k]. Sin embargo, a la hora de evaluar los resultados de la predicción de estructura terciaria existen varias opciones, específicamente CAMEO3D, o CASP[c].

Para este proyecto se utilizará la herramienta CAMEO3D, la cual cuenta con un nivel inferior de aceptación y rigor, a comparación de su contraparte CASP. Sin embargo, el experimento CASP se lleva a cabo de forma bi-anual, por lo que sale del alcance de este proyecto

La herramienta CAMEO3D opera de forma conjunta con la base de datos PDB (Protein Data Bank), la cual actualiza información pertinente sobre las estructuras de las proteínas, que han sido demostradas experimentalmente. Estas estructuras nos permiten determinar las desviaciones y tendencias de los algoritmos que son puestos a prueba con la herramienta. Para calcular la precisión de cada algoritmo se utilizan modelos matemáticos con robustez frente a cambios de dominios (IDDT)[j], lo que implica que minimizan la varianza encontrada por el cambio de matices estructurales típicos de las proteínas.

Esto nos lleva a pensar en que, si bien las mejoras en referencia a precisión existen, otras mejoras como la latencia, o el uso eficiente de recursos, se vuelvan cada vez más pertinentes, por lo que también deben ser incluidos en un análisis completo sobre su funcionalidad.

La información será extraída de la base de datos Uniprot (+85 millones de proteínas)[i] en aras de obtener la información más estandarizada posible.

**Motivación:**

Como se mencionó anteriormente, la predicción de estructuras proteicas es un problema abierto en la biología computacional, lo cual implica que existen amplios márgenes de mejora e indeterminación en las soluciones propuestas actualmente. Más allá de esto, tampoco se cuenta con un conjunto de respuestas predefinidas para cada uno de los casos, lo cual implica un amplio nivel de imprecisión y resultados probabilísticos dentro del paradigma del problema. De estos factores podemos clasificar a este problema como un problema altamente retador, como un proyecto de grado. A esto se le suma el potencial científico y humanitario de los algoritmos que puedan resolver este problema, dada su anterior mencionada utilidad en campos médicos y farmacéuticos (como el reciente virus-corona proveniente de china). Mi interés sobre este tópico, se sostiene específicamente en estos dos pilares ya que siempre he considerado la programación no determinística como un campo de suma importancia, no solo en términos médicos o de biología computacional, sino para la ingeniería de la información como campo más extenso.

**Objetivo General:**

* Implementar una solución algorítmica, capaz de resolver el problema de predicción de estructura terciaria (y por ende secundaria) de proteínas, a fin de determinar su usabilidad en comparación a otras herramientas contemporáneas.

**Objetivos Específicos:**

* Implementar una herramienta de software, basada en HHpred, capaz de resolver el problema de predicción de estructura terciaria proteica.
* Revisar y explicar los algoritmos modernos propuestos para resolver el problema de la predicción de estructura secundaria y terciaria de proteínas.
* Hacer una labor comparativa en referencia al desempeño del algoritmo desarrollado, en contraste con herramientas existentes como I-TASSER en el plano moderno de la biología computacional.

**Métodos:**

Se utilizará el algoritmo TASSER[m], como referencia para comparar los procesos a completar:

* Estructura primaria (input):

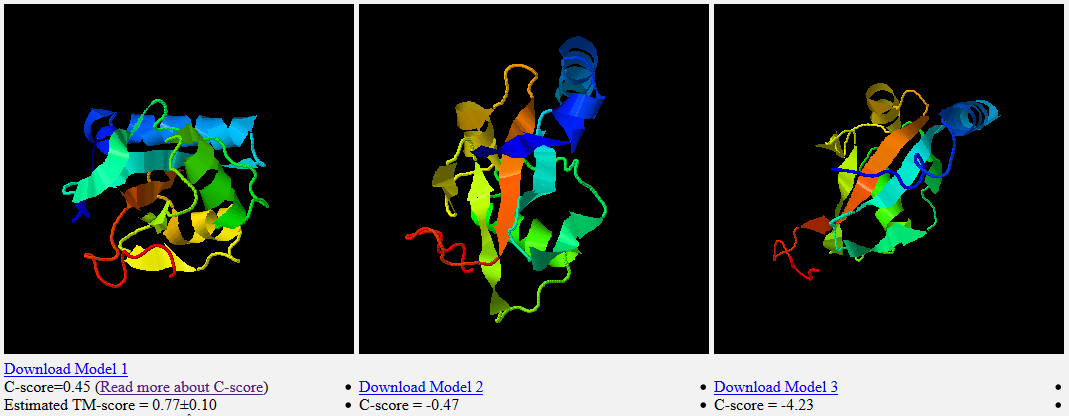
MAKSSFKISNPLEARMSESSRIREKYPDRIPVIVEKAGQSDVPDIDKKKYLVPADLTVGQ  
FVYVVRKRIKLGAEKAIFVFVKNTLPPTAALMSAIYEEHKDEDGFLYMTYSGENTFGSLT  
VA

* Estructura Secundaria (output sub-principal):





* Estructura Terciaria (output principal):

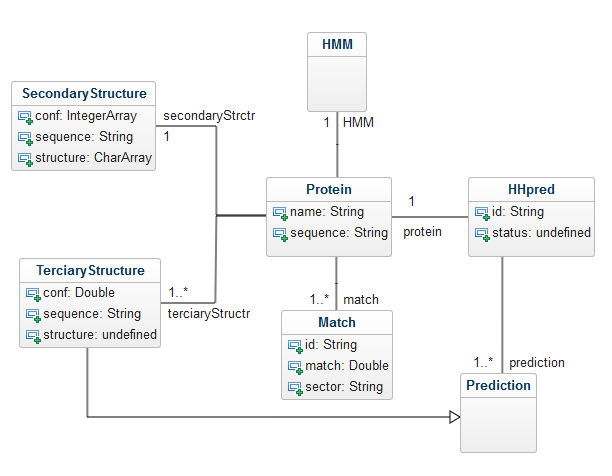


Como podemos observar, este algoritmo segmenta su funcionalidad en 2 sectores clave, justo como se hará en la implementación de este proyecto. También se toman varias medidas y procedimientos intermedios, andes de determinar el output, como medidas de confiabilidad o sectores candidatos. Explicar en detalle estas medidas sale del enfoque del proyecto, simplemente hago referencia a las mismas como una guía para posibles rutas de implementación que se podrán tomar a lo largo del desarrollo, dado que cada una de estas medidas intermedias está construida expresamente para su algoritmo base, por lo que su funcionalidad no será optimizada con una simple transcripción o adaptación.

Los pasos aquí descritos, serán el enfoque principal de los módulos de implementación del proyecto, en los cuales se llevará a cabo una comparación a base de las herramientas EVA y CAMEO3D, para el módulo 2 y 3 respectivamente, de tal forma que se pueda determinar la funcionalidad de la implementación propuesta.

A la hora de la implementación, sin embargo, se utilizará como referencia principal la herramienta líder en el año 2009[c], es decir HHpred[l].

Para la realización del proyecto se utilizará el siguiente diagrama de clases:



Bajo el cual se da a entender el funcionamiento del software. En primer lugar, la clase principal HHpred estará encargada del manejo de input y respuestas, utilizando como estructuras de datos complementarias la clase Protein, la cual a su vez estará encargada del cálculo de predicciones, tanto a nivel secundario como terciario.

**Referencias:**

1. *Xiang Z. (2006). Advances in homology protein structure modeling. Current protein & peptide science, 7(3), 217–227. doi:10.2174/138920306777452312*
2. *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(W1), W296–W303. doi:10.1093/nar/gky427*
3. *Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Tramontano, A. (2014), Critical assessment of methods of protein structure prediction (CASP) — round x. Proteins, 82: 1-6. doi:*[*10.1002/prot.24452*](https://doi.org/10.1002/prot.24452)
4. *Bhattacharya, D., Cao, R., & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics (Oxford, England), 32(18), 2791–2799. doi:10.1093/bioinformatics/btw316*
5. *Cheung NJ, Yu W (2018)* De novo *protein structure prediction using ultra-fast molecular dynamics simulation. PLoS ONE 13(11): e0205819.* [*https://doi.org/10.1371/journal.pone.0205819*](https://doi.org/10.1371/journal.pone.0205819)
6. *Zhang Y. (2008). Progress and challenges in protein structure prediction. Current opinion in structural biology, 18(3), 342–348. doi:10.1016/j.sbi.2008.02.004*
7. *Liam J. McGuffin, Kevin Bryson, David T. Jones, The PSIPRED protein structure prediction server ,* Bioinformatics*, Volume 16, Issue 4, April 2000, Pages 404–405,* [*https://doi.org/10.1093/bioinformatics/16.4.404*](https://doi.org/10.1093/bioinformatics/16.4.404)
8. *Yuan, X., Shao, Y., & Bystroff, C. (2003). Ab initio protein structure prediction using pathway models. Comparative and functional genomics, 4(4), 397–401. doi:10.1002/cfg.305*
9. *Deng, H., Jia, Y., & Zhang, Y. (2018). Protein structure prediction. International journal of modern physics. B, 32(18), 1840009. doi:10.1142/S021797921840009X*
10. *Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics (Oxford, England), 29(21), 2722–2728. doi:10.1093/bioinformatics/btt473*
11. *Ingrid Y. Y. Koh, Volker A. Eyrich, Marc A. Marti-Renom, Dariusz Przybylski, Mallur S. Madhusudhan, Narayanan Eswar, Osvaldo Graña, Florencio Pazos, Alfonso Valencia, Andrej Sali, Burkhard Rost, EVA: evaluation of protein structure prediction servers*, Nucleic Acids Research*, Volume 31, Issue 13, 1 July 2003, Pages 3311–3315,* [*https://doi.org/10.1093/nar/gkg619*](https://doi.org/10.1093/nar/gkg619)
12. *Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research, 33(Web Server issue), W244–W248. doi:10.1093/nar/gki408*
13. *Yang, J., & Zhang, Y. (2015). Protein Structure and Function Prediction Using I-TASSER. Current protocols in bioinformatics, 52, 5.8.1–5.8.15. doi:10.1002/0471250953.bi0508s52*