

Algoritmo de plegamiento de proteínas usando inteligencia de enjambre

Trabajo Terminal No. 2019-B042

Alumnos: *Avilez Benitez Luis Manuel y Cortez Tinoco Alejandro

Directores: Rosaura Palma Orozco y Rosas Trigueros Jorge Luis

*e-mail: lavilezb1500@alumno.ipn.mx

Resumen – El problema de plegamiento de proteínas, que consiste en la determinación de la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos, es un problema de gran importancia en la Bioinformática. En este proyecto se desarrollará una herramienta para la ayuda en la predicción de la estructura secundaria de proteínas. Se aprovecharán para este propósito las ventajas de algoritmos de inteligencia de enjambre.

Palabras clave – Bioinformática, Predicción de estructura de proteínas, Inteligencia de enjambre, Plegamiento de proteínas.

1. Introducción

El plegamiento de proteínas es el proceso físico mediante el cual una cadena de aminoácidos adquiere su estado tridimensional nativo, una estructura de una proteína es usualmente funcional biológicamente de forma reproducible y rápida, es esencial tener una correcta estructura tridimensional de la proteína para que sea funcional, aunque algunas partes de proteínas funcionales permanezcan desplegadas, la dinámica de las proteínas es muy importante debido a que un fallo en el plegamiento nativo de la estructura puede producir proteínas inactivas pero en algunos casos se pueden producir proteínas con funcionalidades tóxicas [1].

Entender el plegamiento de proteínas permitirá tener avances en el diseño de las proteínas comúnmente referido como problema de plegamiento de proteínas inverso, a pesar de los retos en el área de plegamiento de proteínas el diseño de proteínas *de novo* ha demostrado éxitos significativos en el diseño de proteínas y remarcando su utilidad en el área de aplicaciones biotecnológicas [2].

La inteligencia de enjambre se refiere a un tipo de habilidad para resolver problemas específicos que surge de las interacciones sencillas de procesar unidades de información simples. El concepto sugiere la multiplicidad, el proceso estocástico y aleatoriedad, el concepto de inteligencia de enjambre sugiere que este método de resolución de problemas tiene un desempeño destacado cuando el espacio de soluciones es muy amplio [3].

La predicción de estructuras de proteínas es un problema de alta complejidad y uno de los problemas más retadores en la Biología Computacional. Existen dos factores principales en la investigación de la predicción de las proteínas que lo hacen prometedor y necesario. El primero es que la información oculta en las estructuras de proteínas en 3 dimensiones es crítica para explicar el funcionamiento de las proteínas como la catálisis, el almacenamiento, el movimiento y la comunicación de las mismas. El segundo es que existen aproximadamente 60 millones de secuencias de proteínas con estructuras de 3 dimensiones desconocidas en la base de datos de UniProtKB el cual es el centro de colección de información funcional de las proteínas [4].

Tabla 1. Resumen de productos similares.

SOFTWARE	CARACTERÍSTICAS
QUARK [5].	<ul style="list-style-type: none">● Algoritmo informático para la predicción de la estructura de proteínas ab initio y el plegamiento de péptidos de proteínas, que tiene como objetivo construir el modelo 3D de proteínas correcto a partir de la secuencia de aminoácidos solamente.● Mediante la simulación Monte Carlo de intercambio de réplicas bajo la guía de un campo de fuerza basado en el conocimiento a nivel atómico.
Estimation of 3D Protein Structure by Means of Parallel Particle Swarm Optimization [6].	<ul style="list-style-type: none">● Investigación inconclusa realizada en la UNAM aplicando el algoritmo PSO para calcular la estructura tridimensional de las proteínas.

Tabla 1. Resumen de productos similares (continuación).

SOFTWARE	CARACTERÍSTICAS
Minería de datos aplicada a la predicción de proteínas [7] .	<ul style="list-style-type: none">● Aplicación de técnicas de minería de datos.● Explotación de Bases de Datos con secuencias de proteínas. ●● Obtención eficaz de resultados.● Resultados confiables.
Solución Propuesta: Algoritmo de plegamiento de proteínas usando inteligencia de enjambre.	<ul style="list-style-type: none">● Herramienta para ayudar a la predicción de la estructura secundaria de las proteínas aplicando algoritmos de enjambre.● La predicción será sin conocimiento estructurado <i>a priori</i>.

2. Objetivo

Diseñar un algoritmo de predicción de estructuras de proteínas utilizando el paradigma de inteligencia de enjambre para obtener estructuras tridimensionales de proteínas a partir de secuencias de aminoácidos.

Objetivos específicos

- Elegir el tipo de inteligencia de enjambre que ofrezca un mejor costo beneficio para la predicción de estructuras de proteínas.
- Implementar el algoritmo desarrollado usando un lenguaje de alto nivel para evaluar su efectividad. -
- Determinar secuencias de aminoácidos que permitan evaluar la efectividad del algoritmo implementado.

3. Justificación

La determinación de las estructuras de proteínas por métodos experimentales (rayos X, cristalografía y la espectroscopía de resonancia magnética nuclear) ofrecen un costo y tiempo elevados para ser llevados a cabo, por lo cual determinar las estructuras mediante tecnologías computacionales se convierte en algo prometedor e incluso necesario [4].

La inteligencia de enjambre ha sido reconocida por su habilidad de producir soluciones de bajo costo, rápidas y de resultados razonablemente precisos para problemas de alta complejidad en el área de la Bioinformática [8].

La propuesta es utilizar los conocimientos adquiridos a lo largo de la carrera para diseñar una herramienta que analice la predicción de estructuras de proteínas utilizando algoritmos de inteligencia de enjambre sin conocimiento estructurado *a priori*, ya que algunas investigaciones han logrado mejorar la predicción utilizando algoritmos PSO en conjunto con “*levy flight*” logrando que aumente la precisión del algoritmo [9].

4. Productos o Resultados esperados

La herramienta funcionará con la entrada de una secuencia de aminoácidos que tiene la forma del ejemplo de la figura 1 y que representa la estructura primaria de la proteína la cual será introducida en formato FASTA.

```
>1UBQ:A|PDBID|CHAIN|SEQUENCE
MQIFVKLTITGKTIITLEVEPSDTIENVKAKIQDKEGIPPDQQRLLIFAGKQLEDGRTLSDY
NIQKESTLHLVLRGG
```

Figura 1. Representación de ubiquitina en formato FASTA.

Posteriormente la herramienta, mediante la aplicación de algoritmos de inteligencia de enjambre sin conocimiento estructurado *a priori*, intentará determinar su posible estructura secundaria, que será la salida en un formato PDB (Protein Data Bank).

Para evaluar la efectividad de la herramienta se comparará su salida con estructuras ya conocidas.

La arquitectura de la herramienta se muestra en la figura 2.

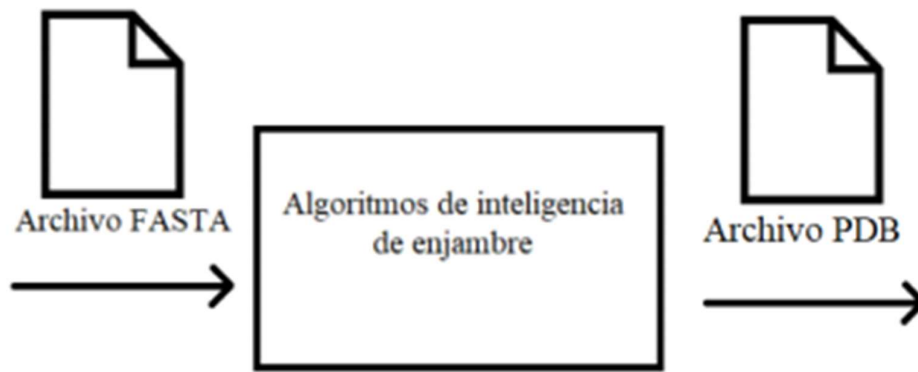


Figura 2. Arquitectura del sistema.

Los productos esperados del TT son:

1. El código.
2. La documentación técnica de la herramienta.
3. El manual de usuario.

5. Metodología

La metodología a emplear en el desarrollo de la herramienta será la metodología Métrica versión 3 en su esquema estructurado, además de ser útil para sistematizar las actividades que dan soporte a todo el ciclo de vida del software.

La metodología Métrica versión 3 [10] contiene todas las actividades y tareas que se deben llevar a cabo para desarrollar la herramienta, cubriendo desde la etapa del estudio de viabilidad hasta la implantación y aceptación del software, conteniendo los siguientes procesos:

- Análisis del sistema de información (ASI).
- Diseño del sistema de información (DSI).
- Construcción del Sistema de Información (CSI).
- Implantación y Aceptación del Sistema (IAS).

6. Cronograma

Ver anexo 1 y 2.

7. Referencias

- [1] "Protein folding", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Protein_folding. [Accessed: 09-Sep- 2019].
- [2] J. Kennedy, *Handbook of Nature-Inspired and Innovative Computing*. 2006, pp. 187.
- [3] "Protein structure prediction", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Protein_structure_prediction. [Accessed: 18- Sep- 2019].
- [4] S. Song, J. Ji, X. Chen, S. Gao, Z. Tang and Y. Todo, "Adoption of an improved PSO to explore a compound multi-objective energy function in protein structure prediction", *Applied Soft Computing*, vol. 72, pp. 539-551, 2018. Available: 10.1016/j.asoc.2018.07.042.
- [5] "De Novo Protein Structure Prediction by QUARK", *Zhanglab.ccmb.med.umich.edu*, 2019. [Online]. Available: <https://zhanglab.ccmb.med.umich.edu/QUARK/?fbclid=IwAR2GsRp3aJOW45mjEfjRh7B0QduBjepjCkI9aH1xKVZ3fmVtqsmQHsFTwY>. [Accessed: 19- Sep- 2019].
- [6] Pérez Germán, Vázquez Katya, Garduño Ramón, "Estimation of 3D Protein Structure by Means of Parallel Particle Swarm Optimization" IIMAS-UNAM, Circuito Escolar, CU, 04510 México D.F., México.
- [7] Salazar Marco, Ruiz Rodrigo, Domínguez Paredes, Santos Raúl, "Minería de datos aplicada a la predicción de proteínas". Trabajo terminal, CDMX.

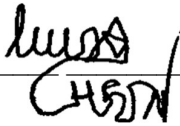
[8] S. Das, A. Abraham and A. Konar, *Computational intelligence in bioinformatics*. Berlin: Springer, 2008, pp. 113-147.

[9] X. Chen, M. Lv, L. Zhao and X. Zhang, "An Improved Particle Swarm Optimization for Protein Folding Prediction", *Mecs press.net*, 2011. [Online]. Available: <http://www.mecs-press.net/ijieeb/ijieeb-v3-n1/IJIEEB-V3-N1-1.pdf>. [Accessed: 18- Sep 2019].

[10] Métrica versión 3, España Ministerio para las Administraciones Públicas, editorial Ministerio para las Administraciones Públicas, 2001

8. Alumnos y Directores

Avilez Benitez Luis Manuel.- Estudiante de la carrera de Ing. en Sistemas Computacionales en ESCOM Boleta: 2016630020, Tel: 7751449246 e-mail: lavilezb1500@alumno.ipn.mx

Firma: 

Cortez Tinoco Alejandro - Estudiante de la carrera de Ing. en Sistemas Computacionales en ESCOM Boleta :2016630436, Tel:: 8112517875 e-mail acortezt1500@alumno.ipn.mx

Firma: Alejandro Cortez Tinoco

Rosaura Palma Orozco.- Dra. en Tecnología Avanzada por el IPN (2012) M en C. en Matemáticas por el CINVESTAV (2004), Ing. en Sistemas Computacionales por la Escuela Superior de Cómputo del IPN (1998). Actualmente es profesora Titular en ESCOM. Áreas de interés: Modelado y Simulación de Sistemas, Sistemas Complejos, Biología Sintética y Optimización Combinatoria. email: rpalma@ipn.mx

Firma: 

Jorge Luis Rosas Trigueros.-Dr. en Ciencias en Biotecnología por el IPN (2012), M. en C. en Ing. Eléctrica por la Universidad de Texas A&M en College Station, Estados Unidos (2002), es Ing. en Sistemas Computacionales por la Escuela Superior de Cómputo del IPN (1998). Actualmente es profesor Titular en ESCOM y sus áreas de interés son: Modelado y Simulación Molecular, Bioinformática y Graficación. e-mail: jlrosas@ipn.mx

Firma: 

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Anexo 1

Nombre de alumno: Avilez Benitez Luis Manuel TT No.

Título del TT: Algoritmo de plegamiento de proteínas usando inteligencia de enjambre

Actividad	ENE	FEB	MAR	ABR	MAY	JUN	AGO	SEP	OCT	NOV	DIC
Determinación del Alcance del sistema											
Identificación del Entorno Tecnológico											
Obtención de requisitos											
Análisis de requisitos											
Determinación de Subsistemas de Análisis											
Elaboración del modelo de datos											
Obtención del Modelo de Procesos del Sistema											
Especificación de Interfaces con otros Sistemas											
Especificación de Principios Generales de la Interfaz											
Definición de Niveles de Arquitectura											
Identificación de Requisitos de Diseño y Construcción											
Especificación de Excepciones											
Diseño de Módulos del Sistema											
Diseño de Comunicaciones entre Módulos											
Evaluación de TT1											
Generación del Código de Componentes											
Generación del Código de los Procedimientos de Operación y Seguridad											
Preparación del Entorno de las Pruebas Unitarias											
Realización y Evaluación de las Pruebas Unitarias											
Elaboración de los Manuales de Usuario											
Realización y evaluación de las Pruebas de Aceptación											
Evaluación de TT2											

Anexo 2

Nombre de alumno: Cortez Tinoco Alejandro TT No.

Título del TT: Algoritmo de plegamiento de proteínas usando inteligencia de enjambre

Actividad	ENE	FEB	MAR	ABR	MAY	JUN	AGO	SEP	OCT	NOV	DIC
Determinación del Alcance del sistema											
Identificación del Entorno Tecnológico											
Obtención de requisitos											
Análisis de requisitos											
Determinación de Subsistemas de Análisis											
Obtención del Modelo de Procesos del Sistema											
Especificación de Principios Generales de la Interfaz											
Definición de Requisitos del Entorno de Pruebas											
Especificación de Formatos de Impresión											
Identificación de Requisitos de Diseño y Construcción											
Especificación de Excepciones											
Diseño de Módulos del Sistema											
Evaluación de TT1											
Generación del Código de Componentes											
Generación del Código de los Procedimientos de Operación y Seguridad											
Preparación del Entorno de las Pruebas Unitarias											
Realización y Evaluación de las Pruebas Unitarias											
Preparación de las Pruebas de Implantación											
Realización de las Pruebas de Integración											
Evaluación del Resultado de las Pruebas de Integración											
Evaluación de TT2											