Una herramienta de propósito general para el plegado de proteínas con técnicas probabilísticas

Luis J. Calva-Rosales, Abraham Sánchez-López, Pablo Camarillo-Ramírez, Juan Carlos Conde-Ramirez

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla, México

{luis.calva, asanchez, pablo.camarillo, juan.conde}@cs.buap.mx

Resumen. La importancia de este trabajo radica en el hecho de que el estudio de plegado molecular tiene el fin de prevenir y entender mutaciones que ponen en peligro la vida de los seres vivos. El plegado molecular de proteínas es de vital importancia para entender los factores fisiológicos externos e internos que provocan que una proteína monomérica pase de un estado a otro. En este articulo se describe el desarrollo de una herramienta que calcula y simula el plegado de proteínas, utilizando técnicas de robótica como PRM (Probabilistic Roadmap Methods). Actualmente la cantidad de proteínas analizadas es escasa dado que las herramientas existentes son caras y necesitan de costosos equipos para realizar esta tarea. Por lo cual se desarrollo esta herramienta utilizando C++ con librerías QT como lenguaje de programación, junto con OpenGL y se consume el servicio Web DSSP para la asignación de estructuras secundarias en bancos de datos de proteínas.

Palabras clave: Plegado molecular, proteínas, PRM, simulación.

1. Introducción

Entre las muchas áreas de la bioinformática, existe la biología computacional estructural, la cual se encarga de estudiar las estructuras biológicas (ADN, Proteínas, etc) con la meta de obtener características, información, verificar comportamientos o calcular energía y poder estudiar sus efectos en los seres vivos [17].

En particular, las proteínas son estructuras fundamentales para todos los seres vivos. Cada proteína consiste de una secuencia de residuos de aminoácidos [5]. A su vez, cada aminoácido en una proteína es llamado *residuo* porque pierde dos átomos de hidrógeno y uno de oxigeno durante la formación de la *cadena péptida* entre dos aminoácidos adyacentes. De esta manera, bajo ciertas condiciones fisiológicas una sola proteína puede llegar a formar una estructura tridimensional compacta y estable; conocida como el estado nativo de una proteína [3]. El proceso para formar un estado nativo se llama plegado de proteínas. Existen dos problemas principales en el plegado de proteínas [16,13,14]:

1. **Predicción de las estructuras del estado nativo de la proteína**. Este problema es normalmente referenciado como la predicción de la estructura nativa de una proteína.
2. **Problema del plegado de la proteína**. Trata del estudio de la secuencia de las transiciones realizadas por los aminoácidos dinámicamente a partir de un estado no estructurado a la estructura nativa (única).

Este último se centra en el proceso de doblado dinámico y temas relacionados con la identificación de los caminos y el cálculo de la cinética de plegamiento.

Para el desarrollo de esta herramienta se utilizaron técnicas de robótica para modelar y configurar las proteínas con el fin de obtener energías y caminos que concuerden con características biológicas.

El contenido de este articulo esta distribuido como sigue. En la Sección 2 se detalla brevemente el estado del arte y se establece la relación con la planificación de movimientos en robótica. Posteriormente, la Sección 3 establece los elementos esenciales a considerar destacados por otros trabajos relacionados. Por su parte la Sección 4 describe el enfoque seguido durante la investigación para construir una herramienta que realiza el cómputo del plegado de proteínas en modo gráfico, así como las técnicas utilizadas. En la Sección 5 se definen los parámetros y las características de las pruebas realizadas con la herramienta propuesta, junto con los resultados medidos en tiempo y posibles caminos obtenidos. Al final, la Sección 6 muestra las conclusiones y el trabajo futuro de esta investigación.

2. Problema del plegado proteínico

Las proteínas y *polipéptidos* son compuestos de enlaces de aminoácidos, esas composiciones son conocidas como la estructura primaria de la proteína. Un aminoácido es una molécula orgánica simple que consiste de un *amino* básico (receptor de hidrógeno), unido a un ácido (donante de hidrógeno) a través de un único átomo de carbono intermedio. Cada aminoácido consiste de un átomo de carbono *tetrahédrico* central conocido como la *alpha* (α) de carbono (C^{α}) el cual tiene cuatro enlaces: un átomo de hidrógeno, un grupo *amino* receptor de átomo (NH_3^+), un grupo ácido pierde un átomo (COO^-), y una cadena lateral distintiva o grupo R que hace la diferencia entre los diferentes aminoácidos (Figura 1).

Entonces el proceso de plegado es importante por varias razones, ya que puede proporcionar la idea de cómo es que se pliegan las proteínas y puede ayudar a entender los factores que controlan este proceso en algunas proteínas (mecanismos). Esta área de investigación es de gran importancia práctica ya que existen algunas enfermedades devastadoras provocadas por plegamientos no nativos de la proteína, como la *encefalopatía espongiforme bovina*[1] [12], en estos casos es importante entender por qué ocurren estos plegados y cómo se podrían prevenir.

[1] La enfermedad de las vacas locas

2.1. Relación con la planificación de movimientos

La planificación de movimientos es un problema fundamental de la robótica que ha sido investigado por más de tres décadas. Se han propuesto una gran variedad de algoritmos para calcular los movimientos factibles de múltiples cuerpos y sistemas en diferentes espacios. En los últimos años, muchos de estos algoritmos han empezado a pasar las barreras de la robótica, encontrando aplicaciones en otros campos como la manufacturación industrial, la animación por computadora y la biología computacional estructural.

Debido a la gran similitud entre las proteínas y los robots manipuladores, en términos de movimiento, se pueden aplicar los algoritmos de planificación utilizados en la robótica en el campo de las proteínas. Esta visión es importante en el sentido de que se pueden obtener diversos caminos que pueden ayudar a entender el comportamiento de la proteína que se esta estudiando.

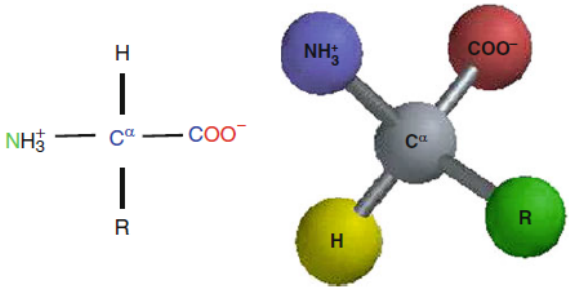


Fig. 1. Forma general para un aminoácido (izquierda), y disposición *tetrahédrica* espacial de un aminoácido (derecha).

3. Trabajo relacionado

Dentro del campo del diseño molecular[2] y de acuerdo a la literatura, existen diferentes maneras en las cuales se puede llegar a modelar una proteína, desde coordenadas cartesianas hasta complejos modelos físicos.

En [10] se muestran los tres diferentes grados de libertad internos de una proteína:

1. El **largo de enlace** que es la distancia de separación entre un par de átomos unidos el cual tiende a variar muy poco.
2. El **ángulo de enlace** que sólo tiene sentido cuando consideramos 3 átomos, y es el ángulo formado por los ejes imaginarios que unen el núcleo del átomo central con los núcleos de los átomos unidos a él.

[2] Ciencia que estudia la estructura y funcionamiento de estructuras moleculares a través de la construcción de modelos físicos o computacionales

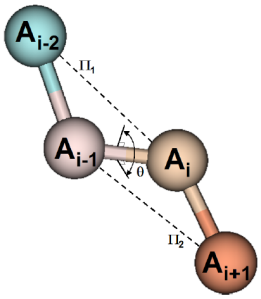


Fig. 2. Ángulo diedro definido como el ángulo más pequeño entre el plano π_1 y π_2.

3. El **ángulos diedros** que es definido por cuatro consecutivos enlaces de átomos como se muestra en la Figura 2.

En la Figura 3 se muestra los ángulos de libertad diedros del aminoácido *Arginina*, dependiendo el aminoácido los ángulos diedros en la cadena lateral pueden variar, pero los ángulos diedros de la cadena *péptida* son constantes(2) en todos los aminoácidos.

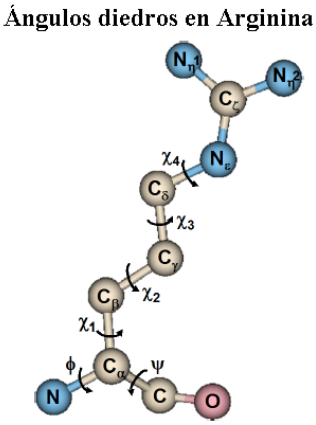


Fig. 3. Las cadenas laterales *diedras* se designan por X y un subíndice. Las cadenas internas *diedras* se designan por Φ y Ψ.

En este trabajo [20] la proteína es modelada como un robot articulado. Solo los ángulos diedros de la cadena *péptida* son utilizados como grados de libertad, los ángulos diedros de las cadenas laterales, el largo de enlace y el grado de enlace son considerados valores constantes. De esta manera una proteína modelada con nuestra aproximación contara con $2k$ grados de libertad.

Para la realización de un modelo como este se utiliza la convención *Denavit-Hartember* que permite modelar y configurar la estructura de una proteína. Se utilizó este enfoque y no uno de agrupación porque permite de una manejar todos los ángulos diedros internos de la proteína [20].

4. Herramienta propuesta

Para el desarrollo de esta herramienta se implementaron diversos módulos que a continuación se explicaran.

Para obtener la información necesaria para modelar una molécula se desarrollo un **interprete** para cargar archivos de tipo PDB que se pueden obtener de la página `www.rcsb.org` (*Protein Data Bank*), la descripción de estos archivos se puede encontrar en [7].

Una vez obtenida la información del archivo PDB que se desea estudiar, se desarrollo un **modulo de modelado y diseño molecular** el cual permite manipular y visualizar la proteína.

El modulo desarrollado es el **modulo que contiene la algoritmia** que permite emplear el algoritmo PRM para resolver el problema de plegado de proteína. En este caso, la Figura 4 permite observar el comportamiento del plegamiento realizado por la proteína.

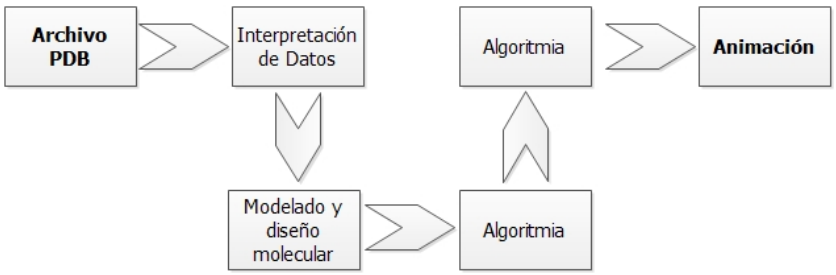


Fig. 4. Una visión general de la arquitectura propuesta del sistema.

4.1. Técnicas de robótica

Ya que la meta de los algoritmos de planificación de movimiento es ayudar a que un robot que se encuentra en un estado inicial encuentre un camino factible a configuración meta [15], los planificadores deben de ser capaces de garantizar una solución válida o determinar si ésta existe o no. La búsqueda de un camino requiere un tiempo exponencial que se puede incrementar dependiendo la cantidad de ángulos de libertad que tenga el robot. En el caso de las proteínas y en nuestro modelo la cantidad de grados de libertad está directamente relacionada al número de aminoácidos que esta contenga.

La herramienta desarrollada resuelve el problema del plegado de proteínas utilizando técnicas basadas en un mapa de ruta probabilístico (*probabilistic roadmap, PRM*) [9]. PRM es un algoritmo basado en muestras aleatorias de un espacio de configuraciones definido como $C - space$, donde si una muestra es factible es agregada a un grafo esta etapa se conoce como **etapa de muestreo**. Posteriormente se realiza la conexión de los mismos utilizando medidas de distancia y de restricción (en el caso de las proteínas restricciones energéticas) conocida como **etapa de conexión**. Una vez que existe un camino entre la configuración inicial y la configuración meta se procede a utilizar un algoritmo de búsqueda como lo es $A*$ o *Dijsktra*, a ésta se le denomina **etapa de consulta**. La meta de este algoritmo es encontrar un camino entre la configuración inicial y meta.

4.2. Consideraciones importantes

El problema de plegado de proteínas tiene diferencias notables de la aplicación usual de PRM. Las colisiones tradicionales son remplazadas por configuraciones con poca energía, en este caso se utilizó el algoritmo de BioCD para realizar esta tarea [4][19]. De esta manera se miden las pequeñas diferencias energéticas que se realizan en cada parte del proceso de PRM. En la aplicación general de PRM es suficiente encontrar un solo camino, en el caso del plegado de proteínas es importante la calidad del camino, se busca el camino más favorable.

En la Figura 5 podemos observar las diferentes etapas del desarrollo de PRM en el campo de las proteínas: (a) etapa de muestreo, (b) etapa de conexión y (c) etapa de consulta [1][2].

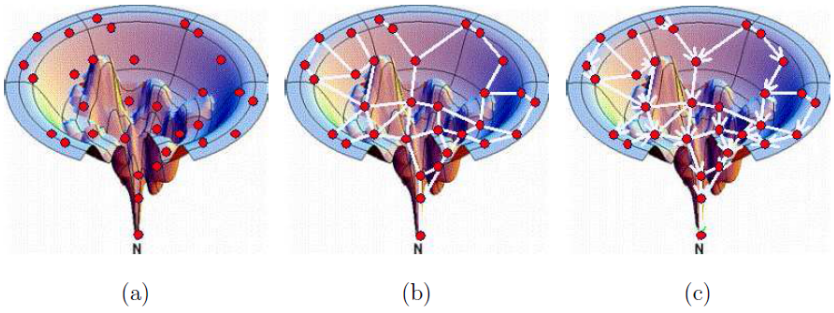


Fig. 5. Aplicación de PRM en el campo de las proteínas.

5. Experimentos y resultados

La herramienta propuesta fue desarrollada utilizando el lenguaje de programación C++ con QT, la librería de gráficos OpenGL, y el servicio Web DSSP

[8]. La herramienta se ejecutó en una computadora con un procesador Intel Core I5 con 16 gigas de RAM.

La herramienta completa se puede observar en las Figuras 6(a) y 6(b). A continuación se describe cada componente que se encuentra en la herramienta desarrollada:

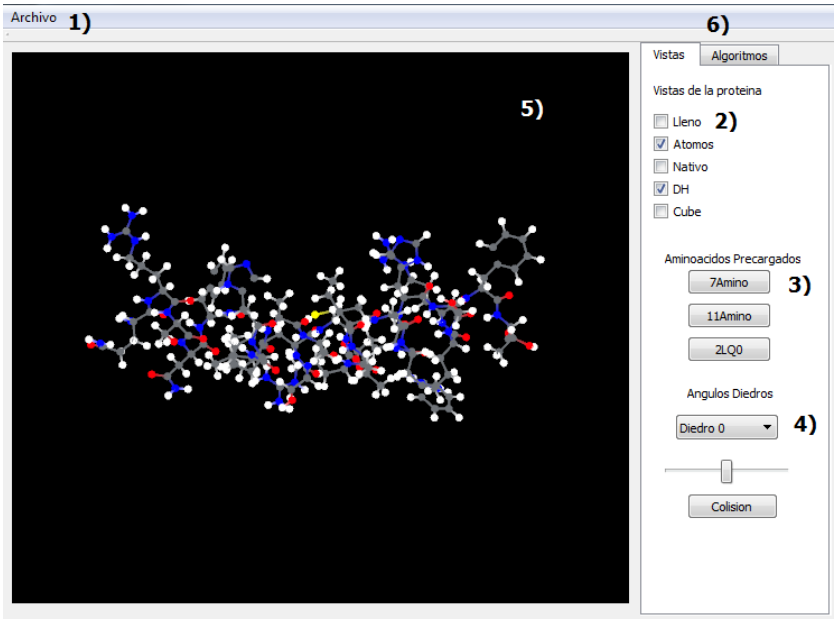
1. En esta sección se pueden abrir proteínas que hayan sido descargadas de la *Protein Data Bank*, permitiendo al usuario utilizar la proteína *monomérica* que desee.
2. Estas casillas de verificación le permiten al usuario modificar la vista actual de la proteína.
3. Estos botones tienen precargadas las proteínas que se utilizaron en este trabajo.
4. En esta sección se puede configurar el estado actual de la proteína modificando los grados de libertad de los distintos ángulos diedros. Con el botón de colisión se puede verificar si la configuración realizada es una configuración correcta.
5. Este lienzo muestra la proteína que está cargada en el sistema dependiendo las casillas que estén verificadas en la sección 2.
6. Estas pestañas nos permite cambiar entre la sección de vistas y algoritmos.
7. Permiten visualizar las configuraciones natural (nativa) y estirada de la proteína, y en el caso del botón invalido encontrar configuraciones no válidas.
8. Son los parámetros básicos necesarios para resolver el problema del plegado de proteínas utilizando PRM.
9. Son los botones que nos permiten inicializar y calcular el *roadmap* con PRM para resolver el problema del plegado de proteínas.
10. Una vez resuelto el *roadmap* se puede visualizar el camino encontrado haciendo clic en el botón animar y modificar la velocidad de la misma modificando la barra horizontal.

Se diseñaron 2 moléculas con 7 aminoácidos ($7Amino$), otra con 11 aminoácidos ($11Amino$) similarmente como se creo la molécula $10 - ALA$ propuesta en [18]. Además se utilizó una molécula más grande de nombre $2LQ0$ que cuenta con 25 aminoácidos.

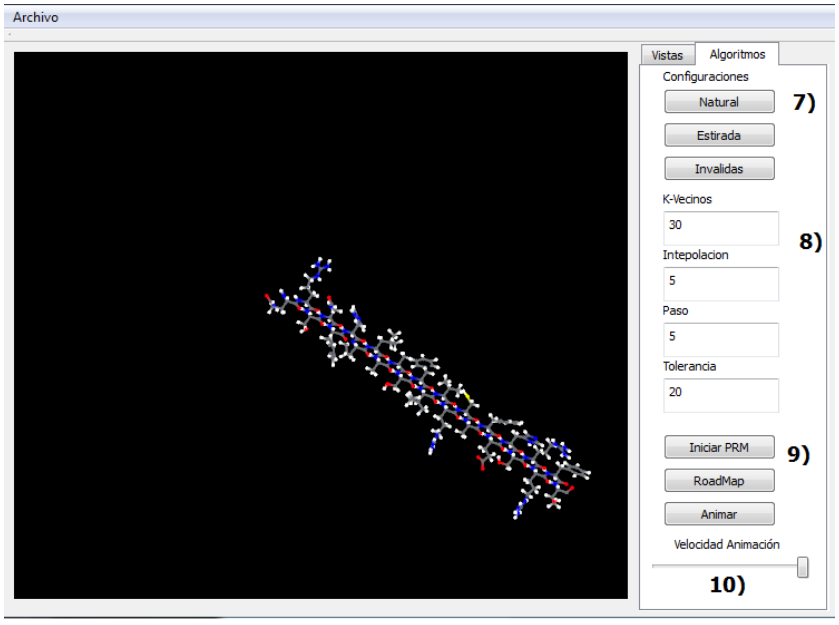
Los parámetros utilizados en el algoritmo de PRM en el problema de plegado de proteína son:

- 30 k-vecinos
- Una interpolación entre configuraciones de 5
- Una distancia euclidiana de 20 para el *Roadmap*[1]
- Un tamaño de paso de 5 para la conexión

Así mismo, se realizaron 5 pruebas con cada una de estas moléculas obteniendo los resultados que se muestran en la Tabla 1.



(a) Esta sección muestra las configuraciones de la vista, y nos muestra la proteína $2LQ0$ en su estado nativo.



(b) Esta sección muestra las configuraciones de algoritmia, se muestra la misma proteína $2LQ0$ en su estado estirado.

Tabla 1. Resultados obtenidos con la herramienta.

Molécula	Nodos	Aristas	Caminos	Tiempo(Horas)
7Amino	722	3107	6	1.5
11Amino	836	47886	4	3.2
2LQ0	2343	107702	3	4.1

6. Conclusiones y trabajo futuro

En este trabajo, se mostró una herramienta para el estudio del plegado de proteínas utilizando técnicas de robótica.

Los resultados al ejecutar dicha herramienta fueron satisfactorios ya que se encontraron diferentes rutas para los procesos de plegado de las diferentes proteínas que se estudiaron. En comparación a las técnicas de Monte Carlo[6,11] que solo obtiene un camino de plegado, las técnicas de PRM [2,18] obtuvieron diferentes caminos de plegado.

Se planea utilizar en trabajos futuros las tecnologías de GPUs para la realización de cálculos y disminuir el tiempo de procesamiento. PRM se adaptó de manera correcta al problema del plegado de proteínas, sin embargo sería importante compararlo con otras aproximaciones basadas en árboles como los RRT.

Actualmente la herramienta permite configurar el diseño molecular de una proteína dada, sería importante en trabajos futuros extender esta visión para trabajar en el ligado de proteínas lo cual es útil en la creación de nuevos fármacos y el diseño de nuevos nanomateriales.

Referencias

1. Amato, N.M., Dill, K.A., Song, G.: Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. Journal of Computational Biology 10(3/4), 239–255 (2003)
2. Amato, N.M., Song, G.: Using motion planning to study protein folding pathways. Journal of Computational Biology 9(2), 149–168 (2002)
3. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science 181, 223–230 (1973)
4. de Angulo, V.R., Cortés, J., Siméon, T.: Biocd : An efficient algorithm for self-collision and distance computation between highly articulated molecular models. In: Robotics: Science and Systems. pp. 241–248 (2005)
5. Brändén, C., Tooze, J.: Introduction to Protein Structure. Introduction to Protein Structure Series, Garland Pub. (1999)
6. Covell, D.G.: Folding protein a-carbon chains into compact forms by monte carlo methods. Proteins: Structure, Function, and Bioinformatics 14(3), 409–420 (1992)
7. Doe, R.: Pdb file format @ONLINE (1996)
8. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–637 (1983 Dec 1983)

9. Kavraki, L., Svestka, P., claude Latombe, J., Overmars, M.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. In: IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION. pp. 566–580 (1996)
10. Kavraki, L.E.: Geometric Methods in Structural Computational Biology. (2007)
11. Kolinski, A., Skolnick, J.: Monte carlo simulations of protein folding. ii. application to protein a, rop, and crambin. Proteins 18, 353–66 (1994 Apr 1994)
12. Lansbury, P.: Evolution of amyloid: What normal protein folding may tell us about fibrillogenesis and disease. Proc. Natl. Acad. Sci. USA 96
13. Levitt, M., Gerstein, M., Huang, E., Subbiah, S., Tsai, J.: Protein folding:the endgame. Annual Review of Biochemistry 66(1), 549–579 (1997), pMID: 9242917
14. Munñoz, V., Eaton, W.A.: A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci U S A 96(20), 113–116 (1999)
15. O'Rourke, J.: Folding and unfolding in computational geometry. In: Akiyama, J., Kano, M., Urabe, M. (eds.) Discrete and Computational Geometry, Lecture Notes in Computer Science, vol. 1763, pp. 258–266. Springer Berlin Heidelberg (2000)
16. Reeke, G.N.: Protein folding: Computational approaches to an exponential-time problem. Annual Review of Computer Science 3(1), 59–84 (1988)
17. Schlick, T.: Molecular Modeling and Simulation: An Interdisciplinary Guide. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2002)
18. Song, G., Amato, N.M.: A motion-planning approach to folding: from paper craft to protein folding. IEEE T. Robotics and Automation 20(1), 60–71 (2004)
19. Thrun, S., Sukhatme, G.S., Schaal, S. (eds.): Robotics: Science and Systems I, June 8-11, 2005, Massachusetts Institute of Technology, Cambridge, Massachusetts. The MIT Press (2005)
20. Zhang, M., Kavraki, L.E.: Solving molecular inverse kinematics problems for protein folding and drug design. In: Currents in Computational Molecular Biology. pp. 214–215. ACM Press, ACM Press (April 2002), book includes short papers from The Sixth ACM International Conference on Research in Computational Biology (RECOM 2002), Washington, DC, 2002