Estrategia de Agrupamiento Rápido para Reducir Trayectorias Largas de Plegamiento de Proteínas

Luis Garreta, Mauricio Martinez, Pedro A. Moreno

Grupo de Investigación en Bioinformática, Facultad de Ingeniería de la Universidad del Valle (Cali, Colombia).

Resumen

Gracias a los avances tanto en software como en hardware, las simulaciones de plegamiento de proteínas están llegando a las escala de los micro y milisegundos, generando trayectorias de plegamiento muy largas con miles y millones de conformaciones de proteínas, lo cual vuelve difícil su procesamiento y análisis. En este artículo presentamos un algoritmo para reducir este tipo de trajectorias, que mediante una estrategia paralela y de agrupamiento rápido, logra obtener rápidamente las conformaciones más representativas de la trayectoria conservando de estas tanto su estructura tridimensional (3D) como su orden temporal. De acuerdo a los resultados, el algoritmo logra reducciones con una compresión muy alta y en tiempos muy cortos, comparado con otros métodos típicos de reducción de plegamiento de proteínas.

Palabras claves: Reducción de trayectorias de plegamiento de proteína; Plegamiento de proteínas; Algoritmos de agrupamiento rápido; bioinformática.

1. Introducción

Las proteínas desempeñan funciones fundamentales en todos los seres vivos, pero para ser funcionales deben a partir de su cadena de aminoácidos plegarse hasta alcanzar una forma 3D única, lo que se conoce como el proceso de plegamiento de las proteínas. Entender los mecanismos y reglas de este proceso es aún un problema vigente y de los más perseguidos dentro de la biología y la computación biológica.

Una herramienta teórica importante para estudiar este proceso son las simulaciones del plegamiento de las proteínas las cuales producen trayectorias de plegamiento, que describen la evolución del plegamiento de una proteína mediante la secuencia de estados que esta atraviesa en función del tiempo. Hoy en día, estas simulaciones están alcanzando unos tiempos de simulación muy grandes en comparación a los que se tenían hace algunos años, del orden de

los milisegundos y microsegundos, y como consecuencia las trayectorias generadas por estas simulaciones son muy extensas y analizarlas trae complicaciones debido a los miles o millones de conformaciones que contienen.

Este problema no es nuevo y se han usado diferentes estrategias basadas principalmente en dos enfoques: la reducción de la dimensionalidad [1] y el agrupamiento [2]. En el primer enfoque se transforma una conformación a un conjunto reducido de variables que la representan lo mejor posible. Para esto se han usado tanto técnicas lineales como no-lineales (e.g. análisis de componentes principales (PCA), escalamiento multi-dimensional [3], Isomap [4], diffusion maps [5]). Sin embargo, aunque se logra la reducción de las conformaciones, se pierde su representatividad como estructuras 3D. Además, estas técnicas consumen mucho tiempo cuando las trayectorias son muy grandes, ya que tienen que transformar todas sus conformaciones.

En este artículo presentamos un algoritmo de reducción para trayectorias largas de plegamiento de proteínas. El algoritmo se caracteriza por dividir la trayectoria en segmentos y mediante una estrategia rápida de agrupamiento, toma de cada uno de ellos a los eventos más disimilares y posteriormente selecciona entre ellos a los k más representativos. El algoritmo aprovecha el orden temporal implícito en la trayectoria para realizar en cada segmento comparaciones locales, entre conformaciones de proteínas vecinas, evitando realizar una comparación de todos contra todos que se vuelve impráctica computacionalmente cuando son miles o millones de conformaciones. De esta manera, el algoritmo reduce muy rápidamente la trayectoria y las conformaciones seleccionadas conservan su orden temporal dentro de la trayectoria. Además, el particionamiento en segmentos permite al algoritmo realizar la reducción por cada segmento de forma independiente y por lo tanto realizar las reducciones de forma paralela, lo que lo vuelve aún más rápido cuando se ejecuta en máquinas multi-core, muy comunes hoy en día.

2. Descripción del Algoritmo

El algoritmo propuesto reduce una trayectoria de plegamiento en tres fases (Figura 1): primero divide la trayectoria en pequeñas subtrayectorias que luego las reduce de manera individual, independiente y paralela. Segundo toma cada subtrayectoria y extrae de forma muy rápida sus conformaciones características y elimina las redundantes utilizando la estrategia de agrupamiento rápido de Hobohm and Sander (1992). Y tercero toma la conformaciones características y selecciona las más representativas mediante una estrategia tipo k-medoides [6], la cual al trabajar sobre pocas conformaciones, mejora sustancialmente su desempeño. Al final, los resultados de cada reducción se unen para obtener la reducción total de la trayectoria.

Además, a diferencia de otras técnicas de reducción de trayectorias, el algoritmo tiene la ventaja de no cambiar la representación de las conformaciones como lo hacen las técnicas de reducción de dimensionalidad, ni de perder el orden temporal como lo hacen las técnicas de agrupamiento. El resultado de nuestro

algoritmo es un conjunto de conformaciones representativas de la trayectoria que siguen conservando tanto su estructura 3D como su orden temporal.

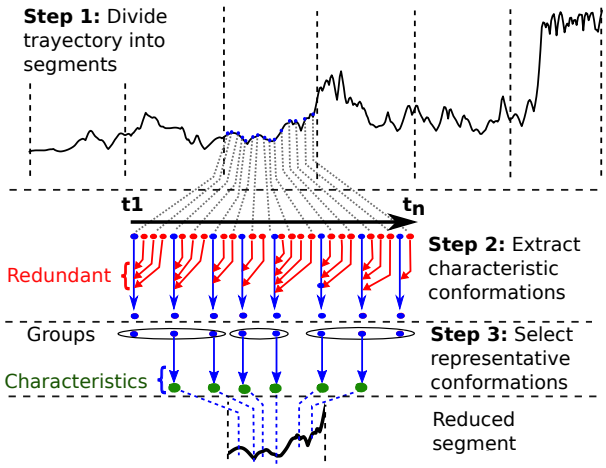

Figura 1: **Algoritmo de reducción.** La trayectoria se divide en subtrayectorias (líneas discontinuas) que se reducen individualmente. Mediante una estrategia de agrupamiento rápido [7] se seleccionan las conformaciones características (puntos rojos) y se ignoran las redundantes (puntos azules). Después, mediante un agrupamiento tipo k-medoides [6] se extraen las representativas que forman la subtrayectoria reducida.

El pseudocódigo del algoritmo propuesto se muestra en el listado 1. La implementación del algoritmo se realizó en lenguaje R excepto la comparación entre conformaciones, que usa la métrica TM-Score para comparar pares de estructuras de proteínas [8], y que por ser la parte que más se repite está implementada en lenguaje Fortran.

Algoritmo 1 Pseudocódigo del algoritmo propuesto de reducción de trayectorias largas de plegamiento.

```
1  Algoritmo de reducción
2  ENTRADA: T: Trayectoria original de plegamiento
3  SALIDA : Trayectoria reducida de plegamiento
4
5  AlgoritmoReduccion (T)
6    Particionar la trayectoria en segmentos
7    Para cada segmento de forma paralela
8       Extraer rápidamente las conformaciones más disimilares
9       Seleccionar de las disimilares a las más representativas
10   Concatenar los resultados de cada segmento
11   Retornar la nueva trayectoria reducida
```

Referencias

[1] Duan, M., Fan, J., Li, M., Han, L., Huo, S.: Evaluation of Dimensionality-reduction Methods from Peptide Folding-unfolding Simulations. Journal of chemical theory and computation **9**(5), 2490–2497 (2013). doi:10.1021/ct400052y

[2] Peng, J.-h., Wang, W., Yu, Y.-q., Gu, H.-l., Huang, X.: Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. Chinese Journal of Chemical Physics **31**(4), 404–420 (2018). doi:10.1063/1674-0068/31/cjcp1806147

[3] Rajan, A., Freddolino, P.L., Schulten, K.: Going beyond clustering in MD trajectory analysis: An application to villin headpiece folding. PLoS ONE **5**(4), 9890 (2010). doi:10.1371/journal.pone.0009890

[4] Das, P., Moll, M., Stamati, H.: Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proceedings of the ... **103**(26) (2006)

[5] Kim, S.B., Dsilva, C.J., Kevrekidis, I.G., Debenedetti, P.G.: Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein. The Journal of Chemical Physics **142**(8), 85101 (2015). doi:10.1063/1.4913322

[6] Kaufman, L., Rousseeuw, P.: Finding Groups in Data. Wiley-Interscience; New York, ??? (1990)

[7] Hobohm, U., Scharf, M., Schneider, R., Sander, C.: Selection of representative protein data sets. Protein Science **1**(3), 409–417 (1992). doi:10.1002/pro.5560010313

[8] Zhang, Y., Skolnick, J.: Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics **68**(4), 1020 (2007). doi:10.1002/prot.21643