Herramienta web para post-análisis de simulaciones de dinámica molecular

Karina Giselle Borgna[1,] M. Laura Fernández[1,2,3] y Marcelo Risk[1,2,3]

[1]*Departamento de Computación, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad de Buenos Aires, Argentina*
[2]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina*
[3]*Departamento de Bioingeniería, Instituto Tecnológico de Buenos Aires, Argentina*

***Resumen*—En este trabajo se busca desarrollar una aplicación web interactiva y personalizada que permita analizar resultados generados mediante simulaciones de dinámica molecular. Con esta herramienta se pretende analizar las propiedades cada átomo o molécula de manera individual o conjunta a partir de una trayectoria, permitiendo segmentaciones tridimensionales que el usuario puede personalizar, así como análisis estadísticos no disponibles en las aplicaciones de post análisis que brindan los paquetes de dinámica molecular. Una herramienta de estas características es de gran ayuda en el estudio de los datos obtenidos permitiendo al usuario proponer el análisis o representación de los mismos que considere necesarios. Para este trabajo, como ejemplo, se tomaron datos generados por el programa Gromacs y fueron incorporados a una base de datos desarrollada en un entorno web. Se seleccionaron simulaciones de una bicapa lipídica expuesta a un campo eléctrico por el gran costo computacional que presenta de modo de generar una aplicación que pueda ser capaz de analizar datos independientemente de la demanda que éstos requieran. El sitio web se desarrolló con el framework Django, el cual permite utilizar Python, mientras que la base de datos se implementó en PostgreSQL. Se obtuvo una aplicación web que permite cargar datos de una trayectoria y generar un análisis estadístico, así como gráficos y representaciones espaciales. Se propone generar modelos 3D que permitan estimar superficie o volumen de la selección de átomos que el usuario considere necesarios.**

***Palabras clave*—simulación; aplicación web; análisis de datos; dinámica molecular; electroporación.**

***Abstract*—This work tries to develop a customized interactive web application that allows to analyze results obtained by molecular dynamics simulations. A tool of these characteristics is a great help in the analysis of the data obtained because allows customizing the analysis or representation of the data. For this study data were generated by the program Gromacs and were incorporated into a database developed in a web environment. For this work, we selected a simulation of a lipid bilayer exposed to an electric field because of the great computational cost in order to generate an application that analyze data regardless of demand that these require. The website was developed using the Django framework, which allows to use Python, while the database was implemented using PostgreSQL. We obtained a web application that allows loading data from a trajectory and generating statistical analysis, graphics and spatial representations. The strength of this application is that it allows to follow each molecule individually but generates individual results and also by groups.**

***Keywords*— Simulation; web user interface; data analysis; molecular dynamics; electroporation.**

I. Introducción

La simulación computacional es una poderosa herramienta que permite investigar aquellos fenómenos producidos durante la experimentación de laboratorio. Las simulaciones *in silico* de moléculas biológicas son un complemento para la experimentación ya sea porque permiten simular fenómenos que por su escala nanométrica son difíciles de analizar en el laboratorio o porque permiten una variedad de condiciones que por su costo no serían viables a la hora de desarrollar un protocolo experimental.

La dinámica molecular (DM) es una tecnología que al resolver de manera numérica las leyes de movimiento de Newton puede describir el movimiento y la trayectoria de átomos, moléculas y conjuntos de moléculas (por ejemplo, membranas celulares). Provee detalles finos del movimiento de partículas individuales en función del tiempo siguiendo las leyes de la física clásica, con una resolución en el tiempo de 2 fs.

Existen una variedad de programas computacionales utilizados para cálculo de mecánica molecular. En este trabajo se utilizó como interfase computacional conjunto de aplicaciones Gromacs [1]-[4]. Este programa es de uso libre y es el más utilizado para la simulación de membranas, se puede utilizar en paralelo llevando a cabo simulaciones para un número muy elevado de partículas.

Permite simular moléculas con campos de fuerza que describen todos los átomos, por ejemplo proteínas, y con campos de fuerza de átomos unidos, por ejemplo lípidos, representando cada carbono alifático y sus hidrógenos asociados como

Dirección de contacto:
Victor Toranzos, victoranzos@|gmail.com

una sola partícula, lo que reduce el costo computacional en simulaciones de elevado número de átomos como ocurre con las membranas biológicas.

El modelado de membranas mediante esta tecnología permite describir los fenómenos que ocurren a escala molecular. Como ejemplo para este trabajo investigamos los efectos de la aplicación de un campo eléctrico sobre la membrana, ya que en la mayoría de los casos resulta en la formación de poros. Este fenómeno, llamado electroporación o electropermeabilización de la membrana celular, está siendo utilizado en medicina para el tratamiento de ciertos tumores ya que permite el ingreso de drogas que normalmente serían impermeables a la membrana, sin embargo los eventos producidos a escala molecular solo se han podido dilucidar mediante dinámica molecular, demostrando la potencia de esta técnica [5]-[10].

En trabajos recientes se sugiere que el inicio del poro está liderado por las aguas que se alinean frente al campo eléctrico y logran introducirse y atravesar la porción hidrofóbica de la membrana, situación que en ausencia de campo eléctrico sería energéticamente desfavorable.

Luego la porción hidrofílica de los lípidos se reacomoda espacialmente para acompañar el movimiento de las aguas dando por resultado un poro que atraviesa de lado a lado la membrana, mientras dure el campo eléctrico [7], [11].

Dado el costo computacional que significa seguir a lo largo de una trayectoria al menos 20000 átomos durante 25 a 100 ns, hemos elegido un sistema de membrana sometida a campo eléctrico para desarrollar una aplicación web que permita realizar un post-análisis de las simulaciones realizadas mediante dinámica molecular. Se eligió un sistema biológico de alta demanda de cómputo de modo de poner a prueba la potencia de la aplicación desarrollada. En este trabajo se implementa una interfaz web que permite el seguimiento de cada átomo o conjunto de átomos a lo largo de una trayectoria. El objetivo es generar una aplicación que permita hacer un análisis estadístico que cada usuario proponga de manera personalizada de las moléculas de interés dentro de la simulación sin que el número sea una limitación. Otro objetivo es obtener gráficos y representaciones tridimensionales de los datos procesados así como modelos 3D personalizadas de los objetos contenidos en la simulación, que no puedan representarse con los programas estándar ya implementados en las aplicaciones de dinámica molecular.

II. Implementación de la Herramienta

A. Descripción de la herramienta y el modelo de datos

Ejecución de la aplicación: El sitio web se desarrolló con el framework Django, en su versión 1.4 [12], el cual utiliza Python 2.7 [13], [14]. La base de datos se implementó en PostgreSQL 8.4 [15]. Para facilitar los cambios en el modelo de datos se utilizó South [16].

La visualización de la información se programó en Python con la biblioteca matplotlib [17]. Los requerimientos del sistema incluyen un mínimo de 4 GB de memoria RAM (testeado en una máquina virtual dentro de una PC Intel i7 con 8 GB de RAM, designando a la máquina virtual 6 GB).

La Fig. 1 muestra el esquema de la herramienta web desarrollada. El primer paso es la selección de un archivo la simulación con las coordenadas y velocidades el cual va ser procesado y cargado en la base de datos. La misma figura muestra otras dos opciones: la primera es el borrado de la base de datos de cualquier archivo GRO cargado previamente y la segunda es el análisis de datos. Para esta última se puede seleccionar analizar un átomo ó un conjunto de ellos, el resultado final es un gráfico y la exportación de los datos a archivos de texto CSV. El objetivo de obtener un archivo CSV es tener un formato universal que se pueda ser leído fácilmente por otro software, por ejemplo, por el lenguaje R para un posterior procesamiento estadístico [18].

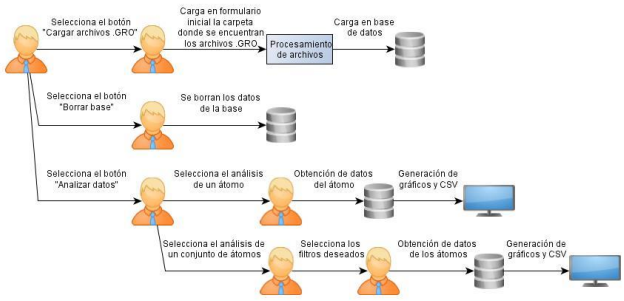

Fig. 1: Esquema de la herramienta web.

Del archivo de base GRO elegido, se selecciona: el nombre de la molécula (POPC, DOPC, CHOL, SOL, etc.[1]), el número, el tipo de átomo perteneciente a cada molécula (C, N, O, H, etc.[2]), el subtipo de átomo (1, 2, W1, etc.), el tiempo de la muestra que representa el momento en la trayectoria, las coordenadas espaciales de cada átomo descriptos como X, Y, Z, las velocidades V_X, V_Y, V_z y finalmente el tamaño de la caja de simulación para cada tiempo determinado.

[1] DOPC: Dioleoylphosphatidyl-Choline, POPC: 1-Palmitoyl-2-oleoylphosphatidylcholine. CHOL: colesterol, SOL: solvente.
[2] C: carbono, N: nitrógeno, O: oxígeno, H: hidrógeno

B. Ejecución de la aplicación

Para ejecutar la aplicación se modifica el archivo de opciones de Python *settings.py*, se configura la base de datos y la carpeta en la que se desean guardar los archivos de donde se obtendrán los datos que se incorporarán a la base. Luego se genera la base de datos de postgres creándose automáticamente las tablas en la base de datos, de acuerdo al modelo de datos existente en el archivo models.py (módulo de software en Python que define los modelos de datos y sus funciones asociada). El ingreso a la aplicación se realiza mediante una dirección IP asignada específicamente.

C. Descripción de la interfaz

En la página principal de la aplicación se encuentran tres opciones: borrar base, cargar archivos de coordenadas y velocidades y analizar datos:

- Borrar base: En esta opción se procede a borrar todos los posibles datos previamente insertados por un usuario anterior en la base para dar lugar a una nueva carga de archivos. Se muestra un mensaje de confirmación antes de borrar definitivamente los datos precargados.

- Cargar archivos GRO: Como se mencionó en la sección anterior, se deben procesar los archivos obtenidos a partir de Gromacs para luego insertar la información en la base de datos. Al llegar a esta opción se deriva al usuario a un campo de texto en el cual debe ingresar la dirección en la que se encuentran los archivos GRO que se desea analizar. La aplicación toma de la carpeta seleccionada sólo los archivos con dicha extensión, los procesa e inserta los datos en la base.

Mientras se cargan los archivos, se muestra un avance del procesamiento en la consola.

Se indica por separado el procesamiento de los archivos iniciales, donde se dividen las columnas y se analiza la información en el formato de archivo de texto. Por otro lado se muestra el avance del ingreso a la base de datos de los datos obtenidos de estos archivos.

- Analizar datos: En esta instancia se muestran dos opciones:

1-Analizar trayectoria de un átomo: Presionando este botón, se muestra un formulario en el que se debe seleccionar de una lista desplegable, el número de átomo que se desea analizar. Luego se mostrarán y se guardarán 3 gráficos correspondientes a las posiciones y 3 gráficos correspondientes a las velocidades V_X, V_Y, V_z (en caso de que el archivo sea con velocidades) respecto a toda la trayectoria del átomo analizado, indicando las variaciones que realiza el átomo en cuanto a su posición y velocidad.

2-Analizar trayectoria de varios átomos: Se muestra un formulario en el que se permite obtener la información de un grupo de átomos en particular, seleccionando los siguientes filtros, los cuales permiten seleccionar una porción del volumen total de simulación y un lapso dentro del tiempo total, como se muestra en la Fig. 2:

Tipo de átomo
Subtipo de átomo
Molécula: lista desplegable con todos los números de molécula
Tipo de molécula
Límites de posiciones en X, Y, Z
Límites de velocidades en X, Y, Z
Límites de t: Se puede analizar toda la trayectoria o solo un período de los datos que ya fueron incorporados a la base.

Luego de seleccionar todas las restricciones descriptas, es decir las coordenadas de los límites y el lapso de la caja dentro del total de la simulación, se mostrarán y se guardarán los seis gráficos correspondientes respecto al período seleccionado de la trayectoria (si no se selecciona ningún período se toma la trayectoria completa). Los gráficos constan de la siguiente información: media de las velocidades, desvío estándar y velocidad máxima y mínima.

Además de los gráficos generados, se muestra en la página un cuadro de doble entrada con la media, el desvío estándar, la máxima, la mínima y la varianza correspondientes a las velocidades de V_X, V_Y y V_z.

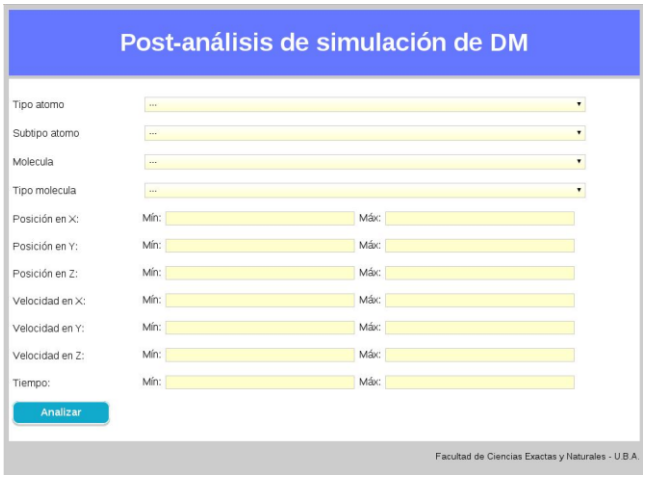

Fig 2. Ventana de opciones para el post-análisis.

Para analizar datos obtenidos de archivos seleccionados (aproximadamente 1 MB por archivo) fue necesario dividir las columnas de estos archivos para agrupar los átomos de acuerdo a varios criterios (tipo de átomo, subtipo de átomo, número de molécula, etc.).

Luego de generar este nuevo conjunto de archivos procesados, se procede a insertar la información en la base de datos. Este paso demora aproximadamente media hora cada diez archivos.

III. Resultados

Para el análisis de un único átomo, se generan los seis gráficos que muestra la Fig. 3, con la variación de la posición y de las velocidades V_X, V_Y y V_z.

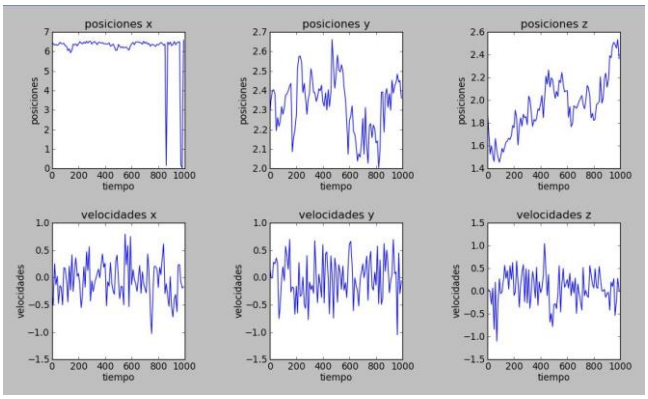

Fig. 3: Resultado del análisis de un átomo a lo largo del tiempo.

Como puede observarse en el gráfico de posiciones de X de la Fig. 3 parecería haber variaciones muy grandes en instantes de la trayectoria. Esto es un artificio de la simulación debido a las condiciones periódicas de contorno, ya que cuando una molécula llega a los límites de la caja vuelve a aparecer desde el otro extremo, situación que es perfectamente detectable al analizar los gráficos.

En el caso del análisis para un grupo de moléculas, también se generan seis gráficos: dos para X, dos para Y y dos para Z. Cada uno tiene la información del desvío estándar, la media, el máximo y el mínimo, como lo muestra las Figs. 4 y 5.

La diferencia entre los dos archivos para cada velocidad, es que uno muestra el procesamiento estadístico y los datos crudos de todas las velocidades (Fig. 4) mientras que el otro solo muestra el procesamiento estadístico (Fig. 5).

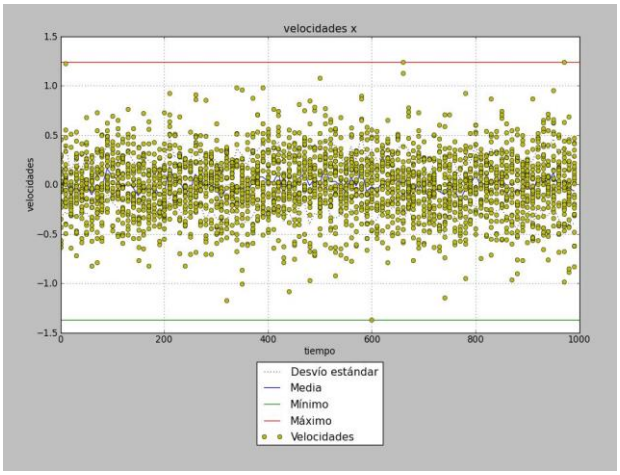

Fig. 4: Resultado del análisis estadístico para un grupo de moléculas incluyendo los datos crudos.

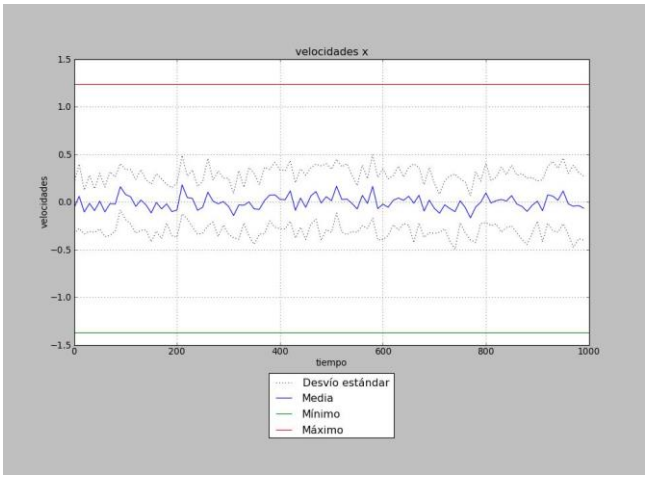

Fig. 5: Resultado del análisis estadístico para un grupo de moléculas

IV. Discusión

La herramienta web presentada en este trabajo permite el seguimiento de átomos o de grupos de átomos a lo largo de una trayectoria simulada por DM, permitiendo un exhaustivo análisis post simulación.

Con esta herramienta se puede analizar una gran cantidad de datos sin embargo el desempeño computacional aún no es el óptimo, esto se debe en parte a la gran cantidad de archivos y la extensión de los mismos que conforman una trayectoria.

Futuras mejoras a la herramienta son la optimización de la base de datos, utilizar una aplicación para Django (django-dowser y Heapy son algunas de las opciones) que ayude a controlar los problemas de memoria para minimizarlos. Otro objetivo es mejorar la interfaz y el análisis y representación de datos tales como gráficos en 3D, modelos 3D de estructuras simuladas, análisis estadísticos avanzados, barras de progreso en los procesamientos que demoran un tiempo considerable

y menús en distintos idiomas (inglés y español). Un objetivo ulterior es que la aplicación pueda leer y procesar cualquier formato de datos obtenido de una simulación independientemente la aplicación de dinámica molecular que le dio origen, universalizando la herramienta.

AGRADECIMIENTOS

Los recursos computacionales fueron provistos por el Centro de Cómputos de Alto Rendimiento (CeCAR) - Facultad de Ciencias Exactas y Naturales – UBA y por el Centro de Cómputos del Instituto Tecnológico de Buenos Aires. La investigación de MLF y MR fue financiada mediante subsidios de la Universidad de Buenos Aires (UBACyT X132/08), CONICET (PIP 112-200801-01080/09). MR agradece la contribución de IBM de Argentina. MLF y MR agradecen al Dr. G. Marshall.

REFERENCIAS

[1] H.J.C. Berendsen, D. van der Spoel y R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation", *Computer Physics Communications*, vol. 91, pp. 43-56, 1995

[2] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark y H. J. C. Berendsen, "GROMACS: Fast, flexible, and free", *J. Comput. Chem.*, vol. 26, pp. 1701-1718, 2005.

[3] B. Hess, C. Kutzner, D. van der Spoel y E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*" J. Chem. Theory Comput*. vol.4, pp.435-447, 2008.

[4] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess y E. Lindah, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit", *Bioinformatics*, vol. 29, pp. 845-854, 2013.

[5] D.P. Tieleman, "The molecular basis of electroporation" BMC Biochemistry, vol. 5, pp. 10, 2004.

[6] R.A. Böckmann, B.L. de Groot, S. Kakorin, E. Neumann y H. Grubmüller, "Kinetics, statistics, and energetics of lipid membrane electroporation studied by molecular dynamics simulations" Biophys. J., vol. 95, pp.1837-1850, 2008,

[7] M.J. Ziegler y P.T. Vernier, "Interface water dynamics and porating electric fields for phospholipid bilayers", *J. Phys. Chem. B*, vol. 112, pp. 13588-13596, 2008.

[8] P.T. Vernier, M.J. Ziegler, Y. Sun, M.A. Gundersen y D.P. Tieleman, "Nanopore-facilitated, voltage-driven phosphatidylserine translocation in lipid bilayers--in cells and in silico", *Phys. Biol,*. vol. 3, pp-233-247, 2006.

[9] M.L. Fernández, G. Marshall, F. Sagués y R. Reigada, "Structural and kinetic molecular dynamics study of electroporation in cholesterol-containing bilayers", *J Phys Chem B*. vol. 114, pp. 6855-6865, 2010.

[10] M.L. Fernández, M. Risk, R. Reigada y P.T. Vernier, "Size-controlled nanopores in lipid membranes with stabilizing electric fields", *BiochemBiophys Res Commun*, vol. 423, pp. 325-330, 2012.

[11] M. Tokman, J.H. Lee, Z.A. Levine, M.C. Ho, M.E.Colvin y P.T. Vernier, "Electric field-driven water dipoles: nanoscale architecture of electroporation", *PLoS One*, vol.8, pp. e61111, 2013.

[12] Django core team. "Django: A Web framework for the Python programming language". Django Software Foundation, Kansas, U.S.A., http://www.djangoproject.com, 2012.

[13] G. van Rossum, "Scripting the Web with Python", en Scripting Languages: Automating the Web, *World Wide Web Journal*, Vol. 2, 1997.

[14] A. Watters, G. van Rossum y J.C. Ahlstrom, "Internet Programming with Python", MIS Press/Henry Holt publishers, New York, 1996.

[15] http://www.postgresql.org/docs/8.4/static/

[16] http://south.aeracode.org/

[17] J.D. Hunter, "Matplotlib: A 2D graphics environment", *Computing In Science & Engineering*, vol. 9, pp. 90-95. 2007.

[18] R Development Core Team , "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org , 2010.