Clasificación vía aprendizaje automático de conformaciones moleculares en estructuras teloméricas

Daniela A Barragán R.[1], Robert A Cazar[1], Miguel A Mendez[2]

[1] Escuela Superior Politécnica de Chimborazo, Ecuador, Dirección de la Universidad Panamericana Sur km 1 1/2, Riobamba, Ecuador, EC060155.

[2] Universidad San Francisco de Quito, Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles sn y Vía Interoceánica, 17-1200-841, Quito, Ecuador.

Autor para correspondencia: daniela.barragan@espoch.edu.ec

Fecha de recepción: 19 de junio del 2016 - Fecha de aceptación: 24 de julio del 2016

ABSTRACT

A great number of studies have been published on the area of small molecules bound to telomeric sequences that fold on a non-cannonical DNA structure known as G-quadruplex structure. The study of these structures has been driven for its potential as targets for drug development. Furthermore, G-quadruplex structures have been identified in other sequences of importance through all the genome. Within the different studies two main models have been used: the human telomeric sequence and the oxytricha, tetrahymena telomeric sequence. Both cases are used as models for computational studies and in vitro essays. Here we study the trajectory generated from an atomistic molecular dynamic simulation to obtain the amount of flexibility and mobility of these model structures. We used data mining and classification tools (i.e. k nearest neighbor method) to identify automatically subpopulations of structures within the simulation. We found the most populated conformations and we discuss the structural findings.

Keywords: DNA, molecular dynamic, G- Quadruplex, classification, data mining, k nearest neighbor.

RESUMEN

Un gran número de estudios han sido publicados en el área de pequeñas moléculas unidas a secuencias teloméricas que se pliegan en estructura no convencionales de AD denominadas estructuras G-Cuádruple. El estudio de estas estructuras ha sido motivado por su potencial utilidad como blancos para desarrollo de fármacos. Además, las estructuras G- Cuádruples han sido identificadas en otras secuencias de importancia a lo largo de todo el genoma. Dentro de los diversos estudios se han utilizado dos modelos: la secuencia telomérica humana y las secuencia telomérica el Oxytricha, Tetrahymena. Ambos han sido empleados como modelos para estudios computacionales y ensayos in vitro. Aquí estudiamos la trayectoria generada a partir de las simulaciones de dinámica molecular atomística para obtener la cantidad de flexibilidad y movilidad de estas estructuras modelos. Utilizamos herramientas de clasificación (es decir, método del vecino cercano k) y minería de datos. Encontramos las conformaciones más pobladas y discutimos los hallazgos estructurales.

Palabras clave: ADN, simulación molecular, G cuádruple, clasificación, minería de datos, vecino cercano k.

1. INTRODUCCIÓN

Tanto en química, biofísica y ramas afines en el estudio de biomoléculas de interés biomédico las herramientas de simulación, específicamente la dinámica molecular ha contribuido a entender en la escala atómica y molecular cómo una molécula existe en distintas conformaciones. Sin embargo, esta

técnica genera gran cantidad de datos en bruto entre las que se cuentan las posiciones de los átomos en el espacio y su evolución en la variable tiempo, denominado la trayectoria de la simulación. Por el principio ergódico, entre mayor es el tiempo de la simulación es equivalente a realizar un muestreo mayor en una hipersuperficie de energía (el espacio de configuraciones del sistema) en estudio. Dado que este espacio es muy grande, la dinámica molecular se ve obligada a realizar una gran cantidad de muestreos (mayor tiempo de simulación) y como consecuencia generar una masiva cantidad de datos (en 1 ns de simulación se muestrean 500 000 configuraciones, y una típica simulación puede variar en el rango de 10 ns a 10000 ns de simulación). De estos datos de trayectoria, para el presente caso, nos interesa reconocer/diferenciar cuales son las estructuras químicas más o menos estables, de los casos cuando la molécula está cambiando de configuración. Esto requiere herramientas de clasificación/reconocimiento de patrones para reconocer la estructura y luego algoritmos automáticos de minería de datos que corran sobre toda la simulación para automáticamente encontrar y clasificar todas las estructuras muestreadas en la simulación.

El caso de estudio en el presente trabajo son secuencias teloméricas que pueden plegarse en estructuras de ADN G-cuádruple - estructuras secundarias de ADN de cuatro hebras caracterizadas por tener secuencias ricas en guanina, que pueden ser de una sola unidad o de orden superior (bi, tri o tetramolecular). Cuatro guaninas pueden adoptar una conformación cuadrada plana, cada base siendo donante y aceptor de dos puentes de hidrógeno vía Hoogsteen y cada piso siendo estabilizado por un catión monovalente (Agarwala *et al*., 2013). Estas estructuras pueden estar ubicadas en regiones teloméricas y no teloméricas (Murat & Balasubramanian, 2014) de sistemas biológicos, son termodinámicamente estables en condiciones fisiológicas (Agarwala *et al*., 2013), presumiblemente debido a que forman 8 enlaces de hidrógeno en lugar de los 2 o 3 en la doble hélice (Bryan & Baumann, 2011).

Los dos modelos a estudiar son moléculas específicamente formadas por secuencias cortas de ADN que contienen un solo bloque contiguo de guaninas (Por ejemplo, cuatro bases guanina contiguas en la secuencia TTGGGGT). Recientes estudios han asociado al G-Cuádruple con inestabilidad genómica y epigenética, cáncer y otras enfermedades fenotípicas (Wu & Brosh, 2010)

Las simulaciones computacionales pueden revelar los orígenes físicos microscópicos de los fenómenos observados experimentalmente, (Maffeo *et al*., 2014) permiten poner a prueba la comprensión de la estructura y la dinámica biomolecular, complementan los experimentos, y hacen predicciones comprobables, por lo que su uso ha aumentado drásticamente esta última década. Se puede utilizar métodos como Minería de datos, algoritmos estadísticos matemáticos, para la generación de resultados interpretables de relevancia experimental. Por ejemplo, los datos procesados pueden servir para analizar las características conformacionales como el polimorfismo y los efectos que ejerce la interacción con otras moléculas en la estructura, estabilidad y energía cinética de las moléculas en estudio. (Miranda & Bringas, 2008)

En este artículo se expone la metodología para el análisis de datos de una simulación, utilizando el método de Vecino Cercano k. En el transcurso de la simulación utilizamos como parámetros el aprendizaje basado en ejemplos los resultados de dos variables, el RMSD y el Radio de Giro de las estructuras. Estas variables fueron seleccionadas porque permiten describir y examinar la estabilidad conformacional de los residuos de guanina en el G-Cuádruple y observar los efectos dinámicos existentes en la molécula de G-Cuádruple.

2. MATERIALES Y MÉTODOS

2.1. Generación de las estructuras

Las coordenadas iniciales de la estructura de ADN G-Cuádruple fueron seleccionadas utilizando la base de datos Protein Data Bank. Todas las estructuras son tétradas de guanina tetramoleculares paralelas. Los sistemas fueron solvatados con el modelo de moléculas de agua TIP3P en cajas periódicas hasta 12 Å de soluto en todas las direcciones. Neutralizamos los sistemas mediante la adición de iones Na^+ y Cl^-.

2.2. Estructuras de ADN G-cuádruple

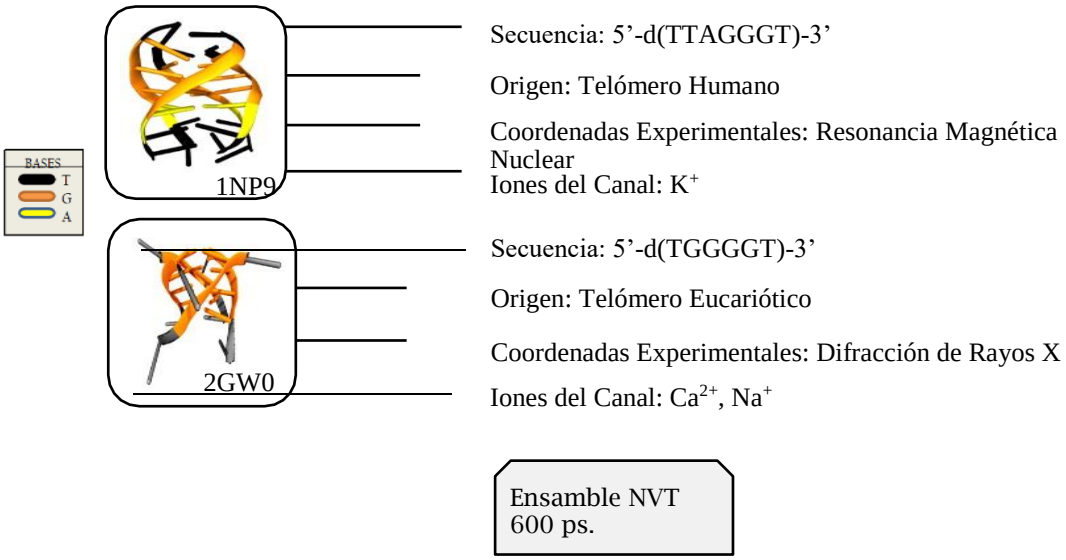

Figura 1. Estructuras estudiadas G-Cuádruples Paralelos Tetramoleculares. Parte superior: 1NP9. Parte inferior: Monómero 2GW0.

Las simulaciones de las estructuras 1NP9 y GW0 se realizaron con el campo de fuerza CHAMM27 usando el software NAMD 2.11. Las trayectorias se visualizaron con VMD 1.9.

2.3. Mecanismo molecular

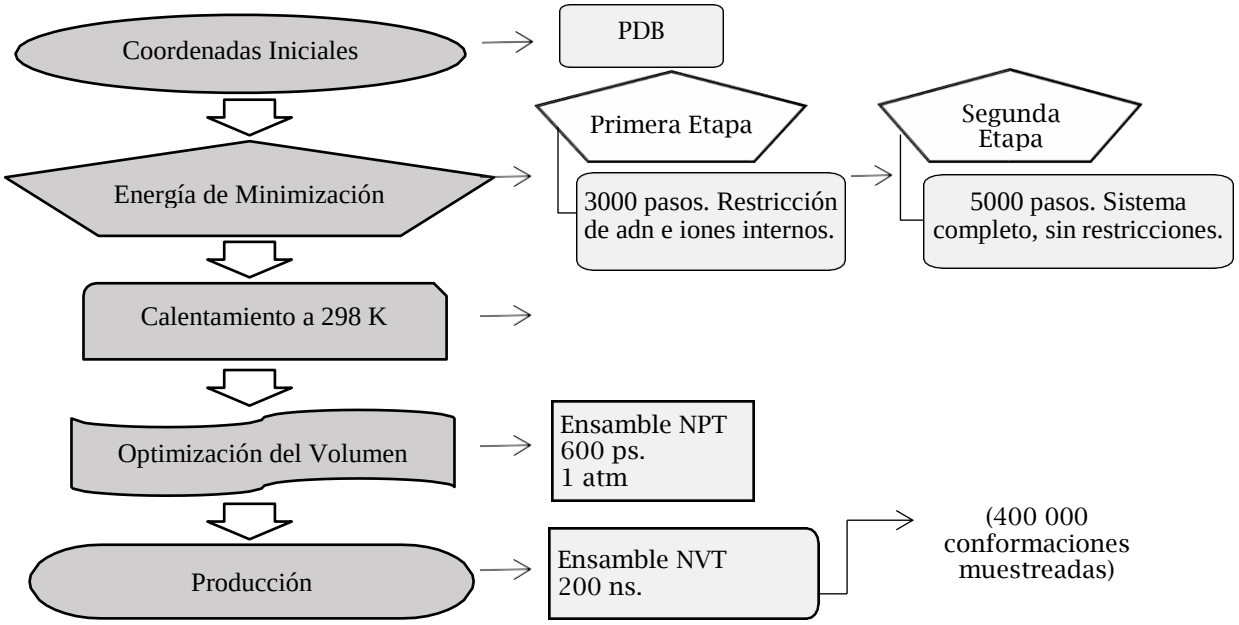

Figura 2. Esquema del Mecanismo Molecular para las estructuras G- Cuádruples.

Para la simulación se siguió un esquema usual de dinámica molecular como el mostrado en la Figura 2. La producción tuvo una trayectoria de 200 ns con el ensamble NVT (corresponde a 100 millones de pasos de simulación). Las coordenadas del sistema fueron guardadas cada 500 fs durante la simulación. El tamaño del "timestep" usado para integrar cada paso de la simulación fue de 2 fs por paso.

2.4 Parámetros estructurales

Para cada estructura, se analizó su geometría utilizando herramientas del Software Gromacs (versión) para calcular las variables Root Media Square Deviation (RMSD) de las posiciones de los átomos en referencia a una estructura inicial y el radio de giro, aproximadamente ½ de la distancia longitudinal de la molécula. Adicionalmente medimos los enlaces de hidrógeno utilizando un script desarrollado en Python.

Para el RMSD y el radio de Giro de las trayectorias analizadas se observaron dos tendencias en los datos:

1) Zonas de la estructura con valores bajos de desviación estándar que oscilaban cercanos a un cierto valor medio aproximadamente constante, a las configuraciones moleculares correspondientes a esas observaciones se les denominó conformaciones estables.
2) Zonas que tienen valores elevados de desviación estándar durante más de 50 pasos consecutivos; a las configuraciones moleculares correspondientes a esas observaciones se les denominó conformaciones de transición.

Valores altos de desviación estándar de RMSD y Radio de Giro indican que la estructura ha cambiado en el tiempo con respecto a una estructura inicial. Ambos parámetros fueron calculados en base al movimiento o cambio de solo las guaninas.

En la Figura 3 se muestra las desviaciones estándar (DS) de los valores de RMSD cada 1000 puntos para solo los residuos de guanina del G-Cuádruple. Para calcular esto se utilizó el software Wolfram Matemática con una función equivalente a calcular la desviación estándar promedio de una región de n puntos y recalcular ese valor conforme nos movemos en cada punto a lo largo de la variable tiempo. De las 40 000 conformaciones estructurales mostradas en la Figura 3 se destacan cuatro zonas que contienen 500 conformaciones cada una; tres fueron clasificadas como zonas estables y una de transición.

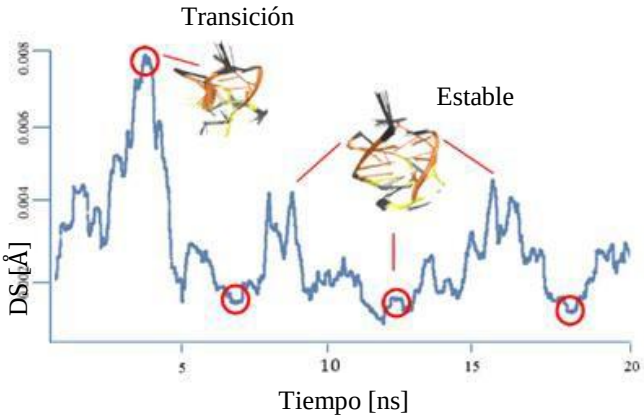

Figura 3. Desviaciones estándar (DS) de los valores de RMSD para el caso de G-Cuádruple Completo. (Realizada en Wolfram Mathematica 8.0).

2.5 Clasificación de las conformaciones estructurales

Las conformaciones estructurales seleccionadas (se utilizó solo 20 ns para encontrar los ejemplos) de los sistemas 1NP9 y del monómero 2GW0 fueron usadas como input o ejemplos para el algoritmo de clasificación de vecino cercano K (k-NN) que permitió evaluar las frecuencias de las conformaciones y determinar las conformaciones más frecuentes y las que estuvieron en menor frecuencia durante toda la simulación (200 ns).

Vecino Cercano (kNN) es un algoritmo de minería de datos que utiliza el concepto de una medición de distancia para clasificar. El valor de k fue 10 después de comprobar la precisión del algoritmo.

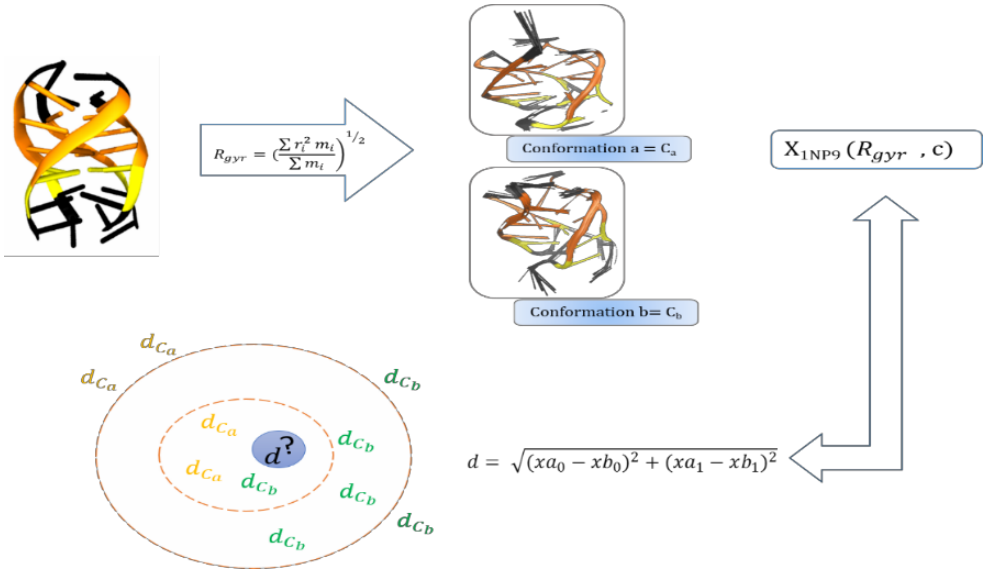

Figura 4. Esquema de Clasificación del algoritmo kNN.

En resumen, utilizamos este algoritmo para clasificar todas las conformaciones obtenidas en las trayectorias de la simulación por dinámica molecular atomística, seleccionando un conjunto existente de los ejemplos que identificados previamente (instances en un modelo de instance-based learning). Este algoritmo kNN nos permitió evaluar las frecuencias de los ejemplos durante la trayectoria.

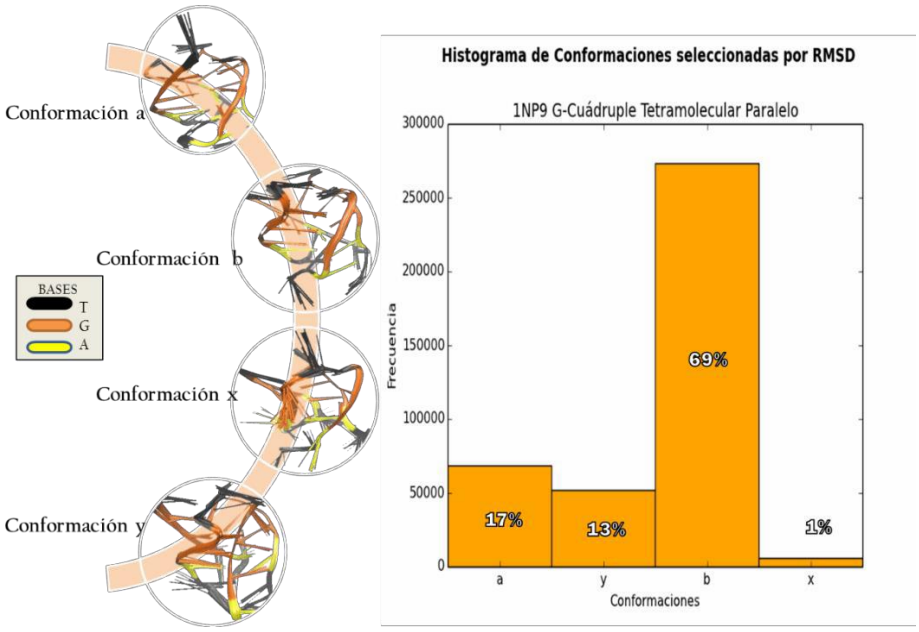

Figura 5. Histograma de las conformaciones seleccionadas por RMSD para los residuos de Guanina de la estructura 1NP9 G-Cuádruple Tetramolecular Paralelo. Conformaciones estables a,b; transición x,y.

3. RESULTADOS Y DISCUSIÓN

3.1. Secuencia Telomérica 5'-d (TTAGGGT)-3'

En la Figuras 5 y 6 se detalla las conformaciones estructurales clasificadas vía el análisis de la

desviación estándar del RMSD y Radio de Giro, respectivamente (a mayor variación en un período fijo de tiempo (time step), significa que más varía la estructura). A estas estructuras durante los periodos de alta variabilidad se les clasificó como estructuras de transición. Vía clasificación con kNN se observó que las conformaciones estables son más frecuentes que las de transición. En las estables

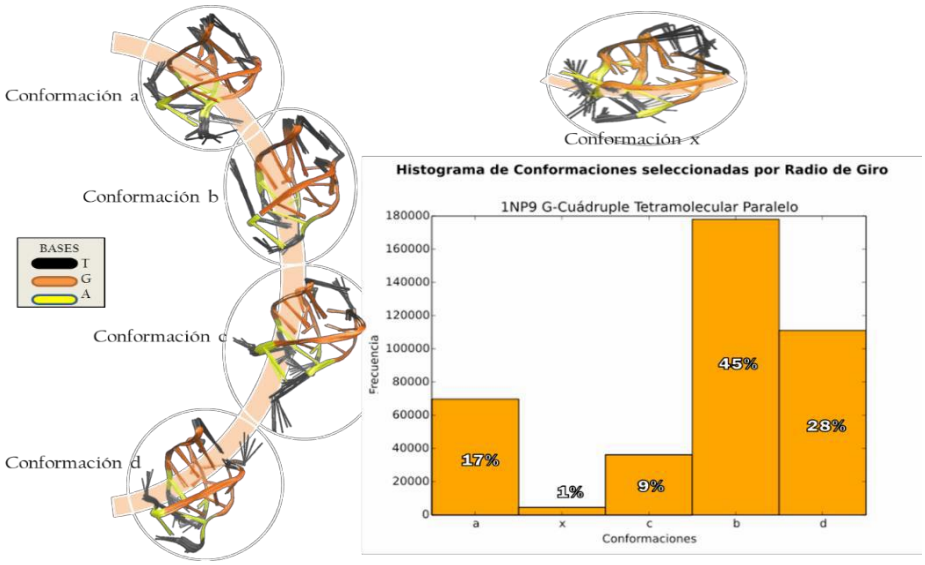

Figura 6. Histograma de las conformaciones seleccionadas por Radio de Giro para los residuos de Guanina de la estructura 1NP9 G-Cuádruple Tetramolecular Paralelo. Conformaciones estables a, b, c, d; transición x.

3.2. Secuencia Telomérica 5'- d(TGGGGT)-3'.

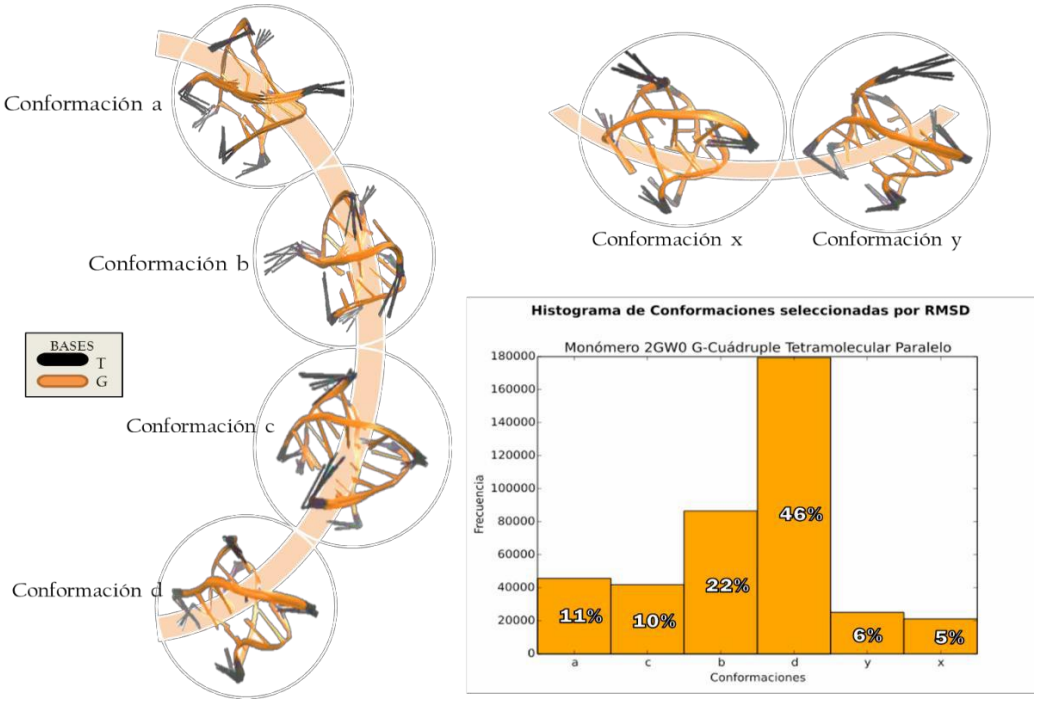

Figura 7. Histograma de conformaciones seleccionadas por RMSD para los residuos de guaninas del monómero 2GW0 G-Cuádruple Tetramolecular Paralelo. Conformaciones estables a, b, c, d. Conformaciones de transición x, y.

se observó cambios importantes en el esqueleto, sin necesariamente afectar la estabilidad de los cuartetos de residuos de G. Sin embargo, las bases nitrogenadas ubicadas en los extremos que no intervienen en la formación de los G-cuartetos presentaron una gran variabilidad conformacional. En el caso de las conformaciones de transición se observó cambios conformacionales incluso en las bases nitrogenadas de los G-cuartetos.

Las estructuras clasificadas por kNN se resumen en las Figura 7 y 8. Las conformaciones se distinguen principalmente por la posición de las bases nitrogenadas y la disposición del esqueleto. Se observó que hay mayor presencia de conformaciones estables que las de transición durante la trayectoria de simulación. Adicionalmente, las conformaciones estables presentaron G-cuartetos con menor variabilidad en comparación con el resto de su estructura. Las bases nitrogenadas ubicadas en los extremos presentaron la mayor variabilidad conformacional. En las conformaciones de transición se presentó un movimiento más libre en las bases T y también existió variabilidad conformacional en algunas bases de las G- tétradas.

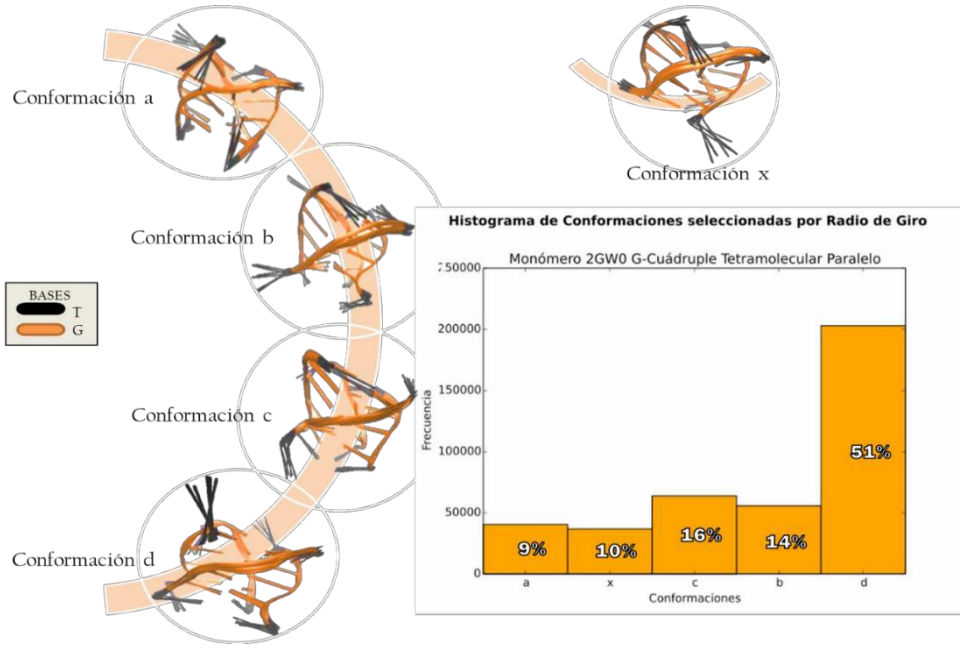

Figura 8. Histograma de conformaciones seleccionadas por Radio de Giro para los residuos de guaninas del monómero 2GW0 G-Cuádruple Tetramolecular Paralelo. Conformaciones estables a, b, c, d. Conformaciones de transición x.

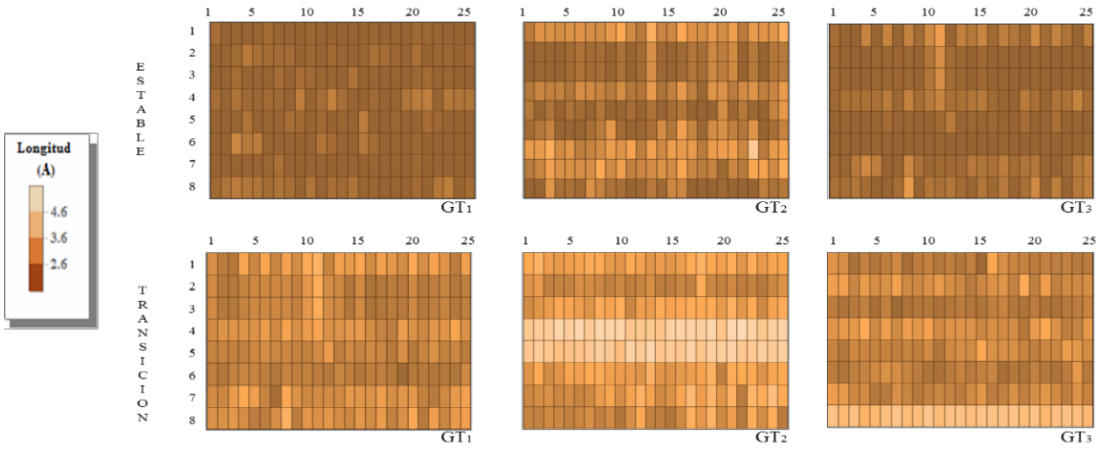

Figura 9. Mapas de Calor de la estructura 1NP9 G-Cuádruple de la longitud de los puentes de hidrógeno seleccionadas por RMSD para los residuos de Guanina. Parte superior conformaciones estables. Parte inferior conformaciones de transición. (Realizada en Wolfram Mathematica 8.0).

3.3 Caracterización estructural

Una vez clasificadas las estructuras existentes en la simulación, se puede proceder a realizar una caracterización estructural por separado de cada una de las estructuras identificadas. Como ejemplo se presenta en la Figura 9 se presenta en formato de Mapas de calor la longitud de los puentes de hidrógeno seleccionadas por RMSD para los residuos de Guanina correspondientes a la estructura 2GW0 G-Cuádruple.

4. DISCUSIÓN

Hemos realizado simulaciones de dinámica molecular en los sistemas 1NP9 y monómero 2GW0 durante una trayectoria de 200 ns para identificar estructuras conformacionales durante la simulación. El principal desafío fue la selección de criterios para identificar una estructura estable versus estructuras que estaban cambiando a otra estructura estable. Adicionalmente se necesitaba una estrategia que permitiera analizar de forma automatizada la gran cantidad de estructuras que se generan en una simulación.

Una vez establecidos los criterios en base a los valores obtenido por el análisis del índice de variación en un período de tiempo de parámetros como RMSD, y radio de giro se procedió a aplicarlos en un fragmento de todos los datos (20 a 60 ns). La diferencia entre las conformaciones estables y de transición se visualizó en el software VMD. Las conformaciones estables tenían un grado mayor de similitud entre ellas.

Esta primera fase permitió identificar los criterios numéricos (i.e. rangos) para vía la utilización del algoritmo kNN, clasificar automáticamente toda la trayectoria de simulación, y generar automáticamente los histogramas de las conformaciones. La versatilidad de Python permitió escribir una rutina para la implementación del procedimiento y que se ejecutara la rutina en un tiempo razonable. Sin embargo, para simulaciones en el rango de microsegundos kNN podría ser demasiado lento requiriéndose algún otro algoritmo o paralelizar la implementación dado que la trayectoria sigue el principio ergódico y se podría partir la trayectoria en fragmentos para su análisis en nodos de computo individuales.

Las conformaciones de transición se identificaron en las dinámicas iniciales siendo indicativo que las estructuras estables aumentan en proporción conforme el tiempo de simulación aumenta (total de la simulación 200 ns). Finalmente, los histogramas de frecuencia revelaron un mayor porcentaje de estructuras estables que de transición al tomar en cuenta toda la trayectoria.

5. CONCLUSIONES

El algoritmo k-NN clasificó un total de 400 000 conformaciones estables y de transición. Este análisis demostró que las estructuras de G-Cuádruple una vez formadas muestran variabilidad conformacional especialmente en las bases que flanquean la estructura y en los átomos del esqueleto de las bases involucradas en la región de G-Cuádruple. Además, se comprobó la estabilidad que confiere a la estructura global la presencia de los residuos de Guanina en las estructuras G-Cuádruples en concordancia con observaciones experimentales sobre este tipo de estructuras. A partir de las conformaciones clasificadas en mayor proporción se realizó un análisis de la longitud los puentes de hidrógeno de Hoogsteen, lo que permitió concluir que la distribución de longitudes de enlace obtenidas se encuentra dentro del rango reportado en la literatura.

AGRADECIMIENTOS

Agradecemos al Instituto de Simulación Computacional (ISC-USFQ), a la Universidad San Francisco de Quito por el uso de HPC - USFQ y a la Escuela Politécnica de Chimborazo por el uso del clúster del laboratorio GetNano-ESPOCH. También se agradece a CEDIA por poner a disponibilidad de estudiantes e investigadores el software Wolfram Mathematica.

REFERENCIAS

Agarwala, P., S. Pandey, K. Mapa, S. Maiti, 2013. The G-quadruplex augments translation in the 5' untranslated region of transforming growth factor β2. *Biochemistry*, 52(9), 1528-1538.

Bharti, S., J. Sommer, J. Zhou, D. Kaplan, J. Spelbrink, J.L. Mergny, J.R. Brosh, 2014. DNA sequences proximal to human mitocondrial DNA deletion breakpoints prevalente in human disease form G-Quadruplexes, a class of DNA structures inefficiently unwound by the mitocondrial Relicate Twinkle Helicase. *J. Bio. Chem.*, 289(43), 29975-29993.

Biffi, G., D. Tannahill, J. Miller, W. Howat, S. Balasubramanian, 2014. Elevated levels of G-Quadruplex Formation in Human Stomach and Liver Cancer Tissues. *Plos One, 10.1371/journal.pone.0102711*. Obtenido de http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102711

Bryan, T., P. Baumann, 2011. G-Quadruplexes: From Guanine gels to Chemotherapeutics. *Mol. Biotech.*, 49(2), 198-208.

Maffeo, C., J. Yoo, J. Comer, D. Wells, B. Luan, A. Aksimentiev, 2014. Close encounters with DNA. *J. Phys. Condens. Matter*, 26(41), 1-35.

Miranda, J., R. Bringas, 2008. Análisis de datos de microarreglos de ADN. Parte II cuantificación y anáslisis de la expresión génica. *Biotecnología Aplicada*, 25(4), 290-300.

Murat, P., S. Balasubramanian, 2014. Existence and consequences of G-quadruplex structures in DNA. *Curr Opin Genet Dev*, 25, 22-31.

Tran, P., A. Cian, J. Gros, R. Moriyama, J-L. Mergny, 2013. Tetramolecular quadruplex stability and assembly. *Top Curr Chem*, 330, 243-273.

Wu, Y., J.R. Brosh, 2010. G-Quadruplex nucleic acids and human disease. *FEBS J.*, 277(17), 3470- 3488.