UN NOVEDOSO POTENCIAL ESTADÍSTICO BASADO EN ÁREAS DE SUPERFICIE ATÓMICA SOBRELAPADAS PARA EL ANALISIS DE MÓLECULAS DE PROTEÍNA, ADN Y ARN

JUDEMIR RIBEIRO

2016-02-03

Lorem ipsum etc ...
Lorem ipsum etc ...

AGRADECIMIENTOS

test

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

Å	Angstrom (1 Å = 10^{-10} m)
AUC	*Area Under Curve*, Área Bajo la Curva. Una de las estadisticas entregadas por el analisis por curva ROC de un clasificador. Valores entre 0.5 (clasificador inútil) y 1.0 (perfecto)
C_α	Carbono alfa.
RMSD	*Root Mean Square Deviation*, Raiz de la desviación cuadrada media.
SASA	*Solvent Accessible Surface Area*, Superfície Accessible al Solvente
BSASA	*Buried Solvent Accessible Surface Area*, Superfície Accesible al Solvente Enterrada
PDB	*Protein Data Bank*, Sitio web donde son publicadadas estructuras moleculares de libre acceso. Tambien puede significar el archivo con la estructura en si.

RESUMEN

La creación y validación de campos de fuerza para el análisis del comportamiento de modelos de moléculas biológicas es una de las metas más importantes en la biofísica. Campos de fuerza basados en conocimiento, también conocidos como potenciales estadísticos o potenciales de fuerza media, utilizan datos experimentales en su construcción. En el caso de las biomoléculas estos datos vienen de estructuras tridimensionales resueltas por cristalografía de rayos X o NMR. Asumiendo que el comportamiento de una molécula o complejo molecular puede ser capturado por una función de energía, que puede ser definida por interacciones entre dos cuerpos, y que las interacciones observadas con mayor frecuencia corresponden a estados de baja energía, es posible crear una función de energía cuyos mínimos corresponden a estados nativos. Adicionalmente, se pueden crear funciones de energía que miden solamente un parámetro de cada cuerpo, como por ejemplo la cantidad de otros átomos cercanos a su alrededor.

De manera estándar estas funciones de energía usan las distancias entre los dos cuerpos como la variable independiente. En el desarrollo de esta memoria de investigación, experimentamos con la utilización del sobrelapamiento de las Superficies Atómicas Accesibles por Solvente (SASA), medido en $Å^2$, en potenciales de interacción intramolecular para proteínas, ADN y ARN. También fueron calculados potenciales de superficie usando el valor crudo de SASA para cada átomo. Nuestra nueva metodología combina estos dos tipos de potenciales para realizar las mediciones.

Para evaluar el desempeño de estos nuevos potenciales en proteína y ARN, se realizaron pruebas previamente validadas. En el caso de las proteínas, se evaluó la capacidad de los nuevos potenciales de detectar errores puntuales en dos sets de modelos, en los cuales los nuevos potenciales mejoraron la AUC de detección de 0.769 a 0.788 y de 0.677 a 0.769 respectivamente. También se evaluó la capacidad de los nuevos potenciales en separar un set de modelos nativos y no nativos, en el cual no lograron mejoras, empeorando la AUC de 0.883 a 0.773. En los potenciales para ARN se utilizaron dos pruebas, una en la cuál se evaluó la capacidad de predecir estructuras no canónicas, donde el nuevo método logró encontrar 13 de los mejores modelos contra 9 para el potencial usando distancias. La segunda prueba consistió en calcular la correlación entre valores de energía y valores de desviación estructural para 85 estructuras con 500 modelos cada una. La nueva metodología logro una correlación de 0.719, mientras que la antigua 0.79. En los potenciales para ADN, se evaluaron 20362 modelos generados a partir de 33 estructuras no redundantes y se comparó la capacidad del potencial en identificar los modelos con menor RMSD. En esta prueba los nuevos potenciales lograron clasificar las estructuras de manera equivalente al método estándar, dado que no hubo diferencias significativas en las distribuciones de RMSD encontrados.

Esta nueva metodología es robusta lo suficiente para ser utilizada en el desarrollo de un futuro potencial para la evaluación de interacciones entre proteínas y ADN/ARN, además de reemplazar el antiguo método.

ABSTRACT

First paragraph.
First paragraph.

INTRODUCCIÓN

HIPÓTESIS Y OBJETIVOS

MATERIALES

3.1 Equipos

Los equipos computacionales utilizados para esta investigación consistieron en cuatro servidores Dell R620, con 16 núcleos y 64 GB de RAM cada uno y un Apple Mac Pro con 12 núcleos y 22 GB de RAM, pertenecientes al laboratorio. Además fue utilizado un laptop personal HP 8740w con 4 núcleos y 20 GB de RAM. Se utilizó el sistema operativo CentOS 6.7 en los servidores Dell y Ubuntu 16.04 tanto en el Apple Mac Pro como en el laptop personal.

3.2 Software

El software utilizado en esta investigación consiste de programas y scripts para manipulación y cálculo de datos escritos en los lenguajes Python 3 y C++, y de programas y librerías de libre acceso para tareas de visualización de datos y generación de gráficos como Scikit (Pedregosa y col. 2012) y para visualización de estructuras 3D, como PyMOL (Schrödinger, LLC 2015).

3.3 Sets de estructuras cristalográficas

3.3.1 Sets utilizados para derivación de potenciales y experimentos en proteínas

El set de datos utilizado para la derivación de todos los potenciales para proteína fue obtenido a partir de un conjunto inicial de 518 estructuras resueltas por medio de cristalografía de rayos X, las cuales no presentaban duplicados, errores o átomos faltantes, poseian más de 100 residuos por estructura, y presentaban entre si una similitud de secuencia menor al 25 % (Ferrada y Melo 2009). Este conjunto inicial fue a su vez filtrado para remover todas las proteínas con más de una cadena, dejando 267 estructuras monoméricas, a fin de simplificar la derivación de los potenciales.

El primer benchmark utilizó el mismo conjunto de prueba utilizado en Ferrada y Melo

2007, que consiste en un set de 152 modelos y 80 estructuras nativas monoméricas. Todos los modelos tenían más de 100 aminoácidos y RMSDs menores a 3.0 Å con más de 90 % de C_α equivalentes respecto a la estructura nativa de la cual fue derivado. Este conjunto fue utilizado para observar la capacidad de los potenciales en clasificar las estructuras en modelos o nativas correctamente.

Para el segundo benchmark en proteínas, reconocimiento de errores en proteínas, se utilizó el conjunto de pruebas usado en Ferrada y Melo 2009. Este consistía de dos sets, uno de 55 modelos, y otro con 57, ambos con estructuras de más de 100 aminoácidos de largo. El primer set de 55 modelos fue nombrado "Clase A", con más de 95 % de C_α equivalentes y RMSDs menor a 1.1 Å respecto a sus estructuras nativas. En total poseía 10295 residuos con 201 de ellos considerados como erróneamente modelados. El segundo set fue identificado por "Clase B", con más de 90 % de C_α equivalentes y RMSDs menores a 1.5 Å. Este contenía un total de 10714 residuos, con 1257 de estos considerados erróneos. Para ambos sets, un residuo modelado es considerado erróneo si este posee un RMSD respecto a su estructura nativa mayor a 1.8 Å para los C_α y mayor a 3.5 Å para átomos de la cadena lateral.

3.3.2 Sets utilizados para derivación de potenciales y experimentos en ARN

Las estructuras cristalográficas utilizadas para derivación de los potenciales para ARN fueron las mismas utilizadas en Capriotti y col. 2011. Estas consisten en 85 monómeros de RNA, que fueron obtenidos al filtrar todas las estructuras de la PDB (Abril 2009) y excluir las estructuras con menos de 20 nucleótidos, resueltas a resoluciones mayores que 3.5 Å, y secuencias redundantes con una identidad mayor al 95 %.

Para el primer benchmark en ARN, correlación entre valores de energía dados por los potenciales y medidas de desviación estructural, se utilizó un set de señuelos también usado y descrito en Capriotti y col. 2011. Estos modelos fueron generados a partir de las 85 estructuras nativas del set de derivación. Para cada una de las estructuras nativas, se generaron 500 modelos, los cuáles a medida eran generados tenían sus restricciones en ángulos dihedrales y de distancia entre ciertos átomos aleatoriamente removidas, con la probabilidad de que ocurra

la remoción aumentando progresivamente, generando así modelos con una desviación respecto a la estructura nativa cada vez más alta.

El segundo benchmark utilizó el set de datos creado por Das y col. 2010. Este consiste en 407 modelos de estructuras representando 32 motivos distintos de RNA con pares de bases no canónicos, elegidos usando el campo de fuerza FARFAR (Das y col. 2010). Estos fueron utilizados para evaluar la capacidad de los potenciales de encontrar los modelos con menor RMSD respecto a su estructura nativa.

3.3.3 Sets utilizados para derivación de potenciales y experimentos en ADN

El set de estructuras cristalográficas utilizado para la derivación de los potenciales en ADN consiste de

MÉTODOS

4.1 Campos de fuerza basados en conocimiento

Los potenciales de fuerza media utilizados y derivados en este trabajo parte del supuesto de que las fuerzas encontradas en sistemas moleculares grandes son excesivamente complejas, por lo tanto la única fuente de información confiable son estructuras resueltas en su estado nativo y en equilibrio. Si la extracción de información es exitosa, el campo de fuerza será capaz de determinar correctamente si un motivo en una molécula es nativo o no. Esta es la llamada aproximación deductiva o *knowledge-based* de un potencial de fuerza media. (Sippl 1993)

Un potencial de fuerza media parte de la ley inversa de Boltzmann:

$$E_{ijkl} = -kT \log(f_{ijkl}) + kT \log Z \tag{1}$$

La función de energía E_{ijkl} es el llamado potencial de fuerza media. La variable f es la frecuencia relativa de un cierto estado al fijar las variables i, j, k, l en los sistemas observados en nuestra base de datos. Z representa la función de partición y no puede ser calculada experimentalmente, y se le da el valor de 1 (Sippl 1993). La ecuación (1) entonces toma la forma:

$$E_{ijkl} = -kT \log(f_{ijkl}) \tag{2}$$

Pero para utilizar exitosamente la ley inversa de Boltzmann es necesario también definir un sistema de referencia apropiado. Este se obtiene promediando un set elegido de variables del sistema, como por ejemplo k y l. Esto nos permite extraer una característica energética general de los sistemas, las cuáles también se definen como un potencial de energía:

$$E_{kl} = -kT \log(f_{kl}) \tag{3}$$

Con esto, ahora podemos obtener el valor neto del potencial de fuerza media:

$$\Delta E_{kl}^{ij} = E_{kl}^{ij} - E_{kl} = -kT \log \left(\frac{f_{kl}^{ij}}{f_{kl}} \right) \tag{4}$$

En el contexto de este trabajo, nuestras variables i y j indican el tipo de interacción entre dos átomos (en el caso de los potenciales SASA, solo se usa la variable i), mientras que k y l indican distancia en la secuencia de residuos y el *bin* de la variable geométrica a analizar, que puede ser la distancia, BSASA o SASA. Se aplica también un factor de corrección para números bajos de observaciones en la base de datos, sugerido en Sippl 1990. Así, cuando en función de l la ecuación final toma la forma:

$$\Delta E_{k}^{ij}(l) = RT \log \left[1 + M_{ijk}\sigma\right] - RT \log \left[1 + M_{ijk}\sigma \frac{f_{k}^{ij}(l)}{f_{k}(l)}\right] \tag{5}$$

Donde M_{ijk} corresponde al número de observaciones de interacciones del par al nivel de separación k, y σ al peso que se le da a cada observación. En este trabajo se utilizó $\sigma = 1/50$. (Melo y Feytmans 1997; Sippl 1990)

4.2 Determinación de tipos atómicos

Para los potenciales en proteínas, se utilizaron 40 tipos atómicos compartidos para los 20 aminoácidos. Esto es debido a que existen 98 tipos atómicos no equivalentes en total, lo que resultaría en una base de datos con muy pocos datos para cada par de interacciones (Melo y Feytmans 1997). Las definiciones se pueden ver en la Tabla 1.

Tipo atómico	Lista de átomos
1	C_α para todos los aminoácidos excepto Glicina
2	C_α Glicina
3	N para todos los aminoácidos excepto Prolina
4	C para todos los aminoácidos
5	O para todos los aminoácidos
6	Ala-C_β, Ile-$C_{\gamma 2}$, Ile-C_δ, Leu-$C_{\delta 1}$, Leu-$C_{\delta 2}$, Thr-C_γ, Val-$C_{\gamma 1}$, Val-$C_{\gamma 2}$
7	Ile-C_β, Leu-C_γ, Val-C_β
8	Arg-C_β, Arg-C_γ, Asn-C_β, Asp-C_β, Gln-C_β, Gln-C_γ, Glu-C_β, Glu-C_γ, His-C_β, Ile-$C_{\gamma 1}$, Leu-C_β, Lys-C_β, Lys-C_γ, Lys-C_δ, Met-C_β, Phe-C_β, Pro-C_β, Pro-C_γ, Trp-C_β, Tyr-C_β
9	Met-S_δ
10	Pro-N
11	Phe-C_γ, Trp-$C_{\delta 2}$, Tyr-C_γ
12	Phe-$C_{\delta 1}$, Phe-$C_{\delta 2}$, Phe-$C_{\varepsilon 1}$, Phe$C_{\varepsilon 2}$, Phe-C_ζ, Trp-$C_{\varepsilon 3}$, Trp-C_ζ, Trp-$C_{\zeta 3}$, Trp-$C_{\eta 2}$, Tyr-$C_{\delta 1}$, Tyr-$C_{\delta 2}$, Tyr-$C_{\varepsilon 1}$, Tyr-$C_{\varepsilon 2}$
13	Trp-C_γ
14	Trp-$C_{\varepsilon 2}$
15	Ser-C_β
16	Ser-O_γ, Thr-O_γ
17	Thr-C_β
18	Asn-$N_{\delta 2}$, Gln-$N_{\varepsilon 2}$
19	Cys-S_γ
20	Lys-N_ζ
21	Arg-C_ζ
22	Arg-$N_{\eta 1}$, Arg-$N_{\eta 2}$
23	His-C_γ
24	His-$C_{\delta 2}$, Trp-$C_{\delta 1}$
25	His-$N_{\varepsilon 2}$
26	His-$C_{\varepsilon 1}$
27	Asp-C_γ, Glu-C_δ
28	Asp-$O_{\delta 1}$, Asp-$O_{\delta 2}$, Glu-$O_{\varepsilon 1}$, Glu-$O_{\varepsilon 2}$
29	Cys-C_β, Met-C_γ
30	Met-$C_{\varepsilon 1}$
31	Tyr-C_ζ
32	Pro-C_δ
33	Asn-C_γ, Gln-C_δ
34	Asn-$O_{\delta 1}$, Gln-$O_{\varepsilon 1}$
35	Lys-$C_{\varepsilon 1}$
36	Arg-N_ε
37	Arg-C_δ
38	His-$N_{\delta 1}$
39	Trp-$N_{\varepsilon 1}$
40	Tyr-O_η

Tabla 1: Definiciones de átomos pesados utilizadas para potenciales en proteínas. Átomos del tipo 10 son convertidos al tipo 3 si son el primer residuo de una cadena proteíca.

En el caso de los potenciales para ADN y ARN, se utilizaron 23 tipos atómicos distintos descritos por Capriotti y col. 2011 para moléculas de ARN. A estos se agregaron dos tipos más, 24 y 25, correspondientes a los carbonos C5 y C7 (nombres IUPAC) del nucleótido timina. Estas definiciones están en la Tabla 2.

Tipo atómico	Lista de átomos (nombres IUPAC)
1	OP1, OP2, OP3 para todos los nucleótidos
2	P para todos los nucleótidos
3	O5' para todos los nucleótidos
4	C5' para todos los nucleótidos
5	C5', C3', C2' para todos los nucleótidos
6	O2', O3' terminales
7	C1' para todos los nucleótidos
8	O4' para todos los nucleótidos
9	N1 pirimidinas; N9 purinas
10	C8 purinas
11	N3, N7 en purinas; N1 en A; N3 en
12	C5 purinas
13	C4 purinas
14	C2 en A
15	C6 en A; C4 en C
16	N6 en A; N4 en C; N2 en G
17	C2 en G
18	C6 en G; C4 en U,T
19	O2 pirimidinas; O6 en G; O4 en U,T
20	C2 pirimidinas
21	C6 pirimidinas
22	C6 pirimidinas
23	N1 en G; N3 en U,T
24	C5 en T
25	C7 en T

Tabla 2: Definiciones de átomos pesados utilizadas para potenciales en ARN y ADN. Se consideran tanto nucleótidos como deoxinucleótidos.

4.3 Derivación de potenciales basados en distancias y conteos de átomos

4.3.1 Derivación de potenciales basados en distancias

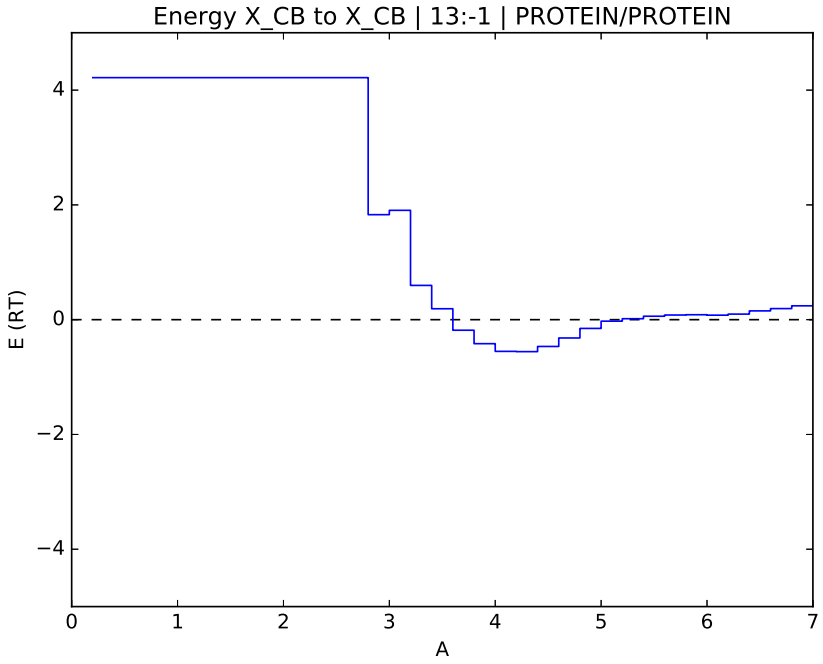
La derivación de los potenciales se hizo utilizando un programa escrito en C++, dada la gran cantidad de datos a procesar. Se utilizaron los mismos parámetros de derivación utilizados en Melo y Feytmans 1998 para los potenciales en proteínas, por lo que solo se consideran interacciones entre átomos a 7 Å de distancia y separados por un mínimo de 13 residuos si los átomos pertenecen a una misma cadena.

Para los potenciales en ARN y ADN, se utilizan los parámetros similares a los utilizados en Capriotti y col. 2011, donde se usan 6 funciones distintas. Todas estas funciones consideran interacciones a 7 Å de distancia en vez de 20 Å. Las primeras 5 funciones solo consideran como interacciones átomos que están exactamente a 1, 2, 3, 4 y 5 residuos de distancia más cualquier interacción en otra cadena. La última función solo considera interacciones a 6 o más residuos de distancia. Las distancias entre los átomos están discretizadas en 35 *bins* uniformes de 0.2 Å, paso necesario para obtener datos de frecuencia. Los pasos descritos en el algoritmo 2 son los mínimos necesarios para la generación del potencial. Las variables *Radius, Lmin, Lmax, Nbins y Sigma* corresponden respectivamente a la distancia máxima de interacción, la distancia mínima entre residuos de una misma cadena que se considera como interacción, la distancia máxima entre residuos de una misma cadena que se considera como interacción, la cantidad de *bins* en que se divide el rango de distancia, y el valor de corrección σ.

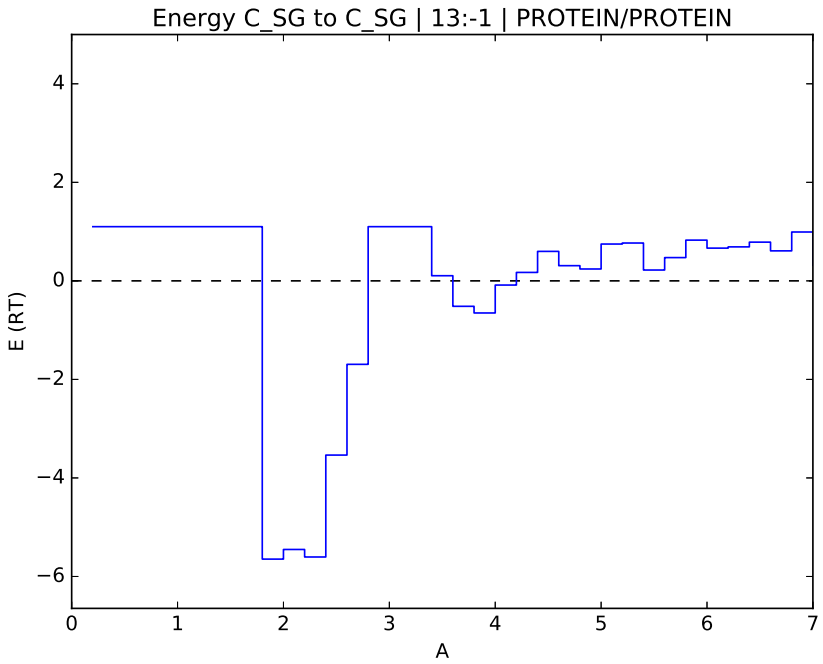
Algoritmo 1 Pasos para la derivación de un potencial a partir de una lista de archivos PDB

procedure GeneratePotential
matrix2D Mij ▷ Tabla de conteo de interacciones de tipo I con tipo J
matrix3D Fij ▷ Tablas de frecuencia de interacciones para cada intervalo de distancia
matrix1D Fxx ▷ Lista de frecuencia de interacciones en cierto intervalo de distancia
list $pdblist \leftarrow$ GetPDBs($argv1$) ▷ Carga estructuras PDB desde lista de archivos en disco
for *pdbstruct* in *PDBlist* **do**
CalculateInteractions($pdbstruct, Radius, Lmin, Lmax$) ▷ Calcula los contactos entre átomos y sus distancias
DDCalculateIntFreq($PDBlist, Fix, Fxx, Mij, Nbins$) ▷ Calcula todas las tablas necesarias para la derivación del potencial
WritePotential($Fij, Fxx, Mij, Nbins, Sigma$) ▷ Escribe el potencial creado a disco

En la figura 1 se pueden observar algunos de los potenciales generados utilizando el método descrito para moléculas de proteína. El archivo en disco contiene la información de estas funciones en un formato de texto, el que se utiliza posteriormente para la evaluación de la energia en otras estructuras. En la figura 2 se observa la matriz de conteo de interacciones (triángulo superior), cuyos datos se usan para el factor de corrección σ usado en la Ecuación 5.



(a)



(b)

Figura 1: Gráficos de las funciones de energía utilizadas en proteínas. En (a) se observa la energía (unidades RT) en función de la distancia en Å para los carbonos beta de todos los aminoácidos. En (b) se tiene la función de energia para los átomos de azufre de císteina, representando los puentes disulfuro.

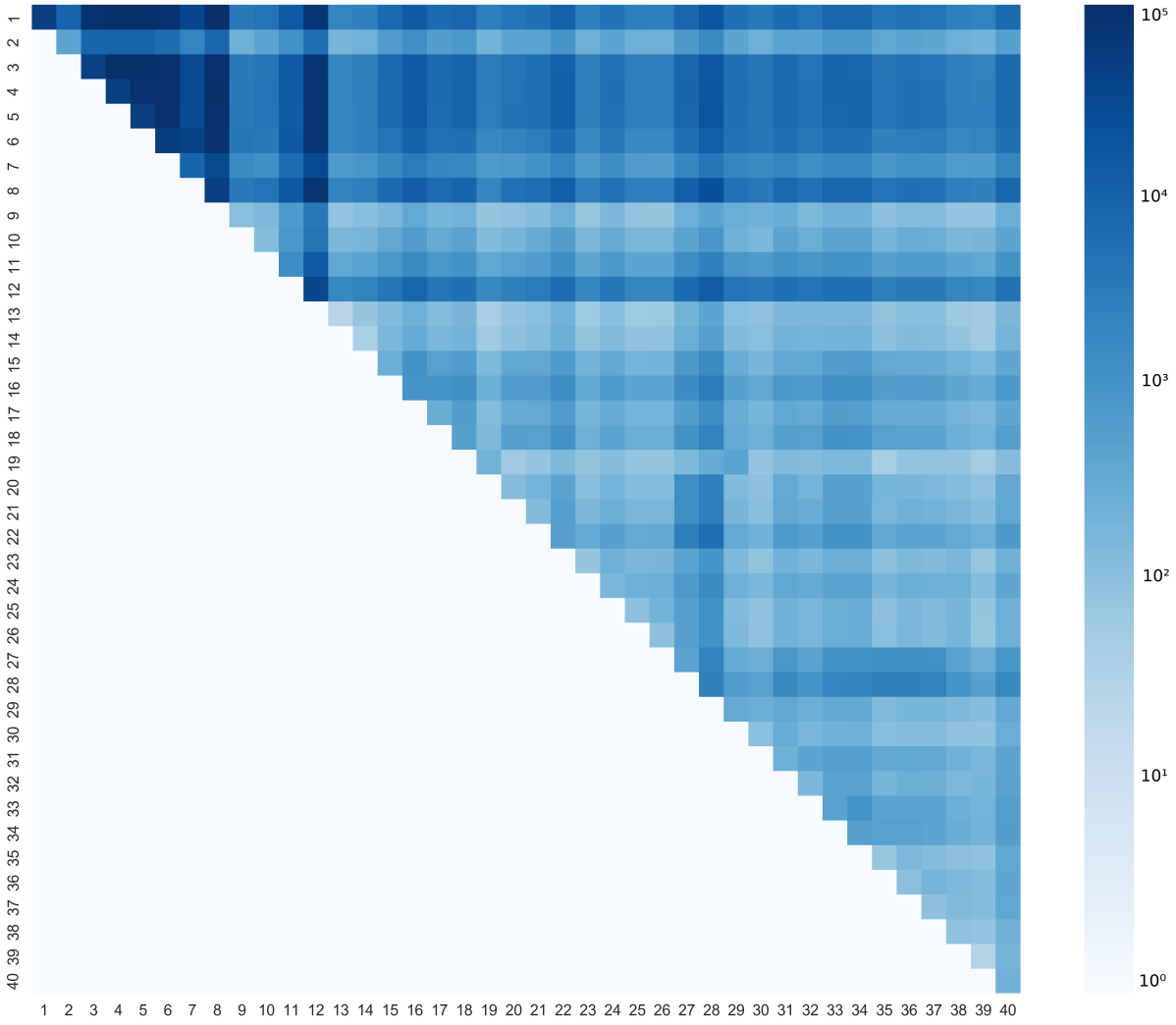


Figura 2: Matriz de conteo de interacciones para los potenciales en proteína, o *Mij*. Solo se usa el triángulo superior de esta estructura, ya que no se considera el orden de las interacciones. Conteos están en escala logaritmíca para facilitar la visualización.

4.3.2 Derivación de potenciales basados en conteos de átomos

Estos potenciales están basados en el conteo de la cantidad de centros atómicos en cierto rango de distancia. Como no dependen de interacciones entre pares de átomos, se debe modificar la Ecuación 5:

$$\Delta E_k^i(l) = RT \log\left[1 + M_{ik}\sigma\right] - RT \log\left[1 + M_{ik}\sigma \frac{f_k^i(l)}{f_{rel}}\right] \tag{6}$$

4.4 Cálculo de la superficie accesible al solvente de una molécula

Para el cálculo de la superfície accesible al solvente o SASA se utilizó el llamado algoritmo de Shrake y Rupley (Shrake y Rupley 1973), descrito en Algoritmo **??**. Este consiste generar para cada átomo de una estructura una nube de puntos con forma esférica que estan a una distancia de radio de Van der Waals más el radio de una molécula de una molécula de agua del centro del átomo. Cada punto representa un área equivalente al área de una esfera con el radio descrito anteriormente dividido por el número de puntos. Al eliminar los puntos que se encuentran en el interior del volumen de las nubes de puntos de otros átomos, es posible obtener la superficie accesible al contar los puntos restantes y multiplicarlos por el valor de superficie que representan.

La nube de puntos debe tener todos sus puntos lo más equidistantes posible en el plano esférico para que el cálculo de superficie no tenga sesgos debido a la distribución de los puntos.

Algoritmo 2 Pasos para la obtención del SASA de una estructura

procedure CALCULATESASA
pdbstruct Mij ▷ Estructura PDB
list unitsphere Fij ▷ Lista de puntos 3D equidistantes en la superficie de una esfera unitaria
matrix1D Fxx ▷ Lista de frecuencia de interacciones en cierto intervalo de distancia
list *pdblist* ← GetPDBs($argv1$) ▷ Carga estructuras PDB desde lista de archivos en disco
for *pdbstruct* in *PDBlist* **do**
CalculateInteractions($pdbstruct, Radius, Lmin, Lmax$) ▷ Calcula los contactos entre átomos y sus distancias
DDCalculateIntFreq($PDBlist, Fix, Fxx, Mij, Nbins$) ▷ Calcula todas las tablas necesarias para la derivación del potencial
WritePotential($Fij, Fxx, Mij, Nbins, Sigma$) ▷ Escribe el potencial creado a disco

4.5 Cálculo de las subsuperfícies de interacción

4.6 Derivación de potenciales basados en BSA

4.7 Derivación de potenciales basados en SASA

4.8 Cálculo del IP (*Information Product*)

RESULTADOS

DISCUSIÓN

CONCLUSIONES

REFERENCIAS

Capriotti E. y col. (2011). «All-atom knowledge-based potential for RNA structure prediction and assessment». *Bioinformatics* 27.8, págs. 1086-1093.

Das R., Karanicolas J. y Baker D. (2010). «Atomic accuracy in predicting and designing non-canonical RNA structure.» *Nature methods* 7.4, págs. 291-4.

Ferrada E. y Melo F. (2007). «Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models.» *Protein science : a publication of the Protein Society* 16.7, págs. 1410-21.

Ferrada E. y Melo F. (2009). «Effective knowledge-based potentials». *Protein Science* 18.7, págs. 1469-1485.

Melo F. y Feytmans E. (1998). «Assessing protein structures with a non-local atomic interaction energy.» *Journal of molecular biology* 277.5, págs. 1141-52.

Melo F. y Feytmans E. (1997). «Novel knowledge-based mean force potential at atomic level.» *Journal of molecular biology* 267.1, págs. 207-22.

Pedregosa F. y col. (2012). «Scikit-learn: Machine Learning in Python». *Journal of Machine Learning Research* 12, págs. 2825-2830. arXiv: `1201.0490`.

Schrödinger, LLC (2015). «The PyMOL Molecular Graphics System, Version 1.8».

Shrake A. y Rupley J. A. (1973). «Environment and exposure to solvent of protein atoms. Lysozyme and insulin». *Journal of Molecular Biology* 79.2, págs. 351-371.

Sippl M. J. (1990). «Calculation of conformational ensembles from potentials of mena force. An approach to the knowledge-based prediction of local structures in globular proteins». *Journal of Molecular Biology* 213.4, págs. 859-883.

Sippl M. J. (1993). «Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures». *Journal of Computer-Aided Molecular Design* 7.4, págs. 473-501.