

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

JUEMIR RIBEIRO

2016-02-03

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

## AGRADECIMIENTOS

test

# ÍNDICE

<b>DEDICATORIA</b>	<b>I</b>
<b>AGRADECIMIENTOS</b>	<b>II</b>
<b>ÍNDICE</b>	<b>III</b>
<b>ÍNDICE DE FIGURAS</b>	<b>IV</b>
<b>ÍNDICE DE TABLAS</b>	<b>V</b>
<b>ABREVIATURAS</b>	<b>VI</b>
<b>RESUMEN</b>	<b>1</b>
<b>ABSTRACT</b>	<b>2</b>
<b>INTRODUCCIÓN</b>	<b>3</b>
<b>HIPÓTESIS Y OBJETIVOS</b>	<b>4</b>
<b>MATERIALES</b>	<b>5</b>
3.1. Equipos . . . . .	5
3.2. Software . . . . .	5
3.3. Sets de estructuras cristalográficas . . . . .	5
3.3.1. Sets utilizados para derivación de potenciales y experimentos en proteínas	5
3.3.2. Sets utilizados para derivación de potenciales y experimentos en ARN .	6
3.3.3. Sets utilizados para derivación de potenciales y experimentos en ADN .	7
<b>MÉTODOS</b>	<b>8</b>
4.1. Campos de fuerza basados en conocimiento . . . . .	8

4.2. Determinación de tipos atómicos . . . . .	9
4.3. Derivación de potenciales basados en distancias y conteos de átomos . . . . .	13
4.3.1. Derivación de potenciales basados en distancias . . . . .	13
4.3.2. Derivación de potenciales basados en conteos de átomos . . . . .	17
4.4. Cálculo de la superficie accesible al solvente de una molécula . . . . .	20
4.5. Cálculo de las subsuperficies de interacción . . . . .	23
4.6. Derivación de potenciales basados en BSA . . . . .	23
4.7. Derivación de potenciales basados en SASA . . . . .	23
4.8. Cálculo del IP ( <i>Information Product</i> ) . . . . .	23
<b>RESULTADOS</b>	<b>24</b>
<b>DISCUSIÓN</b>	<b>25</b>
<b>CONCLUSIONES</b>	<b>26</b>
<b>REFERENCIAS</b>	<b>27</b>

## ÍNDICE DE FIGURAS

1.	Ejemplos de funciones de energía en proteínas . . . . .	15
2.	Cálculo de $\sigma$ . . . . .	16
3.	Ejemplos de funciones de energía para conteo de átomos en proteínas . . . . .	19
4.	Esfera en el software Thomson Applet . . . . .	22

## ÍNDICE DE TABLAS

1.	Definiciones de átomos para proteínas . . . . .	10
2.	Definiciones de átomos para ADN y ARN . . . . .	12

## ABREVIATURAS

Å	Angstrom ( $1 \text{ Å} = 10^{-10} \text{ m}$ )
AUC	<i>Area Under Curve</i> , Área Bajo la Curva. Una de las estadísticas entregadas por el análisis por curva ROC de un clasificador. Valores entre 0.5 (clasificador inútil) y 1.0 (perfecto)
C <sub>α</sub>	Carbono alfa.
RMSD	<i>Root Mean Square Deviation</i> , Raíz de la desviación cuadrada media.
SASA	<i>Solvent Accessible Surface Area</i> , Superficie Accessible al Solvente
BSASA	<i>Buried Solvent Accessible Surface Area</i> , Superficie Accesible al Solvente Enterrada
PDB	<i>Protein Data Bank</i> , Sitio web donde son publicadadas estructuras moleculares de libre acceso. También puede significar el archivo con la estructura en sí.



## 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 solapamiento de las Superficies Atómicas Accesibles por Solvente (SASA), medido en  $\text{\AA}^2$ , 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).

Se utilizaron también nubes de puntos en el plano esférico precalculadas para el cálculo de superficies moleculares accesibles al solvente, adquiridas utilizando el software *Thomson Applet* de la Universidad de Syracuse (Bowick y col. 2002; Saff y Kuijlaars 1997).

El software *Parallel* (Tange 2011) se utilizó en varias ocasiones para paralelizar el procesamiento de datos.

### 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, poseían 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  $C_\alpha$  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  $C_\alpha$  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  $C_\alpha$  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  $C_\alpha$  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, extraídas desde el material suplemental publicado. Estas consisten en 85 monómeros de RNA, que fueron obtenidos al filtrar todas las estructuras de la PDB (Berman, Westbrook y col. 2000) (datos de 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 generado 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 fue el mismo utilizado por Ibarra 2013. Este fue generado a partir de un set inicial de cristales obtenidos desde la PDB (Berman, Westbrook y col. 2000) y NDB (Berman, Beveridge y col. 1996), los cuáles fueron filtrados estructuralmente removiendo las estructuras que tenían otras moléculas distintas de ADN en el cristal, las que tuvieran bases no canónicas o esqueletos de ribosa-fosfato incompletos, y cualquier tipo de estructura cuaternaria que no sea la doble hélice. (triples hélices, cuádruples hélices, etc) y misma cantidad de nucleótidos en las cadenas, resultando en 86 estructuras finales. Luego estas estructuras fueron filtradas por identidad de secuencia utilizando alineamientos globales sin gaps, con un *threshold* de 99% de identidad de secuencia para considerarse idénticas. Las estructuras con menor resolución o menor factor R fueron elegidas como representantes. Con esto, se obtuvo el set final de 34 estructuras no redundantes.

## 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:

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

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:

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

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:

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



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

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

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:

$$\Delta E_k^{ij}(l) = RT \log [1 + M_{ijk}\sigma] - RT \log \left[ 1 + M_{ijk}\sigma \frac{f_k^{ij}(l)}{f_k(l)} \right] \quad (5)$$

Donde  $M_{ijk}$  corresponde al número de observaciones de interacciones del par al nivel de separación  $k$ , y  $\sigma$  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 $_{\alpha}$ para todos los aminoácidos excepto Glicina
2	C $_{\alpha}$ Glicina
3	N para todos los aminoácidos excepto Prolina
4	C para todos los aminoácidos
5	O para todos los aminoácidos
6	Ala-C $_{\beta}$ , Ile-C $_{\gamma 2}$ , Ile-C $_{\delta}$ , Leu-C $_{\delta 1}$ , Leu-C $_{\delta 2}$ , Thr-C $_{\gamma}$ , Val-C $_{\gamma 1}$ , Val-C $_{\gamma 2}$
7	Ile-C $_{\beta}$ , Leu-C $_{\gamma}$ , Val-C $_{\beta}$
8	Arg-C $_{\beta}$ , Arg-C $_{\gamma}$ , Asn-C $_{\beta}$ , Asp-C $_{\beta}$ , Gln-C $_{\beta}$ , Gln-C $_{\gamma}$ , Glu-C $_{\beta}$ , Glu-C $_{\gamma}$ , His-C $_{\beta}$ , Ile-C $_{\gamma 1}$ , Leu-C $_{\beta}$ , Lys-C $_{\beta}$ , Lys-C $_{\gamma}$ , Lys-C $_{\delta}$ , Met-C $_{\beta}$ , Phe-C $_{\beta}$ , Pro-C $_{\beta}$ , Pro-C $_{\gamma}$ , Trp-C $_{\beta}$ , Tyr-C $_{\beta}$
9	Met-S $_{\delta}$
10	Pro-N
11	Phe-C $_{\gamma}$ , Trp-C $_{\delta 2}$ , Tyr-C $_{\gamma}$
12	Phe-C $_{\delta 1}$ , Phe-C $_{\delta 2}$ , Phe-C $_{\epsilon 1}$ , Phe-C $_{\epsilon 2}$ , Phe-C $_{\zeta}$ , Trp-C $_{\epsilon 3}$ , Trp-C $_{\zeta}$ , Trp-C $_{\zeta 3}$ , Trp-C $_{\eta 2}$ , Tyr-C $_{\delta 1}$ , Tyr-C $_{\delta 2}$ , Tyr-C $_{\epsilon 1}$ , Tyr-C $_{\epsilon 2}$
13	Trp-C $_{\gamma}$
14	Trp-C $_{\epsilon 2}$
15	Ser-C $_{\beta}$
16	Ser-O $_{\gamma}$ , Thr-O $_{\gamma}$
17	Thr-C $_{\beta}$
18	Asn-N $_{\delta 2}$ , Gln-N $_{\epsilon 2}$
19	Cys-S $_{\gamma}$
20	Lys-N $_{\zeta}$
21	Arg-C $_{\zeta}$
22	Arg-N $_{\eta 1}$ , Arg-N $_{\eta 2}$
23	His-C $_{\gamma}$
24	His-C $_{\delta 2}$ , Trp-C $_{\delta 1}$
25	His-N $_{\epsilon 2}$
26	His-C $_{\epsilon 1}$
27	Asp-C $_{\gamma}$ , Glu-C $_{\delta}$
28	Asp-O $_{\delta 1}$ , Asp-O $_{\delta 2}$ , Glu-O $_{\epsilon 1}$ , Glu-O $_{\epsilon 2}$
29	Cys-C $_{\beta}$ , Met-C $_{\gamma}$
30	Met-C $_{\epsilon 1}$
31	Tyr-C $_{\zeta}$
32	Pro-C $_{\delta}$
33	Asn-C $_{\gamma}$ , Gln-C $_{\delta}$
34	Asn-O $_{\delta 1}$ , Gln-O $_{\epsilon 1}$
35	Lys-C $_{\epsilon 1}$
36	Arg-N $_{\epsilon}$
37	Arg-C $_{\delta}$
38	His-N $_{\delta 1}$
39	Trp-N $_{\epsilon 1}$
40	Tyr-O $_{\eta}$

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 proteica.

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 1 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  $\sigma$ .

---

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

---

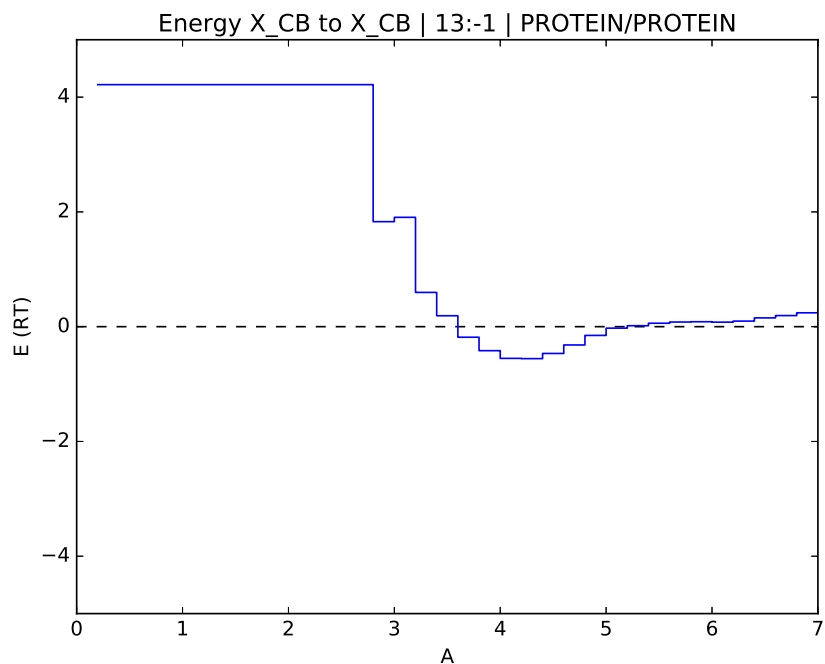
```

1: procedure GENERATEPOTENTIAL
2:   matrix2D  $M_{ij}$  #Tabla de conteo de interacciones de tipo I con tipo J
3:   matrix3D  $F_{ij}$  #Tablas de frecuencia de interacciones para cada intervalo de distancia
4:   matrix1D  $F_{xx}$  #Lista de frecuencia de interacciones en cierto intervalo de distancia
5:   list  $pdblast \leftarrow \text{GetPDBs}(argv1)$  #Carga estructuras PDB desde lista de archivos en disco
6:   for  $pdbstruct$  in  $PDBlist$  do
7:     CalculateInteractions( $pdbstruct, Radius, Lmin, Lmax$ ) #Calcula los contactos entre átomos y sus distancias
8:     DDCalculateIntFreq( $PDBlist, F_{ij}, F_{xx}, M_{ij}, Nbins$ ) #Calcula todas las tablas necesarias para la derivación del potencial
9:     WritePotential( $F_{ij}, F_{xx}, M_{ij}, Nbins, Sigma$ ) #Escribe el potencial creado a disco

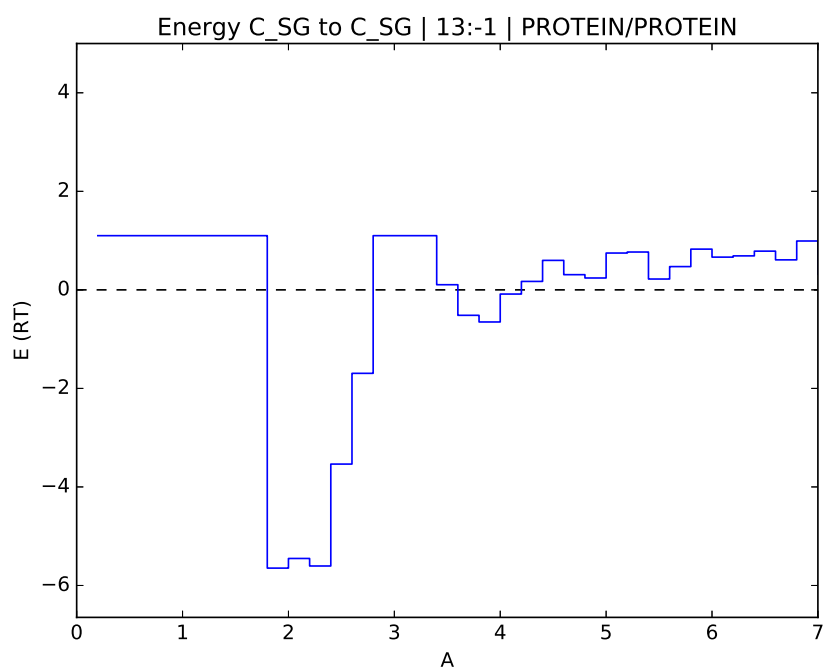
```

---

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 energía 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  $\sigma$  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 energía para los átomos de azufre de cisteína, representando los puentes disulfuro.

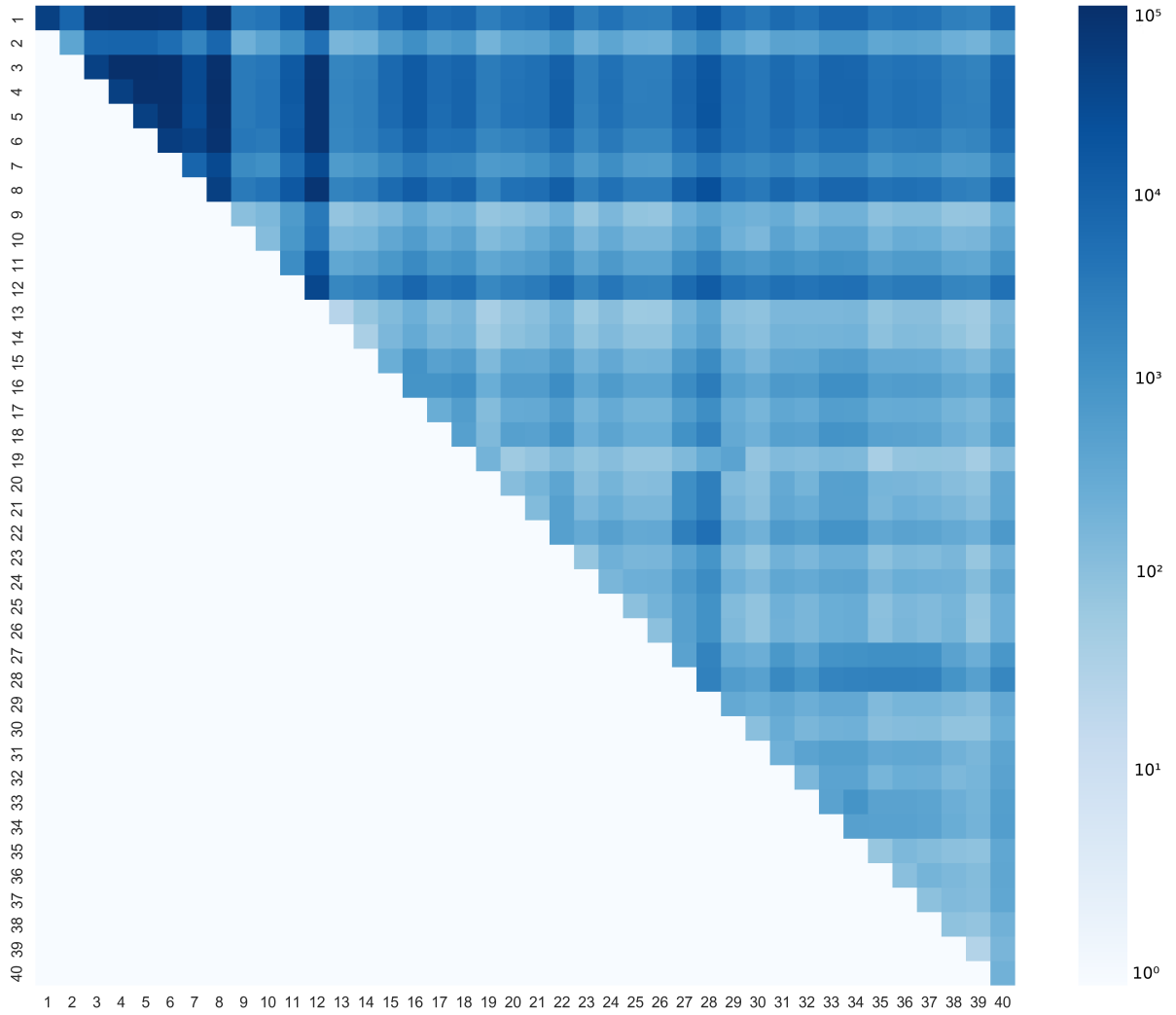


Figura 2. Matriz de conteo de interacciones para los potenciales en proteína, o  $M_{ij}$ . Solo se usa el triángulo superior de esta estructura, ya que no se considera el orden de las interacciones. Conteos están en escala logarítmica para facilitar la visualización.



#### 4.3.2 Derivación de potenciales basados en conteos de átomos

Estos potenciales están basados en el conteo de la cantidad de centros atómicos dentro de un rango de 7 Å de un átomo. Las frecuencias están basadas en *bins* de 20 unidades con un máximo de 100 para todos los tipos de potenciales utilizados (ADN, ARN, proteínas). Como no dependen de interacciones entre pares de átomos o distancias de residuos, se debe modificar la ecuación 5:

$$\Delta E^i(c) = RT \log [1 + M_i \sigma] - RT \log \left[ 1 + M_i \sigma \frac{f^i(c)}{f_{rel}} \right] \quad (6)$$

Donde  $i$  indica el tipo de átomo,  $M_i$  es la cantidad de observaciones del átomo tipo  $i$ ,  $f^i(c)$  es la frecuencia de observaciones en que un átomo  $i$  tuvo  $c$  átomos a su alrededor y  $f_{rel}$  es la frecuencia esperada, equivalente a  $1/(\text{número de bins})$ . El factor de corrección  $\sigma$  se mantiene en  $1/50$ . Un ejemplo de funciones de energía puede ser visto en la figura 3, donde se muestra la energía en función de la cantidad de átomos alrededor para un carbono aromático y para un nitrógeno ácido. La metodología de derivación tiene pequeños cambios aparte de utilizar la ecuación 6. En el algoritmo 2 se tienen los pasos para la derivación, donde podemos notar los cambios respecto al algoritmo 1.  $M_{ij}$  pasa a ser  $M_i$ , una lista con la cantidad de observaciones de cada tipo de átomo,  $F_{ij}$  pasa a ser  $F_i$ , una única tabla con las frecuencias de cada átomo  $i$  para cada *bin* de conteo, y  $F_{xx}$  pasa a ser un único valor,  $F_{rel}$ .

---

**Algoritmo 2** Pasos para la derivación de un potencial de conteo a partir de una lista de archivos PDB

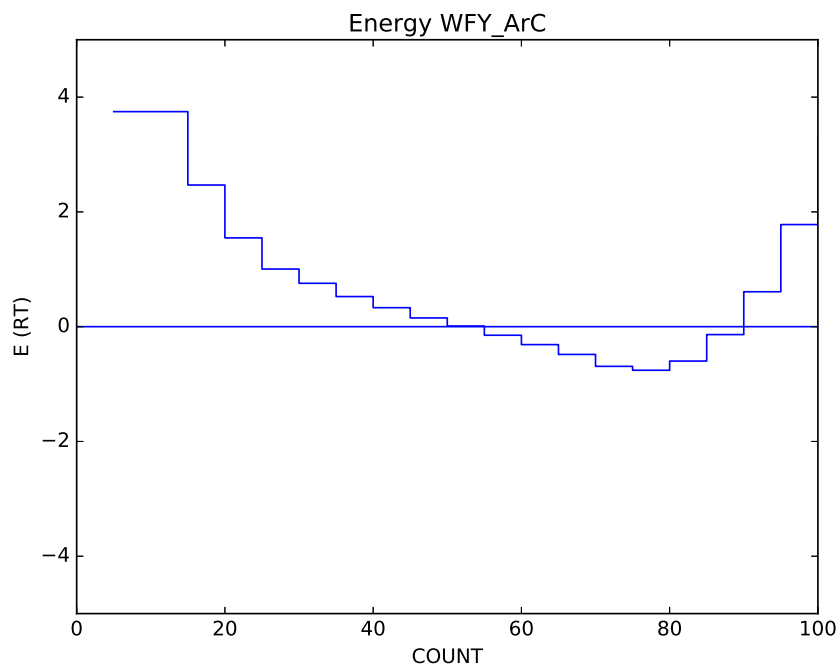
---

```

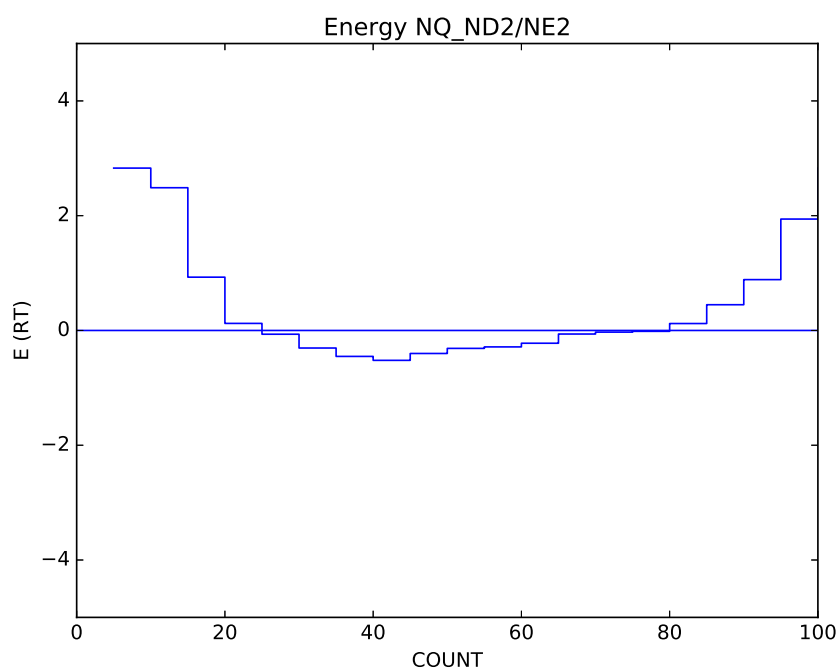
1: procedure GENERATECOUNTPOTENTIAL
2:   matrix1D  $Mi$            # Vector de conteo de átomos del tipo I en la base de datos
3:   matrix2D  $Fi$            # Tabla de frecuencias de conteo para cada tipo atómico y bin
4:   float  $Frel$              # Frecuencia esperada, 1/bins
5:   list  $pdblast \leftarrow$  GetPDBs( $argv1$ )   # Carga estructuras PDB desde lista de archivos en disco
6:   for  $pdbstruct$  in  $PDBlist$  do
7:     CountEnv( $pdbstruct$ ,  $Radius$ ) # Calcula los átomos alrededor de otro átomo dentro del rango Radius
8:     CountCalculateIntFreq( $PDBlist$ ,  $Fi$ ,  $Frel$ ,  $Mi$ ,  $Nbins$ ) # Calcula todas las tablas necesarias para la derivación del potencial
9:     WritePotentialCount( $Fi$ ,  $Frel$ ,  $Mi$ ,  $Nbins$ ,  $Sigma$ ) # Escribe el potencial creado a disco

```

---



(a)



(b)

Figura 3. En (a) se puede observar la energía en función del conteo de átomos en un rango de  $7 \text{ \AA}$  para carbonos aromáticos en proteínas. En (b) se puede observar otra función, pero para nitrógenos ácidos de los aminoácidos asparagina y glutamina.

#### 4.4 Cálculo de la superficie accesible al solvente de una molécula

Para el cálculo de la superficie accesible al solvente o SASA se utilizó el llamado algoritmo de Shrake y Rupley (Shrake y Rupley 1973), descrito en el algoritmo 3. Este consiste en generar para cada átomo de una estructura una nube de puntos con forma esférica que están a una distancia de radio de Van der Waals más el radio 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. En la figura 4 se pueden observar las divisiones del área de una esfera que estos puntos 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 desbalanceada de los puntos. Esto se logró utilizando nubes de puntos precalculadas utilizando el software Thomson Applet (Bowick y col. 2002; Saff y Kuijlaars 1997) de 15092 puntos.

---

**Algoritmo 3** Pasos para la obtención del SASA de una estructura

```

1: procedure CALCULATESASA
2:   pdbstruct pdb #Estructura PDB
3:   matrix2D unitsphere #Matriz de 3xN con las coordenadas 3D de la nube de puntos en una esfera unitaria
4:   for atomI in pdbstruct do
5:     matrix2D points  $\leftarrow$  unitsphere * (atom.vdw + 1,4) + atom.coords #Escala y mueve una copia de la nube de N puntos a las coordenadas del átomo actual
6:     int surfacepoints  $\leftarrow$  N #Contador de puntos expuestos al solvente
7:     float totalSASA  $\leftarrow$  0 #Superficie expuesta final
8:     for atomJ in atomI.interactions do #Itera sobre los átomos J que interactúan con I
9:       for point in points do
10:        if IsPointInVolume(atomJ,point) and point.enabled then #Revisa si el punto está adentro del volumen del átomo J y si no fue contado antes
11:          surfacepoints  $\leftarrow$  surfacepoints - 1 #Decrementa el contador de puntos
12:          point.enabled = False #Desabilita el punto, ya que no es necesario volver a revisarlo
13:          atomI.SASA  $\leftarrow$  surfacepoints * atomI.tAREA / N #Calcula área del átomo
14:          totalSASA  $\leftarrow$  atomI.SASA + totalSASA #Suma el área al área total
15:   return totalSASA #Retorna el SASA final

```

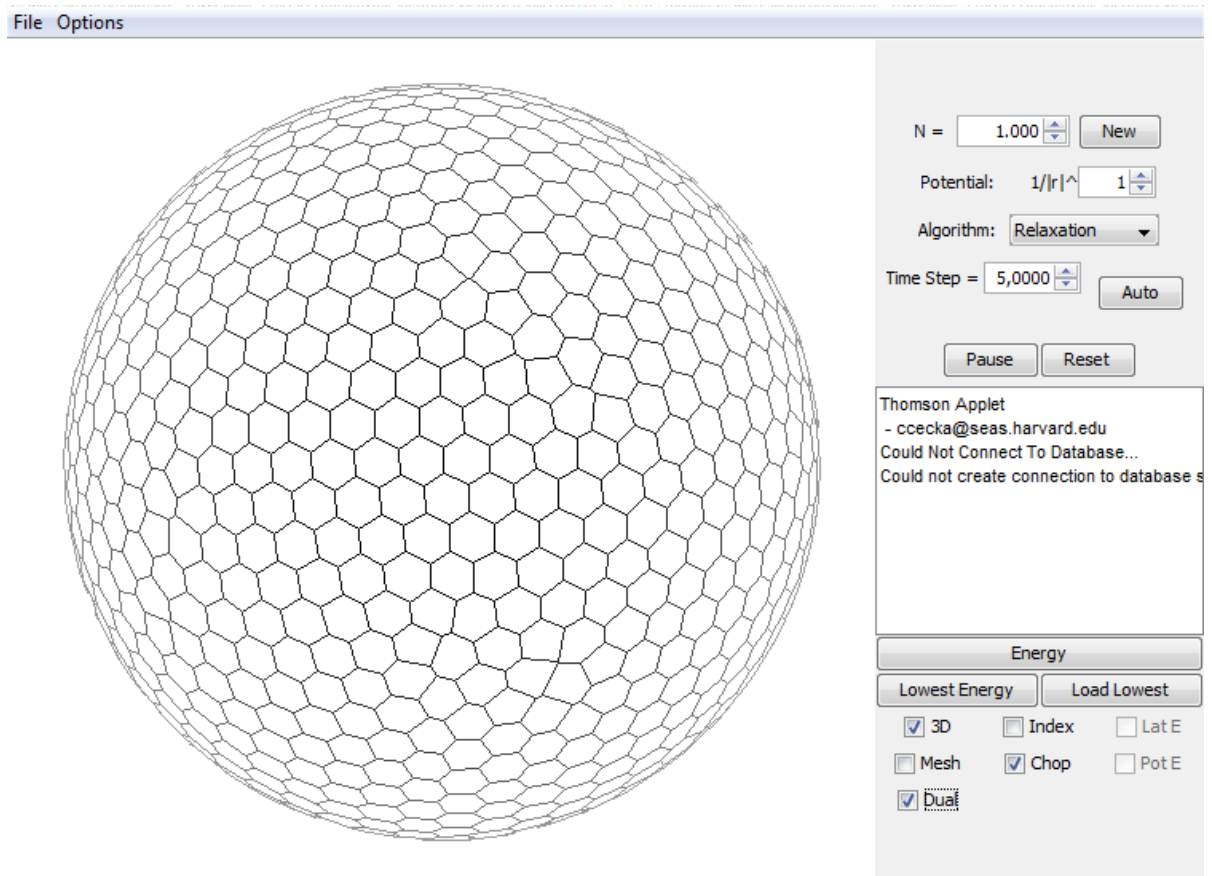


Figura 4. En este ejemplo podemos observar una esfera subdividida en 1000 áreas equivalentes. Inicialmente las subáreas son inicializadas con valores aleatorios, los cuáles son optimizados utilizando un algoritmo de relajación considerando cada punto como una carga eléctrica forzada a moverse sobre una esfera.

#### 4.5 Cálculo de las subsuperficies de interacción

Para calcular las subsuperficies de interacción entre dos átomos, se debe extender el algoritmo 3. Esto se hace re

#### 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

- Berman H. M., Westbrook J. y col. (2000). «The protein data bank.» *Nucleic acids research* 28.1, págs. 235-242.
- Berman H. M., Beveridge D. L. y col. (1996). «The Nucleic Acid Database Project». *Biological Structure and Dynamics* 2.September, págs. 1-13.
- Bowick M. y col. (2002). «Thomson Applet @ S.U.»
- 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.
- Ibarra I. (2013). «Modelamiento Comparativo tridimensional de ADN en doble hebra vía satisfacción de restricciones geométricas espaciales.» *Tesis entregada a la Facultad de Ciencias Biológicas de la Pontificia Universidad Católica de Chile en cumplimiento parcial de los requisitos para optar al Título de Bioquímico.*
- 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.
- Saff E. B. y Kuijlaars a. B. J. (1997). «Distributing many points on a sphere». *The Mathematical Intelligencer* 19.1, págs. 5-11.
- 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 mean 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.
- Tange O. (2011). «GNU Parallel - The Command-Line Power Tool». *login: The USENIX Magazine* 36.1, págs. 42-47.