

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

JUDEMIR RIBEIRO

2016-02-03



## 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>V</b>
<b>ÍNDICE DE TABLAS</b>	<b>VI</b>
<b>ABREVIATURAS</b>	<b>VII</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
4.4. Cálculo de la superficie accesible al solvente de una molécula . . . . .	13
4.5. Cálculo de las subsuperficies de interacción . . . . .	13
4.6. Derivación de potenciales basados en BSA . . . . .	13
4.7. Derivación de potenciales basados en SASA . . . . .	13
4.8. Cálculo del IP ( <i>Information Product</i> ) . . . . .	13
<b>RESULTADOS</b>	<b>14</b>
<b>DISCUSIÓN</b>	<b>15</b>
<b>CONCLUSIONES</b>	<b>16</b>
<b>REFERENCIAS</b>	<b>17</b>

## ÍNDICE DE FIGURAS

## ÍNDICE DE TABLAS

1.	Definiciones de átomos pesados utilizadas para potenciales en proteínas. . . . .	10
2.	Definiciones de átomos pesados utilizadas para potenciales en ARN y ADN. Se consideran tanto nucleótidos como deoxinucleótidos. . . . .	12

## ABREVIATURAS

Å	Angstrom ( $1 \text{ Å} = 10^{-10} \text{ m}$ )
AUC	<i>Area Under Curve</i> , Área Bajo la Curva. Valores entre 0.5 (peor) 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



## 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).

### 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 sí 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. 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:

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

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 por 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.

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
- 4.4 Cálculo de la superficie accesible al solvente de una molécula
- 4.5 Cálculo de las subsuperficies 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. (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».
- 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.