

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

DEDICATORIA	I
AGRADECIMIENTOS	II
ÍNDICE	III
ÍNDICE DE FIGURAS	IV
ÍNDICE DE TABLAS	V
ABREVIATURAS	VI
RESUMEN	1
ABSTRACT	2
INTRODUCCIÓN	3
HIPÓTESIS Y OBJETIVOS	4
MATERIALES	5
3.1. Equipos	5
3.2. Software	5
3.3. Conjuntos de estructuras cristalográficas	5
3.3.1. Conjuntos utilizados para derivación de potenciales y experimentos en proteínas	5
3.3.2. Conjuntos utilizados para derivación de potenciales y experimentos en ARN	6
3.3.3. Conjuntos utilizados para derivación de potenciales y experimentos en ADN	7

MÉTODOS	9
4.1. Campos de fuerza basados en conocimiento	9
4.2. Determinación de tipos atómicos	10
4.3. Derivación de potenciales basados en distancias y conteos de átomos	14
4.3.1. Derivación de potenciales basados en distancias	14
4.3.2. Derivación de potenciales basados en conteos de átomos	18
4.4. Cálculo de la superficie accesible al solvente de una molécula	21
4.5. Cálculo de las subsuperficies de interacción	24
4.6. Derivación de potenciales basados en BSA	27
4.7. Derivación de potenciales basados en SASA	27
4.8. Combinación de los potenciales de distancia y conteo	27
4.9. Combinación de los potenciales BSASA y SASA	29
4.10. Cálculo del IP (<i>Information Product</i>)	29
RESULTADOS	30
5.1. Pruebas en potenciales de proteínas	30
5.1.1. Clasificación de estructuras en nativas o no nativas	30
5.1.2. Detección de resíduos erroneamente modelados	30
5.2. Pruebas en potenciales para ARN	34
5.2.1. Correlación entre energía y desviación estructural	34
5.2.2. Desempeño en encontrar el mejor modelo no canónico	34
5.3. Pruebas en potenciales para ADN	39
5.3.1. Desempeño en la detección del modelo de menor RMSD	39
DISCUSIÓN	42
6.1. Pruebas en proteínas	42
6.2. Pruebas en ARN	44
6.3. Pruebas en ADN	46

CONCLUSIONES **50**

REFERENCIAS **51**

ÍNDICE DE FIGURAS

1.	Ejemplos de funciones de energía en proteínas	16
2.	Matriz de conteo para el cálculo del factor σ	17
3.	Ejemplos de funciones de energía para conteo de átomos en proteínas	20
4.	Esfera en el software Thomson Applet	23
5.	Cálculo de subsuperficies de interacción	26
6.	Ejemplos de funciones de energía SASA en proteínas	28
7.	Curvas ROC para la clasificación de estructuras en nativas y no nativas	31
8.	Curvas ROC para la detección de resíduos mal modelados en proteínas	33
9.	Boxplots de las distribuciones de correlación de desviación estructural en RNA con un potencial.	35
10.	Comparación de las diferencias de RMSD entre el modelo de menor RMSD y el modelo de menor energía encontrado por el potencial.	37
11.	Comparación de las distribuciones de dRMSD entre los modelos con menor RMSD y los modelos con menor energía calculado por cada potencial	40
12.	Boxplots de las distribuciones de la cantidad de contactos encontrados en el conjunto de estructuras usado en la derivación de potenciales para proteínas. .	43
13.	Boxplots de las distribuciones de la cantidad de contactos encontrados en el conjunto de estructuras usado en la derivación de potenciales para ARN.	45
14.	Superposiciones con el cristal original de algunos de los mejores modelos de ADN encontrados según los potenciales, para el conjunto generado con restricciones.	47
15.	Superposiciones con el cristal original de algunos de los mejores modelos de ADN encontrados según los potenciales, para el conjunto generado sin restricciones .	48

ÍNDICE DE TABLAS

1.	Definiciones de átomos para proteínas	11
2.	Definiciones de átomos para ADN y ARN	13
3.	Comparación entre curvas ROC para detección de estructuras nativas y no nativas	30
4.	Comparación entre curvas ROC de los potenciales para proteínas en detección de residuos mal modelados clase A y B	32
5.	Comparación entre distribuciones de coeficientes de Pearson para correlaciones entre energía y desviación estructural en ARN.	34
6.	Tabla con los mejores RMSD encontrados para los 32 casos de estructuras no canónicas	38
7.	Tabla de comparaciones entre distribuciones de dRMSD entre pares de poten- ciales para los conjuntos de modelos de ADN con y sin restricciones.	41

ABREVIATURAS

\AA	Angstrom ($1 \text{ \AA} = 10^{-10} \text{ m}$)
AUC	<i>Area Under Curve</i> , 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	<i>Root Mean Square Deviation</i> , Raiz de la desviación cuadrada media.
GDT	<i>Global Distance Test</i> Medida de la similitud entre dos estructuras con estructura secundaria identica pero con distinta estructura terciaria. Se calcula contando la cantidad de átomos a cierto corte de distancia de la estructura original.
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 publicadas 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 cristalográfica 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 \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 conjuntos 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 conjunto 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 cual 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. En esta prueba no fue observada una mejora significativa del nuevo método en general. Al analizar los componentes de los potenciales por separado observamos que el nuevo potencial de superficie obtiene mejores resultados que el potencial de conteo de vecinos.

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 menor RMSD clasificadas.

Esta nueva metodología es lo suficientemente robusta para ser utilizada en el desarrollo de un futuro potencial para la evaluación de interacciones intermoleculares entre proteínas y ADN o ARN.

ABSTRACT

The creation and validation of force fields for the analysis of the behavior of biological molecules is one of the most important goals in biophysics. Knowledge based force fields, also known as mean force potentials or statistical potentials, use experimental data in their derivation. In the case of biomolecules this data comes tridimensional structures solved by X-ray crystallography or NMR. Assuming that the behavior of a molecule or molecular complex can be captured by an energy function, can be defined by interactions between two bodies, and that the interactions observed with the most frequency correspond to low energy states, it's possible to create energy functions whose minimum match native states. Additionally, energy functions can be created that measure only a single parameter per body, for example the count of atoms inside a volume.

The standard for mean force potentials is using the distances between two bodies as the independent variable. In the development of this research, we experimented with the use of the overlaps of the Solvent Accessible Surface Areas (SASA), measured in Å², in intramolecular interaction potentials for proteins, DNA and RNA. Also surface potentials were generated using the raw SASA values for each atom. Our new method combines both potentials for measurement.

To evaluate the performance of the new potentials in protein and RNA, previously validated tests were used. In the protein's case, the new potentials ability to detect errors in two sets of models, in which the new potentials increased the AUC of detection from 0.769 to 0.788 and from 0.677 to 0.769 respectively. The ability of the new potentials to identify native and non-native models, where there was no improvement, worsening the AUC from 0.883 to 0.773. For the RNA potentials two tests were done, the first evaluated the ability to predict non-canonical structures, were the new method found 13 of the best models versus 9 for the potentials using distances. The second test consisted in calculating the correlation between energy values and structural deviation values for 85 structures, with 500 models each. There was no significant improvement over the standard method in this test. When analyzing the components of the new potentials by themselves we observed that the surface potential has better results than the neighbor count potential.

For the DNA potentials we evaluated 20362 models generated from 33 non-redundant structures and the potential's ability to identify the models with the lowest RMSD was evaluated. In this test the new potential's performance was equivalent to the standard potentials, due to no significant differences between the RMSD distributions found by the potentials.

This new method is sufficiently robust to be used in the development of a future potential for the evaluation of the intermolecular interactions between proteins and DNA or RNA.

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 (Perego 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 Conjuntos de estructuras cristalográficas

3.3.1 Conjuntos utilizados para derivación de potenciales y experimentos en proteínas

El conjunto 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 conjunto 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 conjuntos, uno de 55 modelos, y otro con 57, ambos con estructuras de más de 100 aminoácidos de largo. El primer conjunto 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 conjunto 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 conjuntos, 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 Conjuntos 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 conjunto de señuelos también usado y descrito en Capriotti y col. 2011. Estos modelos fueron generados a partir de las 85 estructuras nativas del conjunto de derivación. Para cada una de las estructuras nativas, se generaron 500 modelos, los cuales 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 conjunto 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 Conjuntos utilizados para derivación de potenciales y experimentos en ADN

El conjunto 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 conjunto inicial de cristales obtenidos desde la PDB (Berman, Westbrook y col. 2000) y NDB (Berman, Beveridge y col. 1996), los cuales 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. La estructuras con menor resolución o menor factor R fueron elegidas como representantes. Con esto, se obtuvo el conjunto final de 33 estructuras no redundantes.

De este conjunto de 33 estructuras no redundantes, también fueron creados los dos

conjuntos de 20362 modelos en el que los potenciales fueron probados. El primer conjunto de modelos fue creado utilizando restricciones geométricas utilizando datos del conjunto no redundante, mientras que el segundo utilizó las restricciones estándar del programa MODELLER. (Ibarra 2013)

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:

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

Para utilizar exitosamente la ley inversa de Boltzmann es necesario también definir un sistema de referencia apropiado. Este se obtiene promediando un conjunto elegido de variables del sistema, como por ejemplo k y l . Esto nos permite extraer una característica energética general de los sistemas, las cuales 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 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 _{γ2} , Ile-C _δ , Leu-C _{δ1} , Leu-C _{δ2} , Thr-C _γ , Val-C _{γ1} , Val-C _{γ2}
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 _{γ1} , 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 _{δ2} , Tyr-C _γ
12	Phe-C _{δ1} , Phe-C _{δ2} , Phe-C _{ε1} , Phe-C _{ε2} , Phe-C _ζ , Trp-C _{ε3} , Trp-C _ζ , Trp-C _{ζ3} , Trp-C _{η2} , Tyr-C _{δ1} , Tyr-C _{δ2} , Tyr-C _{ε1} , Tyr-C _{ε2}
13	Trp-C _γ
14	Trp-C _{ε2}
15	Ser-C _β
16	Ser-O _γ , Thr-O _γ
17	Thr-C _β
18	Asn-N _{δ2} , Gln-N _{ε2}
19	Cys-S _γ
20	Lys-N _ζ
21	Arg-C _ζ
22	Arg-N _{η1} , Arg-N _{η2}
23	His-C _γ
24	His-C _{δ2} , Trp-C _{δ1}
25	His-N _{ε2}
26	His-C _{ε1}
27	Asp-C _γ , Glu-C _δ
28	Asp-O _{δ1} , Asp-O _{δ2} , Glu-O _{ε1} , Glu-O _{ε2}
29	Cys-C _β , Met-C _γ
30	Met-C _{ε1}
31	Tyr-C _ζ
32	Pro-C _δ
33	Asn-C _γ , Gln-C _δ
34	Asn-O _{δ1} , Gln-O _{ε1}
35	Lys-C _{ε1}
36	Arg-N _ε
37	Arg-C _δ
38	His-N _{δ1}
39	Trp-N _{ε1}
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. Las funciones para ADN consideran interacciones a 7 Å de distancia, mientras que las para ARN consideran interacciones hasta 20 Å. Esto último se debe a que el potencial de distancia RASP se usa como base para facilitar las comparaciones. 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 clases 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*, *Nclases* 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

```

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  $pdblist \leftarrow$  GetPDBs( $argv_1$ ) #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( $PDBlist, F_{ij}, F_{xx}, M_{ij}, N_{clases}$ ) #Calcula todas las tablas necesarias para la derivación del potencial
9:     WritePotential( $F_{ij}, F_{xx}, M_{ij}, N_{clases}, Sigma$ ) #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 σ usado en la ecuación 5.

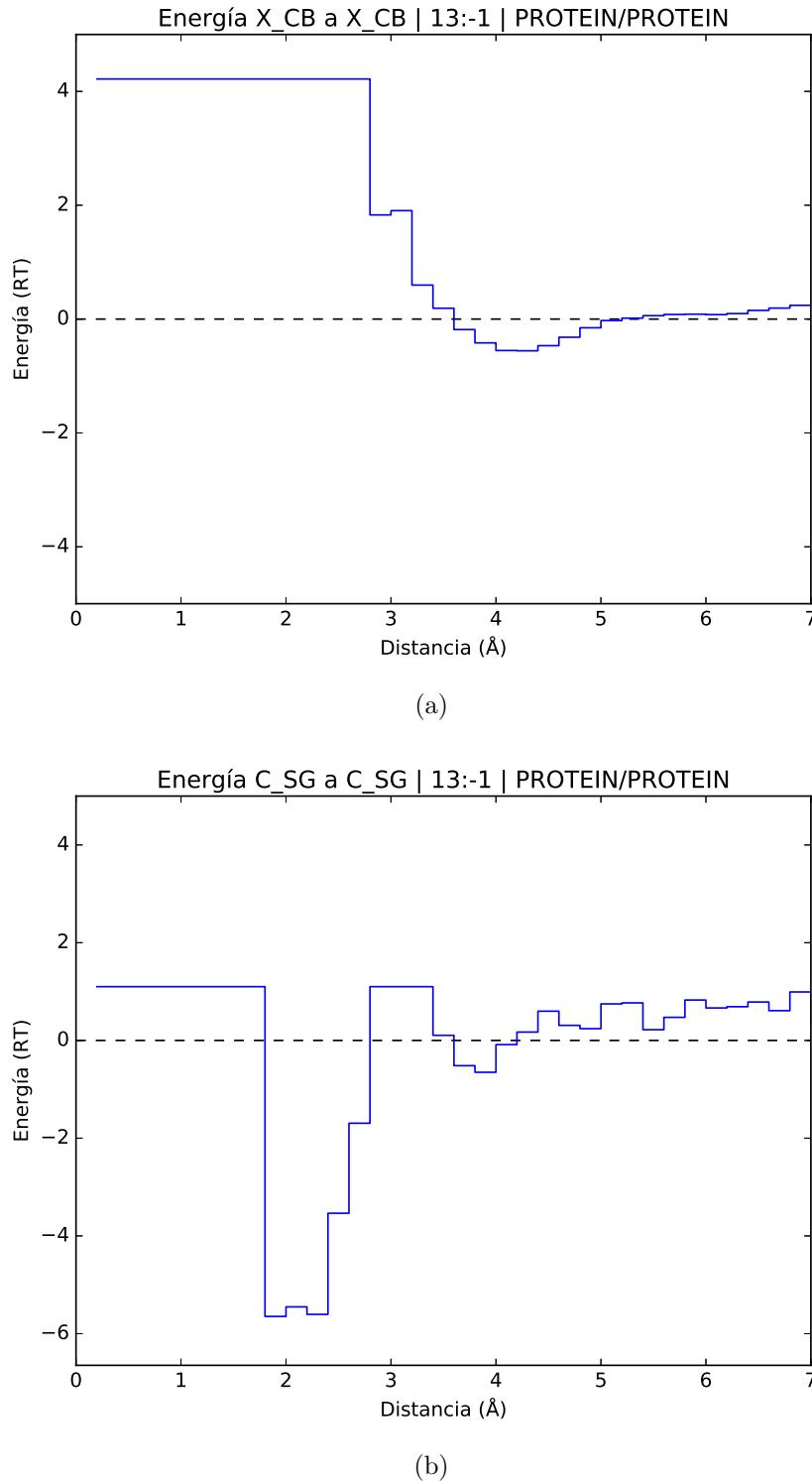


Figura 1. Gráficos de las funciones de energía utilizadas en proteínas. En (a) se observa la energía (unidades RT) en función de la distancia en Å para los carbonos beta de todos los aminoácidos. En (b) se tiene la función de energía para los átomos de azufre de cisteína, representando los puentes disulfuro.

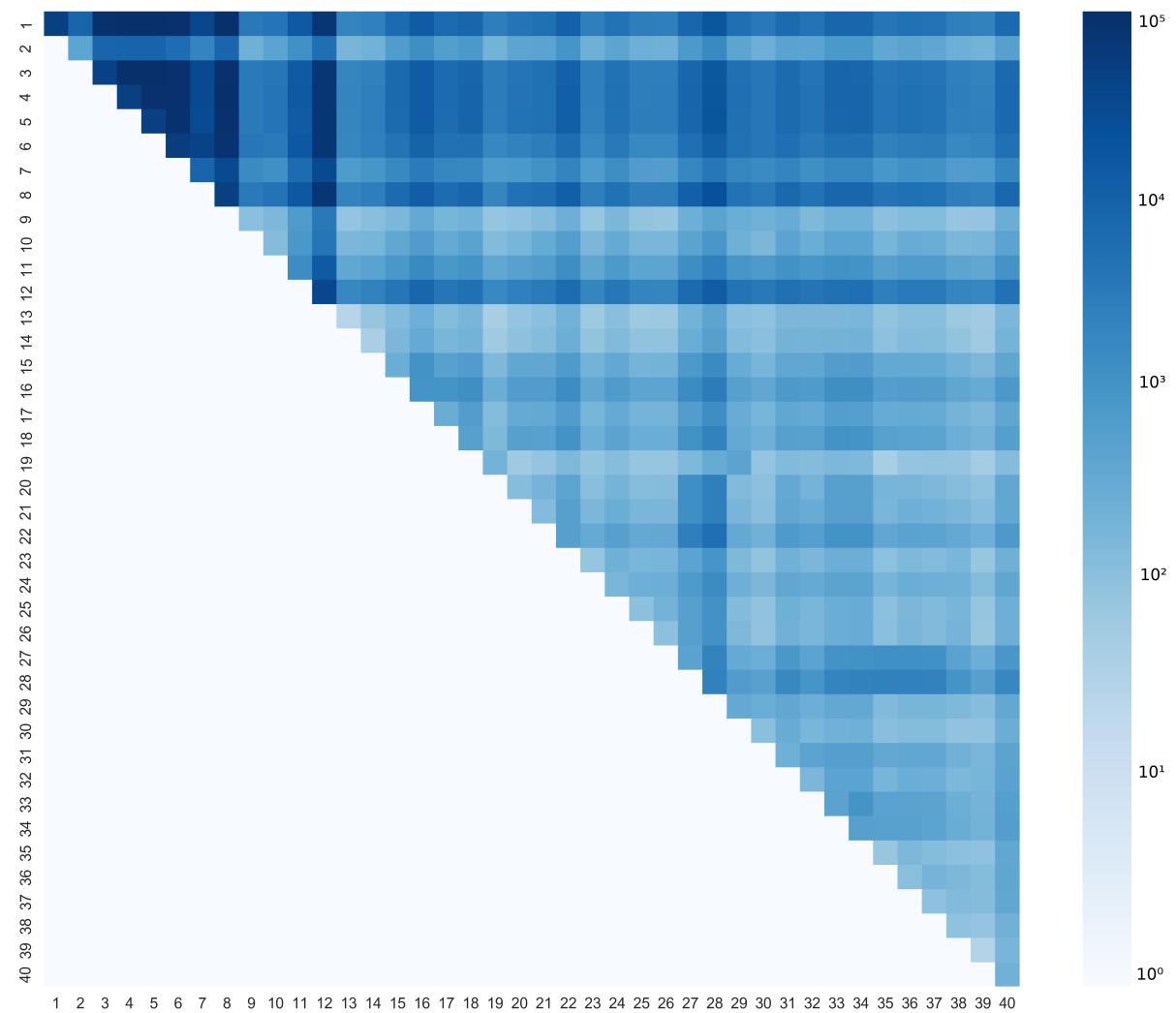


Figura 2. Matriz de conteo de interacciones para los potenciales en proteína, o M_{ij} . Solo se usa el triángulo superior de esta estructura, ya que no se considera el orden de las interacciones. Conteos están en escala logarítmica para facilitar la visualización.

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 clases 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:

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

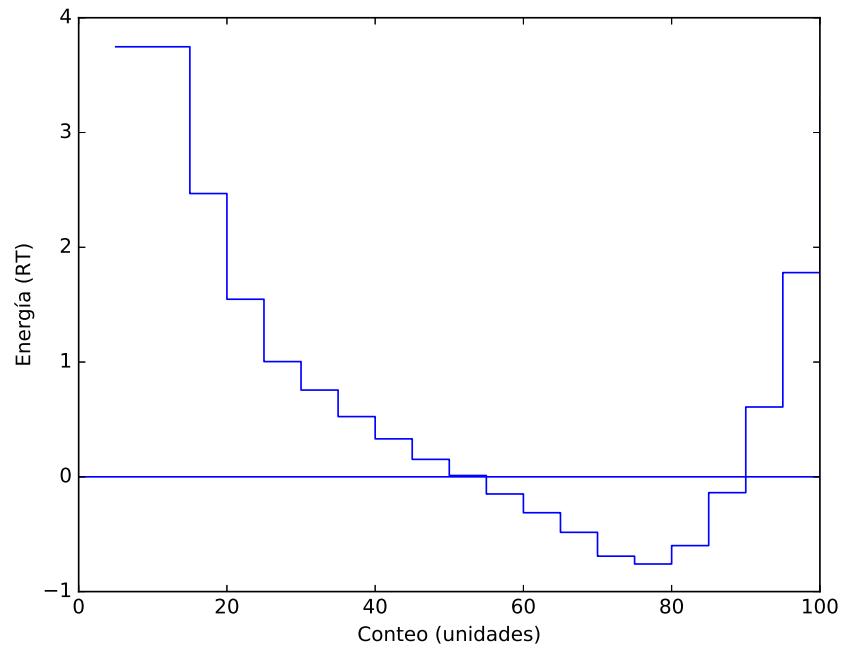
Donde i indica el tipo de átomo, M_i es la cantidad de observaciones del átomo tipo i , $f^i(c)$ es la frecuencia de observaciones en que un átomo i tuvo c átomos a su alrededor y f_{rel} es la frecuencia esperada, equivalente a 1/(número de clases). El factor de corrección σ se mantiene en 1/50. Un 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

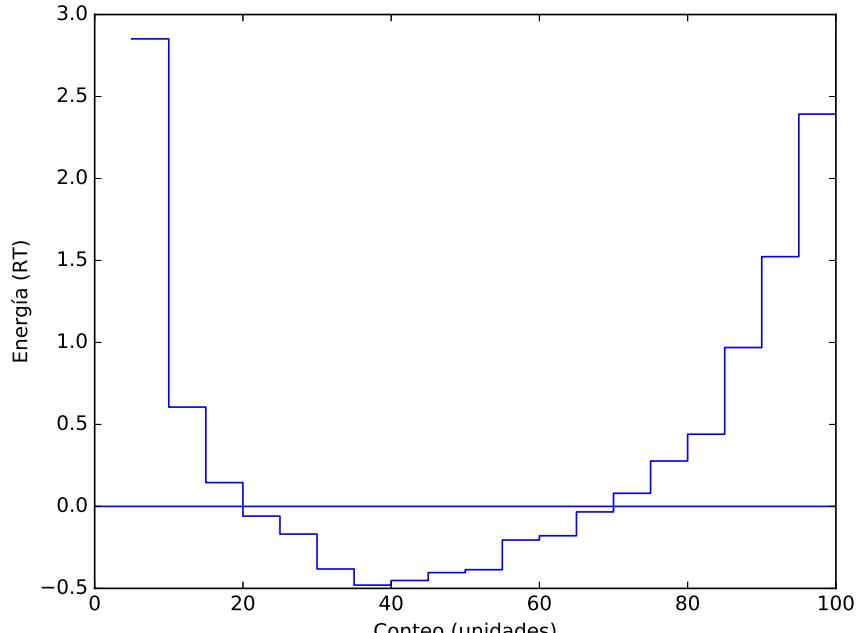
```

1: procedure GENERATECOUNTPOTENTIAL
2:   matrix1D  $M_i$            #Vector de conteo de átomos del tipo I en la base de datos
3:   matrix2D  $F_i$            #Tabla de frecuencias de conteo para cada tipo atómico y bin
4:   float  $F_{rel}$            #Frecuencia esperada, 1/clases
5:   list  $pdblist \leftarrow GetPDBs(argv1)$     #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( $PDBlist, F_i, F_{rel}, M_i, N_{clases}$ )      #Calcula todas las tablas necesarias para la derivación del potencial
9:     WritePotentialCount( $F_i, F_{rel}, M_i, N_{clases}, \Sigma$ )       #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 Å 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 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 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:   pdbstrudct 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 interactuan con
I
9:       for point in points do
10:         if IsPointInVolume(atomJ, point) and point.enabled then      #Revisa si el
punto esta 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  $\leftarrow$  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
  
```

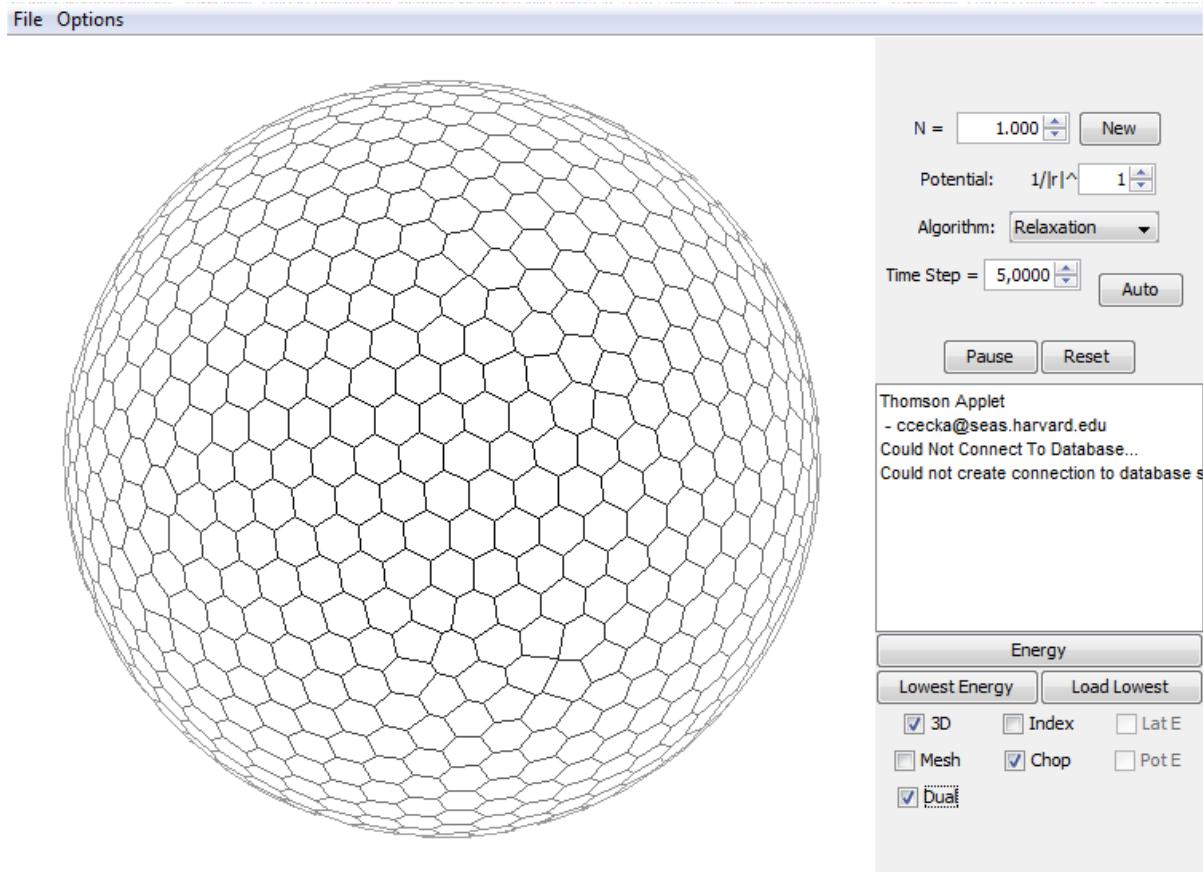


Figura 4. En este ejemplo podemos observar una esfera subdividida en 1000 áreas equivalentes. Inicialmente las subáreas son inicializadas con valores aleatorios, los cuáles son optimizados utilizando un algoritmo de relajación considerando cada punto como una carga eléctrica forzada a moverse sobre una 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 agregando nuevas variables para almacenar la información de que átomos entierran un cierto punto en la superficie de otro. Así es posible encontrar los grupos de puntos de un átomo que son enterrados por otros grupos únicos de átomos, identificando así todos los sobrelapamientos únicos posibles, como indicado en el diagrama de la figura 5. En el algoritmo 4 podemos ver las modificaciones, donde se agregan las estructuras *pBuriedBy* y *AreaBuriedBy*, que permiten almacenar la información de que puntos de la superficie de un átomo son enterrados por cuáles otros grupos de átomos. La estructura final, *AreaBuriedBy* contiene los conjuntos únicos de átomos y la cantidad de puntos que cada uno de los conjuntos entierra.

En la figura 5 tenemos el ejemplo más simple donde ocurren sobrelapamientos, un sistema de 3 cuerpos. Para este sistema, la superficie enterrada del cuerpo 1 por el cuerpo 2 corresponde a la sección en rojo, más la sección en púrpura dividida por 2. Esto se debe a que dos átomos participan en este sobrelapamiento, que incluye el cuerpo 2. Por lo tanto, el valor de superficie enterrada que un cuerpo *B* le causa a un cuerpo *A* es igual a la suma de todos los sobrelapamientos divididos por la cantidad de átomos participantes donde se encuentra el cuerpo *B*.

Algoritmo 4 Pasos para el cálculo de las superficies de interacción

```

1: procedure CALCULATEDBSA
2:   pdbstrudct 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:     map pBuriedBy #Almacena las listas de átomos que entierran un cierto punto en la superficie del átomo I
6:     list AreaBuriedBy #Almacena los conjuntos de átomos únicos encontrados y la cantidad de puntos que entierran
7:     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
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) then #Revisa si el punto está adentro del volumen del átomo J
11:           pBuriedBy[point].add(atomJ) #Agrega átomo a la lista de átomos que entierran este punto
12:           AreaBuriedBy  $\leftarrow$  ParseSets(pBuriedBy) #Procesa los conjuntos encontrados y calcula el área enterrada por cada uno
13:           atomI.AreaBuriedBy  $\leftarrow$  AreaBuriedBy #Almacena las interacciones en el objeto átomo
  
```

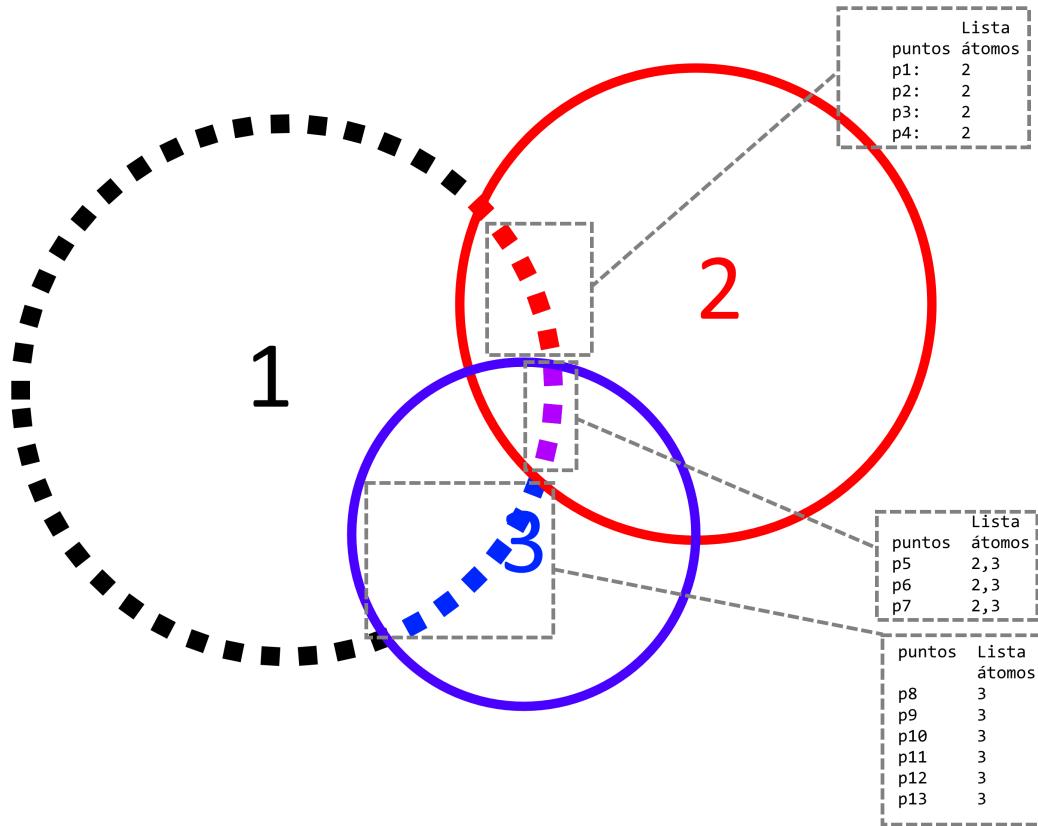


Figura 5. Diagrama en 2D del funcionamiento del algoritmo para el cálculo de las subsuperficies de interacción. En este ejemplo de 3 cuerpos, se observa como el cuerpo 1 tiene su superficie enterrada por los cuerpos 2 y 3. Cada punto en la superficie de 1 representa un valor de superficie equivalente a $(\text{superficie total}) / (\text{número de puntos})$. Los puntos coloreados de rojo representan el área enterrada solo por el cuerpo 2, los puntos azules el área enterrada solo por 3, y los puntos púrpuras el área enterrada por los cuerpos 2 y 3. En los recuadros grises se tienen representaciones del mapa de puntos enterrados hacia lista de átomos, *pBuriedBy* en el algoritmo 4.

4.6 Derivación de potenciales basados en BSA

Al obtener los valores de \AA^2 para cada interacción entre átomos I y J , la derivación de un potencial procede de manera similar a los potenciales que utilizan distancia. Un detalle importante es que dado que los radios de Van der Waals de los átomos I y J pueden ser distintos, lo que puede causar que el valor de superficie que I entierra de J sea distinto al valor que J entierra de I . Para resolver este problema, se derivan dos potenciales al mismo tiempo, uno con los valores de I que entierran a J y otro para los valores de J que entierran a I . Estos dos potenciales se evalúan y suman al evaluar cualquier interacción de I contra J .

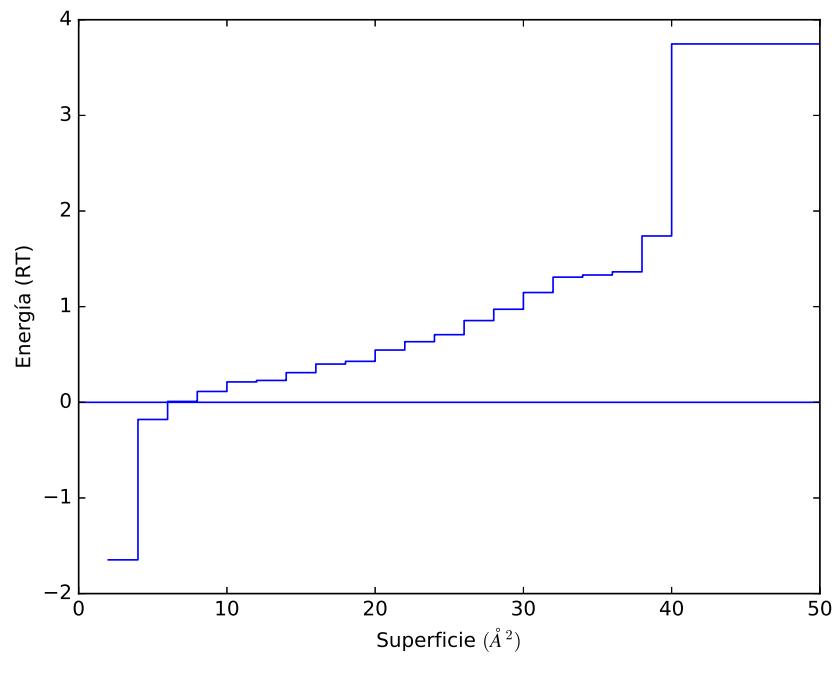
4.7 Derivación de potenciales basados en SASA

Los potenciales basados en SASA se derivan de la misma manera que los potenciales de conteos de átomos. El único cambio el uso del SASA de cada átomo, en vez del número de átomos alrededor, como la variable en todas las ecuaciones y funciones. Los parámetros utilizados para estos potenciales fueron 30 clases de 1.67 \AA^2 con el límite de la última clase en 50 \AA^2 . Estos parámetros se usaron para los potenciales en proteínas, ADN y ARN. En la figura 6 se pueden observar los dos potenciales SASA, uno para átomos de carbono hidrofóbicos y otro para nitrógenos hidrofílicos.

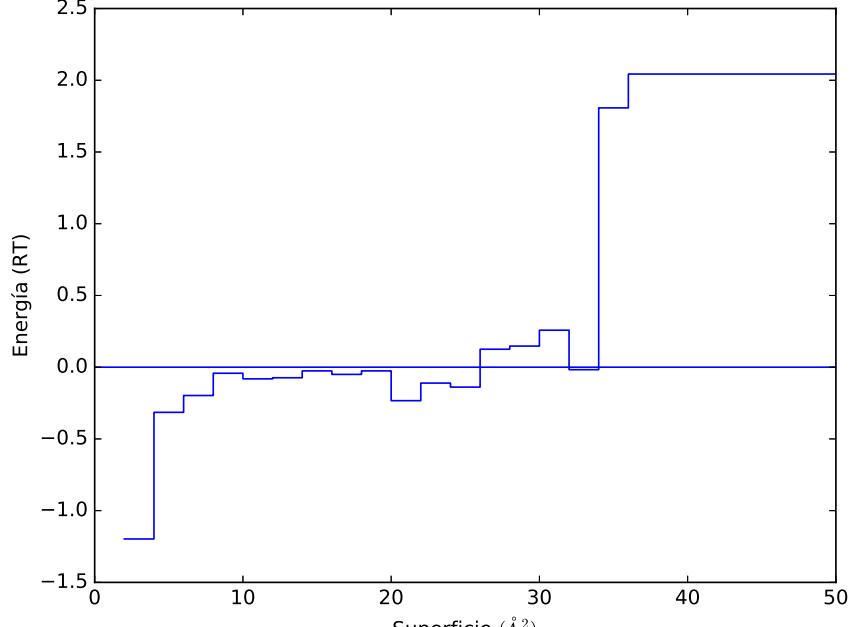
4.8 Combinación de los potenciales de distancia y conteo

Estos potenciales se combinan con la siguiente formula, que es ejecutada con los valores de energía para cada átomo encontrado. Esta suma los valores de los potenciales de distancia y conteo, utilizando un peso W_{ti} que depende del tipo molecular al que pertenece el átomo, de su tipo atómico, y de la cantidad de interacciones con otros átomos que posee. (Melo y Feytmans 1997)

$$E_{ti} = E_i^{DD} + (E_i^{CNT} * W_{ti}) \quad (7)$$



(a)



(b)

Figura 6. En (a) se puede observar la energía en función de la superficie expuesta al solvente para átomos de carbono aromáticos. En (b) se puede observar otra función, pero para nitrógenos terminales de histidina

Donde:

$$W_{ti} = \begin{cases} (Imax_t - C_i - 1)/(Imax_t - Imin_{ti} - 1) & C_i < Imax_t \\ 0 & C_i \geq Imax_t \end{cases}$$

Los términos $Imax$, C , e $Imin$ corresponden respectivamente al número máximo de interacciones encontrado para el tipo molecular correspondiente en el conjunto de entrenamiento, el número de interacciones que posee el átomo evaluado, y el número mínimo de interacciones encontradas para el tipo atómico y molecular evaluado. Los términos E^{DD} y E^{CNT} corresponden a las energías calculados por los potenciales de distancia y conteo respectivamente.

4.9 Combinación de los potenciales BSASA y SASA

Se utiliza la misma fórmula descrita para la combinación de los potenciales de distancia y conteo.

4.10 Cálculo del IP (Information Product)

El IP o *Information Product* es una medida de la información mutua de un potencial (Solis y Rackovsky 2008), que permite la comparación de estos de manera independiente a pruebas usando estructuras. Esta basado en la siguiente ecuación:

$$P = \sqrt{\bar{n}} \cdot \Delta \bar{E}^{ij} \quad (8)$$

Donde \bar{n} es el número promedio de observaciones de interacciones válidas de acuerdo a los parámetros del potencial en el conjunto de estructuras de derivación del mismo mientras que $\Delta \bar{E}^{ij}$ es el valor promedio de energía de estas mismas interacciones. En los potenciales de conteo y SASA el término ij pasa a ser solo i .

RESULTADOS

5.1 Pruebas en potenciales de proteínas

5.1.1 Clasificación de estructuras en nativas o no nativas

Los resultados de todos los potenciales en esta prueba se ven en la figura 7 donde los nuevos potenciales BSASA, SASA y BSASA+SASA no logran superar en performance a los antiguos dependientes de distancia. En la tabla 3 podemos ver la comparación del desempeño de todos los potenciales utilizando el método STaR (Vergara y col. 2008), donde los potenciales BSASA,SASA y combinación BSASA+SASA tienen desempeños estadísticamente equivalentes entre si, y distintos a los potenciales de distancia, combinación distancia más conteo, y conteo.

	DD+CONTEO	BSASA	BSASA+SASA	DD	CONTEO SASA
DD+CONTEO	-	-	-	-	-
BSASA	<0.001	-	-	-	-
BSASA+SASA	<0.001	0.453	-	-	-
DD	<0.001	<0.001	<0.001	-	-
CONTEO	<0.001	0.001	<0.001	<0.001	-
SASA	<0.001	0.284	<0.001	<0.001	<0.001

Tabla 3: Comparación de las curvas ROC para el desempeño de la clasificación de modelos en nativos y no nativos. Valores corresponden al p-value de la prueba no paramétrica utilizada por STaR. Un p-value menor a 0.05 indica que el par no es equivalente.

5.1.2 Detección de resíduos erroneamente modelados

En la figura 8 podemos observar como los nuevos potenciales tienen mejores AUC que sus pares basados en distancias y conteos de átomos tanto como en los modelos clasificados en clase A tanto como en clase B. Al comparar las curvas de la clase A con el método STaR en la tabla 4 (Vergara y col. 2008), solo hay diferencias significativas entre los potenciales de conteo y superficie SASA, considerando a los otros potenciales como equivalentes. Mientras tanto, en los modelos clase B, en los cuales hay más dificultad en encontrar errores, todos los potenciales

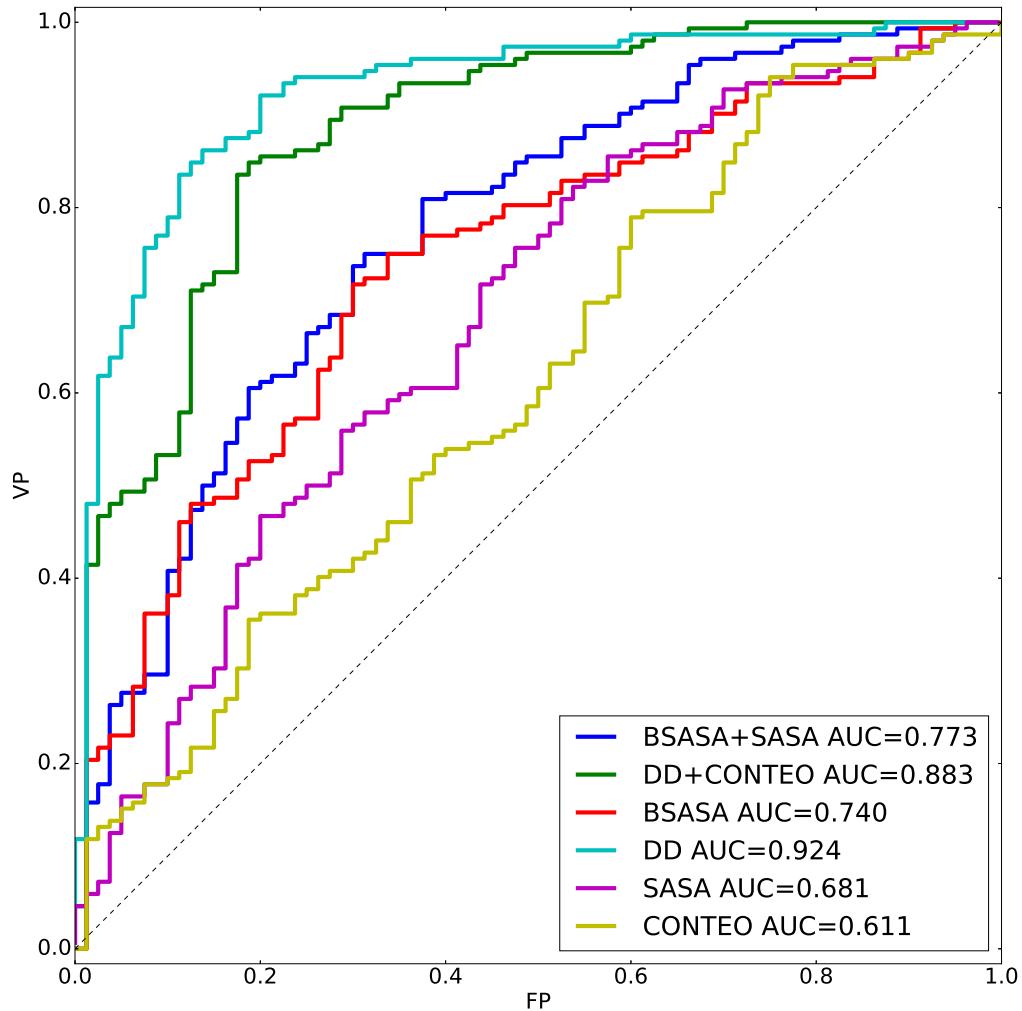


Figura 7. Curvas ROC para el desempeño de la clasificación de modelos en nativos y no nativos.

se consideran distintos excepto los pares DD+CONTEO/CONTEO y BSASA/SASA.

Clase A	DD+CONTEO	BSASA	BSASA+SASA	DD	CONTEO SASA
DD+CONTEO	-	-	-	-	-
BSASA	0.036	-	-	-	-
BSASA+SASA	0.299	<0.001	-	-	-
DD	<0.001	0.726	0.007	-	-
CONTEO	0.041	0.876	<0.001	0.679	-
SASA	0.801	0.004	0.054	0.079	<0.001

Clase B	DD+CONTEO	BSASA	BSASA+SASA	DD	CONTEO SASA
DD+CONTEO	-	-	-	-	-
BSASA	<0.001	-	-	-	-
BSASA+SASA	<0.001	<0.001	-	-	-
DD	<0.001	<0.001	<0.001	-	-
CONTEO	0.051	<0.001	<0.001	<0.001	-
SASA	<0.001	0.732	<0.001	<0.001	<0.001

Tabla 4: Comparación entre curvas ROC para modelos clase A y B. Valores corresponden al p-value de la prueba no paramétrica utilizada por STaR. Un p-value menor a 0.05 indica que el par no es equivalente.

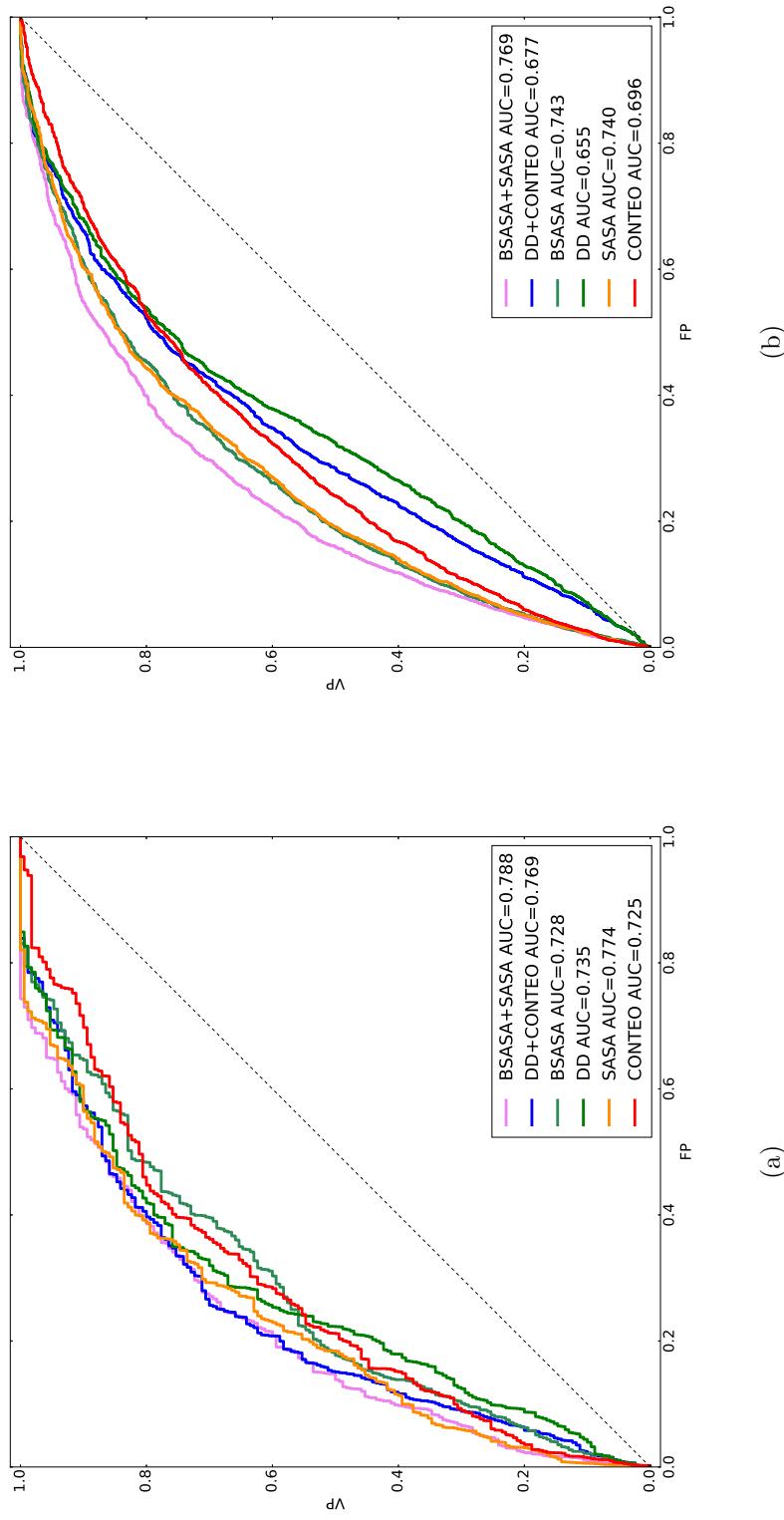


Figura 8. Curvas ROC para el desempeño en la detección de residuos mal modelados. En (a) se tiene la curva para las estructuras de Clase A, descritas anteriormente en métodos, mientras que en la figura (b) se tienen los resultados para las estructuras de clase B.

5.2 Pruebas en potenciales para ARN

5.2.1 Correlación entre energía y desviación estructural

En estas pruebas los nuevos potenciales tienen características de desempeño menores que los potenciales de referencia y sus combinaciones, que son RASP, CONTEO y RASP+CONTEO, como se puede observar en la figura 9. En la tabla 5 se tiene el p-value de las comparaciones entre todos los pares de boxplots vistos en la figura 9, para RMSD tanto para GDT utilizando el test de Wilcoxon, una prueba no paramétrica alternativa al t-test de Student para datos pareados. (Wilcoxon 1945)

RMSD	RASP+CONTEO	BSASA	BSASA+SASA	RASP	CONTEO	SASA
RASP+CONTEO	-	-	-	-	-	-
BSASA	<0.001	-	-	-	-	-
BSASA+SASA	<0.001	<0.001	-	-	-	-
RASP	<0.001	<0.001	<0.001	-	-	-
CONTEO	<0.001	0.4914	<0.001	<0.001	-	-
SASA	<0.001	<0.001	<0.001	<0.001	<0.001	-

GDT	RASP+CONTEO	BSASA	BSASA+SASA	RASP	CONTEO	SASA
RASP+CONTEO	-	-	-	-	-	-
BSASA	<0.001	-	-	-	-	-
BSASA+SASA	<0.001	<0.001	-	-	-	-
RASP	0.1828	<0.001	<0.001	-	-	-
CONTEO	<0.001	<0.001	<0.001	<0.001	-	-
SASA	<0.001	<0.001	<0.001	<0.001	0.0619	-

Tabla 5: Comparación entre distribuciones de coeficientes de Pearson para correlaciones entre energía y desviación estructural en ARN para las medidas de RMSD y GDT. Los valorse en la tabla son los p-value de la prueba de Wilcoxon. Valores de p-value menores a 0.05 indican que la distribución es significativamente distinta.

5.2.2 Desempeño en encontrar el mejor modelo no canónico

En esta prueba se evaluó la capacidad de los nuevos potenciales en encontrar el mejor modelo de RNA con interacciones no canónicas, como fue descrito en la metodología. En la figura 10 se tienen gráficos que indican en cuáles de los 32 casos posibles cada par de potenciales

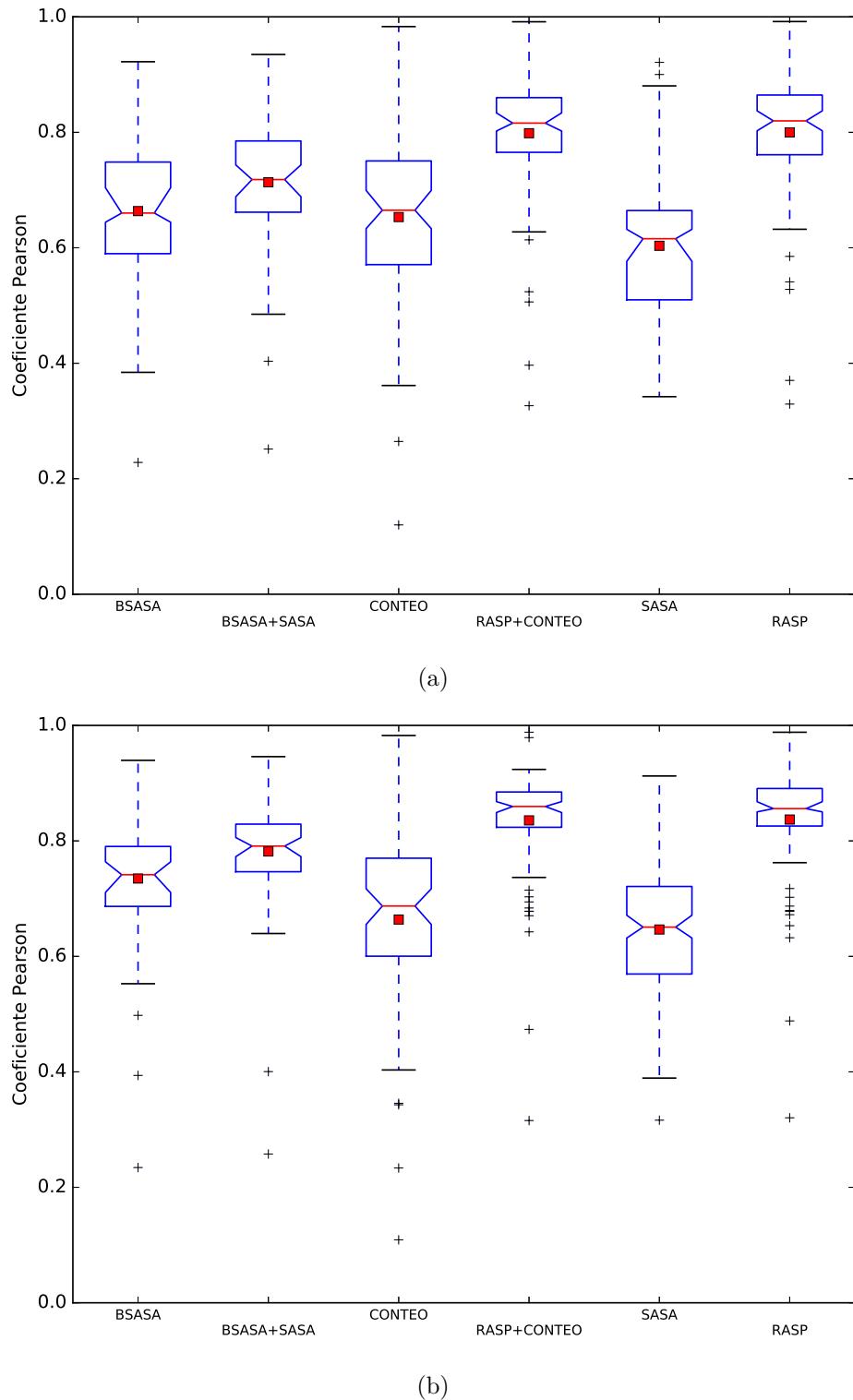


Figura 9. Boxplots de las distribuciones de correlación de desviación estructural en RNA con un potencial. En la figura (a) se tienen las correlaciones entre los valores de energía calculados por los potenciales evaluados y el RMSD, mientras que en (b) se tiene la correlación entre estos valores de energía y el GDT. La línea roja indica la mediana, el cuadrado rojo el promedio y la cintura el intervalo de confianza de la mediana. El N de cada boxplot es de 84.

encuentra un peor o mejor modelo. Se puede observar que el potencial de superficie SASA supera al potencial CONTEO y que el potencial combinado BSASA+SASA supera al potencial utilizando la metodología antigua RASP+CONTEO.

En la tabla 6 tenemos destacado las oportunidades en que un potencial logró encontrar el modelo con el mejor RMSD para cada estructura, cuyo valor está en la última columna. El potencial BSASA+SASA logra 10 aciertos, contra solo 6 del potencial RASP+CONTEO. Entre tanto el potencial RASP solo logra 8 aciertos contra los 7 de BSASA, y en los potenciales de superficie SASA logra 8 aciertos contra solo 4 de CONTEO.

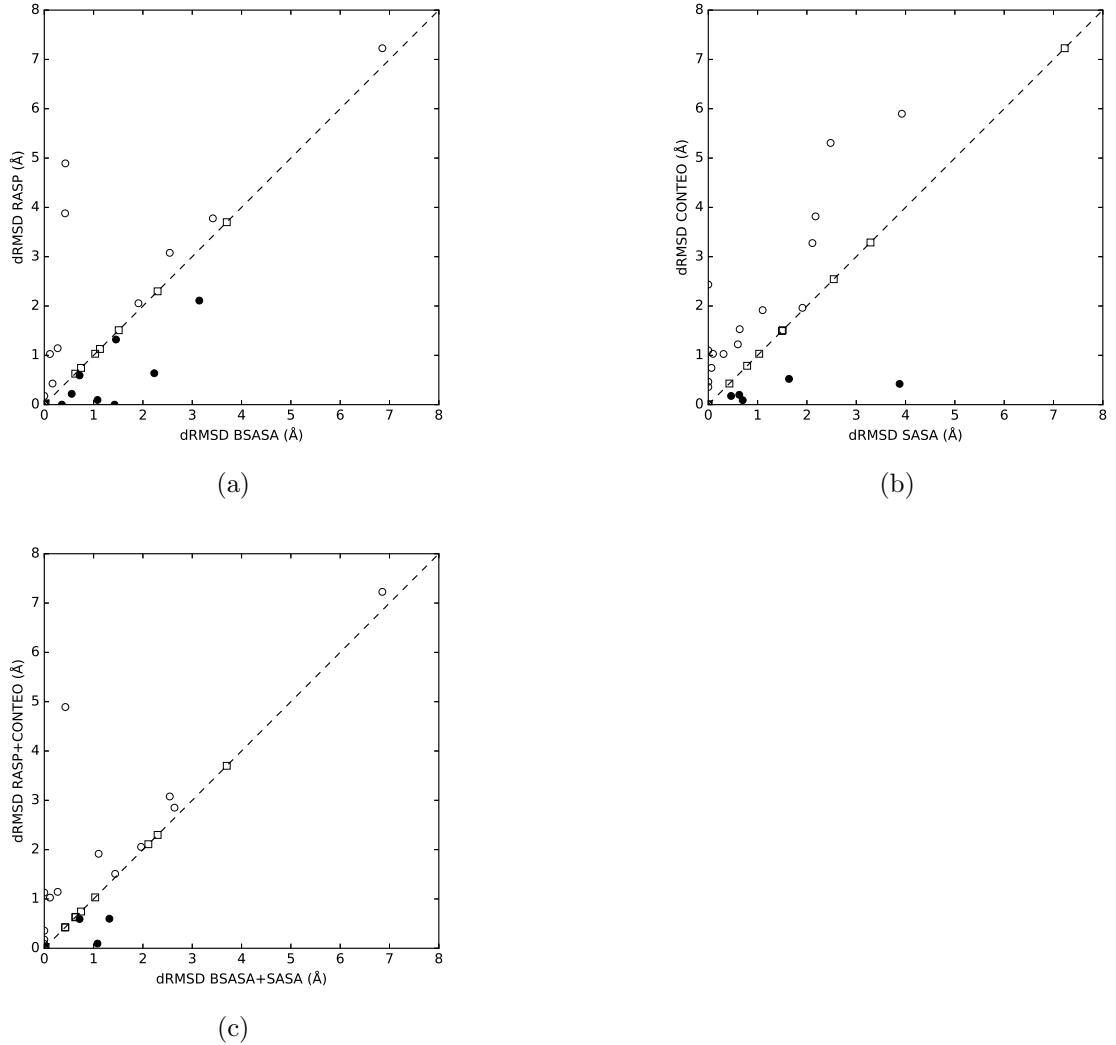


Figura 10. Comparación de las diferencias de RMSD entre el modelo de menor RMSD y el modelo de menor energía encontrado por el potencial. En cada gráfico los puntos negros corresponden a los casos donde el potencial en el eje Y seleccionó un modelo más preciso que el potencial en el eje X. Los puntos blancos indican lo contrario de lo anterior, y los cuadrados blancos indican que ambos potenciales encontraron el mismo resultado. En (a) se tiene la comparación entre los potenciales RASP y BSASA, en (b) entre los potenciales CONTEO y SASA, y en (c) entre los potenciales combinados RASP+CONTEO y BSASA+SASA. Los gráficos utilizan los datos de la tabla 6.

PDB ID	BSASA	RASP	SASA	CONTEO	BSASA+SASA	RASP+CONTEO	MEJOR RMSD
2gis-003	8.075	8.447	8.447	8.447	8.075	8.447	1.22
2oeu-002	9.863	9.863	8.193	7.764	9.863	9.863	7.565
157d-001	1.19	1.19	2.292	3.106	2.292	3.106	1.19
2r8s-003	2.475	2.352	1.759	2.859	2.475	2.352	1.759
1kd5-001	4.044	2.451	2.451	3.344	2.451	2.451	1.814
1xjr-003	4.142	3.109	3.11	4.275	3.109	3.109	0.999
2eew-005	6.214	7.126	6.408	7.126	6.214	7.126	6.099
1d4r-001	1.832	1.832	2.62	2.62	1.832	1.832	1.832
2qwy-003	7.824	12.289	10.685	10.685	7.824	12.289	7.397
1q9a-003	0.819						
2gdi-003	1.995						
1lnt-002	2.346	2.346	1.217	3.652	1.217	2.346	1.217
2r8s-004	10.973	10.973	10.973	10.973	10.973	10.973	9.941
1jj3-006	8.851	9.724	10.073	10.073	8.851	9.724	8.58
2oiu-013	1.214	0.857	0.857	1.214	0.857	1.214	0.857
3b31-099	3.547	3.808	3.808	3.808	3.808	3.808	3.38
1kh6-001	11.485	11.351	10.63	11.254	11.351	10.63	10.031
lopE-099	2.269	2.269	3.278	2.163	2.269	2.269	1.641
1mhk-001	5.136	5.495	5.643	7.616	4.358	4.569	1.718
255d-002	2.095	2.095	2.095	2.559	2.095	2.095	2.095
1d4r-001	4.695	3.27	3.27	3.27	3.27	3.27	3.27
uucg-005	1.148	1.148	1.122	1.122	1.148	1.148	1.122
2qbz-099	4.507	3.523	3.523	4.458	4.507	3.523	3.428
2r8s-004	3.312	6.77	6.77	3.312	3.312	3.312	2.891
2oiu-002	2.24	2.416	2.702	2.416	2.24	2.416	2.24
1q9a-001	5.16	5.16	3.634	5.277	5.16	5.16	1.46
283d-001	1.743	1.743	1.062	1.743	1.743	1.743	0.998
2oeu-003	11.187	11.721	11.187	11.187	11.187	11.721	8.643
359d-001	7.659	7.659	10.139	12.966	7.659	7.659	7.659
1u9s-001	3.293	3.44	3.293	3.347	3.347	3.44	1.384
2eew-003	9.59	9.59	9.59	9.59	9.518	9.59	8.081
1csl-002	4.415	4.078	4.557	3.949	3.859	3.949	3.859

Tabla 6: Tabla con los mejores RMSD encontrados por los potenciales para los 32 casos de estructuras no canónicas. El nuevo potencial BSASA+SASA logra encontrar 10 de los mejores casos, contra 8 de RASP y solo 6 de la combinación de RAS+CONTEO. En los potenciales de superficie, SASA supera a CONTEO con 8 modelos encontrados versus 4 para CONTEO.

5.3 Pruebas en potenciales para ADN

5.3.1 Desempeño en la detección del modelo de menor RMSD

En esta prueba las capacidades de los potenciales en elegir los mejores modelos de ADN generados por el método de Ibarra 2013 fueron comparadas utilizando las distribuciones de dRMSD, la diferencia entre el modelo encontrado por el potencial y el modelo original. Esta prueba fue hecha con un set de modelos que fueron generados utilizando las restricciones espaciales calculadas por Ibarra 2013, y otro en el que estas fueron desactivadas.

En la figura 11 se observan los resultados para los dos conjuntos de modelos, mientras que en la tabla 7 podemos ver los resultados del test de Wilcoxon entre cada distribución de dRMSDs generada por los potenciales. En el primer conjunto, figura 11a y tabla 7a, podemos observar como aunque los potenciales BSASA y DD+CONTEO parecen tener mejor desempeño, de acuerdo al test estadístico todos estos potenciales son equivalentes. Para el caso de la figura 11b y tabla 7b, se pueden observar distribuciones muy parecidas a las anteriores, excepto para los potenciales SASA y CONTEO. Pero de acuerdo al resultado de la comparación los únicos potenciales distintos son los pares DD/SASA y DD+CONTEO/SASA, en los cuáles los potenciales DD y DD+CONTEO tienen un mejor desempeño que SASA. Las comparaciones entre los otros pares de potenciales no tuvieron diferencias significativas.

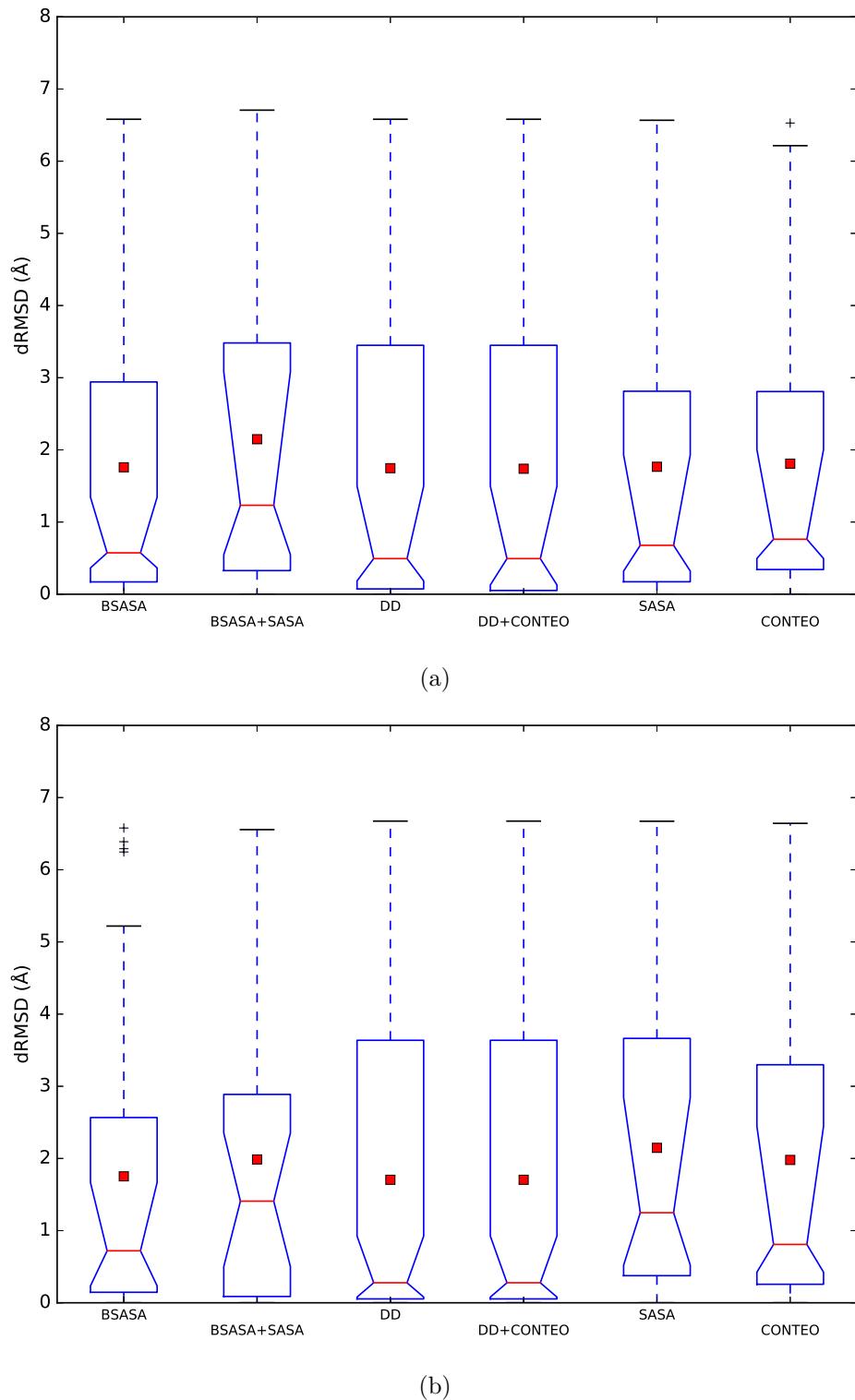


Figura 11. Comparación de las distribuciones de dRMSD entre los modelos con menor RMSD y los modelos con menor energía calculada por cada potencial. En la figura (a) tenemos los resultados calculados con los modelos con las restricciones activadas y en (b) con las restricciones desactivadas. La línea roja indica la mediana, el cuadrado rojo el promedio y la cintura el intervalo de confianza de la mediana. El N de cada boxplot es de 32.

A	DD+CONTEO	BSASA	BSASA+SASA	DD	CONTEO SASA
DD+CONTEO	-	-	-	-	-
BSASA	0.793	-	-	-	-
BSASA+SASA	0.085	0.187	-	-	-
DD	1.00	0.824	0.085	-	-
CONTEO	0.168	0.757	0.410	0.249	-
SASA	0.418	0.976	0.368	0.516	0.338

B	DD+CONTEO	BSASA	BSASA+SASA	DD	CONTEO SASA
DD+CONTEO	-	-	-	-	-
BSASA	0.423	-	-	-	-
BSASA+SASA	0.134	0.285	-	-	-
DD	1.0	0.424	0.134	-	-
CONTEO	0.081	0.285	0.771	0.081	-
SASA	0.010	0.059	0.285	0.010	0.147

Tabla 7: Tabla de comparaciones entre distribuciones de dRMSD entre pares de potenciales para los conjuntos de modelos de ADN con y sin restricciones. Los valores en la tabla son el p-value de la prueba de Wilcoxon. Valores menores a 0.05 indican que las distribuciones comparadas son significativamente distintas.

DISCUSIÓN

6.1 Pruebas en proteínas

La primera prueba que se le hace a los nuevos potenciales BSASA y SASA es la clasificación de estructuras en nativas y no nativas, un problema considerado ya resuelto por una gran cantidad de software y métodos publicados. (citar metodos) En la figura 7 podemos observar como los potenciales de la metodología nueva utilizando subsuperficies y la combinación con potenciales de superficie tienen peores resultados que los que utilizan distancias y la combinación de conteo de átomos y distancia. Esto se debe al rango de interacción más corto que tienen los potenciales BSASA, que impide el reconocimiento de posibles interacciones no favorables más distantes (figura 12) que no es capaz de observar ya que el rango máximo de distancia es equivalente a dos veces el radio de Van der Waals más 1.4 Å de distancia, siendo 1.4 Å el radio de una molécula de agua. En la figura 12 podemos ver las distribuciones de la cantidad de interacciones entre pares de átomos para las estructuras utilizadas en la derivación de los potenciales en proteína. La distribución para BSASA muestra que el potencial evalúa una menor cantidad de interacciones que los potenciales de distancia, que usan un rango de máximo de 7 Å para evaluar interacciones. De acuerdo a la ecuación 8 esto implica directamente que el potencial BSASA tiene un menor contenido de información disponible para la evaluación de la estructura, dado el menor número de interacciones pareadas observadas, que es el término n en la ecuación. Entre tanto, el potencial SASA logra un mucho mejor desempeño que conteo, dado que es mucho más fino en registrar si un átomo está enterrado o expuesto, ya que se mide directamente la superficie en contacto con el solvente, en vez del método indirecto del potencial de conteo.

En la segunda prueba, la capacidad de los nuevos potenciales BSASA y SASA en detectar errores aislados y expuestos en la superficie, tanto como errores más internos y difíciles de detectar fue evaluada utilizando perfiles de energía, en vez de usar el valor promedio de energía como en las pruebas anteriores. Estos perfiles de energía contienen valores de energía

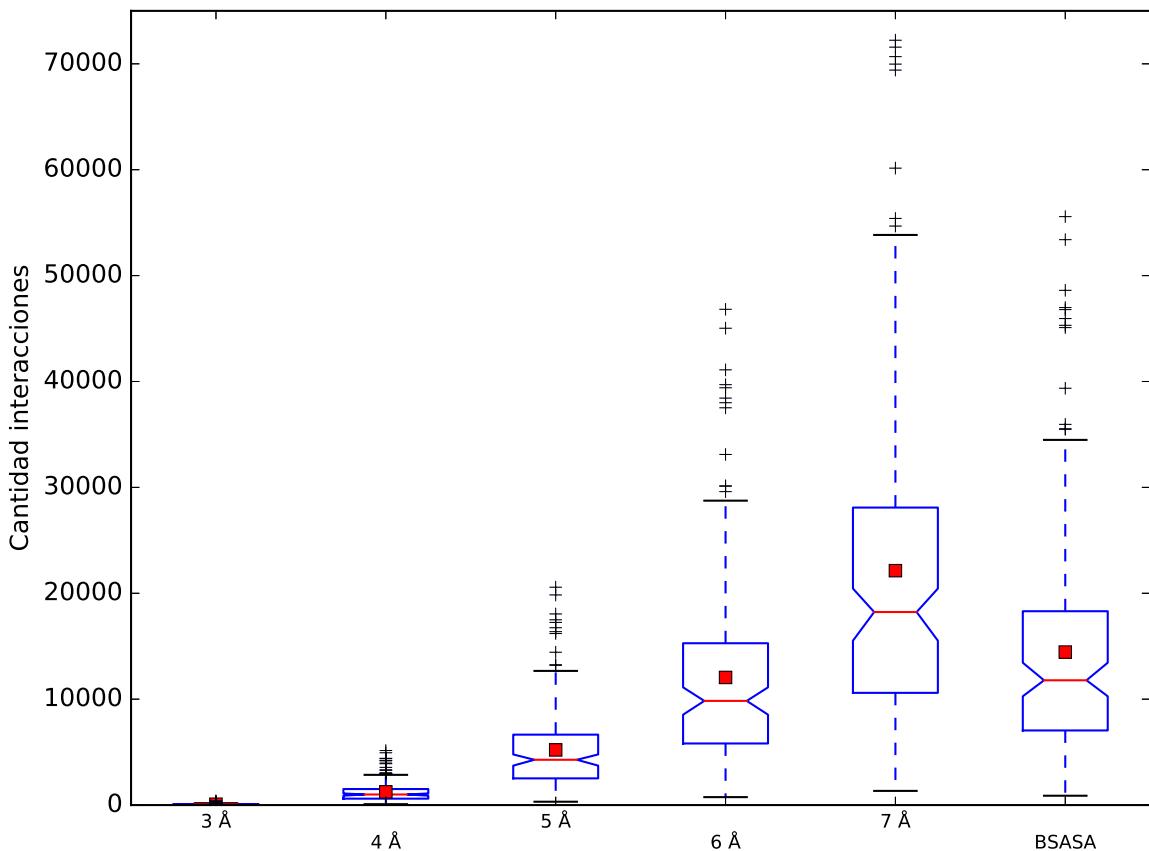


Figura 12. Boxplots de las distribuciones de la cantidad de contactos encontrados en el conjunto de estructuras usado en la derivación de potenciales para proteínas. En la figura se tienen las distribuciones de la cantidad de contactos para potenciales dependientes de distancia variando el radio máximo de interacción de 3 a 7 Å, más el potencial BSASA. La línea roja indica la mediana, el punto rojo el promedio, y la cintura el intervalo de confianza.

promediados en una ventana deslizante de 7 residuos lo que permite eliminar variaciones bruscas en los valores de energía de cada residuo. En esta prueba el valor final de energía de cada residuo se evalúa como un elemento independiente. Al evaluar los errores de clase A, notamos que el desempeño de los pares de potenciales comparables no es distingible, excepto en el caso de los potenciales SASA y CONTEO. Dado que los errores en los residuos clase A son en residuos aislados y expuestos, el mejor desempeño de SASA versus CONTEO era esperado. A su vez, los potenciales DD y BSASA y sus respectivas combinaciones DD+CONTEO y BSASA+SASA no muestran diferencias significativas en estos modelos. Pero en los modelos clase B, cuyos residuos poseen errores más internos en la estructura afectando también residuos correctamente modelados su alrededor o entorno local, los potenciales BSASA y BSASA+SASA logran un desempeño superior a los potenciales DD y DD+CONTEO. Esto se debe a la capacidad del potencial BSASA de excluir naturalmente los contactos que están escondidos detrás de otros átomos, algo que los potenciales de distancia no pueden detectar normalmente sin utilizar algoritmos para la detección de esos casos. (Ferrada y Melo 2007, 2009)

6.2 Pruebas en ARN

En la primera prueba se observó la correlación del valor de energía calculado por los potenciales con las medidas de desviación estructural de más de 80 estructuras con 500 modelos cada una. Esta prueba es muy similar a la primera prueba en proteínas, que buscaba separar modelos entre nativos y no nativos, que no es aún posible realizar dada la falta de una base de datos de estructuras y modelos de ARN ya clasificados en nativos y no nativos. Al igual que en la primera prueba en proteínas, los potenciales BSASA y su combinación BSASA+SASA tienen un desempeño menor al potenciales de referencia RASP. El potencial RASP utiliza un rango de 20 Å para la detección de interacciones, lo que significa que es capaz de detectar y evaluar motifs de estructuras terciarias estadísticamente probables de estar en proximidad. En la figura 13 apreciamos como el potencial RASP mide casi una unidad de magnitud más contactos que el potencial BSASA.

Pero en la segunda prueba, la cual se realiza utilizando fragmentos de RNA de menor

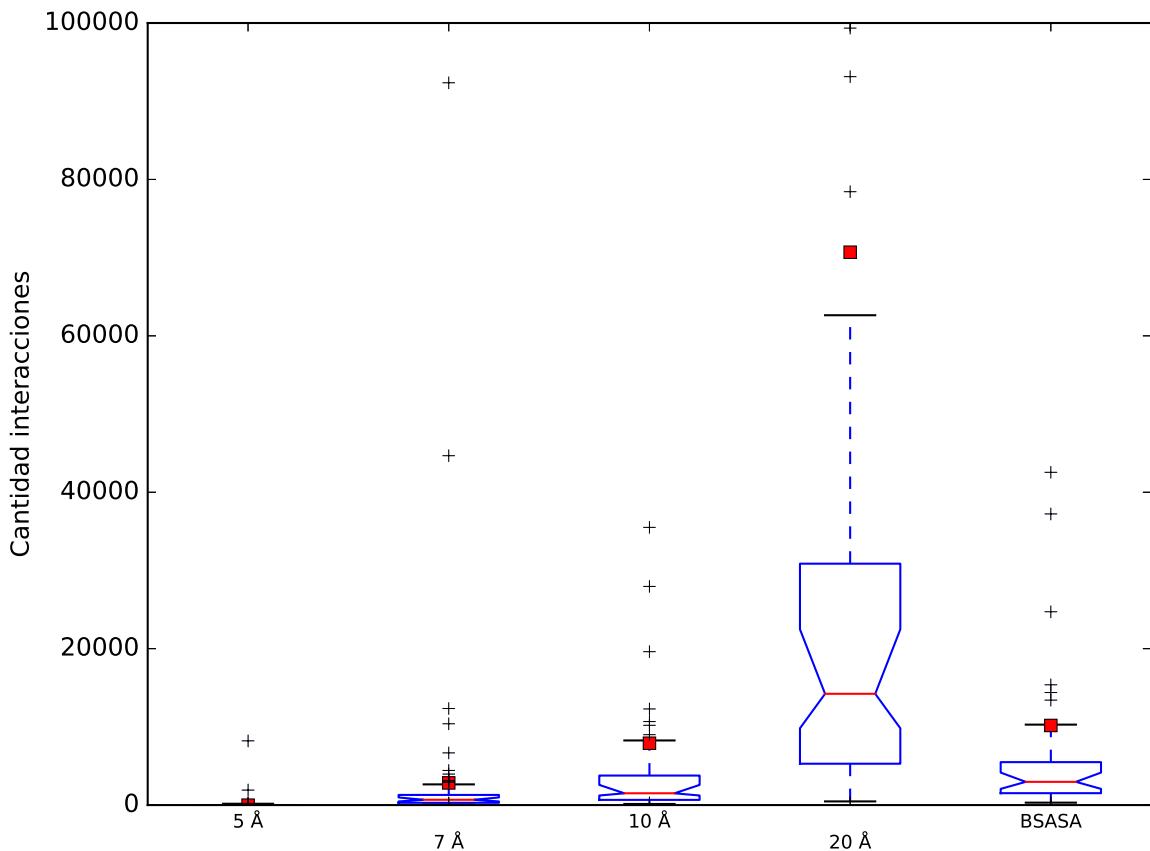


Figura 13. Boxplots de las distribuciones de la cantidad de contactos encontrados en el conjunto de estructuras usado en la derivación de potenciales para ARN. En la figura se tienen las distribuciones de la cantidad de contactos para potenciales dependientes de distancia variando el radio máximo de interacción de 5 a 20 Å, más el potencial BSASA. La línea roja indica la mediana, el punto rojo el promedio, y la cintura el intervalo de confianza.

tamaño, podemos apreciar como el potencial BSASA tiene mejor desempeño que RASP. Dado que se usan fragmentos pequeños de ARN, ocurre un efecto similar a lo que ocurre en la utilización de residuos en vez de la estructura completa, como en el segundo experimento en proteínas. Estos resultados indican que el potencial BSASA tiene una mejor desempeño al evaluar estructuras pequeñas (fragmentos) o errores locales en estructuras completas.

6.3 Pruebas en ADN

La prueba realizada en se enfoca en la capacidad del potencial en identificar el modelo más cercano a uno nativo. Esta prueba se asemeja al primer experimento realizado en ARN. En este caso se utilizan los modelos ya generados por el método con restricciones adicionales a MODELLER creado por **Ilibarra2013** en los cuales se generan modelos utilizando para una determinada secuencia de ADN derivando parámetros a partir de otras estructuras, excluyendo la estructura con la secuencia pedida del conjunto de datos. El potencial utilizado en este caso también posee las mismas restricciones, por lo que la estructura original no está en su conjunto de entrenamiento, lo que nos permite eliminar sesgos en la evaluación de los modelos. En la primera prueba, en donde se utilizan las restricciones de modelaje generadas a partir del set no redundante de estructuras de ADN, no se observaron diferencias significativas en el desempeño de los potenciales para encontrar los mejores modelos. Entretanto, para los modelos generados utilizando los parámetros por defecto del software MODELLER, solo se encontraron diferencias significativas entre los potenciales DD, DD+SASA y SASA. Al analizar algunas de las estructuras encontradas y los datos de la figura 11, podemos ver como los potenciales fallan en encontrar los modelos de menor RMSD, en algunos casos encontrando modelos con diferencias mínimas, como en la figura 14 y en otros seleccionando modelos con más de 4 Å comparados con el modelo de menor RMSD encontrado como en la figura 15.

Esto indica que los parámetros utilizados para los potenciales en ARN no son los óptimos para estructuras de ADN, o que el conjunto de entrenamiento utilizado no posee las características necesarias para generar potenciales capaces de una mejor discriminación entre los modelos específicos utilizados en esta prueba.

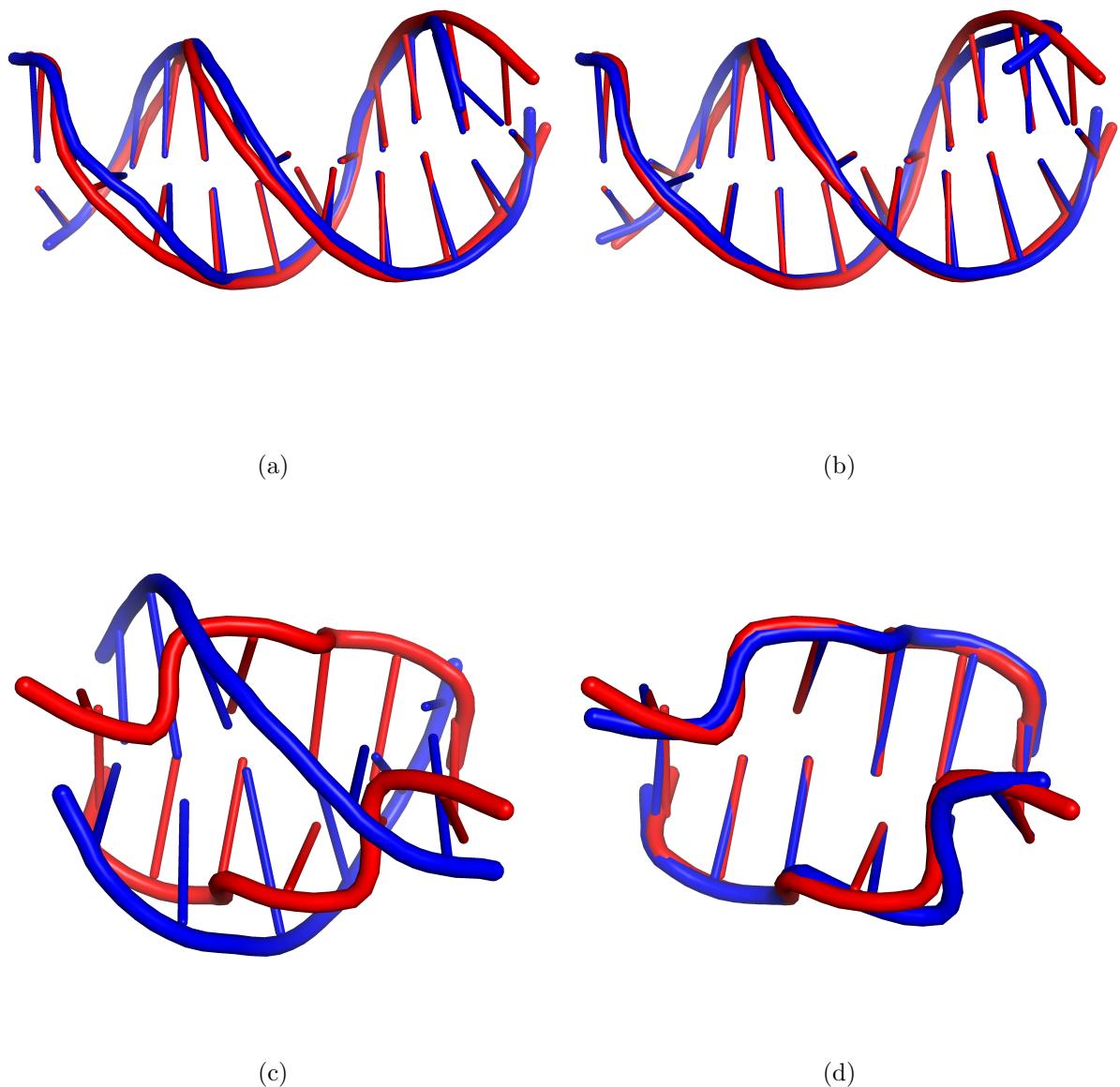


Figura 14. Superposiciones de algunos modelos generados para una secuencia específica con su cristal original para el conjunto sin restricciones de modelamiento. En todas las figuras los modelos de color rojo son el cristal original, mientras que superpuesto en azul está el modelo generado. En (a) se tiene el mejor modelo para la secuencia de la estructura 1D65 encontrado por el potencial BSASA, y en (b) el modelo con el menor RMSD, que en este caso fueron el mismo. En (c) y (d) se tiene uno de los peores resultados, con (c) mostrando el mejor modelo encontrado para el potencial BSASA para la estructura 181D con un dRMSD de 6.582 Å contra el modelo de menor RMSD encontrado en (d).

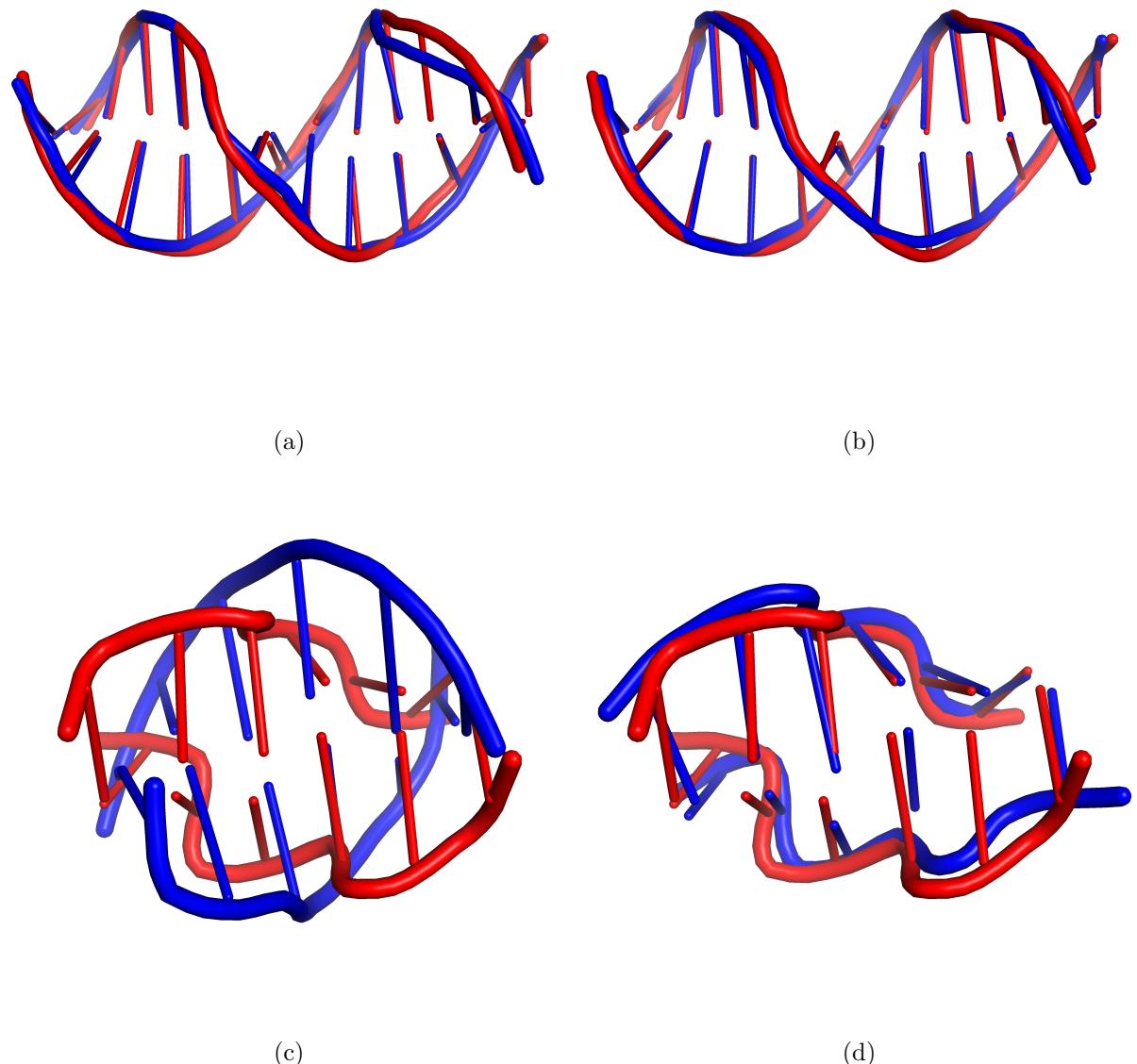


Figura 15. Superposiciones de algunos modelos generados para una secuencia específica con su cristal original para el conjunto sin restricciones de modelamiento. En todas las figuras los modelos de color rojo son el cristal original, mientras que superpuesto en azul está el modelo generado. En (a) se tiene el mejor modelo para la secuencia de la estructura 424D encontrado por el potencial BSASA, y en (b) el modelo con menor RMSD. En (c) tenemos el caso de 3EW9, donde el potencial BSASA no logra encontrar el mejor modelo, y en (d) tenemos el modelo de menor RMSD con el cristal.

Otro efecto observado fue que la combinación DD+SASA entregó resultados casi idénticos en ambas pruebas, lo que indica que la fórmula utilizada para ajustar la proporción del valor de energía que entrega cada potencial puede no ser óptima al evaluar estructuras de ADN.

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 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.
- Solis A. D. y Rackovsky S. (2008). «Information and discrimination in pairwise contact potentials». *Proteins: Structure, Function and Genetics* 71.3, págs. 1071-1087.
- Tange O. (2011). «GNU Parallel - The Command-Line Power Tool». ;*login: The USENIX Magazine* 36.1, págs. 42-47.
- Vergara I. A. y col. (2008). «BMC Bioinformatics». 5, págs. 1-5.
- Wilcoxon F. (1945). Individual Comparisons by Ranking Methods.