Análisis Estadístico de interacciones

# Intro

Previamente se generó una tabla de todos los pdbs del dataset y se contaron los tipos de interacciones de cada uno. Esto generó 8 columnas de acuerdo con cómo se están definidas estas interacciones:

* HBOND\_LIG\_SC
* HBOND:LIG\_MC
* HBOND:SC\_LIG
* HBOND:MC\_LIG
* VDW:LIG\_SC
* VDW:LIG\_MC
* VDW:SC\_LIG
* VDW:MC\_LIG

En reuniones previas se determinó que la definición de estas interacciones no indica direccionalidad, por lo que HBOND\_LIC\_SC es el mismo tipo de enlace que HBOND:SC\_LIG. Por ello se agruparon las columnas con mismo tipo de enlace y partes involucradas y se sumaron, dejando 4 categorías:

* HBOND:SC (puentes de hidrógeno entre ligando y side chain de aminoácidos)
* HBOND:MC (puentes de hidrógeno entre ligando y main chain de aminoácidos)
* VDW:SC (interacciones de Van der Waal entre ligando y side chain de aminoácidos)
* VDW:MC (interacciones de Van der Waal entre ligando y main chain de aminoácidos)

Una vez sumadas se normalizaron siguiendo la ecuación

Antes de normalizar:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PDBID** | **HBOND:SC** | **HBOND:MC** | **VDW:SC** | **VDW:MC** |
| 3JBN | 0 | 189 | 188 | 16 |
| 4D5Y | 0 | 276 | 501 | 46 |

Después de normalizar:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PDBID** | **HBOND:SC normal** | **HBOND:MC normal** | **VDW:SC normal** | **VDW:MC normal** |
| 3JBN | 0 | 0.480916031 | 0.478371501 | 0.040712468 |
| 4D5Y | 0 | 0.335358445 | 0.608748481 | 0.055893074 |

La tabla de pdbs con sus interacciones normalizadas se utilizó para el análisis estadístico. Cabe destacar que en este proceso se descubrió que los siguientes pdbs tenían vacíos los campos de la tabla, lo que implica que en la transformación de los datos del RING al conteo de interacciones algo falló o el RING no logró resultados con éstas.

5AJ4, 1NJM, 1NJP, 1VFG, 2AZ0, 2KMJ, 4A3K, 4B3P, 7ACE, 7XLT.

# Interacciones de puentes de hidrógeno

Estas interacciones son las que se encontraron entre el RNA ligando y la cadena identitaria de cada aminoácido.

|  |  |
| --- | --- |
| Side chain | **HBOND:SC normal** |
| mean | 0.0599 |
| std | 0.1810 |
| min | 0.0000 |
| 25% | 0.0000 |
| 50% | 0.0000 |
| 75% | 0.0000 |
| max | 1.0000 |
| var | 0.0327 |
| skew | 2.9520 |
| kurt | 7.3888 |

La estadística dio rara. El sd bajo indicaría que hay poca dispersión de datos. Aun así, parecería estar muy concentrada en el último cuartil ya que los límites del primer y segundo cuartil dan 0.

El skewness y kurtosis evalúan la simetricidad y concentración de datos en los extremos de la curva de datos, respectivamente. Estos dos valores que tenemos con este tipo de interacción parecerían indicar dispersión (una curva simétrica como la de una distribución normal se espera que tenga un skewness cercano a 0 y un kurtosis cercano a 3).

### Plots

* Los plots me confuden bastante. Por un lado el gráfico de densidad de datos da una forma muy similar a una distribución normal. Pero al verlo con atención hay un segundo mini-pico entre 0.5 y 0.75, que va a aumentar bastante el kurtosis (las concentraciones de datos en las “colas” del gráfico suben mucho el kurtosis, es muy sensible) y puede subir el skew porque le quita simetría a la data.
* Al boxplot no le logro sacar ninguna info. Es como si hubiera definido que la gran mayoría de las mediciones como outliers. No tiene sentido; no debería ser un problema de correr el script porque es un copypaste del mismo que usé para las otras interacciones.

A blue line graph with numbers

Description automatically generated

A graph with a line in the middle

Description automatically generated

## Main chain

|  |  |
| --- | --- |
|  | **HBOND:MC normal** |
| mean | 0.3134 |
| std | 0.2155 |
| min | 0.0000 |
| 25% | 0.1591 |
| 50% | 0.3013 |
| 75% | 0.4291 |
| max | 1.0000 |
| std | 0.2155 |
| skew | 0.9422 |
| kurt | 1.5133 |

Las interacciones entre ligando las uniones peptídicas. Con los valores de los cuartiles parecería que hay no hay una concentración alta de datos en ninguno de ellos. El sd es bajo, por lo que no debería haber tanta dispersión.

El skew y kurt cambiaron bastante respecto a las interacciones con la side chain. En este caso son mucho más bajos, pero no llegan a los valores esperados de una distribución normal.

### Plots

* Estos plots tienen un poco más de sentido. El de densidad muestra que hay tres valores más altos que el resto muy cercanos entre sí, con un pico muy leve en 1 (eso significaría que hay ciertos pdbs que su total de interacciones son de este tipo?).
* El boxplot tiene mucho más sentido. Hay varios outliers del lado del máximo (que deberían ser los que generan el pico cercano a 1 en el density plot) pero los cuartiles son más parejos, con el cuarto teniendo mucha más dispersión que los otros 3. Parecería que el segundo cuartil tiene mayoría de interacciones.

A graph with a line

Description automatically generated

A graph with a blue rectangle and a green rectangle

Description automatically generated

# Interacciones de Van der Waals

## Side chain

|  |  |
| --- | --- |
|  | **VDW:SC normal** |
| mean | 0.5923 |
| std | 0.2368 |
| min | 0.0000 |
| 25% | 0.5000 |
| 50% | 0.5955 |
| 75% | 0.7500 |
| max | 1.0000 |
| var | 0.0561 |
| std | 0.2368 |
| skew | -0.5632 |
| kurt | 0.3697 |

Este tipo de interacción también parecer tener distribución de datos entre todos los cuartiles. Hay mayor distribución que entre los puentes de hidrógeno con el side chain. Tiene un sd similar a las de las interacciones de puente de hidrógeno, por lo que tendría que haber una dispersión de datos similar a estas.

El skew negativo indica un que los datos están sesgados hacia la izquierda de la distribución. El kurtosis tan bajo indicaría que casi no hay datos en los extremos del rango.

### Plots

A blue line graph on a white background

Description automatically generated

Los datos se concentran en un pico máximo cerca de 0.5 junto con dos picos chicos en cercanos a 0 y a 1. No parece haber mucho spread pero no es una distribución normal en lo absoluto.

El boxplot definió varios outliers cerca del mínimo, que indicaría al pico cercano a 0. El primer cuartil parece tener mayor spread que el resto.

A graph with a blue rectangle and green line

Description automatically generated

## Main chain

|  |  |
| --- | --- |
|  | **VDW:MC normal** |
| mean | 0.0319 |
| std | 0.0537 |
| min | 0.0000 |
| 25% | 0.0000 |
| 50% | 0.0163 |
| 75% | 0.0396 |
| max | 0.6667 |
| var | 0.0029 |
| skew | 3.5972 |
| kurt | 19.4505 |

Parecería que casi no hay datos en el primer cuartil. El sd calculado es el más bajo de los 4 tipos de interacciones, lo que lleva a pensar que casi no casi todos los datos están cerca del mean.

El skew cercano a 3 indica que este tipo de interacciones es muy similar a una distribución normal. Sin embargo el kurtosis es el más alto de todos, lo que muestra que hay datos en los extremos del rango que prolongan las colas de la curva.

### Plots

A graph of a graph

Description automatically generated

* Se muestra una clara concentración en la región cercana al 0, con un semipico justo al lado. No parece haber datos en los extremos; no sé porqué el kurtosis dio tan alto.
* El boxplot es similar al de puentes de hidrógeno de side chain aunque esta vez además de una gran cantidad de outliers por el lado del máximo del rango, sí hay datos en los cuartiles 3 y 2.

A graph with lines and dots

Description automatically generated

# A graph with numbers and a bar Description automatically generatedHistogramas

A blue graph with white text

Description automatically generated

A graph with a long straight line

Description automatically generated

A blue graph with white text

Description automatically generated