LabJournal

Carlos Vigil Vásquez

Integrando información química, estructural y relacional para predecir interacciones proteína-ligando utilizando redes moleculares

Resumen

El proceso de identificación de nuevas interacciones droga-blanco es importante en el desarrollo y descubrimiento de fármacos. Determinación experimental de estas es costoso y laborioso, por ende el uso de métodos computacionales capaces de predecir este tipo de interacciones de manera precisa pasa a ser un paso crucial dentro del procedimiento de descubrimiento farmacológico. Métodos *in silico* generalmente optan por la minería de datos y/ aprendizaje de máquina para generar este tipo de predicciones, obviando la naturaleza relacional propia del proceso bioquímico implicado dentro de la interacción de moléculas pequeñas y proteínas. Presentamos un método para la prediccion de interacciones proteína-ligando basado en formalismo de redes. Este método usa una reimplementación del algoritmo *nearest neighbours* para generar predicciones sobre una red de multiples capas, la cual se construye a partir de información relacional entre blancos y ligandos y la similitud estructural para estos elementos. Esta red junto con sus propiedades matemáticas servirán de base para generar las predicciones, permitiendo mejorar la eficiencia y precisión de las predicciones obtenidas para interacciones proteína-ligando.

Introducción

Hipótesis

La integración de información bioquímica, estructural y relacional obtenida a partir de un modelo relacional de la información disponible por medio de formalismo de redes permitirá mejorará un método existente para la predicción de interacciones proteína-ligando.

Objetivos

Objetivo general

Crear método predictivo *in silico* basado en formalismo de redes capaz de generar predicciones proteína-ligando novedosas a partir de un set de datos obtenido de una base de datos de actividad farmacológica.

Objetivo específico

- Generar set de datos para interacciones proteína-ligando a partir de la base de datos ChEMBL 23.
- Construir red de interacciones bipartita proteína-ligando a partir del set de datos.

- Contruir red de similitud para los ligandos del set de datos por medio de la comparación estructural utilizando *molecular fingerprints* y calculo del coeficiente de Tanimoto.
- Contruir red de similitud para los blancos del set de datos por medio del calculo del porcentaje de identidad para cada par de blancos.
- Implementar algoritmo de *Nearest Neighbours* para generar predicciones.
- Validar método basado en algoritmo de *nearest neighbours* por medio de curvas ROC y PR.
- Mejorar el rendimiento del método predictivo por medio de la adición de información relacional y/o estructural.

Métodos

Selección de datos

El set de datos de interacciones proteína-ligando utilizado para crear la red de consulta se obtiene a partir de la ChEMBL, base de datos química manualmente curada de moleculas bioactivas con propiedades de droga, empleando los siguientes filtros:

- Especie debe ser igual a *Homo Sapiens*.
- Ensayo debe ser de unión proteína-ligando.
- IC50 en un rango de 0 a 10.000 nM.
- Molécula debe tener como máximo 80 átomos pesados.
- Comentario de actividad debe denotar que el ligando es activo para dicho blanco anotado.
- Valor de actividad debe ser "menor", "mucho menor", "menor o igual", "igual" o "igual igual" (representado como "<", "≪", "<=", "=" y "==", respectivamente) al valor obtenido experimentalmente.
- Ligando debe tener una fase clínica máxima igual o mayor a 1.

A este set de datos lo llamaremos *Candidatos Clínicos & Drogas* ("CC&D"), nombre que se empleará por el resto del informe.

Construcción del *espacio de interacción proteína-ligando.*

A partir del set de datos *CC&D* se procede a construir una red conformada por 2 capas de información, donde los nodos y uniones corresponden a diferentes elementos considerados dentro de la interacción proteína-ligando. Esta red de dos capas consiste de 3 partes diferentes (Figura):

- El *espacio de ligandos*, construido a partir de los nodos correspondiente a los ligandos del set de datos y unidos entre sí por su similitud estructural calculada a partir de la comparación de un descriptor molecular para cada par de nodos.
- El *espacio de blancos*, construido a partir de los nodos correspondientes a los blancos del set de datos y unidos entre sí a por su similitud <++> calculada a partir de la identidad de secuencia para cada par de nodos.

- El *espacio de interacción proteína-ligando*, construido a partir de todos los nodos considerados dentro del set de datos, unidos entre sí por anotaciones de interacción proteína-ligando. Este espacio considera a los otros dos mencionados anteriormente, por lo que en este nivel se consideran las uniones de similitud entre nodos también.

Esquema de red bicapa

Para el calculo de los valores de similitud para cada elemento de la red (ligandos y proteínas) se realizó el siguiente procedimiento:

- Para los ligandos; (i) se seleccionan 2 ligandos del set de datos, (ii) luego se calculan los descriptores moleculares Obabel FP2 a partir de una reimplementación *in-house* del algoritmo y, finalmente, (iii) se calcula el coeficiente de Tanimoto para el par de descriptores estructurales. El resultado es una matriz de similitud de tamaño $N_{Ligandos}\ X\ N_{Ligandos}$.
- Para los blancos; (i) se seleccionan 2 blancos del set de datos, (ii) luego se alinean ambas secuencias aminoacídicas utilizando el algoritmo de Needleman-Wunsch utilizando la matrix BLOSUM62 y, finalmente, (iii) se calcula el porcentaje de identidad para el par de blancos. El resultado es una matriz de similitud de tamaño $N_{Blancos}\ X\ N_{Blancos}$.

Método predictivo

Con la finalidad de cuantificar el posible rendimiento de un modelo predictivo basado en redes, se procede a reimplementar el método de *nearest neighbour* de A. Schuelles usando la red bicapa descrita anteriormente para generar predicciones del tipo proteína-ligando utilizando la siguiente metodología:

- Se selecciona un par proteína-ligando de la red.
- Se reduce la red bicapa a un subgrafo conpuesto de el ligando de origen, el blando protéico y los ligandos vecinos del ligando de origen.
- A partir de dicho subgrafo, se procede a extraer la unión entre el ligando de origen y su ligando vecino con el peso más alto (en este caso, aquella unión con el valor de Tanimoto máximo), valor utilizado como métrica de puntuación de la predicción.

Este algoritmo se repite para cada par proteína-ligando posible dentro de la red bicapa, obteniendo un total de predicciones de $N_{Ligandos}\ X\ N_{Blancos}$. Para validar el modelo propuesto, se realiza una validación cruzada de 10 iteraciones para el set de datos, graficando curvas ROC y *Precision-Recall*.

Resultados y discusión

Selección de datos

En la Tabla 1 se declara la cantidad de ligandos, proteínas e interacciones proteína-ligando obtenidas para cada paso de filtrado aplicado.

Tabla 1: Preparación de set de datos *Candidatos Clínicos & Drogas*. El signo de suma ("+") denota que el filtro se aplica sobre el set de datos obtenido con el/los filtro anterior(es).

Filtro aplicado	Ligandos	Blancos	Interacciones proteína-ligando
Homo Sapiens	<++>	<++>	<++>
+ Ensayo de unión proteína-ligando	<++>	<++>	<++>
+ $0 \leqslant IC50 \leqslant 10000$	<++>	<++>	<++>
+ Átomos pesados $\leqslant 80$	<++>	<++>	<++>
+ Comentario de actividad indica que es "activo"	<++>	<++>	<++>
+ Valor de actividad es "<", "$\ll$", "$\leqslant$", "=" o "=="	<++>	<++>	<++>
+ Fase clínica máxima $\geqslant 1$	1.232	897	9.641

Construcción del *espacio de interacción proteína-ligando.*

Método predictivo

Conclusiones

Proyecciones / Ideas

- Incorporación de información estructural al método predictivo introducirá "atajos" a la hora de generar la prediccion vía el algoritmo *Nearest-Neighbour*.

Descriptor estructural a partir de resultados de *docking molecular* masivo del set de datos BioLip

Idea

A partir de los datos obtenidos de la simulación de *docking molecular* masivo de la base de datos obtenida de BioLip, se propone generar un vector estructural para los ligandos y blancos representados en este set de datos.

Este descriptor se podra emplear para predecir interacciones proteína-ligando utilizando los métodos conocidos dentro del laboratorio y para validar el experimento de *docking molecular* realizado a finales de año del año 2019.

Objetivos

Objetivo general

Generar descriptor estructural basado en resultados de *docking molecular* que permita comparar tanto ligandos como blancos del set de datos de BioLip.

Objetivos especificos

- Convertir archivos `bestranking.lst` obtenidos de la simulación de *docking* en una matriz de valores de tamaño $N_{Ligandos}$ X $N_{Blancos}$.
- Limpiar matriz de valores obtenida de todos los blancos y ligandos que presentan valores anomalos a partir de un criterio de selección basado en la puntuación obtenida del *docking*.
- Generar "perfil de *docking* molecular" para todos los ligandos utilizando las puntuaciones obtenidas en la simulación de *docking* molecular masivo.
- Generar matriz de similitud todos contra todos de los ligandos utilizando el descriptor creado.

Métodos

Selección de datos

Simulación de *docking* molecular

Creación de perfil de *docking*

Para cada blanco considerado dentro de la simulación anterior se obtiene un archivo de salida, el cual nos entrega una serie de valores asociados a la simulación realizada. De aquí se extraen 2 columnas: (i) *ChemPLP score* y (ii) identificador del ligando *dockeado*. Con esta información se procede a armar una matriz donde cada columna corresponde a un blanco, cada fila corresponde a un ligando y cada valor para el par ligando-blanco corresponde al *ChemPLP docking* score. Este procedimiento se repite con todos los blancos considerados dentro del set de datos, para así construir finalmente una matriz de tamaño $N_{Ligandos}$ x $N_{Blancos}$

Posteriormente, se procede a filtrar de la matriz todos aquellos ligandos que no pudieron dockearse en un 10 % de los blancos, es decir, que no presentan un valor de *ChemPLP docking score* para el 10 % de los blancos.

- Se filtran los blancos que no poseen su ligando cocristalizado (filtrado en el paso anterior).
- Se reduce la redundancia de los ligandos del set de datos, dejando solo 1 valor de *ChemPLP docking score* para cada par ligando-blanco.

Resultados

Selección de datos

Simulación de docking molecular

Creación del perfil de *docking*