

## El uso del Algoritmo PageRank en Redes Metabólicas

Amelia Sánchez Reyes

### Introducción

Mi trabajo parte del artículo "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks" de Gábor Iván y Vince Grolmusz, el cual explora el uso del algoritmo PageRank para entender Redes Metabólicas y Redes de Interacción Proteína-Proteína.

El análisis de grandes redes de interacción de proteínas (PPI) y redes metabólicas es esencial debido a la creciente cantidad de datos biológicos disponibles. El problema de identificar los nodos más importantes en una gran red es común en varias disciplinas, incluida la biología. PageRank, un algoritmo de clasificación utilizado en motores de búsqueda web como Google, se destaca por su capacidad para identificar nodos importantes en una red basada en enlaces.

El objetivo de este trabajo es estudiar y entender el Algoritmo de PageRank. Así como, visualizar cómo puede ser útil en el análisis de redes metabólicas o y redes de interacción proteína-proteína.

### Métodos

Se hizo una investigación que explica el funcionamiento del algoritmo de PageRank y una Red Metabólica. De igual forma, se evidencian las similitudes que existen entre una Red Metabólica y la World Wide Web (WWW). De esta manera se puede ver como el algoritmo que dio éxito a Google puede ser aplicado perfectamente a redes biológicas.

A continuación, se enuncian las principales características:

	<b>PageRank</b>	<b>WWW</b>	<b>Red Metabólica</b>
<b>Definición</b>	Algoritmo desarrollado por los creadores de Google para clasificar web en los resultados de búsqueda.	Red informática mundial que funciona a través de internet.	Representación estructural y funcional de las interacciones bioquímicas que ocurren dentro de una célula.
<b>Características</b>	Se basa en la idea de que un nodo es importante si muchos otros nodos lo enlazan. No todos los enlaces tienen el mismo valor; un enlace desde un nodo con un alto PageRank cuenta más que un enlace desde un nodo con un bajo PageRank.	En la WWW los nodos son las páginas web y los enlaces son las urls que tienen dentro de su contenido, de modo que enlazan.	En una red metabólica los nodos son los Metabolitos (pequeñas moléculas involucradas en las reacciones metabólicas como la glucosa); los enlaces son las reacciones químicas, esto es, las transformaciones químicas que convierten un metabolito en otro. Estas reacciones son catalizadas por enzimas.

<b>Propiedades</b>	<p><b>Distribución estacionaria:</b> PageRank se basa en la distribución estacionaria de una cadena de Markov.</p> <p><b>Robustez:</b> Debido a que los enlaces desde nodos con altos PageRank tienen más peso, no es fácil manipular el PageRank simplemente añadiendo muchos enlaces de baja calidad.</p> <p><b>Escalabilidad:</b> PageRank es escalable y puede calcularse eficientemente incluso para la enorme escala de la web.</p>		<p><b>Conectividad:</b> Descubre cómo los metabolitos están interconectados a través de reacciones bioquímicas. Algunas reacciones pueden tener una alta conectividad si participan en muchas reacciones diferentes.</p> <p><b>Modularidad:</b> Las redes metabólicas pueden estar organizadas en módulos o subredes que presentan rutas metabólicas específicas (Ejemplo: Ciclo de Krebs, Vía de las Pentosas Fosfato). Esto permite una organización funcional donde diferentes rutas pueden ser reguladas de manera independiente.</p> <p><b>Robustez:</b> La capacidad de la red para mantener su funcionamiento general a pesar de perturbaciones o fallos en componentes individuales.</p>
--------------------	---	--	--

### *Explicación del algoritmo*

El algoritmo de PageRank puede imaginarse como un "caminante" que navega por una red (grafo), siguiendo enlaces al azar. Este caminante sigue un enlace en un nodo con cierta probabilidad (por ejemplo, el 85% del tiempo) y, con una probabilidad restante (por ejemplo, el 15% del tiempo), "teletransporta" a un nodo seleccionado aleatoriamente. La probabilidad de que el caminante esté en un nodo determinado en un momento dado se convierte en el PageRank de ese nodo.

La formula para actualizar el PageRank de un nodo es la siguiente:

$$PR(P) = \frac{1 - d}{N} + d \sum_{i=1}^k \frac{PR(P_i)}{L(P_i)}$$

donde

- PR(P) es el PageRank de la página P.
- d es el factor de amortiguación (por lo general, 0.85)
- N es el número total de páginas en la web
- Pi son las páginas que enlazan a P.
- L(Pi) es el número de enlaces salientes de la página Pi.

La fórmula se aplica repetidamente para cada página hasta que los valores de PageRank convergen, es decir, cambian muy poco entre iteraciones sucesivas.

### *Resultados: Análisis de Redes Metabólicas con Page Rank*

Utilizando algoritmos como PageRank, se pueden identificar metabolitos o reacciones que son particularmente importantes para la red, similar a cómo se identifican páginas web importantes en internet.

El artículo "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks" de Gábor Iván y Vince Grolmusz explora el uso del algoritmo PageRank, para analizar redes de interacción de proteínas y redes metabólicas en biología:

#### **Red de Interacción Proteína-Proteína (PPI)**

- **Representación usual:** Las redes PPI suelen representarse con grafos no dirigidos.
- **Limitación de PageRank:** Para estos grafos, el PageRank es proporcional al grado de los nodos (el número de conexiones), lo que no ayuda mucho en diferenciar la importancia de los nodos más allá de contar las conexiones.

#### **Redes Metabólicas**

- **Grafos dirigidos:** Las redes metabólicas son grafos dirigidos, donde los nodos representan reacciones bioquímicas y las aristas dirigidas conectan nodos si una reacción usa el producto de otra.
- **Ventaja de PageRank:** En este contexto, el PageRank puede revelar propiedades profundas y robustas de la red, más allá de simplemente contar las conexiones.

#### **PageRank personalizado para redes PPI**

##### *1. PageRank Personalizado*

- **Desarrollo original:** Desarrollado para predecir preferencias personales en la valoración de contenido en la Web.
- **Funcionamiento:** El caminante aleatorio tiene una probabilidad de  $c + c'$  de teletransportarse;  $c'$  corresponde a intereses personales y  $c$  al resto de los nodos.
- **Aplicación en redes:** Este método puede evaluar la importancia de los nodos en relación a otros nodos relevantes conocidos.
- 

##### *2. Aplicabilidad en PPI*

- **Robustez y Escalabilidad:** El cálculo de PageRank personalizado es robusto y escalable, útil para redes grandes.
- **Análisis de datos de melanoma:** Se utilizó para analizar datos de proteómica en pacientes con melanoma, personalizando el PageRank a 13 proteínas detectadas con niveles elevados en plasma.
- **Resultados:** La red PPI humana contiene 27,801 nodos y 38,806 aristas. Muchos de los nodos con mayor PageRank están claramente relacionados con el melanoma.

- **Selectividad del método:** De 27,800 nodos, sólo 0.57% están relacionados con melanoma según la base de datos UniProt, lo que demuestra la selectividad y poder del PageRank personalizado.

### *Discusión*

Lo sorprendente de estos resultados es la robustez que tiene PageRank para definir cuales son los nodos más importantes, y como esto ayuda a evitar los falsos positivos o falsos negativos. Tener esta claridad ayuda a resolver problemas como la identificación de proteínas relacionadas con el cáncer, así como, identificación de nodos importantes en redes metabólicas de seres vivos (como la bacteria que produce la Tuberculosis).

### *Conclusión*

- La integración de la biología y la informática, utilizando algoritmos como PageRank, permite un análisis más robusto y significativo de grandes conjuntos de datos biológicos.
- La robustez del PageRank es especialmente valiosa para manejar errores en los datos de interacción de proteínas.
- La personalización de PageRank mejora la capacidad de identificar nodos de interés en redes complejas, lo que puede ser útil para estudios proteómicos y la investigación de enfermedades como el cáncer.