Descubrimiento de motivos conservados en proteínas mediante métodos de Monte-Carlo y cadenas de Markov

Sánchez Vázquez Raúl, Ponce de León Sentí Eunice Esther,

Martin Álvarez Tostado Eduardo Mauricio

Noviembre 2020

Introducción

El presente proyecto pretende generar un método para la búsqueda de motivos conservados en familias de proteínas, con lo cual será posible obtener las características que prevalecen en las entidades biológicas, tales como los virus y así poder entender su mecanismo de funcionamiento.

Para tal propósito se hará uso de herramientas computacionales y estadísticas.

- Del lado computacional se trabajará con el algoritmo de colonia de hormigas el cual permite el estudio de grandes conjuntos de datos cuya eficiencia en tiempo computacional es mayor que la de métodos exhaustivos. (Dorigo y col., 2006)
- Por otro lado para la modelación de las características derivadas se hará uso de las cadenas de Markov como método de inferencia. (Ortega, 2010)

Objetivo del proyecto

Diseñar e implementar un método basado en métodos de Monte-Carlo y cadenas de Markov para el problema de descubrimiento de motivos conservados en secuencias de proteínas.

Materiales y métodos

Conceptos

- Los aminoácidos son un conjunto de 20 tipos distintos de moléculas y constituyen las piezas básicas para construir proteínas. Las proteínas constan de una o más cadenas de aminoácidos; estas cadenas se llaman polipéptidos. La secuencia de la cadena de aminoácidos determinará cómo se pliega tridimensionalmente el polipéptido, pues la forma que adquiera es muy importante para que sea biológicamente activo.(Brody, s.f.)
- Un aspecto importante de la caracterización de secuencias biológicas son los motivos. Un motivo es un elemento conservado en la secuencia de aminoácidos o nucleótidos, que habitualmente se asocia con una función concreta. Los motivos se generan a partir de alineamientos múltiples de secuencias con elementos funcionales o estructurales conocidos, por lo que son útiles para predecir la existencia de esos mismos elementos en otras proteínas de función y estructura desconocida funciones de proteínas desconocidas.(Fernández, 2013)
- Una Cadena de Markov a tiempo discreto es una sucesión de variables aleatorias $X_n, n \geq 1$ que toman valores en un conjunto finito o numerable ϵ, conocido como espacio de estados, y que

satisface la siguiente propiedad:

$$P(X_{x+1} = j | X_0 = i_0, ..., X_{n-i} = i_{n-1}, X_n = i_n) = P(X_{x+1} = j | X_n = i_n) \quad (1)$$

para todo n y cualesquiera estados $i_0, i_1, ..., i_n, j$ en ϵ. La propiedad (1) se conoce como propiedad de Markov.(Ortega, 2010)

Con base en la definición se puede entender una cadena de Markov como un autómata donde la transición de un estado a otro depende de una probabilidad, conocida como probabilidad de transición.

Con base en lo anterior es posible representar las secuencias de aminoácidos mediante cadenas de Markov, donde cada estado refiere a un aminoácido y la transición implica la probabilidad de sucesión de un aminoácido a otro.

A groso modo se infiere que es posible generar una secuencia de aminoácidos mediante una cadena de Markov, siempre y cuando se conozca la matriz de probabilidades de transición, la cual se obtiene mediante un análisis exploratorio de los datos.

Método propuesto

1. Realizar el alineamiento mediante el software clustal Ω (Mallawaarachchi, 2017)
2. Análisis descriptivo del alineamiento: Mediante técnicas de análisis exploratorio de datos se obtienen los principales estadísticos del alineamiento buscando la consistencia del conjunto respecto a cada secuencia.
3. Obtener la probabilidad de transición entre aminoácidos:
   - *a*) Realizar el muestro de las secuencia mediante un muestreo aleatorio simple.
   - *b*) Iterará las secuencias del muestreo para obtener el valor de transición.
   - *c*) Generar matriz de transición considerando la propiedad de Markov.
4. Realizar la búsqueda de motivos mediante algoritmo de inteligencia colectiva.
   - Para tal efecto se utilizará el algoritmo de colonia de hormigas, cuya entrada será la matriz de transición, la cual será abstraída como un grafo donde cada nodo representa un aminoácido y la arista entre ellos será la probabilidad de transición. (Dorigo y col., 2006)
   - Para evaluar la solución se empleará la probabilidad condicional mediante la regla de bayes, dicha funciones será aplicada sobre la secuencia de aminoácidos obtenida de la caminata aleatoria de las hormigas artificiales.
   - La salida del algoritmo será una cadena de aminoácidos (posible motivo) de tamaño n
5. Introducir la cadena en el algoritmo de Markov y obtener su probabilidad de ocurrencia.

La implementación del método propuesto fue una conjunción computacional de los lenguajes de programación Julia y Python.

- Lectura de datos: Una vez que se realizó el alineamiento de la secuencia (mediante clustalΩ), dicha secuencia se carga en el programa convirtiéndola (Kronopt, 2020) a un dataframe (Leong, 2019).
- Análisis de frecuencias: Se estudia la distribución de cada secuencia y se compara de forma gráfica.
- Probabilidad de transición: Para la obtención de la matriz de transición se realizó un muestreo aleatorio simple, para iterar y encontrar la probabilidad transición desde un aminoácido hasta el siguiente.

- Búsqueda de motivo. Para encontrar un posible motivo dentro del espacio de solución (secuencias de aminoácidos) se implementó el algoritmo de colonia de hormigas mediante el lenguaje Julia, donde para la entrada se utilizo la matriz de transición generando así un grafo completo pesado y dirigido (donde los nodos representan los aminoácidos y las aristas la probabilidad de transición), el cual sería recorrido por las hormigas artificiales buscando una subsecuencia (motivo) basado en el criterio (función objetivo) de que la probabilidad conjunta del recorrida fuera mayor o igual aun umbral dado.
- Cadena de Markov: La caminata aleatoria de las hormigas artificiales solo arrojo una cadena con un valor de probabilidad, por lo cual se implementó un modelo de cadena de Markov el cual es alimentado por la matriz de transición y la subcadena (motivo) arrojada por el algoritmo de colonia de hormigas.(Ortega, 2010)
- Búsqueda posicional: El motivo encontrado se convirtió en una expresión regular de tal forma que se pudieran encontrar las posiciones (columnas del dataframe) de los componentes del motivo, mostrando así la estructura y comportamiento del motivo.

Resultados

Una vez realizada la implementación se procedió a probar el método con diversos conjuntos de datos pudiendo distinguir entre cuatro resultados relevantes

- Gráfico de frecuencias de la familia: Define la distribución que posee cada secuencia en la familia pudiendo distinguir la consistencia de los datos dentro de una familia al visualizar una tendencia
- Aminoácidos altamente conservados: Es un gráfico que resalta la conservación de los aminoácidos de tal forma que aquellos que muestran una frecuencia alta son candidatos para formar parte del motivo conservado.
- Solución de colonia de hormigas: Dicho algoritmo proporciona como salida un *objeto* compuesto por la subsecuencia de aminoácidos y su probabilidad condicional.
- Solución de cadena de Markov: Dicho algoritmo proporciona como salida la probabilidad de que la subsecuencia provista sea un motivo conservado.

Es importante notar que la probabilidad condicional arrojada por la colonia de hormigas y la probabilidad arrojada por la cadena de Markov son el parámetro que define el resultado final, pues al compararlos debe existir una aproximación para decir que la subsecuencia tiene n probabilidad de ser motivo. Más aún, si la subcadena es mayor o igual al umbral definido se puede decir o no si se trata de un motivo conservado.

En la Figura 1 se muestra un resultado de la ejecución del algoritmo donde es posible ver la semejanza de las probabilidades, así como el hecho de que están por encima del umbral.

Conclusiones

Como se planteó en un inicio el objetivo del proyecto era diseñar un método para la búsqueda de motivos conservados en proteínas, lo cual fue posible gracias a la integración de las cadenas de Markov y la optimización por colonia de hormigas, logrando así la generación de una herramienta de software capaz de encontrar motivos teniendo en cuenta una probabilidad de que estos lo sean, posibilitando así el estudio de las estructuras proteicas y sus propiedades estadísticas.

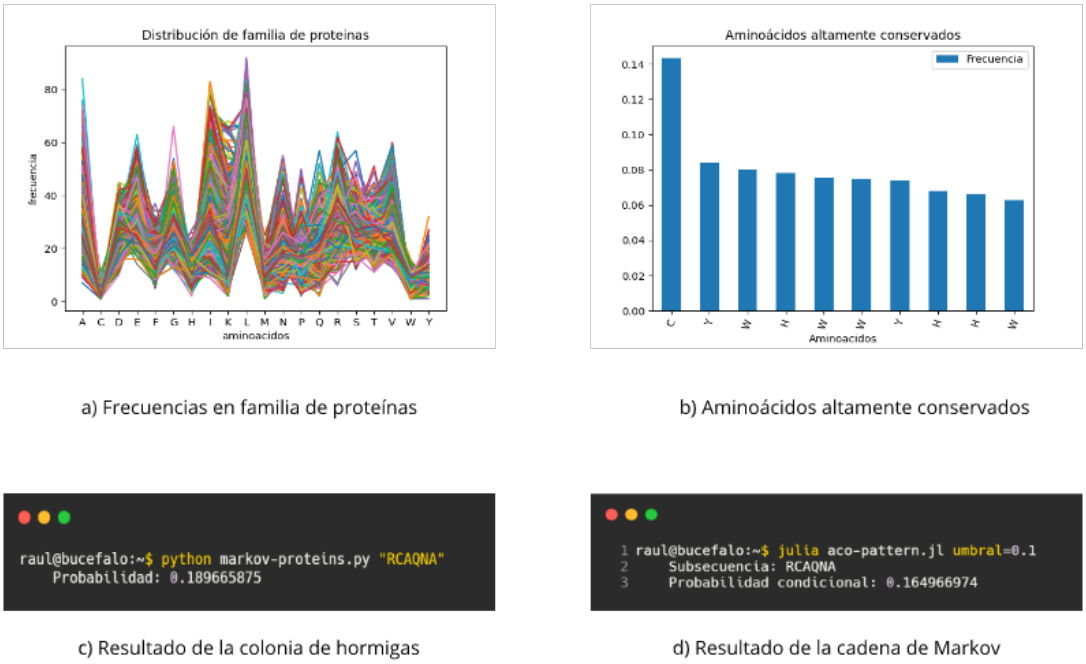

Figura 1: Resultados del algoritmo

Bibliografía

Dorigo, M., Birattari, M. & Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*, *1*(4), 28-39. https://doi.org/10.1109/MCI.2006.329691

Ortega, J. (2010). *Modelos Estocasticos I*. CIMAT.

Fernández, J. (2013). *Motivos y Dominios*. https://www.tamps.cinvestav.mx/~ertello/bioinfo/sesion09.pdf (accessed: 04.11.2020)

Mallawaarachchi, V. (2017). *Multiple Sequence Alignment using Clustal Omega and T-Coffee*. https://towardsdatascience.com/multiple-sequence-alignment-using-clustal-omega-and-t-coffee-3cc662b1ea82 (accessed: 12.11.2020)

Leong, N. (2019). *Python for Data Science- A Guide to Pandas*. https://towardsdatascience.com/python-for-data-science-basics-of-pandas-5f8d9680617e (accessed: 05.11.2020)

Kronopt. (2020). *FastaParser*. https://fastaparser.readthedocs.io/en/latest/history/ (accessed: 05.11.2020)

Brody, L. C. (s.f.). *Aminoácido*. https://www.genome.gov/es/genetics-glossary/Aminoacido (accessed: 04.11.2020)