MA4402 Simulación Estocástica: Teoría y Laboratorio.
Profesor: Joaquin Fontbona T.
Auxiliares: Bruno Hernández P. y Pablo Araya Z.
Integrantes: Felipe Hernández y Sebastian Bustos

MCMC en alineación de segmentos de ADN

El proyecto se vincula al paper "DNA motif alignment by evolving a population of Markov Chains" de Chengpeng Bi ([1]). Se desea estudiar la implementación de algoritmos estocásticos para encontrar secuencias de ADN recurrentes, llamadas motifs (Figura 1), las cuales se asocian a alguna función biológica importante en el proceso de transcripción del ADN.

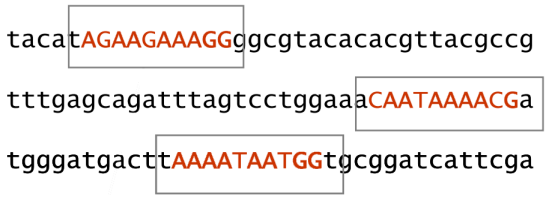

Figura 1: Ejemplo de motifs

Los motifs pueden ser representados mediantes distribuciones de probabilidad, estas distribuciones se modelan a través de la normalización por columnas de las matrices de frecuencias, las cuales cuentan las veces que está presente el nucleótido en toda la secuencias que se desea estudiar (Figura 2).

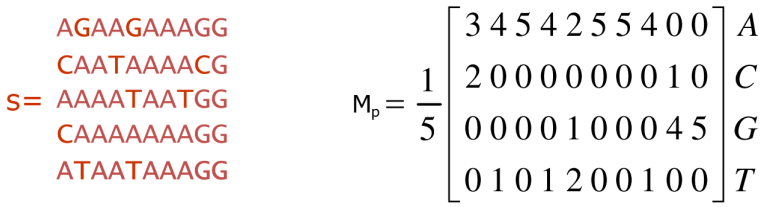

Figura 2: Ejemplo de matrices de probabilidad

Haciendo el supuesto que los motifs distribuyen de manera distinta a las zonas que no contienen motifs es posible aplicar Metropolis Hasting Sampler (MHS) para buscar zonas de alta probabilidad que esté asociadas a los motifs.

En la presentación se trabajarán dos tipos de algoritmos de MCMC. El primero, llamado IMC, consiste en aplicar Metropolis Hasting repetidas veces de manera secuencial e independiente. El segundo, llamado PMC, consiste en aplicar el algoritmo Metropolis Hasting de manera paralela permitiendo intercambio de información entre las cadenas. Se presentarán resultados sobre tres conjuntos de secuencias obtenidas del dataset JASPAR ([2]) para ambos métodos. Por un lado, se compararán resultados con el paper y por otro se estudiará el efecto de utilizar ADN de distintos tipos de organismos (insectos y plantas).

Referencias

1. Bi, C. DNA motif alignment by evolving a population of Markov chains. BMC Bioinformatics 10, S13 (2009). https://doi.org/10.1186/1471-2105-10-S1-S13
2. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, and Mathelier A JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles Nucleic Acids Res.