Instituto Politécnico Nacional

Escuela Superior de Cómputo

Bioinformatics

Lab Session 1. Applications of Bioinformatics

González Bocio Erik Alexander 2020630163

Jorge Luis Rosas Trigueros

Fecha de realización: 14 de febrero de 2022

Fecha de entrega: 21 de febrero de 2022

Marco teórico.

La Bioinformática es una subdisciplina de la biología y las ciencias computacionales que se encarga de adquirir, almacenar, analizar y diseminar la información biológica, en gran parte correspondiente a las secuencias de ADN y aminoácidos. La Bioinformática usa programas informáticos que tienen muchas aplicaciones, como, por ejemplo: determinar las funciones de genes y proteínas, establecer relaciones evolutivas y predecir la conformación tridimensional de las proteínas.

La bioinformática, en sentido amplio, se podría definir como la disciplina científica que utiliza la tecnología de la información para organizar, analizar y distribuir información biológica, con la finalidad de responder preguntas complejas en biología10, es decir, una disciplina que engloba métodos matemáticos, estadísticos y computacionales para solucionar problemas biológicos usando ADN, ARN, secuencias de aminoácidos e información relacionada11.

Las principales aplicaciones de la bioinformática son la gestión, la simulación, la minería de datos y el análisis de la información generada en el PGH, con aplicación también en la predicción de estructuras proteicas, estudios de secuencias y otras actividades derivadas de la investigación en biología.

En esta primera práctica hablaremos de las aplicaciones desarrolladas en 3 países diferentes.

Material y equipo.

- Navegador Web
- Computadora de escritorio/laptop
- Word o herramienta de edición de texto

Desarrollo de la práctica.

According to the last digit of your Student ID, search for applications of Bioinformatics developed in:

- Alemania

  Investigadores bioinformáticos del Instituto Leibniz de Genética Vegetal e Investigación de Plantas de Cultivo (IPK), de la Universidad Martin Luther Halle-Wittenberg (MLU) y del Instituto Leibniz de Bioquímica Vegetal (IPB) desarrollaron “Kmasker plants” que permite la identificación de secuencias repetitivas y facilita el análisis de los genomas de las plantas.

  En bioinformática, el término k-mero se utiliza para describir una secuencia de nucleótidos de una cierta longitud “k”. Al definir y contar dichas secuencias, los investigadores pueden cuantificar secuencias repetitivas en el genoma que están estudiando y asignarlas a las posiciones correspondientes.

  El uso de NGS es cada vez más importante, pero la composición libre de errores de genomas complejos a partir de los resultados de NGS sigue siendo un desafío. Bajo la dirección del Dr. Thomas Schmutzer, científicos de la MLU, el IPK, la Universidad de Wageningen & Research y el IPB Halle trabajaron en estrecha cooperación en el rediseño y desarrollo de “Kmasker plants”. Esta colaboración fue apoyada en gran medida por los dos centros de servicio "GCBN" y "CiBi" de la Red Alemana de Infraestructura Bioinformática "de.NBI".

  Permiten el filtrado rápido y libre de referencias de secuencias de nucleótidos utilizando k-meros derivados de genomas. También permite estudios comparativos entre diferentes cultivares o especies estrechamente relacionadas, y apoya la identificación de secuencias adecuadas como sondas fluorescentes para hibridación in situ (FISH) o ARN guía específicos de CRISPR/Cas9. Además, “Kmasker plants” se ha publicado con un servicio web que contiene los índices precalculados para plantas de cultivo de importancia económica, como la cebada o el trigo.

  El servicio web “Kmasker plants” está ahora disponible como parte de IPK Crop Analysis Tool Suite (CATS) y por lo tanto, como un servicio de la de. Plataforma de servicio NBI.

Alternativamente, se puede acceder directamente e instalar el código fuente del “Kmasker plants” a través de GitHub.

- Israel

  ENViz (Enrichment Analysis and Visualization) es una aplicación de Cytoscape que realiza un análisis de enriquecimiento conjunto de dos tipos de conjuntos de datos coincidentes de muestra en el contexto de anotaciones sistemáticas. Dichos conjuntos de datos pueden ser de expresión génica o cualquier otro dato de alto rendimiento recopilado en el mismo conjunto de muestras. El análisis de enriquecimiento se realiza en el contexto de la información de la vía, la ontología génica o cualquier anotación personalizada de los datos. Los resultados del análisis consisten en asociaciones significativas entre los elementos perfilados de uno de los conjuntos de datos con los términos de anotación (por ejemplo, miR-19 se asoció al proceso del ciclo celular en muestras de cáncer de mama). Los resultados del análisis de enriquecimiento se visualizan como una red interactiva de Cytoscape.

  El enfoque de ENViz para el análisis integrado de datos utiliza el poder de las estadísticas de enriquecimiento combinadas con bases de datos de anotaciones genómicas para asignar estadísticamente anotaciones de funciones relevantes a elementos perfilados explorados. Por lo tanto, proporciona una mejor subestimación de la relación entre diferentes entidades moleculares en células u organismos.

  A pesar de que el desarrollo de ENViz fue motivado por las mediciones biológicas modernas disponibles, el análisis conjunto de dos conjuntos de datos coincidentes de muestra y las anotaciones sistemáticas pueden aplicarse a otros estructurados de manera similar.

  La dirección la llevo MeMed Diagnostics Ltd. una empresa de diagnóstico personalizado, centrada en la prevención del uso indebido de antibióticos de Haifa, Israel.

  Sigue un enfoque de análisis de enriquecimiento, impulsado por tres matrices de entrada: (i) la matriz de datos primarios (por ejemplo, la medición de la expresión de genes en un conjunto de muestras), (ii) la matriz de anotación que proporciona anotación binaria en cada uno de los elementos de la matriz de datos primarios [por ejemplo, anotación de la vía u ontología génica (GO)] y (iii) la matriz de datos pivote que proporciona información sobre el mismo conjunto de muestras de la matriz de datos primarios (por ejemplo, medición de la expresión de miRNAs). Para cada elemento de datos dinámico, ENViz realiza estos pasos:

  - Calcular la correlación con cada elemento de los datos primarios.
  - Clasificar los elementos de datos primarios en función de las correlaciones anteriores.
  - Calcular el enriquecimiento estadístico de los elementos anotados (por ejemplo, vías) en la parte superior de la lista clasificada anteriormente en función de una estadística hipergeométrica mínima (mHG).

  Brevemente, dado un vector de anotación binario clasificado, calculamos el enriquecimiento de esta anotación en los k elementos clasificados superiores en función de la estadística hipergeométrica, donde se selecciona k para optimizar este enriquecimiento. Finalmente, se informa de la puntuación mHG [−log(mHG P-value)] para la asociación pivot-anotación. El nivel de significancia calculado es válido para cada par de anotaciones de pivote individuales, pero no se corrige para el número de pares probados.

  Los resultados significativos se representan en Cytoscape como una red de enriquecimiento: un gráfico bipartito con nodos correspondientes a elementos pivote y de anotación, y aristas

correspondientes a asociaciones significativas de anotación de pivote, donde el umbral de significación está definido por el usuario.

- México

  Desde 2002 el CICESE cuenta con un grupo pequeño en bioinformática y biocomputación que lidera el doctor Carlos Alberto Brizuela Rodríguez, investigador del Departamento de Ciencias de la Computación, y en el que participan también el doctor Israel Martínez, especialista en computación biomolecular, así como cuatro estudiantes de doctorado y cuatro de maestría.
  El grupo se compone de ingenieros electrónicos, computólogos y matemáticos, y sus líneas de trabajo abarcan el desarrollo de nuevas herramientas computacionales que permiten profundizar en el conocimiento biológico, particularmente la predicción de las estructuras de las proteínas.

  Actualmente se conocen cerca de 8 millones de secuencias de proteínas. Sin embargo, hay menos de 100 mil estructuras resueltas. La diferencia que hay entre la cantidad de secuencias conocidas y estructuras resueltas es gigantesca, y el resolver estructuras experimentalmente es muy costoso en tiempo y dinero lo que hace que la tasa de resolución de proteínas al año no llegue a 10 mil estructuras.
  La forma en la que abordan su investigación es hacer un mapeo (mapping) de las cadenas de aminoácidos que conforman alguna proteína, para luego preguntarse cómo será la estructura tridimensional específica que adoptará dicha proteína. Ello es un problema biológico que no ha tenido solución en 40 años. ¿Y esto por qué es importante? Porque la función que realiza la proteína está estrechamente relacionada con su forma. Si los investigadores quieren explicar cómo se realizan las interacciones entre proteínas diferentes, o entre una proteína y un medicamento, o entre proteína y ADN, entonces es preciso conocer la estructura tridimensional. De esta forma, al conocer mecanismos de interacción que sean específicos de la bacteria se pueden diseñar fármacos que los ataquen; por ello será de enorme valor conocer las estructuras de las proteínas involucradas en cada proceso.

  Dos tesis de maestría realizadas por David Omar Rodríguez e Ivetth Corona, y una tercera por Cristian Lezcano, todas ellas dirigidas por el doctor Brizuela Rodríguez, demostraron que se puede llegar a una precisión de hasta 97 por ciento en predecir estructuras de proteínas con sistemas informáticos mediante el método conocido como empaquetamiento de la cadena lateral en proteínas.

  La precisión de predicción de este método era de hasta 85 por ciento, pero las investigaciones en el CICESE han permitido establecer que la función de energía de la proteína es la clave en el proceso de predicción del empaquetamiento, por lo que actualmente tratan de abordar este aspecto para casos específicos.

Diagramas, gráficas y pantallas.

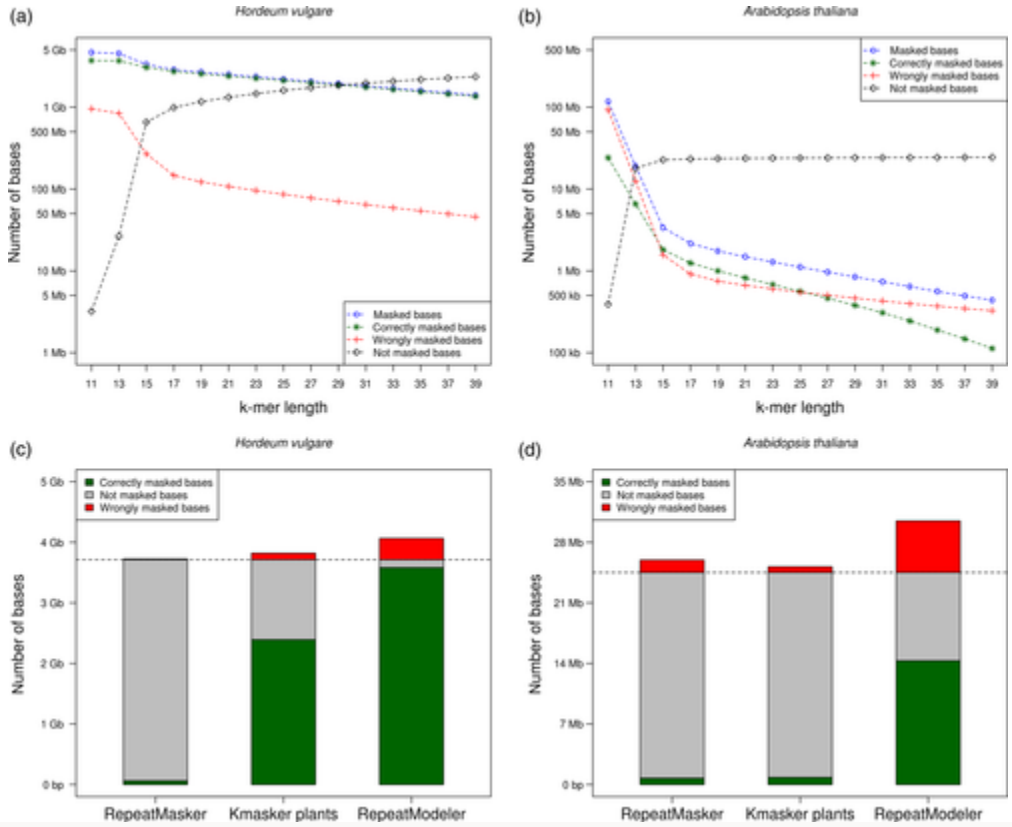

Fig 1. Evaluación de las capacidades de enmascaramiento repetido de las plantas Kmasker.

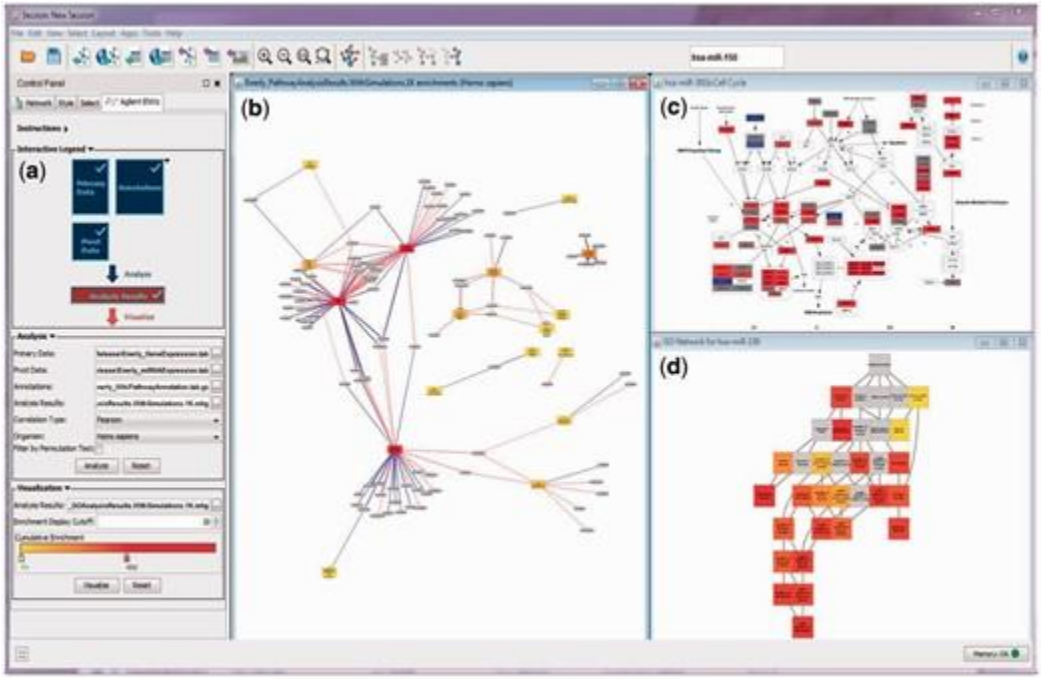

Fig 2. Pantalla de la aplicación ENViz

Conclusiones y recomendaciones.

La bioinformática es muy importante para las investigaciones biológicas, tiene una infinidad de usos, tanto para encontrar hasta para crear, y como vemos en esta primera práctica la creación de aplicaciones es una parte importante de la bioinformática, que nos ayuda a adquirir y

almacenar información a través de la tecnología y más específico con las ciencias computacionales, todas estas desarrolladas alrededor del mundo, poniendo como ejemplo los 3 países citados en esta práctica.

Bibliografía

- Be, H. (2020). *Herramientas bioinformáticas permiten el análisis de genomas de plantas complejas*. Ingenieria.uner.edu.ar. Retrieved 14 February 2022, from http://ingenieria.uner.edu.ar/boletin/index.php/lo-ultimo-en-cyt/359-herramientas-bioinformaticas-permiten-el-analisis-de-genomas-de-plantas-complejas.
- *Bioinformática | NHGRI*. Genome.gov. (2022). Retrieved 20 February 2022, from https://www.genome.gov/es/genetics-glossary/Bioinformatica.
- *El CICESE es pionero en Bioinformática, una ciencia emergente en México*. México Ciencia y Tecnología. (2022). Retrieved 21 February 2022, from http://www.cienciamx.com/index.php/tecnologia/tic/165-el-cicese-es-pionero-en-bioinformatica-una-ciencia-emergente-en-mexico.
- *La bioinformática en la práctica médica: Integración de datos biológicos y clínicos*. Scielp. (2008). Retrieved 21 February 2022, from https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872008000500015.
- Steinfeld, I., Navon, R., L. Creech,, M., Yakhini, Z., & Tsalenko, A. (2015). *https://academic.oup.com/bioinformatics/article/31/10/1683/176511*. Oxford Academy. Retrieved 19 February 2022, from https://academic.oup.com/bioinformatics/article/31/10/1683/176511.