DESARROLLO DE TÉCNICAS DE INTELIGENCIA COMPUTACIONAL PARA EL ANÁLISIS DE DATOS GENÓMICOS

Pablo Javier Vidal[a,b], Jessica Andrea Carballido[c], Ana Carolina Olivera [a,b], Matías Gabriel Rojas[a] y Mariel Denise Volman Stern[b]

[a]*Instituto para las Tecnologías de la Información y las Comunicaciones. Consejo Nacional de Investigaciones Científicas y Técnicas. Universidad Nacional de Cuyo (ITIC-UNCuyo)*

[b]*Facultad de Ingeniería - Universidad Nacional de Cuyo (FING-UNCuyo)*

[c]*Instituto de Ciencias e Ingeniería de la Computación, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur (ICIC-CONICET-UNS)*

contacto: pjvidal@conicet.gov.ar,pablo.vidal@ingenieria.uncuyo.edu.ar

CONTEXTO

Esta presentación corresponde a las tareas de investigación que se llevan a cabo en el ITIC-UNCuyo en conjunto con la Facultad de Ingeniería de la UNCuyo y la Universidad Nacional del Sur en el marco del proyecto 06/B081-B (2019-2021) y una Beca Doctoral del CONICET.

RESUMEN

Esta línea de investigación se centra en el diseño y desarrollo de técnicas de Inteligencia Computacional en combinación con otros métodos con el propósito de contribuir al área de Ciencias de la Computación aplicando el conocimiento desarrollado a problemas de bioinformática, en particular aquellos de las áreas de genómica estructural.

Palabras clave: Bioinformática, Genómica Estructural, Inteligencia Computacional.

1 INTRODUCCIÓN

El conocimiento de las secuencias del genoma humano y de otras especies permite que se realicen investigaciones del genoma como un todo. Para analizar el conjunto de genes la bioinformática trabaja con diversas técnicas computacionales que están a disposición de los científicos debido a que pueden procesar e interpretar una gran cantidad de datos en n-dimensiones y reconstruir y separar estos conjuntos para tareas de clasificación o generar regresiones numéricas para tareas de predicción. Esto permite determinar automáticamente si un caso de estudio presenta una patología especifica o bien, si ciertas características del DNA (Deoxyribonucleic Acid) pueden permitir una clasificación de nuevos casos en forma rápida y eficiente, entre otras posibilidades. Nuestro grupo de investigación está abocado al diseño y aplicación de técnicas de Inteligencia Computacional para el análisis de datos genómicos está formado por investigadores del Instituto para las Tecnologías de la Información y las Comunicaciones de la Universidad Nacional de Cuyo y una Investigadora de la Universidad Nacional del Sur especialista en Bioinformática. Asimismo, participan del grupo una alumna de la Facultad de Ingeniería de la UNCuyo y un Becario Doctoral del CONICET.

Entre los problemas abordados por el grupo de investigación se encuentra el ensamblado de fragmentos de cadenas de DNA (Deoxyribonucleic Acid Fragment Assembly Problem, DNA-FAP). El cual consiste en encontrar dado un conjunto de cientos o miles de fragmentos de DNA, que pueden contener errores, la secuencia de DNA original a partir de las permutaciones de los fragmentos que mejor representen a dicha secuencia [Pev00].

Una vez secuenciada la cadena de DNA de interés, independientemente del objetivo, es indispensable compararla con las

secuencias disponibles en las diferentes bases de datos. Este análisis permite inferir algunas métricas de similitud entre las secuencias y llevarse a cabo diferentes tipos de análisis (filogenéticos, evaluación de la conservación de los dominios proteicos, las estructuras terciarias y secundarias, entre otros).

Existen dos tipos de alineamientos, el de pares de secuencias (Pairwise Sequence Alignment, PSA) y el alineamiento múltiple de secuencias (Multiple Sequence Alignment, MSA). El primero, consiste en la alineación de dos secuencias, mientras que el segundo implica alinear tres o más. El enfoque comúnmente adoptado para abordar el problema del MSA es el alineamiento progresivo que consiste en dividir el problema en subproblemas y resolverlos mediante PSA. Este enfoque es susceptible a quedar atascado en óptimos locales, debido a que una falla en la alineación de los primeros pares puede afectar a la solución final.

Nuestro grupo ha obtenido resultados promisorios en estos temas que se detallan en la Sección 3.

2 LINEAS DE INVESTIGACIÓN y DESARROLLO

- Diseño de novedosas técnicas de Inteligencia Computacional.
- Aplicación de Inteligencia Computacional a problemas de Bioinformática.

3 RESULTADOS OBTENIDOS/ESPERADOS

3.1 Avances en bioinformática.

En lo que respecta a secuenciamiento, alineamiento y ensamblado de fragmentos de cadenas de DNA se ha abordado el problema del secuenciamiento de las cadenas de DNA. Se diseñó un algoritmo híbrido basado con Cuckoo Search para el secuenciamiento de cadenas de DNA [RCOV20a]. En este trabajo, se híbrida al *Cuckoo Search* con dos búsquedas locales diferentes para mejorar las capacidades de búsqueda del algoritmo canónico. Se lleva a cabo una evaluación numérica de ambas propuestas y la versión canónica utilizando un conjunto de datos de referencia bien conocido. Los resultados demuestran que la hibridación mejora la búsqueda y transforma la búsqueda del cuco en un procedimiento robusto con el potencial de tratar también con secuencias más largas o secuencias de longitud desconocida. Considerando el problema de MSA el grupo ha evaluado la capacidad de un Algoritmo Genético Celular Memético (Memetic Cellular Genetic Algorithm, MCGA) para realizar el alineamiento múltiple de secuencias [RCOV20c]. Se propone un algoritmo basado en el CGA combinado con un algoritmo de búsqueda local basada en la inserción, eliminación y reubicación de espacios en la secuencia. Con esto, se busca integrar la lenta difusión de la mejor solución y la rápida convergencia a un óptimo del CGA con la capacidad de identificar a los mejores vecinos de cada solución que posee el algoritmo de búsqueda local.

Asimismo, se ha trabajado con la caracterización multi-objetivo de la selección de características para microarrays de datos de cáncer y su impacto en las soluciones [DPOV20]. Se ha diseñado un algoritmo híbrido que combina el algoritmo genético celular (*Cellular Genetic Algorithm*, CGA) con una búsqueda de vecindario variable (*Variable Neighborhood Search*, VNS) diseñada para este problema. Esta técnica se comparó con otros métodos del estado del arte y los resultados experimentales indicaron que nuestra propuesta supera numéricamente a las otras evaluadas [ROCV20].

3.2 Avances teóricos en Inteligencia Computacional

La capacidad de precisión de una técnica de aprendizaje automático debe ser lo suficientemente robusta ante la aparición de diversos muestras o instancias en diferentes tipos de problemas. Cada una de estas técnicas presentan parámetros ajustables que mejoran o empeoran la capacidad de clasificación de los mismos [EMS19]. La construcción de un modelo de aprendizaje automático eficiente es un proceso complejo y lento que implica determinar el algoritmo apropiado y obtener una arquitectura de modelo óptima mediante el ajuste de sus hiperparámetros [KJ+13]. El ajuste incorrecto lleva una precisión inexacta y computacionalmente más costosa. Teniendo en cuenta que el ajuste de los hiperparámetros de una técnica de ML (*Machine Learning*) es un problema de optimización difícil, se pueden utilizar métodos de inteligencia computacional como enfoque para su resolución. Con respecto al ajuste de hiperpárametros para un mejor desempeño de un SVM (*Support Vector Machine*), el grupo ha comenzado a explorar modificaciones a las configuraciones por defecto utilizadas usualmente en el SVM.
En [RCOV20b] se propuso la utilización de meta-heurísticas para la optimización de los hiperparámetros del SVM utilizando un *kernel Wavelet*. Se observó que todas las meta-heurísticas lograron alcanzar valores altos de precisión en comparación a la configuración por defecto del SVM utilizada comúnmente en la literatura. Por otro lado, el algoritmo genético celular utilizado se destacó en el tiempo de ejecución promedio que le demandó alcanzar el criterio de parada, lo que sugiere una mayor eficiencia con respecto a las otras propuestas. De la misma forma, se evidenció que el *kernel Wavelet* es capaz de lograr una distribución adecuada de los datos, siempre y cuando posea una correcta configuración de sus parámetros. El caso de estudio utilizado fue un conjunto de datos relacionados a la enfermedad retinopatía diabética.

4 FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo está formado por: Un Profesor Adjunto con Dedicación Simple de la Facultad de Ingeniería de la UNCuyo (Investigador Asistente del CONICET), una Profesora Adjunta con Dedicación Exclusiva de la Universidad Nacional del Sur (Investigadora Adjunta del CONICET), Una Profesora Titular con Dedicación Semiexclusiva de la Facultad de Ingeniería de la UNCuyo (Investigadora Adjunta del CONICET), una Becaria EVC-CIN de la UNCuyo y un Becario doctoral del CONICET

El Ing. Rojas es dirigido por el Dr. Vidal y la Dra. Carballido en su doctorado que se encuentra realizando en la Facultad de Ciencias Exactas y Naturales de la UNCuyo en el tema "*Desarrollo de Metaheurísticas Aplicadas a Genómica Funcional y Estructural*".

La Srta. Volman Stern es dirigida por los Dres. Vidal y Olivera en su Beca EVC-CIN en el tema *procesamiento de imágenes utilizando inteligencia artificial* [VOV21].

5 BIBLIOGRAFIA

[DPOV20] Dussaut, J., Ponzoni, I., Olivera, A., y Vidal, P. Algoritmos evolutivos multiobjetivo aplicados a la selección de características en microarrays de datos de cáncer. Entre Ciencia e Ingeniería 14, 28 (dic. 2020),40–45, 10.31908/19098367.2014.

[EMS19] Elshawi, R., Maher, M., y Sakr, S. Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287 (2019)

[KJ+13] Max Kuhn, Kjell Johnson, et al. Applied predictive modeling, volume 26. Springer, 2013.

[Pev00] Pevzner, P. Computational Molecular Biology: An Algorithmic Approach. MIT Press, 2000.

[POV19] Patrana, S., Olivera, A. C., y Vidal, P. J. Análisis del algoritmo genético celular para el problema de ensamblado de cadenas de ADN. Informes Científicos Técnicos - UNPA 3, 11 (2019), 236–248, 10.22305/ict-unpa.v11.n3.804.

[RCOV20a] Rojas, M. G., Carballido, J. A., Olivera, A. C., y Vidal, P. J. Hybrid cuckoo search for solving DNA fragment assembly problem. In IV Congreso Internacional de Ciencias de la Computación y Sistemas de Información (2020).

[RCOV20b] Rojas, M. G., Carballido, J. A., Olivera, A. C., y Vidal, P. J. Optimización de support vector machine mediante metaheurísticas para clasificación de retinopatia diabética. In Simposio Argentino de Inteligencia Artificial de JAIIO (2020).

[RCOV20c] Rojas, M. G., Carballido, J. A., Olivera, A. C., y Vidal, P. J. A memetic cellular genetic algorithm for multiple sequence alignment. In Proceedings of the 2020 IEEE Biennial Congress of Argentina (November 2020).

[ROCV20] Matías Gabriel Rojas, Ana Carolina Olivera, Jessica Andrea Carballido, and Pablo Javier Vidal. A memetic cellular genetic algorithm for cancer data microarray feature selection. IEEE Latin America Transactions, 2020.

[VOV21] Volman Stern, M. D., Olivera, A. C., y Vidal, P. J. Paralelización del filtro convolución para imágenes digitales. In Anales del V Congreso Internacional de Ciencias de la Computación y Sistemas de Información (2021).