Técnicas metaheurísticas avanzadas aplicadas a la resolución de problemas bioinformáticos

Gabriela Minetti[1], Carolina Salto[1], Hugo Alfonso[1] y Fernando Sanz Troiani[2]
Laboratorio de Investigación en Sistemas Inteligentes (LISI)
Facultad de Ingeniería - Universidad Nacional de La Pampa
Calle 110 Esq. 9 (6360) General Pico - La Pampa - Rep. Argentina
Te. / Fax: (02302) 422780/422372, Int. 6302
e-mail: [1]{minettig, saltoc, alfonsoh@ing.unlpam.edu.ar}, [2]fstnando@gmail.com

Resumen La finalidad de esta línea de investigación es el estudio y resolución de problemas del area Bioinformática mediante la utilización de métodos inteligentes. Particularmente, nuestro trabajo se enfoca en la resolución de problemas de secuenciamiento de un genoma por medio del diseño e implementación de nuevas técnicas metaheurísticas ya sean basadas en trayectoria como en población. También consideramos la posibilidad de hibridar y/o distribuir estos métodos dependiendo de la complejidad del problema a resolver.

Palabras claves: Bioinformática, ADN, metaheurísticas, secuenciamiento de un genoma, ensamblado de fragmentos, métodos de búsqueda híbrida y distribuida, optimización combinatoria.

Contexto

Esta línea de investigación se desarrolla en el marco del proyecto de investigación "Resolviendo problemas complejos con técnicas metaheurísticas avanzadas" dirigido por la Dra. Carolina Salto y llevado a cabo en el Laboratorio de Investigación de Sistemas Inteligentes (LISI), de la Facultad de Ingeniería de la Universidad Nacional de La Pampa. Este proyecto mantiene desde hace varios años una importante vinculación con investigadores de la Universidad Nacional de San Luis (Argentina) y de la Universidad de Málaga (España), con quienes se han realizado varias publicaciones conjuntas.

I. Introducción

En las últimas décadas, los avances en el campo de la Biología Molecular y en las tecnologías de genómica han provocado un crecimiento muy fuerte en la información biológica generada por la comunidad científica. La secuenciación de genomas y de proteomas, la identificación de genes, la generación de perfiles de expresión genética y otras áreas genéticas han demostrado la necesidad de la participación de expertos matemáticos, ingenieros y físicos para obtener resultados en corto tiempo y de mayor calidad en estas áreas.

La Bioinformática es, entonces, un campo interdisciplinar que involucra a los expertos antes mencionados para: analizar la secuencia genómica, identificar y predecir las estructuras moleculares, determinar el perfil de expresión genética, etc. Estas actividades, en su mayoría, necesitan ser formuladas como problemas de optimización para poder llevarse a cabo. En algunos de ellos encontramos analogías con problemas de optimización combinatoria clásicos, como es el caso de, uno de los problemas tratados aquí, el ensamblado de fragmentos de ADN (*Fragment Assembly Problem*, FAP) con el problema del viajante de comercio (*Travelling Salesman Problem*, TSP).

El conjunto de técnicas bioinformáticas utilizadas en las distintas áreas de la Biología Molecular, en particular en el ensamblado de fragmentos de ADN, es extenso y de componentes heterogéneos. Es posible distinguir dos grandes grupos de técnicas algorítmicas. El primero está formado por algoritmos especialmente diseñados para un uso bioinformático específico, por ejemplo para alinear un par de secuencias de ADN. Por otro lado, el segundo subconjunto está conformado por un grupo de técnicas modernas de uso generalizado, denominadas metaheurísticas.

En el primer caso, los algoritmos han sido diseñados y modelados específicamente para manejar información biológica. Es el caso de herramientas como: CAP3 (ensambla fragmentos de ADN) [1], CLUSTALW-pairwise (compara un par de secuencias de ADN) [2], CLUSTAL-MSA (compara múltiples secuencias de ADN) [2], FASTA (permite identificar proteínas relacionadas entre sí) [3], [4], BLAST y sus variantes (conjunto de herramientas de búsqueda local para alinear secuencias) [5], [6], Vector de momentos de composición (permite predecir la estructura secundaria de la proteína) [7], Modelado de la dinámica molecular (identifica más de un tipo de estructura proteíca) [8], IRAP (determina cómo interactúan las proteínas entre sí, acoplamiento de proteínas) [9], entre otras.

En el segundo caso, las técnicas metaheurísticas han sido intensamente usadas en diferentes campos y se han adaptado a muchos usos bioinformáticos. La razón es que pueden resolver problemas de grandes dimensiones con fuertes restricciones de manera eficiente. Ejemplos de esto son: los algoritmos evolutivos, los métodos de Montecarlo guiados, la optimización basada en colonias de hormigas, los algoritmos meméticos, la optimización basada en cúmulo de partículas, entre otras.

Aunque este tipo de algoritmos tiene un uso más generalizado, resultan eficaces y eficientes cuando la complejidad del problema y su respectivo espacio de soluciones son extensos o crecen continuamente. Esta es una característica muy importante y extremadamente necesaria a la hora de manipular enormes cantidades de información biológica. Además estas metaheurísticas presentan otra ventaja significativa, en el área de la Bioinformática, ya que resultan sumamente eficientes en la resolución de problemas de optimización combinatoria; como por ejemplo los problemas de: alineamiento de secuencias, ensamblado de fragmentos, análisis proteínico, entre muchos otros. Las características y ventajas mencionadas anteriormente son difíciles de encontrar o de incorporar en las técnicas del primer grupo.

Los objetivos de esta línea de investigación son: estudiar problemas bioinfórmaticos, especialmente los relacionados con el secuenciamiento de cadenas de ADN en un genoma, analizar, diseñar y desarrollar algoritmos metaheurísticos para resolverlos eficientemente. Como dijimos anteriormente, la formulación de varios de estos problemas es análoga a la de problemas de optimización clásicos, por ende en ciertos casos será necesario primero analizar el comportamiento de las diferentes metaheurísticas para resolver los problemas clásicos y luego adoptarlas o no, según sea el caso.

II. Desarrollo

En esta sección describimos los desarrollos que se llevan a cabo en esta línea de investigación, pero primero introducimos uno de los principales problemas de secuenciamiento de un genoma como es el problema de ensamblado de fragmentos de ADN y su analogía con el problema del viajante de comercio.

FAP es un problema resuelto en las primeras fases del proyecto del genoma y por lo tanto muy importante, ya que los demás pasos dependen de su precisión. El proceso de ensamblado de fragmentos consiste en: una primera fase de superposición (calcula el *puntaje de solapamiento* entre los fragmentos), una segunda de distribución (encuentra el orden de los fragmentos basado en el puntaje de similitud computado) y una tercera de consenso (deriva la secuencia de ADN a partir de la distribución anterior). Una resolución óptima del problema se produce cuando el algoritmo escapaz de ensamblar un determinado conjunto de fragmentos en un solo *contig*. Un *contig* es una secuencia en la que la solapamiento entre los fragmentos adyacentes es mayor o igual a un umbral predefinido (parámetro de corte denominado $cutoff$).

Desde el punto de vista de la optimización combinatoria, la construcción de un consenso es similar a la de un recorrido en un problema del viajante de comercio. Esto es porque cada fragmento tiene una ubicación específica en la formación de una secuencia en la etapa de consenso. Aunque los puntos terminales de un recorrido de TSP sean irrelevantes ya que su solución es un recorrido circular de ciudades, en el caso de FAP estos puntos son importantes ya que ellas representan los extremos opuestos de la secuencia original de ADN. En TSP el ordenamiento de las ciudades es la solución final al problema. En cambio para FAP, el ordenamiento de fragmentos es sólo un resultado intermedio que será utilizado en la fase de consenso.

Por un lado, nuestro trabajo consiste en profundizar el estudio de las distintas variantes de algoritmos de optimización basados en colonia de hormigas (*Ant Colony Optimization* -ACO-) [10] considerando diferentes configuraciones paramétricas. Para ello utilizamos un problema tradicional de optimización combinatoria, como es el caso del problema del viajente de comercio asimétrico (*Asymmetric Traveling Salesman Problem*, ATSP), con el objeto de determinar cuál de las variantes ACO (*Ant System* -AS- [11], *Elitist Ant System* -EAS- [11], *Ant Colony System* -ACS- [12], *Max-Min Ant System* -MMAS- [13], [14], [15], *Rank-Based Ant System* -ASrank- [16]) realiza una mejor manipulación de un gran número de ciudades. De esta forma, adecuaríamos la variante ACO seleccionada a FAP.

Por otra parte, analizamos y comparamos el comportamiento de ensambladores de fragmentos de ADN metaheurísticos, (*Inversion Simulated Annealing* -ISA- [17], *Problem Aware Local Search* -PALS- [18] y *Genetic Algorithms* con estrategias de inicio -GAG_{50}- [19]) al resolver instancias de FAP con ruido. De esta manera podemos estudiar la robustez de los mismos y proponer nuevos y mas eficientes algoritmos para desarrollar esta tarea. Estudiar la robustez de un ensamblador significa analizar las diferencias entre las soluciones encontradas para las instancias sin y con ruido. Si no se detectan diferencias (estadísticamente significativas), el ensamblador muestra un comportamiento neutro (insensible, indistinto) a pequeñas variaciones en los datos de entrada (ruido). Consecuentemente, este ensamblador se considera robusto para resolver instancias ruidosas.

III. Resultados obtenidos/esperados

En esta sección presentamos los resultados obtenidos de nuestra investigación en el transcurso del año 2011.

El análisis de las variantes algorítmicas y paramétricas de ACO [20] arrojó que, las distintas variantes sólo son susceptibles a cambios en los valores paramétricos si el espacio de búsqueda es grande. En el caso de ATSP esto sucede cuando el número de ciudades es mayor a 100. Por otra parte, la variante que se comporta mejor ante un elevado número de ciudades es ASrank, aunque el esfuerzo computacional de MMAS es menor al de ASrank .

A partir del estudio realizado a los diferentes ensambladores de fragmentos, en [21], surge que PALS es la metaheurística más robusta a la hora de solucionar las instancias de FAP con ruido. Además, es uno de los algoritmos que mejores soluciones encuentra. Aunque, al igual que los otros ensambladores (ISA y GAG_{50}), PALS no logra obtener soluciones finales con el número óptimo de contigs, especialmente cuando resuelve instancias de gran tamaño. También se detecta que la falencia de PALS es la rápida convergencia a óptimos locales.

En un intento de lograr un ensamblador robusto que solucione eficientemente instancias ruidosas de gran tamaño, proponemos una nueva metaheurística que aproveche las fortalezas de PALS y mitigue sus debilidades. Este nuevo

ensamblador, al igual que PALS, estima el número de contigs y el *fitness* de cada posible movimiento de fragmentos pero, a diferencia de PALS, evita la convergencia prematura a óptimos locales. Esto resulta en nuevo ensamblador metaheurístico híbrido y paralelo, denominado PH-PALS [22] capaz de escapar de los óptimos locales y reducir el número de contigs en las soluciones correspondientes a las instancias de mayor tamaño.

Existen distintos puntos de interés que merecen una investigación más profunda. Uno de ellos tiene que ver con ajustes en el algoritmo PH-PALS y en su parametrización, con el objetivo de mejorar su eficiencia. Otro está relacionado con la incorporación de procesos evolutivos en cada una de las islas que genera dicho algoritmo. Por otra parte, se prevé adoptar y adaptar ASrank y MMAS al problema de ensamblado de fragmentos de ADN.

IV. Formación de recursos humanos

Durante el año 2011 se ha concluido la escritura de una tesis doctoral y se ha realizado la correspondiente presentación y defensa para obtener el título de Doctor en Ciencias de la Computación (UNSL) [23]. Además, a lo largo de ese mismo año uno de los becarios del proyecto ha presentado su tesis para alcanzar el título de Ingeniero en Sistemas, a partir de las actividades desarrolladas en el LISI [20].

En tanto que, en el LISI se trabaja con alumnos avanzados en la carrera Ingeniería en Sistemas en temas relacionados a la resolución de problemas de optimización usando técnicas inteligentes, con el objeto de guiarlos en el desarrollo de sus tesinas de grado y, también, de formar futuros investigadores.

References

[1] W. Huang and A. Madan, "CAP3: A DNA Sequence Assembly Program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[2] J. Thompson, D. Higgins, and T. Gibson, "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, pp. 4673–4680, 1994.

[3] W. Pearson, "Comparison of methods for searching protein sequence databases," *Protein Sci.*, vol. 4, pp. 1145–1160, 1995.

[4] W. Pearson and D. Lipman, "Improved tools for biological sequence analysis," *Proc. Natl Acad. Sci. USA,85*, pp. 2444–2448, 1998.

[5] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, no. 1990, pp. 403–410, 1990.

[6] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, no. 25, pp. 3398–3402, 1997.

[7] J. Ruan, K. Wang, J. Yang, L. Kurgan, and K. Cios, "Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences," *Artificial Intelligence in Medicine*, no. 35, pp. 19–35, 2005.

[8] D. York, T. Darden, L. Pedersen, and M. Anderson, "Molecular dynamics simulation of hiv-1 protease in a crystalline environment and in solution," *Biochemestry*, vol. 32, no. 6, pp. 1143–1153, 1993.

[9] J. Chen, W. Hsu, M. Lee, and S. Ng, "Systematic assessment of high-throughput experimental data for reliable protein interactions using network topology," *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, pp. 368–372, 2004.

[10] M. Dorigo, "Optimization, learning and natural algorithms," Ph.D. dissertation, Politecnico di Milano, Italy, 1992.

[11] V. M. y. A. C. M. Dorigo, "The ant system: Optimization by a colony of cooperative agents," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. Part B, 26, no. 1, pp. 29–41, 1996.

[12] M. D. y L. M. Gambardella., "Ant colony system: A cooperative learning approach of cooperating agents," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 54–66, 1997.

[13] T. Stützle, "Local search algorithms for combinatorial problems: Analysis, improvemnts, and new applications, volume 220 of diski," *Infix, Sankt Agustin, Germany*, 1999.

[14] T. Stützle and H. Hoss., "Improving the ant system: A detailed report on the max-min ant system." FG Intellektik, FB Informatik, TU Darmstadt, Germany, Tech. Rep. AIDA-96-12, Aug. 1996.

[15] T. Stutzle and H. Hoss., "MAX-MIN Ant System," *Future Generation Computer Systems*, vol. 16, no. 8, pp. 889–914, 2000.

[16] B. Bullnheimer, R. F. Hartl, and C. Strauß, "A new rank based version of the ant system - a computational study," *Central European Journal for Operations Research and Economics*, vol. 7, pp. 25–38, 1997.

[17] G. Minetti, G. Luque, G. Leguizamón, and E. Alba, "A new Hybrid SA for Solving the DNA Fragment Assembly Problem," in *XXVIII Internacional Conference of the Chilean Computing Science Society (SCCC)*, Santiago, Chile, November 2009, pp. 109 – 116.

[18] E. Alba and G. Luque, "A New Local Search Algorithm for the DNA Fragment Assembly Problem," in *Evolutionary Computation in Combinatorial Optimization, EvoCOP'07*, ser. Lecture Notes in Computer Science. Valencia, Spain: Springer, 2007, vol. 4446, pp. 1–12.

[19] G. Minetti, E. Alba, and G. Luque, "Seeding strategies and recombination operators for solving the DNA fragment assembly problem," *Information Processing Letters*, vol. 108, no. 3, pp. 94–100, October 2008.

[20] F. Sanz, "Optimización basada en colonias de hormigas: un análisis paramétrico y comparativo," Facultad de Ingeniería, UNLPam, Tesina de grado, November 2011.

[21] G. Minetti, G. Leguizamón, and E. Alba, "Assembling DNA Sequences Containing Noisy Information With Metaheuristic Algorithms," *Journal of Information Sciences, Elsevier (en evaluación)*, 2011.

[22] G. Minetti and G. Leguizamón and E. Alba, "A new Parallel and Hybrid Metaheuristic for Solving Noisy DNA Strands," *Journal of Information Sciences, Elsevier (en evaluación)*, 2011.

[23] G. Minetti, "Problema de ensamblado de fragmentos de ADN resuelto mediante metaheurísticas y paralelismo," Ph.D. dissertation, Universidad Nacional de San Luis, November 2011.