ÁP DỤNG THUẬT TOÁN ACO VÀO VIỆC GIẢI CÁC BÀI TOÁN TỐI ƯU TRONG SINH HỌC PHÂN TỬ

*Trần Quốc Chiến, Đặng Đức Long, Đoàn Duy Bình**

TÓM TẮT

Bài toán tối ưu trong sinh học phân tử là một trong những lĩnh vực khoa học tính toán được nghiên cứu nhiều hiện nay; trong đó có vấn đề dự đoán cấu trúc chuỗi RNA bằng những thuật toán tối ưu. Thuật toán ACO (Ant Colony Optimization)- tối ưu đàn kiến – là phương pháp nghiên cứu lấy cảm hứng từ việc mô phỏng hành vi của đàn kiến trong tự nhiên nhằm mục đích giải quyết các bài toán tối ưu phức tạp trong thực tế. Các cá thể kiến trao đổi thông tin trên đường đi thông qua vết mùi (Pheromone) để lại trên đường đi. Các đường đi có nồng độ mùi ít hơn sẽ được loại bỏ, cuối cùng tất cả đàn kiến sẽ đi trên con đường có khả năng trở thành con đường ngắn nhất từ tổ đến nguồn thức ăn. Trong bài báo này chúng tôi giới thiệu thuật toán ACO (Ant Colony Optimization) là một phương pháp mới giải bài toán tối ưu tìm cấu trúc bậc 2 của phân tử RNA có tổng năng lượng bền vững nhất.

Từ khóa: tối ưu hóa, thuật toán tối ưu đàn kiến, RNA, sinh học phân tử, Tin sinh học

1. Đặt vấn đề:

Tối ưu hóa là tìm trạng thái tối ưu của một hệ thống sao cho đạt được mục tiêu mong muốn về chất lượng theo một tiêu chuẩn nào đó.

Dang tổng quát của bài toán được cho bởi (xem [5,[3]]):

$f(x) \rightarrow \text{Min (max)}$

$$\begin{cases} g_i(x) \geq b_i , & i=\overline{1, m_1} , \\ g_j(x) \leq b_j , & j=\overline{m_1+1, m_2} , \\ g_k(x) = b_k , & k=\overline{m_2+1, m} , \end{cases} \qquad (1)$$

Trong đó f(x) được gọi là hàm mục tiêu; $g_i(x)$, ($i = \overline{1,m}$), được gọi là các hàm ràng buộc. Mỗi một đẳng thức hay bất đẳng thức được gọi là một ràng buộc. Gọi:

$$D=\left\{ x \in R^n \middle| \begin{array}{ll} g_i(x) \geq b_i, & i=\overline{1, m_1} \\ g_j(x) \leq b_j , & j=\overline{m_1+1, m_2} \\ g_k(x) = b_k , & k=\overline{m_2+1, m} \end{array} \right\} , \qquad (2)$$

gọi là miền ràng buộc hay miền chấp nhận được. Mỗi một vector

$x = (x_1, x_2,..., x_n) \in D$ được gọi là phương án của bài toán (hay lời giải chấp nhận được). Phương án $x^* \in D$ được gọi là phương án tối ưu của bài toán nếu thỏa mãn điều kiện sau:

$f(x^*) \leq f(x), \forall x \in D$ (đối với bài toán tìm Min)

$f(x^*) \geq f(x), \forall x \in D$ (đối với bài toán tìm Max)

khi đó tương ứng $f(x^*)$ gọi là giá trị tối ưu.

Ngày nay, có một chuyên ngành khoa học là Tin sinh học (Bioinformatics), với nhiệm vụ là ứng dụng và phát triển các phương pháp của công nghệ thông tin để xử lý

và khai thác lượng thông tin vô cùng lớn trong sinh học phân tử, đang phát triển rất nhanh. Trong lĩnh vực nghiên cứu này, các bài toán tối ưu hóa rất phổ biến, từ việc so sánh các chuỗi phân tử sinh học, dự đoán cấu trúc bền vững của các phân tử DNA (deoxyribo nucleic acid), RNA (ribo nucleic acid), protein, đến dự đoán tương tác của các phân tử (xem [2]). Một bài toán như thế là vấn đề dự đoán cấu trúc bậc 2 tối ưu của một phân tử RNA nhất định dựa trên thông tin về chuỗi của nó sao cho cấu trúc này có năng lượng tự do tổng cộng là âm nhất (ở trạng thái bền vững nhất). Việc giải bài toán này một cách chính xác trong nhiều trường hợp là chưa thể thực hiện được (thuộc lớp bài toán NP-hard). Do đó, hiện nay đang có nhiều thuật toán được đưa ra để giải quyết vấn đề này.

Thuật toán ACO (Ant Colony Optimization) là một thuật toán tối ưu hóa hiện đại được ứng dụng trong các bài toán tối ưu phức tạp; cơ sở của thuật toán là dựa trên sự di chuyển của đàn kiến trong quá trình tìm kiếm nguồn thực ăn thông qua việc phát ra các pheromone. Trong khuôn khổ bào báo này chúng tôi giới thiệu nguyên lý của việc sử dụng thuật toán ACO vào việc dự đoán cấu trúc bậc 2 tối ưu của một chuỗi RNA.

2. Cấu trúc của RNA (xem5.[1]5.[3])

Các RNA được cấu tạo từ các đơn phân là các ribonucleotide; các ribonucleotide này nối kết với nhau bằng các liên kết 3',5'-phosphodiester tạo thành các chuỗi polyribonucleotide - cấu trúc sơ cấp (bậc 1) của các phân tử RNA. Trong phân tử RNA có bốn loại đơn phân chính (dựa trên sự khác nhau về gốc base) là adenine (A), uracil (U), guanine (G) và cytosine (C). Về mặt thông tin, một phân tử RNA với cấu trúc bậc 1 có thể được biểu diễn dưới dạng một chuỗi ký tự của bốn chữ cái: A, U, G, và C.

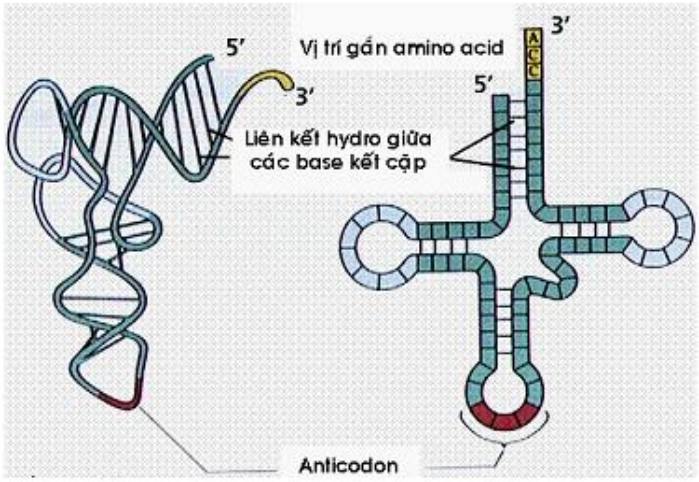
Ví dụ như phân tử 5S ribosome RNA có cấu trúc bậc 1 thể hiện như sau:

$^{(1)}$GUCUACGGCCAUACCACCCUGAACGCGCCCGAUCUCGUCUGAUCUCGGA
AGCUAAGCAGGGUCGGGCCUGGUUAGUACUUGGAUGGGAGACCGCCUGG
GAAUACCGGGUGCUGUAGGCUU$^{(120)}$

với con số trong ngoặc đơn đánh dấu vị trí thứ tự của các đơn phân trong chuỗi.

Trong cơ thể sinh vật có nhiều loại phân tử RNA khác nhau như: RNA thông tin (mRNA), RNA vận chuyển (tRNA), ribosome RNA (rRNA), RNA nhỏ trong nhân (snRNA), RNA điều khiển như siRNA, miRNA, v.v. … Các phân tử này có vai trò thiết yếu trong quá trình sống và phát triển của mọi sinh vật (xem 5.[4]). Để thực hiện các vai trò trong sự sống như vậy, các phân tử RNA phải tồn tại với các cấu trúc không gian ba chiều nhất định (trạng thái cấu trúc bền vững nhất). Cơ sở của các cấu trúc không gian ba chiều đó là việc hình thành các cấu trúc bậc hai của phân tử RNA. Cấu trúc bậc 2 của RNA được hình thành khi sợi đơn của chuỗi các đơn phân của chúng uốn cong và gập khúc trong không gian để đưa một số các đơn phân lại gần nhau, tạo ra các liên kết không hóa trị (chủ yếu là liên kết hydro giữa các cặp đơn phân như G-C và A-U) và các cấu trúc con như vòng, cặp tóc, v.v…, dẫn đến trạng thái bền vững (làm năng lượng tự do âm hơn) của cả chuỗi RNA. Do đó việc dự đoán cấu trúc bậc 2 của RNA có vai trò quan trọng trong việc xác định cấu trúc trong không gian ba chiều (bậc 3) cũng như xác định tính chất, sự hoạt động của phân tử RNA. Ví dụ về cấu trúc bậc 2 và bậc 3 của một

phân tử tRNA được mô tả trong ở Hình 1.



Hình 1. Cấu trúc bậc ba (trái) và bậc hai của một phân tử tRNA.

3. Thuật toán ACO (xem 5.[2]**)**

ACO (Ant Colony Optimization) – là phương pháp nghiên cứu lấy cảm hứng từ việc mô phỏng hành vi của đàn kiến trong tự nhiên nhằm mục đích giải quyết các bài toán tối ưu phức tạp trong thực tế.

Vào năm 1991 A.Colorni và M. Dorigo, đã giới thiệu Giải thuật kiến và đã nhận được sự chú ý đông đảo nhờ vào khả năng tối ưu của nó trong nhiều lĩnh vực khác nhau. ACO lấy cảm hứng từ việc quan sát hành vi của đàn kiến trong quá trình tìm kiếm nguồn thực ăn. Từ sự qua sát đấy, người ta đã khám phá ra rằng, đàn kiến luôn tìm được những nguồn thức ăn cách tổ của chúng với đường đi ngắn nhất. Các cá thể kiến trao đổi thông tin trên đường đi thông qua vết mùi (Pheromone) để lại trên đường đi. Vết mùi sẽ bay hơi dần và mất đi theo thời gian, nhưng nó cũng có thể được củng cố nếu những con kiến khác tiếp tục đi trên con đường đó lần nữa. Cứ như vậy, các con kiến sau sẽ lựa chọn con đường nào có nồng độ vết mùi dày đặc hơn và chúng lại tiếp tục gởi thêm mùi trên con đường mà chúng đã chọn. Các đường đi có nồng độ mùi ít hơn sẽ được loại bỏ, cuối cùng tất cả đàn kiến sẻ đi trên con đường có khả năng trở thành con đường ngắn nhất từ tổ đến nguồn thức ăn.

Bắt nguồn từ những con đàn kiến trong tự nhiên, thông qua các hành vi của chúng, Dorigo xây dựng các con kiến nhân tạo (Artificial ants) cũng có những đặc trưng như kiến tự nhiên, tức là có khả năng sản sinh ra mùi để lại trên đường đi, có khả năng lần theo nồng độ mùi để lựa chọn con đường có nồng độ mùi cao hơn để đi. Gắn với mỗi đường đi (i,j) (cạnh) là nồng độ mùi τ_{ij} và thông số heuristic η_{ij} trên cạnh đó.

Ban đầu, nồng độ mùi trên mỗi cạnh (i,j) được khởi tạo bằng một hằng số c, hoặc được xác định bằng công thức:

$$\tau_{ij} = \tau_0 = \frac{\mathrm{m}}{\mathrm{C}^{nn}}, \tag{3}$$

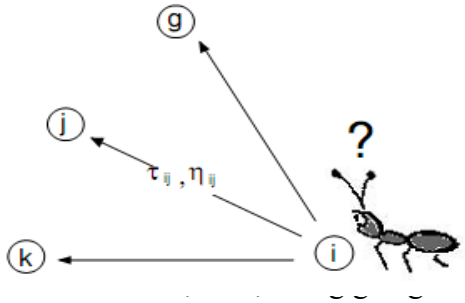
trong đó:

τ_{ij} : Nồng độ vết mùi trên cạnh (i,j)

m: Số lượng kiến,

C^{nn}: Chiều dài hành trình cho bởi phương pháp tìm kiếm gần nhất.

Tại đỉnh i, một con kiến k sẽ chọn đỉnh j chưa được đi qua trong tập láng giềng của i,



theo một quy luật phân bổ xác xuất được xác định theo công thức sau:

$$p_{ij}^{k} = \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l \in N_i^k}[\tau_{il}]^{\alpha}[\eta_{il}]^{\beta}}, j \in N_i^k \quad (4)$$

trong đó:

p_{ij}^{k} : Xác suất con kiến k lựa chọn cạnh (i,j),

α : Hệ số điều chỉnh ảnh hưởng của τ_{ij},

η_{ij} : Thông tin heuristic giúp đánh giá chính xác sự lựa chọn của con kiến khi quyết định đi từ đỉnh i qua đỉnh j, được xác định theo công thức:

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (5)$$

trong đó d_{ij} là khoảng cách giữa đỉnh i và đỉnh j,

β : Hệ số điều chỉnh ảnh hưởng η_{ij},

N_{ij}^{k} : tập các đỉnh láng giềng của i mà con kiến k chưa đi qua.

4. Giải quyết bài toán dự đoán cấu trúc bậc 2 của RNA bằng thuật toán ACO

4.1. Phân tích bài toán

Để dự đoán cấu trúc bậc 2 của một chuỗi RNA, chúng tôi đưa về bài toán tìm cấu trúc (tập hợp các liên kết không hóa trị) có năng lượng tự do nhỏ nhất cho chuỗi này. Năng lượng tự do của một cấu trúc bậc 2 được xác định bởi tổng các đóng góp năng lượng của các phần tử cặp đôi, vòng, cặp tóc, v.v…. Giá trị của các đóng góp về năng lượng của các phần tử cấu trúc con riêng lẻ đã được xác định bằng thực nghiệm (xem 5.[1]5.[4]). Mô hình được thể hiện như sau:

$$\min E(S) = \sum_{i,j \in S} e(r_i, r_j) \quad (6)$$

với ***E(S)*** là năng lượng của toàn bộ chuỗi, $e(r_i,r_j)$ là năng lượng của các

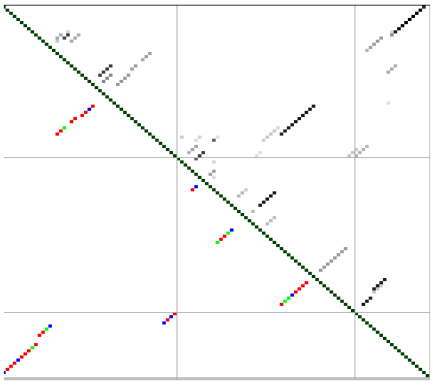
phần tử giữa đơn phân *i* và *j* trong chuỗi *S*.

Phương pháp tổ hợp (combinatorial) có thể giải quyết bài toán tối ưu này. Khi đó, các cấu trúc bậc 2 được tạo ra bằng các tổ hợp tất cả các cặp đôi có thể theo tất cả các cách có thể và tạo nên một dãy các cấu trúc. Năng lượng tự do của các cấu trúc của dãy này được tính toán cụ thể từ dữ liệu thực nghiệm và công thức (6), sau đó chọn ra cấu trúc có năng lượng nhỏ nhất. Nhưng trên thực tế, nếu chúng ta thực hiện phương pháp tổ hợp đơn giản như vậy thì khối lượng tính toán và yêu cầu bộ nhớ sẽ tăng theo cấp số mũ của chiều dài chuỗi. Trên lý thuyết, số các cấu trúc bậc 2 có thể của một chuỗi sẽ lớn hơn hay bằng 1.8^N, trong đó N là số đơn phân (ký tự) của chuỗi. Do đó phương pháp tổ hợp đơn giản chỉ thực hiện được với các chuỗi ngắn (nhỏ hơn 200 đơn phân) (xem 5.[1]5.[4]) . Thuật toán ACO sẽ được sử dụng ở đây để giải quyết bài toán tối ưu tổ hợp này trong thời gian và với bộ nhớ cho phép trong thực tế.

Từ ví dụ phân tử 5S ribosome RNA có cấu trúc bậc 1 thể hiện như sau:

$^{(1)}$GUCUACGGCCAUACCACCCUGAACGCGCCCGAUCUCGUCUGAUCUCGGA
AGCUAAGCAGGGUCGGGCCUGGUUAGUACUUGGAUGGGAGACCGCCUGG
GAAUACCGGGUGCUGUAGGCUU$^{(120)}$

Ta có thể biểu diễn các đơn phân của chuỗi RNA trong ma trận cỡ (NxN). Khi đó các cặp đôi có thể sẽ là các điểm trên đường chéo của ma trận đó (Hình 3).



Hình 3. Đồ thị tạo cặp trong chuỗi biểu thị trong các đơn phân tử RNA

Bài toán đặt ra là làm sao tìm ra được một tập hợp cặp đôi và các khu vực giữa chúng trong toàn chuỗi (đường đi từ đầu đến cuối ma trận) sao cho năng lượng tổng hợp của cả chuỗi RNA (trên cả đường đi) là nhỏ nhất (âm nhất).

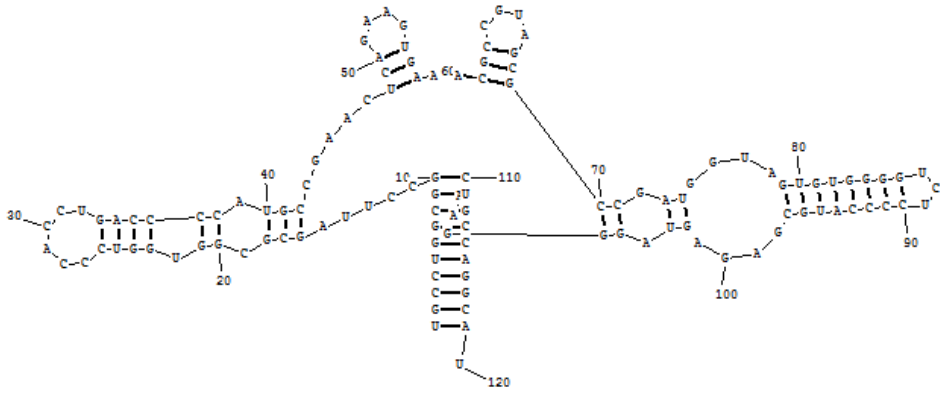
Bài toán được quy về bài toán tối ưu trên đồ thị, nối các cặp điểm trên đồ thị. Cho đồ thị **G=(C, L)** trong đó **C** là tập các phần tử trong chuỗi RNA, với chuỗi RNA gồm các phần tử A, U, G, C; và **L** là tập tất cả kết nối chấp nhận được giữa các phần tử của chuỗi RNA.

4.2. Giải quyết bài toán

Từ việc phân tích trên, chúng tôi triển khai và tìm được một đồ thị các năng lượng kết giữa các phân từ của chuỗi RNA. Từ đó chúng tôi dùng thuật toán ACO để thực hiện việc tìm đường đi giữa các tập đỉnh của chuỗi RNA với hàm mục tiêu là:

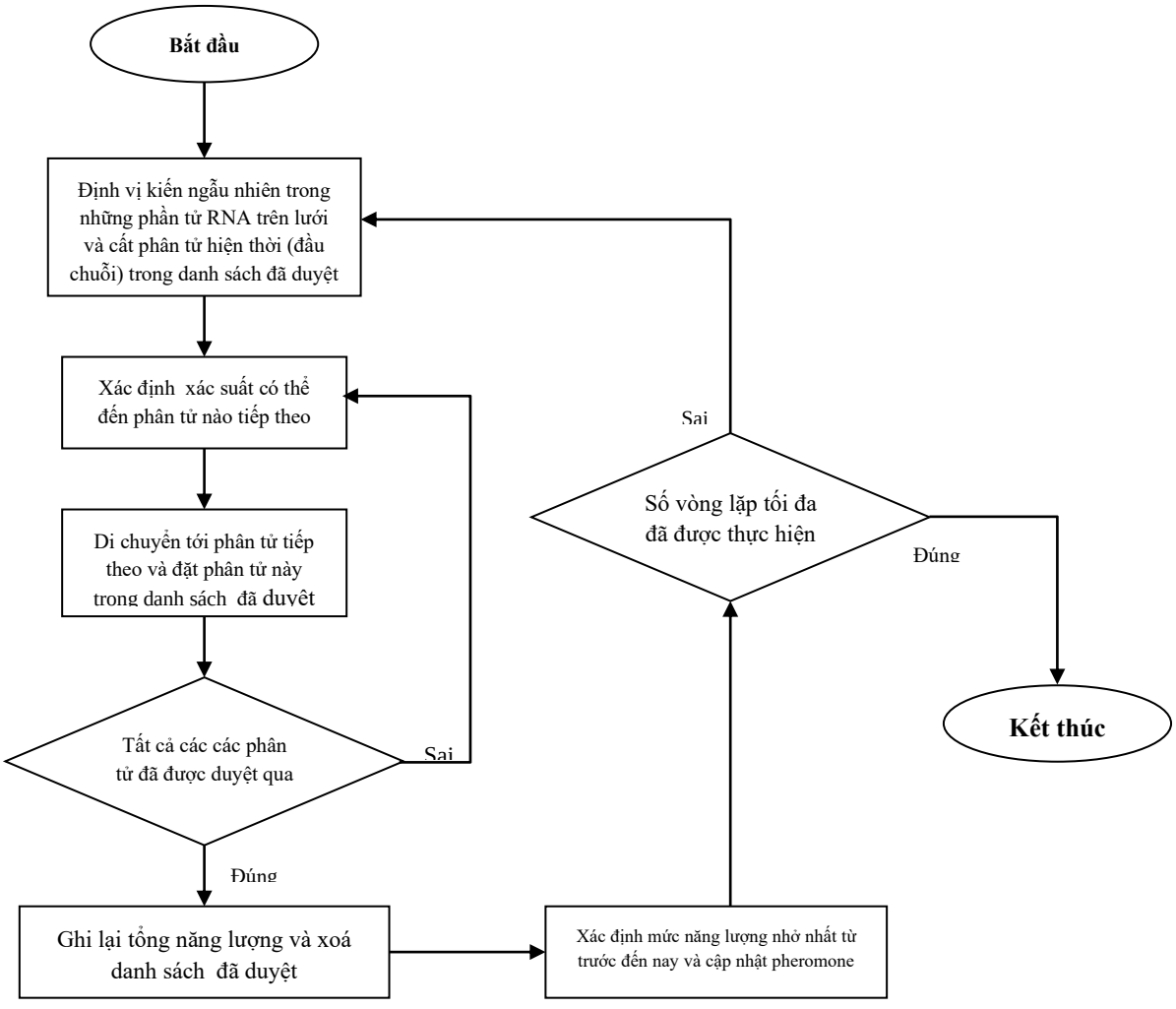
$f(x) \rightarrow \text{Min}$

Mô hình cấu trúc bậc hai của chuỗi RNA tương ứng với ví dụ trên:



Hình 4. Cấu trúc bậc hai của chuỗi RNA.

4.3. Sơ đồ thuật toán ACO áp dụng cho bài toán



Hình 5. Sơ đồ thuật toán ACO áp dụng cho bài toán.

5. Kết luận

Với thuật toán ACO đã thể hiện nhiều ưu điểm trong việc giải bài toán tối ưu tổ

hợp, nhưng nó chưa được sử dụng để giải quyết vấn đề dự đoán cấu trúc bậc 2 của phân tử RNA. Ở đây chúng tôi đã phân tích đặc điểm của vấn đề này và của thuật toán ACO và chứng minh rằng thuật toán ACO phù hợp và có những ưu điểm tốt để giải quyết vấn đề ở đây. Chúng tôi đã xây dựng sơ đồ giải thuật căn bản sử dụng thuật toán ACO để dự đoán được cầu trúc bậc hai của chuỗi RNA. Tiếp theo chúng tôi sẽ xây dựng phần mềm căn cứ trên sơ đồ giải thuật tạo ra để dự đoán một cách chính xác và thuận tiện hơn cấu trúc bậc 2 của phân tử RNA.

TÀI LIỆU THAM KHẢO

[1] Baxevanis A.D., Francis Ouellette B. F. (Eds). 2005. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd edition*. CRC Press, Taylor & Francis Group.

[2] Marco Dorigo, Thomas Stűtzle. 2004. *Ant Colony Optimization*, Massachusetts Instituteof Technology

[3] Nguyễn Hải Thanh. 2006. *Tối ưu hóa*, NXB Bách Khoa Hà Nội

[4] Trần Thị Xô, Nguyễn Thị Lan. 2004. *Cơ sở di truyền và công nghệ gen*, NXB Khoa học Kỹ thuật.

APPLICATION OF ACO ALGORITHMS TO DEALING WITH MOLECULAR BIOLOGY PROBLEMS

Tran Quoc Chien, Doan Duy Binh[1]
Dang Duc Long[2]

[1] The University of Da Nang - University of Science and Education
[2] The University of Da Nang - University of Technology

ABSTRACT

Optimization problems in molecular biology is one of the most investigated fields in computer science today; one notable case is the prediction of RNA structures by optimizing algorithms. ACO (Ant Colony Optimization) algorithm is the research method inspired from the simulation of the behavior of ants in nature for the solution to optimization problems. The communication among ants or between ants and the environment is based on the use of chemicals produced by the ants; these chemicals are called pheromones. Roads with fewer pheromones will be gradually removed; eventually all ants will go on the road having the potential to become the shortest path from their nest to a food source. This paper introduces the ACO (Ant Colony Optimization) algorithm as a new way to solve the problem of predicting the optimal secondary structures of RNAs that have the most stable total energy.

Key words: optimization, Ant Colony Optimization, RNA, molecular biology, BioInfomatics

* PGS.TSKH. Trần Quốc Chiến, ThS. Đoàn Duy Bình, Email: doanduybinh@gmail.com Trường Đại học Sư Phạm, Đại học Đà Nẵng

TS. Đặng Đức Long, Trường Đại học Bách khoa, Đại học Đà Nẵng