UED Journal of Sciences, Humanities & Education – ISSN 1859 - 4603

TẠP CHÍ KHOA HỌC XÃ HỘI, NHÂN VĂN VÀ GIÁO DỤC

SỬ DỤNG KỸ THUẬT TÍNH TOÁN MỀM DỰ ĐOÁN CẤU TRÚC BẬC HAI CỦA RNA

Đoàn Duy Bình

Nhận bài:
21 – 09 – 2015
Chấp nhận đăng:
30 – 11– 2015
http://jshe.ued.udn.vn/

Tóm tắt: Dự đoán cấu trúc của RNA đóng vai trò quan trọng trong nghiên cứu các qúa trình của tế bào. Nhiều thuật toán đã được phát triển trong hai thập kỷ qua để dự đoán cấu trúc của chuỗi RNA đã biết trình tự sắp xếp nucleotide, nhưng đến nay vẫn còn nhiều vấn đề tồn tại. Phương pháp tiếp cận bằng toán mềm đã nhận được sự quan tâm của các nhà khoa học trong việc giải quyết các trường hợp phức tạp của chủ đề này. Ở đây, chúng tôi mô tả các khái niện cơ bản của RNA và các yếu tố khác biệt về cấu trúc, cũng như một số kỹ thuật tính toán mềm được phát triển để dự đoán cấu trúc bậc hai của RNA. Trong bài báo này, chúng tôi trình bày các kết quả nghiên cứu về việc sử dụng thuật toán ACO (Ant Colony Optimization) đã cải tiến để dự đoán cấu trúc bậc hai RNA, đồng thời đưa ra hướng nghiên cứu tiếp theo cần giải quyết.

Từ khóa: Cấu trúc của RNA; axit ribonucleic; quá trình tế bào; thuật toán tối ưu đàn kiến; tính toán mềm.

1. Đặt vấn đề

Trong suốt vài thập kỷ qua, việc xác định cấu trúc RNA đóng vai trò rất quan trọng, vì nó là cơ sở cho việc tìm hiểu bệnh di truyền và tìm ra các loại thuốc mới [1]. Bài toán dự đoán cấu trúc bậc hai của RNA là một trong những vấn đề quan trọng trong lĩnh vực nghiên cứu về sinh học phân tử. Phương pháp nhiễu xạ tia X có thể được sử dụng để xác định trực tiếp cấu trúc bậc hai của RNA. Tuy nhiên, phương pháp này khó thực hiện, tốn nhiều thời gian và giá thành cao. Vì vậy, việc phát triển các phương pháp toán học để tính toán, dự đoán cấu trúc bậc 2 của RNA là rất cần thiết.

Bài viết này đưa ra một cái nhìn tổng quan nhất định về kỹ thuật tính toán mềm dựa trên những kỹ thuật đã được phát triển trong những năm qua cho bài toán dự đoán cấu trúc bậc hai của RNA. Đầu tiên, chúng tôi mô tả các vấn đề cơ bản, liên quan đến sinh học cùng với những công việc cơ bản trong dự đoán cấu trúc. Tiếp theo, chúng tôi trình bày những công cụ tính toán mềm, đặc biệt là thuật toán ACO, từ đó đưa ra hướng nghiên cứu và phát triển các thuật toán áp dụng cho bài toán dự đoán cấu bậc hai RNA tối ưu trong tương lai.

*** Liên hệ tác giả**
Đoàn Duy Bình
Trường Đại học Sư phạm, Đại học Đà Nẵng
Email: doanduybinh@gmail.com

2. Cấu trúc bậc hai của RNA

Cấu trúc bậc hai của phân tử RNA là sự sắp xếp bền vững trong không gian (2 chiều) của các nucleotide cơ bản dựa trên việc cuộn của mạch phân tử polymer và cặp đôi (tạo liên kết không hóa trị) giữa các nucleotide trong mạch đó. Cấu trúc bậc hai của RNA là nền tảng để tạo thành cấu trúc bậc ba hoàn chỉnh trong không gian 3 chiều của phân tử này và là yếu tố quyết định tính chất, chức năng của nó. Người ta đã chứng minh được rằng đối với các phân tử RNA có chức năng giống nhau thì cấu trúc bậc hai của chúng được bảo tồn [1].

Mỗi phần tử RNA biểu diễn một chuỗi dài các đơn phân gọi là các nucleotide và mỗi nucleotide chứa một base (bất kỳ trong các loại sau: A (Adenine), C (Cytosine), G (Guanine) và U (Uracil). Theo truyền thống, cấu trúc bậc hai RNA được mô hình hóa như một cây. Sau đó, cấu trúc RNA được xem như một chuỗi đặc biệt và được gọi là mô hình chuỗi [2, 6]. Một dãy cụ thể

của các base dọc theo chuỗi được gọi là cấu trúc chính của phân tử. Các cấu trúc thường được mô phỏng như một từ qua các chữ cái A, U, G và C. Thông qua việc tạo ra hai nhóm các liên kết hydro của các cặp base bổ sung A-U và C-G dạng cặp base ổn định và được gọi là các cặp Watson-Crick, trong khi cặp A-U hình thành hai liên kết hydro, cặp C-G hình thành ba liên kết hydro và có xu hướng ổn định hơn những cặp A-U. Những base khác cũng đôi khi ghép cặp, đặc biệt là G-U. Các cặp G-U được gọi là các cặp base chao đảo và hình thành chỉ là một liên kết hydro.

Để mô tả rõ bài toán cấu trúc bậc hai RNA, cần thiết tìm hiểu một số định nghĩa của cấu trúc RNA [2].

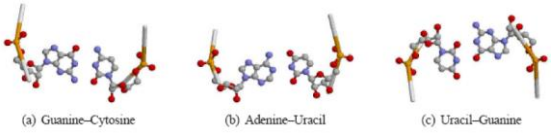
2.1. Đinh nghĩa 1

Bốn chữ cái được sử dụng để biểu diễn cho một chuỗi RNA, đó là cấu trúc chính của RNA:

$S= s_1s_2…s_n$ với $s_i \in \{A, U, G, C\}$ và i=1, 2,… n

2.2. Đinh nghĩa 2 (Các cặp base chính tắc)

Trong một cấu trúc bậc hai của RNA, các cặp base được hình thành như là một trong ba cặp: C-G (G-C), A-U (U-A) và G-U (U-G). Các cặp base ∈{(A, U), (U, A), (C, G), (G, C)} gọi là các cặp Watson-Crick. Cặp base {(G, U), (U, G)} được gọi cặp base lắc lư (Wobble).



Hình 1. *Các cặp base chính tắc*

2.3. Định nghĩa 3

Với (i, j) được biểu diễn cho cặp base hình thành bởi các base tại vị trí thứ i và base tại vị trí thứ j, sao cho một tập con của s = {(i, j), 1 ≤ i ≤ j ≤ n} gọi là cấu trúc bậc hai RNA nếu s thỏa mãn các điều kiện sau:

1.(i, j) là một cặp base chính tắc

2.Cho (I, j)∈ s, (i', j') ∈ s, nếu i ≤ i' ≤ j ≤ j' thì i=i'

3. Nếu (i, j) ∈ s, thì j-i> 3

2.4. Định nghĩa 4

Chúng ta có thể gọi hai cặp base (i, j) và (i', j'), tương thích nếu:

1.i=i' và j=j' (chúng cùng một cặp base)

2.i<j<i'<j' ((i, j) trước (i',j')) hoặc

3.i<i'<j'<j ((i, j) bao gồm (i',j'))

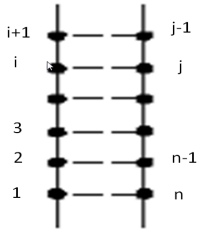
2.5. Định nghĩa 5

Cấu trúc bậc hai RNA không có các nút thắt (pseudoknot): - là một cấu trúc bậc hai RNA trong đó không có hai cặp khác biệt (i, j) và (k, l) thỏa mãn i≤k≤j≤l. Hình 3 biểu diễn các cấu trúc bậc hai RNA không có nút thắt.

2.6. Định nghĩa 6

2.6.1. Xếp chồng cặp base

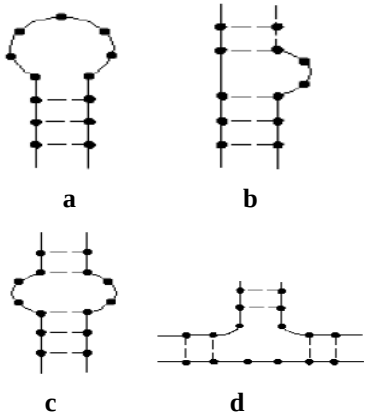
Nếu một cặp base (i, j) ∈ P và (i+1, j-1) ∈ P sẽ tạo thành xếp chồng như được biểu diễn ở Hình 2.



Hình 2. *Các cặp base xếp chồng*

2.6.2. Vòng lặp

Là một bộ *(i, i+1,… k),* trong đó $\forall i \le j \le k$, các base S_j không tạo thành cặp với các base còn lại thì sẽ hình thành vòng lặp. (Hình 3 là các kiểu vòng lặp)



Hình 3. *Các kiểu liên kết tạo thành vòng:*
a. Vòng kẹp tóc (Hairpin), b. Vòng lồi ra (Bulge),
c. Vòng bên trong (Internal),
d. Vòng nhiều nhánh (Branch)

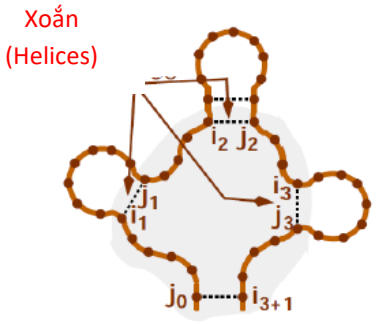
2.6.3. Vòng nhiều nhánh

Bao gồm nhiều cặp base $(i_1, j_1)…(i_k, j_k) \in P$ và một cặp base đóng $(i_0, j_{k+1}) \in P$, với thuộc tính sau:

$$0 \le l \le k : (j_l < i_l+1)$$

$\forall$ 0 ≤ l, l' ≤ k là đúng thì không có một cặp (i', j') ∈ P với $i' \in [j_l \dots i_{l+1}]$ và $j' \in [j_{l'} \dots i_{l'+1}]$

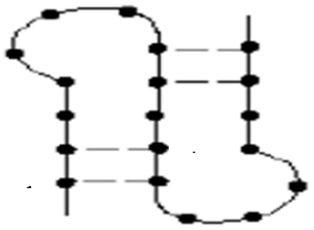
$(i_1,j_1)\ldots(i_k,j_k)$ được gọi là các xoắn (helices) của vòng nhiều nhánh



Hình 4. *Vòng nhiều nhánh*

2.7. Định nghĩa 6

Cấu trúc bậc hai RNA có các nút thắt (pseudoknot): là cấu trúc bậc hai mà ở đó tồn tại ít nhất hai cặp base *(s, t)* và *(u, v)*, mà $s<u<t<v$ (chúng thường gọi là các cặp base chéo), với $1<s<u<t<v<n$



Hình 5. *Cấu trúc RNA có nút thắt (pseudoknot)*

3. Bài toán dự đoán cấu trúc bậc hai chuỗi RNA

Bài toán tìm cấu trúc bậc hai của RNA được định nghĩa bởi bài toán tìm $\hat{P}$ của chuỗi RNA *S* với $S \in \{A, U, G, C\}^*$ có chiều dài là $n = |S|$, sao cho tổng năng lượng liên kết E đạt đến mức tối ưu nhất (tức năng lượng tự do có giá trị nhỏ nhất).

Năng lượng trên chuỗi S chính là năng lượng đóng góp của các thành phần trong cấu trúc như sau:

- Vòng kẹp tóc (Hairpin Loop): eH (i,j)
- Xếp chồng (Stack): eS (i,j, i+1, j-1)
- Vòng bên trong (Internal loop): eI(i,j,i',j')
- Vòng lồi ra (Bulge): eB(i,j)
- Vòng nhiều nhánh (MultiLoop): eM (j_0, i_1, j_1, i_k, j_k, i_{k+1})

Năng lượng eM được tính như sau:

$$eM = a + bk + ck' \quad (1)$$

a, b, c là các trọng số, với:

a = năng lượng của cặp base đóng vòng

k = số lượng xoắn (helices)

k'= số lượng base không ghép cặp trong vòng.

b, c = năng lượng tương ứng với k và k'.

Năng lương của các thành phần trong cấu trúc được tính như sau:

$$E_{i,j}^{P} = \sum_{(i,j)\in P} e(i,j) \quad (2)$$

Trong đó e (i, j) là số năng lượng tự do liên quan đến cặp base (i, j). e(i,j) được xác định bởi phương pháp nhiệt động học phân tử.

Năng lượng của cấu trúc bậc hai trên chuỗi RNA S và tập cấu trúc P sẽ là:

$$E(P) = \sum_{(i,j)\in P} E_{i,j}^{P} \quad (3)$$

Giá trị năng lượng tự do được dự đoán (kcal/mol tại 37°C) cho các cặp base trong xếp chồng được thể hiện ở Bảng 1 [10].

Bảng 1. *Giá trị năng lượng các cặp trong xếp chồng*

	A/U	**C/G**	**G/C**	**U/A**	**G/U**	**U/G**
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

Giá trị năng lượng tự do được dự đoán (kcal/mol tại 37°C) cho các thành phần trong cấu trúc bậc hai, bởi kích thước vòng lặp [10]:

Bảng 1. *Giá trị năng lượng cho các kiểu vòng với số lượng base tương ứng*

Kích thước	Vòng bên trong (Internal)	Vòng lồi ra (Bulge)	Vòng kẹp tóc (Hairpin)
1	-	3.8	-
2	-	2.8	-
3		3.2	5.4
4	1.1	3.6	5.6
5	2.0	4.0	5.7
6	2.0	4.4	5.4
7	2.1	4.6	6.0
8	2.3	4.7	5.5
9	2.4	4.8	6.4
10	2.5	4.9	6.5
11	2.6	5.0	6.6
12	2.7	5.1	6.7
13	2.8	5.2	6.8
14	2.9	5.3	6.9
15	2.9	5.4	6.9
16	3.0	5.4	7.0
17	3.1	5.5	7.1
18	3.1	5.5	7.1
19	3.2	5.6	7.2
20	3.3	5.7	7.2
21	3.3	5.7	7.3
22	3.4	5.8	7.3
23	3.4	5.8	7.4
24	3.5	5.8	7.4
25	3.5	5.9	7.5
26	3.5	5.9	7.5
27	3.6	6.0	7.5
28	3.6	6.0	7.6
29	3.7	6.0	7.6
30	3.7	6.1	7.7

4. Tính toán mềm

Tính toán mềm (Soft Computing) [11] khác với tính toán cứng truyền thống (Hard Computing) ở chỗ tính toán mềm cho phép sự không chính xác, tính bất định, gần đúng, xấp xỉ trong tính toán. Các mô hình tính toán mềm thường dựa vào kinh nghiệm của người thực hiện trong việc sử dụng dung sai cho phép, tính bất định, gần đúng, xấp xỉ để tìm lời giải hiệu quả. Trong thực tế cuộc sống, các bài toán liên quan đến hoạt động nhận thức, trí tuệ của con người đều hàm chứa những đại lượng, thông tin mà bản chất là không chính xác, không chắc chắn, không đầy đủ. Ví dụ: sẽ chẳng bao giờ có các thông tin, dữ liệu cũng như các mô hình toán đầy đủ và chính xác cho các bài toán dự báo thời tiết. Nhìn chung, con người luôn ở trong bối cảnh không có thông tin đầy đủ và chính xác cho các hoạt động để ra quyết định của bản thân mình.

Trong lĩnh vực khoa học kỹ thuật cũng vậy, các hệ thống phức tạp trên thực tế thường không thể mô tả đầy đủ và chính xác bởi các phương trình toán học truyền thống. Kết quả là những cách tiếp cận kinh điển dựa trên kỹ thuật tính toán, phân tích và tính toán bằng các phương trình toán học nhanh chóng tỏ ra không còn phù hợp. Vì vậy, kỹ thuật tính toán mềm sẽ giúp giải quyết những bài toán mà bằng phương pháp tính toán thông thường không giải quyết được.

Một số đặc điểm của tính toán mềm:

- Tính toán mềm căn cứ trên các đặc điểm, hành vi của con người và tự nhiên để đưa ra các quyết định hợp lý trong điều khiển không chính xác và không chắc chắn.

- Các thành phần của tính toán mềm có sự bổ sung, hỗ trợ lẫn nhau.

- Tính toán mềm là một hướng nghiên cứu mở, bất kỳ một kỹ thuật mới nào được tạo ra từ việc bắt chước trí thông minh của con người đều có thể trở thành một thành phần của tinh toán mềm.

Tính toán mềm bao gồm 3 thành phần chính:

1. Điều khiển mờ;
2. Mạng nơ-ron nhân tạo;
3. Lập luận xác suất.

Những giải thuật trong tính toán mềm liên quan đến lập luận xác xuất bao gồm giải thuật luyện kim (simulated annealing - SA), giải thuật di truyền (genetic algorithm - GA)[8,9], giải thuật đàn kiến (Ant Colony Optimization - ACO),… SA xuất phát từ phương thức xác suất và kỹ thuật luyện bao gồm việc nung và điều khiển làm nguội các kim loại để đạt được trạng thái năng lượng nhỏ nhất. Trong khi đó, giải thuật di truyền dựa trên ý tưởng từ cơ chế di truyền trong sinh học và tiến trình tiến hóa trong cộng đồng các cá thể của một loài. Giải thuật đàn kiến sử dụng chiến lược của kiến trong thế giới thực để giải bài toán tối ưu.

5. Tính toán mềm trong dự đoán cấu trúc bậc hai RNA

5.1. Thuật toán ACO

ACO (Ant Colony Optimization) [6] – 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.

Bắt nguồn từ những con kiến trong tự nhiên, thông qua 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 từ đỉnh i đến đỉnh 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:

$$\tau_{ij} = \tau_0 = \frac{m}{C^{nn}}, \quad (4)$$

trong đó:

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

- m: số lượng kiến,

- C^{nn}: chiều dài của đường đi, được tạo ra bởi láng giềng 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 suất được xác định theo công thức sau:

$$p_{ij}^{k} = \frac{\left[\tau_{ij}\right]^{\alpha}\left[\eta_{ij}\right]^{\beta}}{\sum_{l \in N_i^k}\left[\tau_{il}\right]^{\alpha}\left[\eta_{il}\right]^{\beta}}, j \in N_i^k \quad (5)$$

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:

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (6)$$

trong đó dij là khoảng cách giữa đỉnh i và đỉnh j,

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

N_i^k : tập các đỉnh láng giềng của i mà con kiến k chưa đi qua.

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

Thuật toán kiểm tra xem một chuỗi có phù hợp không được thể hiện như sau: xác định một tập tất cả các thân (stem) từ chuỗi chứa các nucleotide. S là chiều dài tối thiểu của mỗi thân (stem) (thường là 3), L là chiều dài tối thiểu của một vòng lặp được hình thành bởi một thân (stem) (thường là 3). Chú ý rằng trong khi vòng lặp làm việc thì kiểm tra giới hạn của k, và thoát khỏi vòng lặp khi chỉ số k vượt ra khỏi hai đầu của dãy.

Mã giả cho thuật toán xác định thân như sau:

Stems ← Khởi tạo rỗng

for *i* = 0 …n **do** {n chiều dài chuỗi}

for *j* = *i* + 2S + L – 1 … n **do**

Khởi tạo *k* = 0 {Biến đếm cho kích thước của thân}

while cặp base hợp lệ giữa (*i*+*k*) và (*j*-*k*) **do** Inc (*k)*

if *k* >= *S* **then**

Stems ← Chèn thân (*i, j, k*)

endif

endwhile

endfor

endfor

Sau khi xác định thân (stem) trong vùng chứa nó là cần thiết để xác định một đại diện thích hợp của bài toán để mô tả làm thế nào những con kiến có thể tạo các tổ hợp tốt nhất của các thân mà không loại trừ lẫn nhau. Có thể đưa ra hai giải pháp cho bài toán dự đoán cấu trúc bậc hai của chuỗi RNA trong bài báo này là: Chuyển tất cả các phần tử của chuỗi đã được kiểm tra vào một ma trận năng lượng, sau đó chuyển bài toán thành bài toán tìm đường đi ngắn nhất trên đồ thị, với mỗi đỉnh là các gốc base, cạnh là các liên kết có thể chấp nhận được giữa các base. Sau đó, ứng dụng thuật ACO để áp dụng vào bài toán.

Một giải pháp cho tất cả các con kiến là xác định được cấu trúc bậc hai từ việc xây dựng trực tiếp các tập thân chấp nhận được. Kiến sẽ liên lục thêm xác suất các gốc phù hợp vào trong cấu trúc cho đến khi không thể bổ sung những thân có thể. Kết quả cuối cùng là không chứa các pseudoknot và không chứa các thân loại trừ lẫn nhau.

Xây dựng một tập các thân phụ thuộc vào các thiết lập đã được xác định. Xác suất để một con kiến k chọn gốc i được thể hiện qua công thức sau:

$$p_k(i,j) = \frac{[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}}{\sum_{g \in N_i^k}[\tau(i,g)]^{\alpha}.[\eta(i,g)^{\beta}} \quad (7)$$

Trong đó:

- N_i^k là láng giềng của con kiến k, được định nghĩa là một tập các gốc mà đáp ứng được các điều kiện sau:

1. Thân không thực sự có trong một phần giải pháp của k,

2. Thân không xung đột với bất kỳ gốc nào đã thực sự có trong giải pháp cục bộ của k,

3. Thân không tạo thành một giả nút (pseudo-knot) khi thêm vào cấu trúc.

Mã giả để kiểm tra tính tương thích của thân đích t với thân trong cấu trúc k. Trong phạm vi của thân có nghĩa là chỉ số thấp nhất và cao nhất của nucleotide tạo thành cặp base trong thân.

For chọn một gốc s trong giải pháp cục bộ của k **do**

if s = t then

return sai, trùng lặp

else if bất kỳ cặp base nào nằm giữa s và t **then**

return sai, xung đột

else if (chồng chéo khoảng giữa các gốc s và t) và (s không lồng trong t và t không lòng trong s) **then**

return sai, giả nốt

endif

endfor

return success

5.3. Cài đặt thuật toán cho bài toán

5.3.1. Dữ liệu vào và kết quả ra của bài toán

Dữ liệu vào (Input):

- Chuỗi RNA, bao gồm các phân tử (A, U, G, C)

- Bảng năng lượng được xây dựng thông qua các file.

- Số lượng đàn kiến cho thuật toán

- Hằng số bay hơi: ρ

- Lượng pheromone và các hệ số điều chỉnh ảnh hưởng α , β

Kết quả đầu ra (Output):

- Cấu trúc bậc hai của chuỗi RNA với mức năng lượng nhỏ nhất (âm nhất)

5.3.2. Các thông số đầu vào của thuật toán

Thuật toán ACO được thực hiện với những mô tả trong bài báo. Trừ khi có điều ngược lại được nêu ra thì tất cả thí nghiệm đều được chạy bằng cách sử dụng các tham số sau:

Hằng số bay hơi: $\rho = 0.6$

Ma trận pheromone được khởi tạo với các giá trị ngẫu nhiên từ 1 đến 6.

Lượng pheromone thêm vào là 9

α : hệ số điều chỉnh ảnh hưởng của pheromone, có giá trị là 1

β : hệ số điều chỉnh ảnh hưởng của heuristic, có giá trị là 1

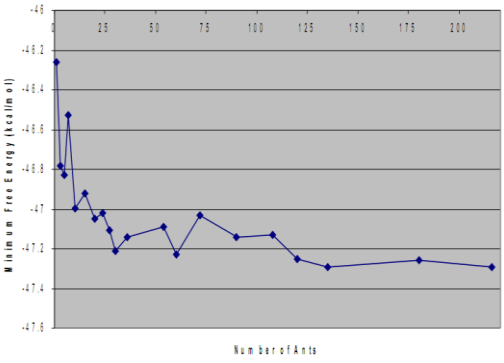
Kích thước nhỏ nhất của một vòng lặp là 3

Các chuỗi được sử dụng là tất cả các ribosomal RNA từ các sinh vật khác nhau.

Những tham số này đã được chọn lọc trong quá trình thực nghiệm với thuật toán ACO cho bài toán dự đoán cấu trúc bậc hai RNA.

5.3.3. Số kiến

Với thuật toán ACO một nơi chắc chắn để bắt đầu điều tra là thay đổi số lượng kiến nhằm xác định tính chất quan trong của thuật toán là dựa trên mật độ của kiến. Số lượng kiến sử dụng trong bài toán thay đổi từ 1 đến đến 220 và số lần lặp thay đổi từ 1.100 xuống 5 để bù đắp cho số lượng của kiến.



Hình 6. *Số lượng kiến biến đổi từ 1 đến 220. Chuỗi là: E.Coli*

So sánh kết quả khi sử dụng phương pháp quy hoạch động (Dynamic Programming – DP) [7].

Một hai thí nghiệm được thực hiện bằng cách sử dụng cùng một chuỗi RNA và hai phương pháp dự đoán cấu trúc bậc hai. Tất cả ACO chạy bằng hiệu suất của thuật toán quy hoạch động về năng lượng tự do nhỏ nhất, mặc dù có những thay đổi nhỏ về chất lượng của base phù hợp và cặp base phù hợp. Nhìn chung, thuật toán ACO ngang tầm với thuật toán quy hoạch động để xác định cấu trúc tối ưu của chuỗi E.Coli, dài 120 nucleotide. Tuy nhiên, với một số cấu trúc khác thì vẫn chưa tối ưu (Bảng 3).

Bảng 3. *Kết quả so sánh khi sử dụng thuật toán ACO và DP*

Cấu trúc	Thuật toán	Độ dài	Kích thước của Pun gốc	Năng lượng tự nhiên (kcal/ mol)	Dự đoán năng lượng tự do (kcal/ mol)	Phần trăm cặp base phù hợp	Thời gian (s)
S. Cerevisiae	**ACO**	118	582	-44.1	**-52.8**	67.0	**12**
	DP				**-54.1**	66.7	**0.1**
E. Coli	**ACO**	120	550	-47.0	**-51.5**	25.0	**5**
	DP				**-51.5**	25.0	**0.1**
H. Rubra	**ACO**	543	16289	-114.7	**-190.7**	27.3	**598**
	DP				**-204.5**	41.3	**.04**
T. Thermophila	**ACO**	506	10296	-97.2	**-159.3**	39.6	**371**
	DP				**-177.4**	67.8	**0.4**
C. Elegans	**ACO**	697	34617	Không biết	**-107.5**	15.8	**1640**
	DP				**-142.5**	18.6	**2**

6. Kết luận

Thuật toán ACO đã thể hiện nhiều ưu điểm trong việc giải bài toán tối ưu tổ hợp; nó đã được sử dụng để giải quyết vấn đề dự đoán cấu trúc bậc 2 của RNA. Trong nghiên cứu này, chúng tôi đã áp dụng thuật toán ACO vào bài toán dự đoán cấu trúc bậc 2 chuỗi RNA nhưng mới chỉ dừng lại với những cấu trúc không xoắn và đã thu được kết quả thực nghiệm khá tốt. Hướng nghiên cứu tiếp theo của chúng tôi là tối ưu hóa thuật toán ACO để thực hiện với những cấu trúc xoắn, cũng như kết hợp với một số thuật toán tiến hóa khác để xây dựng thuật toán ACO mới, áp dụng cho lớp các bài toán sinh học phân tử.

Tài liệu tham khảo

[1] M. Neethling and A.P. Engelbrecht (2006), "Detennining rna secondary structure using set-based particle swann optimization", IEEE Congress on Evolutionary Computation, pp. 6134-41.

[2] Q. Liu, X. Ye, and Y. Zhang (2006), "A hopfield neural network based algorithm for rna secondary structure prediction", Proc. oj the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06), pp. 1-7.

[3] 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.

[4] V. Batenburg, A. P. Gultyaev, and C. W. A. Pleij (1995), "An APL-programmed genetic algorithm for the prediction of RNA secondary structure", Journaloj Theoritical Biology, vol. 174, no. 3, pp. 269-280.

[5] Doan Duy Binh (2010), "Application of Meta-heuristic algorithm for a search of shortest path", University of Da Nang Journal of Science and Technology, 5(40)/2010 (1): 9-16

[6] Marco Dorigo, Thomas Stűtzle (2004), Ant Colony Optimization, Massachusetts Instituteof Technology.

[7] Rivas E, Eddy SR. (1999), "A dynamic programming algorithm for RNA structure prediction including pseudoknots". J. Mol. Biol 1999, 285:2053-2068.

[8] Shapiro B., Navetta J. (1994), "A massively parallel genetic algorithm for RNA secondary structure prediction". The Journal of Supercomputing.vol.8,195–207

[9] Shapiro B.A., Wu J.C., Bengali D. and Potts M.J. (2001) "The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation". Bioinformatics. 17. 137-148.

[10] Freier, S. M., Kierzek, R., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986), Biochemistry 25, 3209-3213.

[11] Mahmoud ElHefnawi and Mohamed Mysara (2012), Recurrent Neural Networks and Soft Computing, Janeza Trdine 9, 51000 Rijeka, Croatia.

USING A SOFT COMPUTING TECHNIQUE TO PREDICT THE RNA SECONDARY STRUCTURE

Abstract: Prediction of an RNA structure plays an important role in studying cellular processes. Over the last two decades, many algorithms have been developed to predict the structure of an RNA sequence with a known nucleotide order; however, problems have still remained until now. The soft computing approach has gained attention of researchers in solving complex cases of this topic. Here we describe the basic concepts of RNA and its distinctive structural elements, as well as some of the soft computing-based techniques developed for RNA secondary structure prediction. In the paper, we present the results of our research on the use of the Ant Colony Optimization (ACO) algorithm which has been improved to predict the RNA secondary structure, then introduce approaches for further research.

Key words: RNA structure; ribonucleic acid; cellular processes; Ant Colony Optimization; soft computing.