Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.0007

CẢI TIẾN THUẬT TOÁN DI TRUYỀN VÀ ÁP DỤNG DỰ ĐOÁN CẤU TRÚC BẬC HAI RNA

Đoàn Duy Bình[1], Phạm Minh Tuấn[2], Đặng Đức Long[3]

[1] Khoa Tin học, Trường Đại học Sư phạm, Đại học Đà Nẵng
[2] Khoa Công nghệ Thông tin, Trường Đại học Bách khoa, Đại học Đà Nẵng
[3]Viện Nghiên cứu và Giáo dục Việt Anh, Đại học Đà Nẵng

doanduybinh@gmail.com, pmtuan@dut.udn.vn, long.dang@vnuk.edu.vn

***TÓM TẮT**: Cấu trúc RNA là một lĩnh vực nghiên cứu quan trọng. RNA là trung tâm trong nhiều giai đoạn tổng hợp protein, cũng có vai trò cấu trúc và chức năng trong tế bào. Dự đoán cấu trúc RNA có thể vượt qua được nhiều vấn đề với việc xác định cấu trúc vật lý. Hiện tại các phương pháp vật lý cho dự đoán cấu trúc RNA là tốn nhiều thời gian và chi phí, do đó các phương pháp cho việc dự đoán tính toán được ưa chuộng. Các thuật toán khác nhau đã được sử dụng để dự đoán cấu trúc RNA bao gồm cả quy hoạch động và các phương pháp so sánh. Trong bài báo này chúng tôi giới thiệu thuật di truyền là một phương pháp được áp dụng cho bài toán dự đoán cấu trúc RNA. Thuật toán di truyền mà chúng tôi áp dụng là đã được cải tiến một số tham số trong các toán tử di truyền như: chọn lọc, lai ghép và đột biến. Cuối cùng nhóm nghiên cứu đã so sánh và đánh giá với những thuật toán liên quan cho bài toán dự đoán cấu trúc bậc hai, đó là thuật toán quy hoạch động.*

***Từ khóa**: RNA; Thuật toán di truyền; Quy hoạch động; Cấu trúc bậc hai; Dự đoán;*

I. GIỚI THIỆU

Ribonucleic acid (RNA) là một phân tử sinh học quan trọng. Nó đóng một vai trò chính trong quá trình tổng hợp protein từ deoxyribonucleic acid (DNA). Nó cho phép hiểu về vai trò cấu trúc và xúc tác trong tế bào [1]. Mỗi phân tử RNA bao gồm một chuỗi ribonucleotide được liên kết với nhau bởi liên kết hóa học cộng hóa trị và mỗi ribonucleotide chứa một trong bốn base: adenine (A), guanine (G), cytosine (C) và uracil (U), nó có thể tự gấp lại để hình thành cấu trúc bậc hai với các cặp base A≡U, G=C và G-U, các đối xứng của chúng là U≡A, C=G và U-G. Những cặp base này gọi là những cặp chính tắc. Một chuỗi RNA có thể gấp lại để hình thành nhiều cấu trúc bậc hai khác nhau. Việc xác định một cấu trúc bậc hai chính xác được gọi là bài toán dự đoán cấu trúc bậc hai [2].

Một trong những phương pháp đơn giản dự đoán cấu trúc bậc hai của một chuỗi RNA là dựa trên số lượng các cặp base. Waterman và các cộng sự đã thiết kế hai thuật toán quy hoạch động đơn giản tìm số lượng tối đa các cặp base trong một chuỗi RNA [3]. Một số nhà nghiên cứu đã cố gắng căn chỉnh một vài chuỗi RNA với thông tin cấu trúc bậc hai. Các phương pháp căn chỉnh và gấp chuỗi được tiến hành cùng lúc [4]. Một số cấu trúc bậc hai có nút thắt (pseudoknots) là phức tạp hơn và khó khăn trong việc phân tích. Để mà dự đoán cấu trúc bậc hai RNA có nút thắt, một số nhà nghiên cứu đã thiết kế ra mô hình cây ngữ pháp để biểu diễn cho cấu trúc này [6], [7], [8]. Akutsu đã sử dụng quy hoạch động giải quyết bài toán dự đoán cấu trúc bậc hai có nút thắt đơn giản, trong $\Theta(n^4)$ thời gian. Tabaska và Stormo đã điều chỉnh chuỗi RNA với cấu trúc có nút thắt trong thời gian đa thức. Đáng tiếc, các phương pháp này có thể chỉ dự đoán cấu trúc của một chuỗi RNA có chiều dài không quá 200 phân tử, trong thời gian chấp nhận được.

Thuật toán di truyền đã được áp dụng trong lĩnh vực dự đoán cấu trúc bậc hai RNA bắt đầu vào đầu những năm 90 của thế kỷ trước [9]. Kể từ đó, có rất nhiều cải tiến trong việc sử dụng thuật toán di truyền (genetic algorithm - GA) vào bài toán dự đoán cấu trúc bậc hai RNA. Một thuật toán di truyền nhanh cho dự đoán cấu trúc bậc hai RNA là GArna và được giới thiệu trong [10]. Đây là thuật toán được sử dụng cho việc tính toán nhanh cấu trúc bậc hai RNA, kết quả của nó được sử dụng để giải thích cấu trúc bậc hai RNA ảnh hưởng như thế nào đến sự khởi tạo dịch mã và sự điều chỉnh biểu diễn.

Với bài toán dự đoán cấu trúc bậc hai RNA, thuật toán di truyền là thiết thực hơn vì nó có thể giải quyết cho bài toán với kích thước lớn. Trong bài báo này, chúng tôi giải quyết bài toán dự đoán cấu trúc bậc hai RNA với thuật toán di truyền đã được cải tiến. Chúng tôi sẽ thêm vào một số tham số trong toán tử di truyền để tăng độ chính xác của việc tìm ra cấu trúc bậc hai RNA tối ưu. Kết quả thực nghiệm của chúng tôi cho thấy các toán tử mới của chúng tôi có những cải tiến lớn so với các toán tử truyền thống khác.

Trong bài báo này chúng tôi trình bày những vấn đề sau, đầu tiên là các khái niệm và định nghĩa về cấu trúc bậc hai RNA được trình bày ở mục II. Mục III chúng tôi giới thiệu về mô hình năng lượng tự do và các mô hình khác. Giới thiệu về thuật toán GA ở mục IV. Mục V kết quả của việc áp dụng thuật toán GA vào bài toán dự đoán cấu trúc bậc hai RNA. Cuối cùng, mục 6 là những đánh giá về kết quả và trình bày ngắn gọn các công việc sẽ tiếp tục thực hiện trong tương lai.

II. CẤU TRÚC BẬC HAI RNA

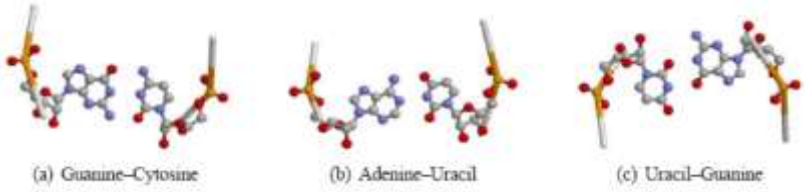
Các phân tử RNA được mô tả đặc điểm bởi một chuỗi của bốn loại ribonucleotide hoặc là các base: Adenine (A), Cytosine (C), Guanine (G) và Uracil (U). Một chuỗi tuyến tính các base của sợi RNA tạo thành cấu trúc chính hoặc chuỗi. Dưới đây mà một số định nghĩa làm sáng tỏ về chuỗi RNA và cấu trúc bậc hai RNA như sau:

A. Định nghĩa 1

Một chuỗi RNA có chiều dài là n ribonucleotide là một chuỗi $X = x_1 x_2 ... x_n$, trong đó $x_i \in \{A, C, G, U\}, \forall i \in \{1, ..., n\}$

B. Định nghĩa 2

Trong một cấu trúc bậc hai của RNA, các cặp base được hình thành là một trong ba cặp như sau: 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 không bền vững (Wobble). Sự ổn định của các cặp base được cho bởi mối quan hệ sau đây: C-G> A-U ≥ G-U [19], [20].



Hình 1. Các cặp base chính tắc

C. Định nghĩa 3

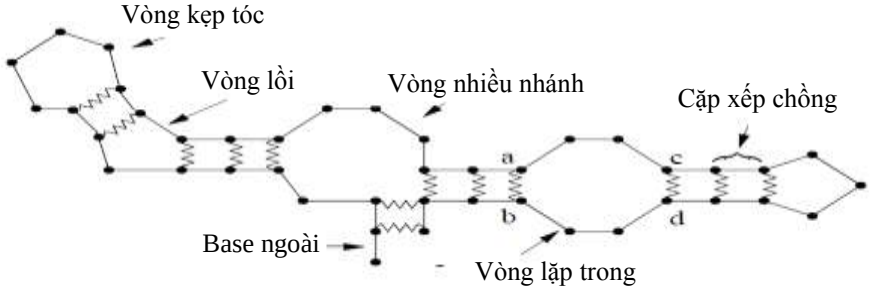
Với (i, j) được biểu diễn cho cặp base hình thành bởi base tại vị trí thứ i và base tại vị trí thứ j, sao cho một tập con của $y = \{(i, j), 1 \le i \le j \le n\}$ gọi là cấu trúc bậc hai RNA nếu s thoả mãn các điều kiện sau:

- (i, j) là một cặp base chính tắc
- Cho $(i, j) \in y$, $(i', j') \in y$, nếu $i \le i' \le j \le j'$ thì $i = i'$
- Nếu $(i, j) \in y$, thì $i - j > 3$
- Mỗi base chỉ ghép cặp với một và chỉ một base khác trong cấu trúc.

D. Định nghĩa 4

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

- *i=i'* và *j=j'* (chúng cùng một cặp base)
- *i<j<i'<j'* (*(i, j)* trước *(i', j')*) hoặc
- *i<i'<j'<j* (*(i, j)* bao gồm *(i', j')*)



Hình 2. Cấu trúc RNA không có nút thắt (pseudoknot -free)

Bước đầu tiên trong việc tìm hiểu các cấu trúc bậc hai RNA là xác định các cấu trúc con mà chúng được tạo ra, chúng ta gọi là các hoạ tiết cấu trúc RNA. Các hoạ tiết cấu trúc chúng ta xem xét trong công việc này là như sau:

- Vòng kẹp tóc (Hairpin loop) chứa đúng một cặp base,
- Vòng lặp trong (Internal loop) chứa đúng 2 cặp base,
- Vòng lồi (Bulge loop) là trường hợp đặc biệt của vòng lặp trong mà không có base tự do ở một bên và có ít nhất một base tự do ở phía bên kia,
- Cặp xếp chồng (Stack pair) là một vòng được hình thành bởi hai cặp base (i, j) và (i+1, j-1) liên tiếp nhay, vừa là kết thúc vừa là liên kết trên trục chính,

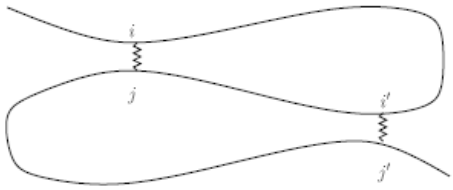
- Vòng nhiều nhánh (Multibranched loop) là vòng lặp chứa nhiều hơn hai cặp base
- Base bên ngoài (External base) là không nằm trong bất kỳ vòng nào.

E. Định nghĩa 5

Cấu trúc bậc hai RNA không có các nút thắt (pseudoknot free) y tương ứng với chuỗi RNA x có độ dài n là cấu trúc bậc hai RNA trong đó bất kỳ hai cặp base (i, j) và $(i', j') \in y$, chúng đều đang lồng nhau, tức là $i < i' < j' < j$, hoặc là liên tiếp nhau, tức là $i < j < i' < j'$. Ở đây chúng ta giả định mà không mất tính tổng quát rằng $i < j,\ i' < j'$ và $i < i'$.

F. Định nghĩa 6

Cấu trúc bậc hai RNA có các nút thắt (pseudoknotted) S tương ứng với chuỗi RNA s có độ dài n là cấu trúc bậc hai RNA tồn tại ít nhất hai cặp base (i, j) và $(i', j') \in S$, mà $i < i' < j < j'$ (đây thường là các cặp base giao nhau). Ở đây chúng ta giả định mà không mất tính tổng quát rằng: $i < j,\ i' < j'$ và $i < i'$.



Hình 3. Cấu trúc RNA có nút thắt (pseudoknotted)

III. MÔ HÌNH NĂNG LƯỢNG TỰ DO VÀ CÁC MÔ HÌNH KHÁC

A. Nhiệt động học và mô hình năng lượng tự do RNA

Sự ổn định của cấu trúc bậc hai RNA được định lượng bằng sự thay đổi năng lượng tự do ΔG, được đo theo đơn vị kcal/mol. Năng lượng tự do ΔG cho biết hướng của một sự thay đổi tức thời và được giới thiệu bởi J. W. Gibbs vào năm 1878 [11]. Sự thay đổi năng lượng tự do ΔG lượng hóa sự khác biệt về năng lượng tự do giữa trạng thái gấp lại của các phân tử và trạng thái không gấp. Một RNA được gấp có sự thay đổi năng lượng tự do âm và càng thấp hơn là cấu trúc càng ổn định. Sự thay đổi năng lượng tự do là một hàm của thay đổi enthalpy ΔH, thay đổi entropy ΔS và nhiệt độ T (độ Kelvin), theo Gibbs có hàm như sau:

$$\Delta G = \Delta H - T.\Delta S \quad \textbf{(1)}$$

Enthalpy (H) là một thước đo của dòng nhiệt năng xuất hiện trong một quá trình. Sự thay đổi enthalpy ΔH cho một phản ứng toả nhiệt, chẳng hạn như RNA gấp (tức là các dòng nhiệt năng từ hệ thống vào môi trường xung quanh) là âm. Enthalpy được đo bằng kcal/mol. Sự hình thành của những thân (stem/helix) RNA là nhân tố chi phối enthalpy, thông qua liên kết hydro và các tương tác xếp chồng.

Entropy (S) được chấp nhận rộng rãi như là một hàm nhiệt động học để đo sự hỗn loạn của hệ thống. Như vậy, sự thay đổi ΔS dùng để đo sự thay đổi của mức độ hỗn loạn. Nếu ΔS là dương, nó có nghĩa là sự gia tăng về mức hỗn loạn. Giá trị âm cho thấy một sự giảm sút sự hỗn loạn.

Một mô hình năng lượng tự do RNA là xây dựng một lý thuyết miêu tả được các quy tắc và các biến tùy theo các mô hình cấu trúc (bậc hai) của các chuỗi RNA. Chúng ta xem xét một mô hình năng lượng tự do RNA có ba thành phần chính:

- Một tập các đặc điểm cấu trúc $(f_1, f_2, ..., f_p)$, trong đó p là số lượng các đặc điểm của mô hình. Một đặc điểm là một mảnh của cấu trúc bậc hai RNA mà nhiệt động lực học được xem là quan trọng cho việc gấp RNA. Ví dụ: hãy xem xét một mô hình đơn giản với p = 3 đặc điểm: f_1 là đặc điểm của cặp base C-G, f_2 là đặc điểm của cặp base A-U và f_3 là đặc điểm của cặp base G-U.

- Một tập các tham số năng lượng tự do $(\theta_1, \theta_2, ..., \theta_p)$, với tham số năng lượng tự do θ_i tương ứng với đặc điểm f_i. Tham số θ_i đôi khi được ký hiệu bởi $\Delta G(f_i)$. Trong ví dụ với mô hình đơn giản có ba đặc điểm chúng tôi đưa ra ở trên, ta có thể có các giá trị tương ứng cho ba tham số như sau: θ_1 = -2.0 kcal/mol, θ_2 = -1.0 kcal/mol và θ_3 = -0.8 kcal/mol.
- Một hàm năng lượng tự do giúp xác định tính ổn định nhiệt động lực học của một chuỗi X gập lại thành cấu trúc bậc hai cụ thể S, tức là phải phù hợp với X.

Hầu hết các mô hình cho dự đoán cấu trúc bậc hai không có nút thắt (pseudoknott) với hàm năng lượng tự do của chuỗi X và cấu trúc S là tuyến tính trong các tham số θ_i, có dạng:

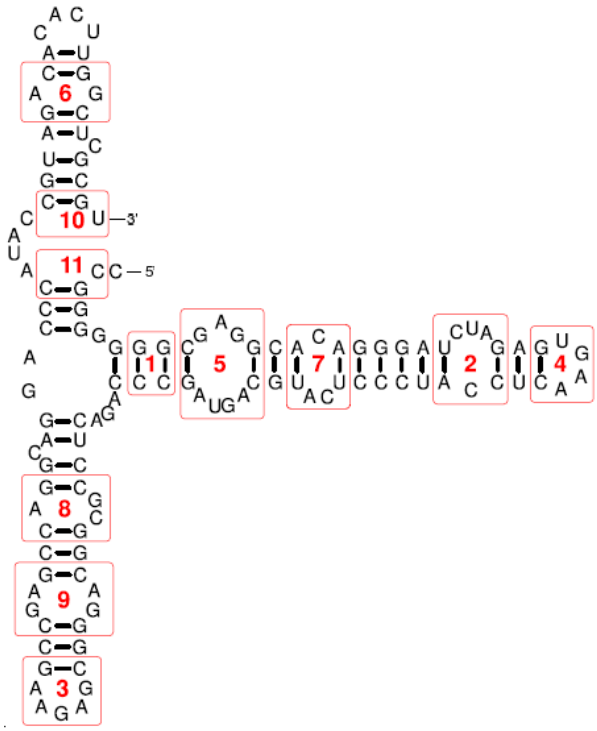
$$\Delta G(X, S, \theta) := \sum_{i=1}^{p} c_i(X, S)\theta_i = c(X, S)^{\mathrm{T}}\theta \quad \textbf{(2)}$$

trong đó,

- $\theta := (\theta_1, \theta_2, ..., \theta_p)$: biểu thị vecto các giá trị tham số θ_i
- $c_i(X,S)$ là số lần đặc điểm f_i xuất hiện trong cấu trúc bậc hai S của chuỗi X
- $c(X, S) := (c_1(X, S), ..., c_p(X, S))$ biểu thị vecto của số lượng đặc điểm $c_i(X, S)$.

B. Mô hình Turner

Với bài toán dự đoán cấu trúc bậc hai không có thắt nút (pseudoknot), thì mô hình Turner là mô hình năng lượng được sử dụng rộng rãi nhất hiện nay. Đây là mô hình được công nhận là thực tiễn về mặt sinh học, dựa trên các thí nghiệm tán quang, phổ biến nhất là sử dụng phương pháp thực nghiệm để xác định sự thay đổi năng lượng tự do của cấu trúc RNA ngắn, với sai số chuẩn từ 2-5%.



Hình 4. Cấu trúc bậc hai của chuỗi RNA tùy ý. Được đánh dấu trong các hộp màu đỏ là các hoạ tiết cấu trúc RNA, bao gồm xếp chồng (1), vòng kẹp tóc (3, 4), vòng lặp trong (2, 5, 6, 7, 8 và 9) và base ngoài (10 và 11)

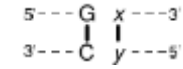
Mô hình Turner chứa các giá trị năng lượng tự do tại 37°C. Ngoài ra, các mô hình Turner chứa giá trị enthalpy và entropy cho mỗi đặc điểm của mô hình. Điều này cho phép dự đoán cấu trúc bậc hai năng lượng tự học tối thiểu hóa (hoặc chưa tối ưu) ở nhiệt độ khác nhau từ 37°C, bằng cách sử dụng công thức 1.

Năng lượng tự do cho những vòng xếp chồng (Stack loop).

Bảng 1. (a). Ví dụ của năng lượng cho những vòng xếp chồng có kiểu như (b). (c) là một giá trị cụ thể từ bảng, trong đó x=G và y=C

x \ y	A	C	G	U
A	-	-	-	-2.40
C	-	-	-3.40	
G	-	**-3.30**	-	-1.50
U	-2.20	-	-2.50	-

(a)



(b)

(c)

Những năng lượng bất ổn bởi kích thước vòng.

Bảng 2. (a) Năng lượng tự do cho từng vòng với kích thước cụ thể. (b) Một ví dụ của vòng lặp trong có chiều dài là 4

Size	Internal	Bulge	Hairpin
1	-	3.80	-
2	-	2.80	-
3	-	3.20	5.70
4	**1.70**	3.60	5.60
5	1.80	4.00	5.60
.	.	.	.
30	3.70	6.10	7.70

(a)

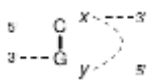
(b)

Năng lượng tự do cho các vòng kẹp tóc tổng quát.

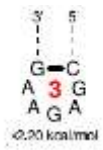
Bảng 3. (a). Bảng năng lượng tự do cho vòng kẹp tóc của kiểu trong (b). (c) là một ví dụ, trong đó $x = G$ và $y = A$

x \ y	A	C	G	U
A	-1.50	-1.50	-1.40	-1.80
C	-1.00	-0.90	-2.90	-0.80
G	-2.20	-2.0	-1.60	-1.10
U	-1.70	-1.40	-1.80	-2.00

(a)



(b)



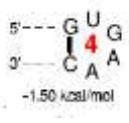
(c)

Năng lượng tự do cho vòng kẹp tóc với chiều dài là 4.

Bảng 4. (a) Ví dụ của giá trị năng lượng cho vòng kẹp tóc có độ dài 4. (b) một ví dụ cụ thể cho những vòng kẹp tóc.

Chuỗi	Năng lượng
GGGGAC	-3.00
GGUGAC	-3.00
...	…
CGAAGG	-2.50
CUACGG	-2.50
...	…
CGAGAG	-2.00
...	…
GUGAAC	-1.50
UGGAAA	-1.50

(a)



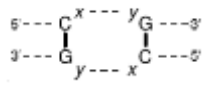
(b)

Năng lượng tự do cho vòng lặp trong tổng quát.

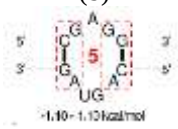
Bảng 5. (a) Bảng năng lượng cho vòng lặp trong với cặp base đóng là (C, G). Đây là kiểu được hiển thị trong hình (b). (c) là ví dụ mà ở đó có giá trị được tìm thấy trong bảng ở hình (a), ở đó $x = G, y = A$ và $x = A, y = G$

x \ y	**A**	**C**	**G**	**U**
A	00.0	0.00	**-1.10**	0.00
C	0.00	0.00	0.00	0.00
G	**-1.10**	0.00	0.00	0.00
U	0.00	0.00	0.00	0.00

(a)



(b)



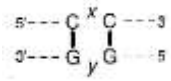
(c)

Năng lượng tự do cho vòng lặp trong đối xứng với kích thước 2.

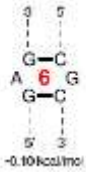
Bảng 6. (a) Năng lượng tự do cho vòng lặp trong đối xứng với kích thước 2, và có kiểu được hiển thị ở (b). Năng lượng cho ví dụ với vòng lặp trong, trong đó $x = G$ và $y = A$, được hiển thị ở (c)

x \ y	A	C	G	U
A	0.40	0.30	**-0.10**	0.40
C	-0.40	0.50	0.40	0.00
G	0.40	0.40	-1.70	0.40
U	0.40	0.50	0.40	-0.30

(a)



(b)



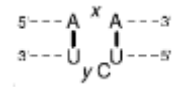
(c)

Năng lượng tự do cho vòng lặp trong không đối xứng có kích thước 3.

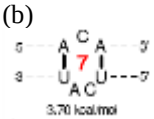
Bảng 7. (a) Năng lượng tự do cho vòng lặp trong không đối xứng với kích thước 3, với kiểu thể hiện ở (b), năng lượng tương ứng cho ví dụ ở (c), với $x = C$ và $y = A$; (d) Năng lượng tự do cho vòng lặp trong không đối xứng với kích thước 3, với kiểu thể hiện ở (e), năng lượng tương ứng cho ví dụ ở (f), với $x = A$ và $y = C$

x \ y	A	C	G	U
A	3.60	3.20	3.10	5.50
C	**3.70**	4.00	5.50	3.70
G	5.50	5.50	5.50	5.50
U	5.50	3.70	5.50	2.80

(a)



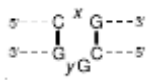
(b)



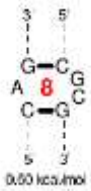
(c)

x \ y	A	C	G	U
A	1.00	**0.60**	0.40	4.00
C	4.00	4.00	4.00	4.00
G	0.80	4.00	2.20	4.00
U	4.00	4.00	4.00	4.00

(d)



(e)



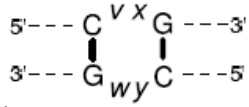
(f)

Năng lượng tự do cho vòng lặp trong đối xứng với kích thước 4.

Bảng 8. (a) Giả sử một phần của bảng năng lượng tự do cho vòng lặp trong đối xứng với kích thước 4, kiểu vòng thể hiện là (b). Năng lượng cho ví dụ là (c), với $v = A,\ w = A,\ x = G$ và $y = G$

x \ y	AA	AC	AG	AU	AC	…	GG	GU	UA	UC	UG	UU
AA	1.30	1.20	0.30	2.00	1.60	…	**1.00**	-0.40	2.00	1.90	1.10	1.40
.	.	.	.	.	.	…	.	.	.	.	.	.
.	.	.	.	.	.	…	.	.	.	.	.	.
GG	**1.00**	0.90	0	2.00	1.40	…	0.80	-0.70	2.00	1.70	0.90	1.20

(a)



(b)

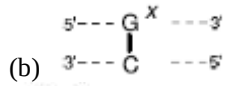
(c)

Năng lượng tự do cho nhưng điểm cuối bên ngoài.

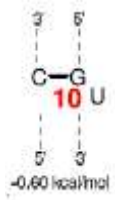
Bảng 9. (a) Năng lượng tự do cho những điểm cuối lủng lẳng với chiều kết thúc là 3', với kiểu là (b), tương ứng năng lượng cho vi dụ là (c); (d) Năng lượng tự do cho những điểm cuối lủng lẳng với chiều kết thúc là 5', với kiểu là (e), tương ứng năng lượng cho vi dụ là (f)

x	A	C	G	U
	-1.10	-0.40	-1.30	**-0.60**

(a)



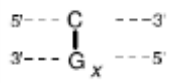
(b)



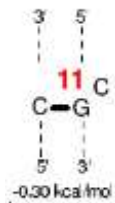
(c)

x	A	C	G	U
	-0.20	**-0.30**	0	0

(d)



(e)



(f)

Các quy tắc năng lượng tự do hỗn hợp.

Bảng 10. Các quy tắc năng lượng tự do hỗn hợp

STT	Quy tắc	Tên	Giá trị
1	Phép ngoại suy cho các vòng lớn của vòng lặp trong, kẹp tóc hoặc lồi lớn hơn 30	*Len_Par*	1.079
2	Vòng lặp trong không đối xứng: hiệu chỉnh giá trị cực đại	*Asym_Max*	3.00
3	Vòng lặp trong không đối xứng: mảng Ninio	*Asym _Par*	0.5 0.5 0.5 0.5
4	Vòng đa nhánh – phân nhánh	*Multi_a*	3.40
5	Vòng đa nhánh – tổn thất trên xoắn	*Multi_b*	0.40
6	Vòng đa nhánh – tổn thất trên base tự do	*Multi_c*	0
7	Tổn thất cho sự kết thúc không là GC	*Non_GC_terminal*	0.50
8	Sự tăng cho vòng kẹp tóc GGG	*bonusGGG*	-2.20
9	Sườn kẹp tóc nhiều C	*C_Hairpin_1*	0.30
10	Phân đoạn kẹp tóc nhiều C	*C_Hairpin_2*	1.60
11	Kẹp tóc 3 nhiều C	*C_Hairpin_3*	1.40
12	Năng lượng tự do khởi tạo giữa các phân tử	*Intermol*	4.10

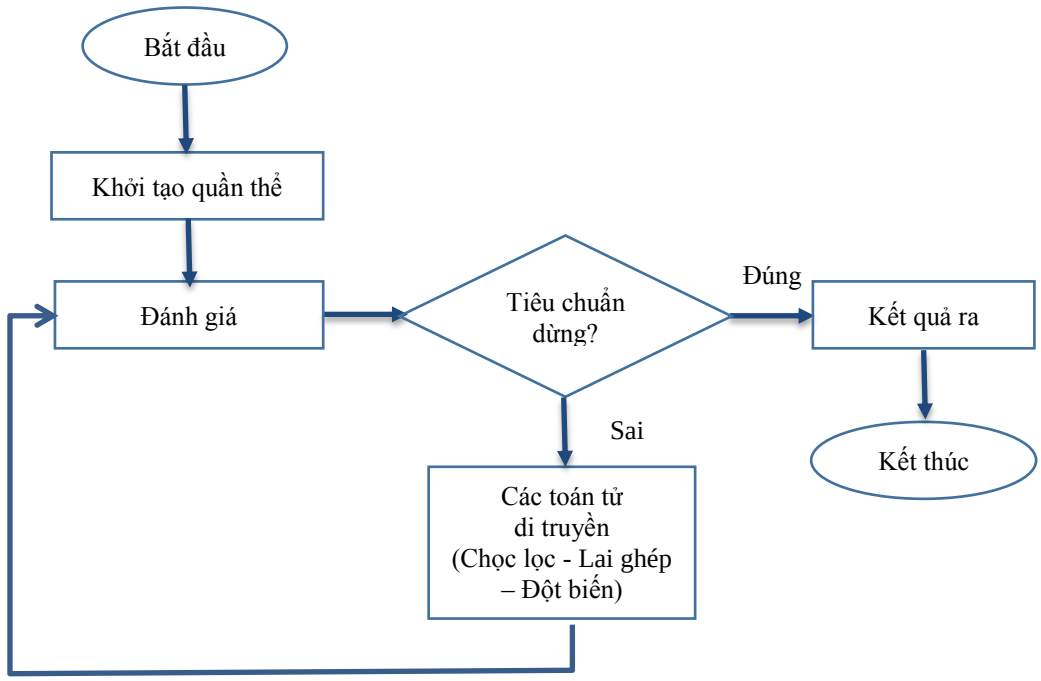
IV. THUẬT TOÁN DI TRUYỀN

Các nguyên lý và tính năng cơ bản

Thuật toán di truyền (Genetic Algorithms – GA) được giới thiệu đầu tiên bởi Holland [21] vào năm 1975 và trở thành một trong những kỹ thuật được sử dụng chủ yếu để giải quyết các bài toán tối ưu tổ hợp, tuyến tính và phi tuyến. Ngày nay cũng có những thuật ngữ tính toán mềm và thuật toán thông minh thường được sử dụng để chỉ đến lĩnh vực nghiên cứu này.

Thuật toán di truyền (GA) là kỹ thuật tìm kiếm ngẫu nhiên toàn cục với việc giả lập các quy luật tiến hóa và di truyền học để cố gắng tìm ra giải pháp tối ưu cho vấn đề tối ưu hóa phức tạp. GA đạt được về mặt lý thuyết và thực nghiệm đã được chứng minh nhằm cung cấp cho việc tìm kiếm mạnh mẽ trong không gian phức tạp và chúng được áp dụng rộng rãi trong kỹ thuật, kinh doanh và giới khoa học.

Sơ đồ thuật toán GA:



Hình 5. Lưu đồ thuật toán di truyền

Những thuật ngữ và những toán tử của thuật toán di truyền

a) Cá thể

Một cá thể là một giải pháp riêng biệt. Các nhóm cá thể có cùng hai dạng giải pháp sau:

- Nhiễm sắc thể, đây chính là thông tin di truyền thô (kiểu gen) mà GA giải quyết.
- Kiểu hình, đấy chính là cách biểu hiện của nhiễm sắc thể trong điều kiện của mô hình.

b) Gen

Gen là những chỉ dẫn cơ bản để xây dựng một thuật toán di truyền. Một nhiễm sắc thể là một chuỗi của những gen. Gen có thể được mô tả một giải pháp khả thi của bài toán, mà không thực sự là giải pháp. Một gen là một chuỗi số bit có độ dài tùy ý. Chuỗi bit là một biểu diễn nhị phân của số các khoảng từ một giới hạn dưới. Một gen là một biểu diễn của GA về một giá trị nhân tố duy nhất cho một yếu tố kiểm soát, mà ở đó yếu tố kiểm soát phải có cận trên và cận dưới. Phạm vi này có thể chia thành một số khoảng mà có thể được biểu diễn bởi chuỗi bit của gen. Một chuỗi bít có độ dài n có thể tương ứng với $2^n - 1$ khoảng.

c) Quần thể

Một quần thể là một tập hợp các cá thể. Một quần thể bao gồm một số cá thể đang được kiểm tra, những tham số kiểu hình xác định những cá thể và một số thông tin về không gian tìm kiếm.

d) Mã hóa

Mã hóa là quá trình biểu diễn gen riêng lẻ. Quá trình đó có thể được thực hiện bằng cách sử dụng các bít, các số, cây, mảng, danh sách hoặc những đối tượng khác. Việc mã hoá phụ thuộc chủ yếu vào bài toán cần giải quyết. Ví dụ, người ta có thể mã hóa trực tiếp số thực hoặc số nguyên.

e) Chọn lọc

Chọn lọc là quá trình lựa chọn hai cha mẹ từ quần thể cho lai ghép. Sau khi quyết định mã hoá, bước tiếp theo là quyết định làm thế nào để thực hiện chọn lọc như là làm thế nào để chọn những cá thể trong quần thể sẽ tạo ra con cái cho thế hệ tiếp theo và bao nhiêu con mà mỗi cá thể sẽ tạo ra. Mục đích của việc chọn lọc là nhấn mạnh vào các cá thể thích hợp trong quần thể với hy vọng rằng những con cái của chúng có khả năng thích nghi cao hơn.

f) Lai ghép

Lai ghép là quá trình lấy hai giải pháp cha mẹ để sinh sản một đứa con. Sau quá trình chọn lọc, quần thể là được làm phong phú thêm với những cá thể tốt hơn. Sự sinh sản thu được những bản sao của chuỗi tốt nhưng không tạo ra cá thể mới. Toán tử lai ghép được áp dụng cho vùng lưu trữ ghép đôi với hy vọng rằng nó tạo ra một con tốt hơn.

g) Đột biến

Sau khi lai ghép, các chuỗi sẽ bị đột biến. Đột biến truyền thống được coi như mà một toán tử tìm kiếm đơn giản. Nếu lai ghép được cho là khai thác các giải pháp hiện tại để tìm ra những con tốt hơn, đột biến sẽ giúp cho việc khám phá toàn bộ không gian tìm kiếm. Đột biến được xem như một toán tử cơ sở để duy trì sự đa dạng di truyền trong quần thể. Nó đưa ra các cấu trúc di truyền mới trong quần thể bằng cách sửa đổi ngẫu nhiên một số khối căn bản của nó. Đột biến giúp thoát khỏi cái bẫy của cực tiểu cục bộ và duy trì sự đa dạng trong quần thể.

Mối quan tâm khi thiết kế thuật toán

Khi một GA được sử dụng để giải quyết các bài toán tối ưu hóa, tuân theo các điểm lưu ý sau đây nó rất có giá trị:

- Lựa chọn một bài toán tối ưu hóa và hiểu những gì bạn có thể thay đổi để đạt được các giải pháp;
- Chọn một sự biểu diễn thích hợp. Biểu diễn được trình bày chi tiết quá làm gia tăng độ phức tạp tính toán, trong khi biểu diễn kém quá làm giảm độ chính xác của bài toán;

Hãy thử sử dụng các toán tử di truyền khác, điều này có nghĩa rằng hầu hết cá thể phù hợp sẽ được sao chép vào thế hệ tiếp theo mà không bị nhiễu loạn bởi lai ghép hoặc đột biến. Điều này thường dẫn đến một tốc độ hội tụ nhanh.

V. ÁP DỤNG THUẬT TOÁN GA VÀO BÀI TOÁN DỰ ĐOÁN CẤU TRÚC BẬC HAI RNA

A. Thuật toán sinh helix

Trong mô hình của nhóm nghiên cứu, một helix được xác định bởi ba ràng buộc. Thứ nhất, mỗi helix phải có ít nhất ba cặp base chính tắc xếp chồng lên nhau. Thứ hai, mỗi chuỗi hoặc cặp kết nối hai thân phải có chiều dài ít nhất là 3 ribonucleotide. Thứ ba, mỗi helix không được chia rẽ các base lẫn nhau. Thuật toán sinh helix như sau:

Sinh tập có thể của các cặp base (r_i, r_j) từ chuỗi RNA đã cho;

Khởi tạo helix h;

For mỗi cặp (r_i, r_j) **do**

While (helix h là đúng) và (helix h chưa hoàn thành) **do**

If ((r_i, r_j) là cặp base chính tắc) và ((r_i, r_j) không là phần của một helix đã tồn tại) **then**

Thêm cặp base (r_i, r_j) vào trong helix h

i++;

j--;

else

if helix h chứa ít hơn 3 cặp base **then**

helix h là không hợp lệ;

else **if** helix có ít hơn 3 base ở giữa cặp base cuối **then**

helix h là không hợp lệ;

else

helix h là hoàn thành;

```
                    endif
                endif
            endif
        endwhile
        if (helix h là hợp lệ) then
            chèn helix h vào trong tập các helix H
        endif
endfor
```

Hình 6. Mã giả sinh helix

B. Không gian tìm kiếm

Một nhận xét chung về kích thước của không gian tìm kiếm là kết quả từ thuật toán tạo helix nên được thực hiện. Nếu chúng ta coi như rằng mỗi helix thành lập trong cấu trúc hoặc hiện diện hoặc vắng mặt, thì toàn bộ số lượng các cấu trúc được tạo ra từ một chuỗi cho trước là 2^H, trong đó H là tập các helix.

Bảng 11. Số lượng các helix thành lập của mỗi chuỗi tương ứng

Tên sinh vật	Chiều dài chuỗi	Tổng số Helix thành lập
Saccharomyces cerevisiae (Nấm bánh mì)	118	175
Caenorhabditis elegans	697	6074
Acanthamoeba griffini	556	3637
Hildenbrandia rubra	543	3933
Haloarcula Marismortui	122	198

Ví dụ, với chiều dài 697 của chuỗi Caenorhabditis elegans sẽ thành lập được 6074 helix và không gian tìm kiếm của 2^{6074} cấu trúc có thể, một không gian tìm kiếm cực lớn.

C. Thuật toán di truyền (GA) cho dự đoán cấu trúc bậc hai RNA

Các tham số và toán tử cho thuật toán GA

Các toán tử và tham số cho GA áp dụng với bài toán dự đoán cấu trúc bậc hai RNA như sau:

- Thế hệ,
- Kích thước quần thể,
- Xác suất lai ghép (P_c),
- Xác suất đột biến (P_m),
- Toán tử lai ghép,
- Chiến lược chọn lọc,
- Lựa chọn tối ưu (On/Off),
- Mô hình nhiệt động học,
- Nút thắt (On/Off),
- Gieo ngẫu nhiên.

Thế hệ là số lượng thế hệ mà GA sẽ thực hiện.

Kích thước quần thể là số lượng cá thể trong một quần thể.

P_c là xác suất của lai ghép xảy ra trên một cặp bố mẹ.

P_m là xác suất của đột biến xảy ra trên một con đã cho.

Lai ghép là kiểu lai ghép được dùng.

Chiến lược chọn lọc là tham số cung cấp hai loại chiến lược quản lý đầu ra của lai ghép. Chọn lọc mẫu (STDS) chỉ đơn giản là cho hai con lai ghép trực tiếp để sinh ra thế hệ tiếp theo, không quan tâm đến tính thích nghi của chúng. Sao chép tốt nhất (KBR) lựa chọn cha mẹ phù hợp nhất và con phù hợp nhất và chuyển chúng sang thế hệ tiếp theo.

Lựa chọn tối ưu nếu được áp dụng, giữ lại những cá thể tốt nhất từ thế hệ trước cho thế hệ tiếp theo.

Mô hình nhiệt động học là tham số xác định mô hình nào được sử dụng trong GA.

Toán tử nút thắt xác định xem nút thắt có được phép xuất hiện trong cấu trúc được dự đoán.

Gieo ngẫu nhiên là toán tử cho phép người sử dụng chọn gieo ngẫu nhiên để khởi tạo một số ngẫu nhiên của GA.

Mã giả của thuật toán GA cho bài toán như sau:

```
Sinh tập có thể của các cặp base từ chuỗi RNA đã cho;
Sinh tập có thể các helix sử dụng tập các cặp base;
Sinh quần thể ngẫu nhiên ban đầu.
For tất cả thế hệ do
    For kích thước của quần thể/2 do
        Lựa chọn hai cha mẹ;
        If giá trị ngẫu nhiên < P_c then
            Lai ghép cha mẹ để tạo ra hai con;
        else
            Chuyển cha mẹ sang thế hệ sau;
        endif
        for mỗi con do
            if giá trị ngẫu nhiên < P_m then
                Đột biến ngẫu nhiên con được chọn này;
            endif
        endfor
        if chiến lược chọn lọc KBR then
            Giữ lại cha mẹ và con tốt nhất dựa trên thích nghi;
        else if chiến lược chọn lọc là STDS then
            Luôn giữ lại con ;
        endif
        Chèn những cá thể được giữ lại vào trong quần thể mới ;
        Ap dụng lựa chọn tối ưu =1và giữ lại cá thể tốt nhất từ thế hệ trước ;
    endfor
endfor
Đưa ra cấu trúc tốt nhất ;
```

Hình 7. Mã giả áp dụng GA cho bài toán tìm cấu trúc RNA

VI. ĐÁNH GIÁ VÀ SO SÁNH KẾT QUẢ

A. Đánh giá

Khi thực hiện thuật toán di truyền cho bài toán dự đoán cấu trúc bậc hai RNA cần phải thiết lập các thông số ban đầu cho các tham số của thuật toán. Các thông số được thể hiện dưới đây:

Bảng 12. Giá trị các tham số cho GA

Tham số	Giá trị
Thế hệ,	700
Kích thước quần thể,	800
Xác suất lai ghép (P_c),	70%
Xác suất đột biến (P_m),	80%
Toán tử lại ghép,	CX, OX2, PPX
Chiến lược chọn lọc,	STDS, KBR
Lựa chọn tối ưu (On/Off),	1
Mô hình nhiệt động học,	INN-HB
Nút thắt (On/Off),	0
Gieo ngẫu nhiên	20

Kết quả thực hiện với một số chuỗi như sau:

Chuỗi Caenorhabditis elegans với toán tử lai ghép là OX2

Bảng 13. Các tham số được thiết lập để kiểm tra chuỗi Caenorhabditis elegans với toán tử lai ghép OX2

Năng lượng tự do (ΔG kcal/mol)	Toán tử chọn lọc	Toán tử lai ghép	P_c	P_m	Mô hình nhiệt động học
-299.88	STDS	OX2	70%	80%	INN-HB
-299.78	STDS	OX2	70%	80%	INN-HB
-299.73	STDS	OX2	70%	80%	INN-HB
-297.86	STDS	OX2	70%	80%	INN-HB
-297.67	STDS	OX2	70%	80%	INN-HB
-296.78	STDS	OX2	70%	80%	INN-HB
-295.78	STDS	OX2	70%	80%	INN-HB
-295.68	STDS	OX2	70%	80%	INN-HB
-295.88	STDS	OX2	70%	80%	INN-HB
-294.98	KBR	OX2	70%	80%	INN-HB
-294.78	KBR	OX2	70%	80%	INN-HB
-292.18	KBR	OX2	70%	80%	INN-HB
-291.28	STDS	OX2	70%	80%	INN-HB
-290.18	KBR	OX2	70%	80%	INN-HB
-286.45	KBR	OX2	70%	80%	INN-HB
-285.78	KBR	OX2	70%	80%	INN-HB
-283.69	KBR	OX2	70%	80%	INN-HB
-277.25	KBR	OX2	70%	80%	INN-HB
-274.91	KBR	OX2	70%	80%	INN-HB

Chuỗi Caenorhabditis elegans với toán tử lai ghép là CX.

Bảng 14. Các tham số được thiết lập để kiểm tra chuỗi Caenorhabditis elegans với toán tử lai ghép CX

Năng lượng tự do (ΔG kcal/mol)	Toán tử chọn lọc	Toán tử lai ghép	P_c	P_m	Mô hình nhiệt động học
-298.85	STDS	CX	70%	80%	INN-HB
-298.78	STDS	CX	70%	80%	INN-HB
-298.63	STDS	CX	70%	80%	INN-HB
-297.68	STDS	CX	70%	80%	INN-HB
-297.63	STDS	CX	70%	80%	INN-HB
-295.58	STDS	CX	70%	80%	INN-HB
-294.78	STDS	CX	70%	80%	INN-HB
-294.68	STDS	CX	70%	80%	INN-HB
-293.88	STDS	CX	70%	80%	INN-HB
-292.98	KBR	CX	70%	80%	INN-HB
-291.78	KBR	CX	70%	80%	INN-HB
-288.12	KBR	CX	70%	80%	INN-HB
-287.48	STDS	CX	70%	80%	INN-HB
-286.28	KBR	CX	70%	80%	INN-HB
-284.65	KBR	CX	70%	80%	INN-HB
-284.88	KBR	CX	70%	80%	INN-HB
-281.68	KBR	CX	70%	80%	INN-HB
-272.56	KBR	CX	70%	80%	INN-HB
-271.45	KBR	CX	70%	80%	INN-HB

Chuỗi Caenorhabditis elegans với toán tử lai ghép là PPX.

Bảng 15. Các tham số được thiết lập để kiểm tra chuỗi Caenorhabditis elegans với toán tử lai ghép PPX

Năng lượng tự do (ΔG kcal/mol)	Toán tử chọn lọc	Toán tử lai ghép	P_c	P_m	Mô hình nhiệt động học
-276.55	STDS	PPX	70%	80%	INN-HB
-276.78	STDS	PPX	70%	80%	INN-HB
-275.63	STDS	PPX	70%	80%	INN-HB
-273.68	STDS	PPX	70%	80%	INN-HB
-273.63	STDS	PPX	70%	80%	INN-HB
-272.58	STDS	PPX	70%	80%	INN-HB
-271.78	STDS	PPX	70%	80%	INN-HB
-270.68	STDS	PPX	70%	80%	INN-HB
-269.88	STDS	PPX	70%	80%	INN-HB
-269.98	KBR	PPX	70%	80%	INN-HB

Năng lượng tự do (ΔG kcal/mol)	Toán tử chọn lọc	Toán tử lai ghép	P_c	P_m	Mô hình nhiệt động học
-267.78	KBR	PPX	70%	80%	INN-HB
-267.12	KBR	PPX	70%	80%	INN-HB
-266.42	STDS	PPX	70%	80%	INN-HB
-265.38	KBR	PPX	70%	80%	INN-HB
-264.45	KBR	PPX	70%	80%	INN-HB
-263.84	KBR	PPX	70%	80%	INN-HB
-262.66	KBR	PPX	70%	80%	INN-HB
-261.76	KBR	PPX	70%	80%	INN-HB
-260.65	KBR	PPX	70%	80%	INN-HB

Với việc áp dụng GA vào bài toán dự đoán cấu trúc bậc hai RNA cho thấy rằng với toán tử lai ghép OX2, CX và toán tử chọn lọc STDS sẽ tìm ra những cấu trúc có năng lượng thấp hơn.

B. So sánh

Quy hoạch động (DPA) [12], [13] [14], [15] [16] [17] [18] được áp dụng cho dự đoán cấu trúc bậc hai RNA. Đặc biệt được thể hiện qua phần mềm mfold. Phần mềm này sử dụng mô hình nhiệt động học cho đánh giá năng lượng tự do của cấu trúc. Thuật toán quy hoạch động, song song với mô hình này, cố gắng tìm ra một cấu trúc có năng lượng cực tiểu. Mfold sử dụng mô hình nhiệt động học chuẩn INN-HB, nhưng bổ sung mô hình cho các cấu trúc RNA con chung. Năng lượng được tính chính là năng lượng các thành phần đã được trình này ở mục III.B.

Khi thực hiện mfold thì với cấu trúc được dự đoán cho những chuỗi ngắn tốt hơn so với nhưng chuỗi dài. Khi áp dụng GA thì năng lượng thấp nhất và các cặp base chính xác nhiều hơn.

Nhìn chung khi áp dụng GA với mô hình nhiệt động học đơn giản cũng giống như DPA với mô hình nhiệt động học phức tạp.

Các kết quả được so sánh qua các bảng sau:

Bảng 16. Xenopus laevis với chiều dài chuỗi là 945 - Số lượng cặp base được biết là 251

Mfold - DPA (ΔG -kcal/mol)	GA (ΔG - kcal/mol)	Cặp base được dự đoán	Cặp base được dự đoán chính xác	Tỷ lệ (%) được dự đoán chính xác
-250.6	-222.85	249	92	36.7
-249.6	-219.75	251	71	28.3
-248.0	-216.51	245	**113**	**45.0**
-244.3	**-223.49**	246	86	34.3
-242.5	-202.27	251	81	32.3

Bảng 17. Drosophila virilis với chiều dài chuỗi là 784 - Số lượng cặp base được biết là 233

Mfold - DPA (ΔG -kcal/mol)	GA (ΔG - kcal/mol)	Cặp base được dự đoán	Cặp base được dự đoán chính xác	Tỷ lệ (%) được dự đoán chính xác
-146.3	-124.43	236	37	15.9
-146.2	-124.07	246	37	15.9
-142.8	-120.26	252	**82**	**35.2**
-142.2	**-131.55**	254	33	14.2
-139.0	-122.10	238	22	9.4

C. Kết luận

Thuật toán di truyền cho phép áp dụng dự đoán những cấu trúc bậc hai RNA với những chuỗi lớn. Trong mô hình nhiệt động học đơn giản thì thuật toán di truyền tính toán năng lượng tự do của cấu trúc được thực hiện tốt hơn khi áp dụng DPA với mô hình nhiệt động học phức tạp.

D. Hướng nghiên cứu tiếp theo

Cải thiện thêm về thuật toán GA dựa trên những cải tiến hiện có và lai ghép với thuật toán đàn kiến (ACO). Áp dụng những cải tiến này để dự đoán những cấu trúc bậc hai RNA tối ưu, bao gồm cả những cấu trúc có nút thắt.

TÀI LIỆU THAM KHẢO

[1] P. G. Higgs, "RNA Secondary Structure: Physical and Computational Aspects," Quarterly Rev. of Biophysics, vol. 33, pp. 199-253, 2000.

[2] M. Zuker and D. Sankoff, “RNA secondary structures and their prediction,” Mathematical Bioscience, Vol. 46, pp. 591{621, 1984.

[3] M. S. Waterman and T. F. Smith, “RNA secondary structure: A complete mathematical analysis," Mathematical Bioscience, Vol. 42, pp. 257-266, 1978.

[4] J. Kim, J. Cole, and S. Pramanik, “Alignment of possible secondary structures in multiple RNA sequences using simulated annealing," Computer Applications in the Biosciences, Vol. 12, Aug. 1996.

[5] D. Sankoff, “Simultaneous solution of the RNA folding, alignment and protosequence problems," SIAM Journal on Applied Mathematics, Vol. 45, No. 5, pp. 810-825, 1985.

[6] B. Knudsen and J. Hein, “RNA secondary structure prediciton using stochastic context-free grammars and evolutionary history," Bioinformatics, Vol. 15, pp. 446-454, 1999.

[7] S. Kobayashi and T. Yokomori, “Modeling RNA secondary structures using tree grammars," In Proceedings of Genome Informatics Workshop V, pp. 29-38, 1994.

[8] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori, “Tree adjoining grammars for RNA structure prediction," Theoretical Computer Science, Vol. 210, pp. 277-303, Jan. 1999.

[9] B.A. Shapiro and J. Navetta, “A Massively-Parallel Genetic Algorithm for RNA Secondary Structure Prediction,” J. Supercomputing, vol. 8, pp. 195-207, 1994.

[10] I. I. Titov, D. G. Vorobiev, V. A. Ivanisenko, and N. A. Kolchanov, “A Fast Genetic Algorithm for RNA Secondary Structure Analysis,” Russian Chemical Bull. vol. 51, no. 7, pp. 1135-1144, 2002.

[11] Müller, Ingo (2007). A History of Thermodynamics - the Doctrine of Energy and Entropy. Springer

[12] M. Zuker, “Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction,” Nucleic Acids Research, vol. 31, no. 13, pp. 3406-3415, 2003.

[13] M. Zuker and P. Stiegler, “Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information,” Nucleic Acids Research, vol. 9, pp. 133-148, 1981.

[14] M. Zuker, “On Finding All Suboptimal Foldings of an RNA Molecule,” Science, vol. 244, pp. 48-52, 1989.

[15] M. Zuker, “Prediction of RNA Secondary Structure by Energy Minimization,” Computer Analysis of Sequence Data, A. M. Griffin and H. G. Griffin, eds., pp. 267-294, Humana Press, July 1994.

[16] M. Zuker, D. H. Mathews, and D. H. Turner, “Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide,” RNA Biochemistry and Biotechnology, series NATO ASI Series, J. Barciszewski and B. Clark, eds., Kluwer Academic Publishers, 1999.

[17] M. Zuker, “Calculating Nucleic Acid Secondary Structure,” Current Opinion in Structural Biology, vol. 10, pp. 303-310, 2000.

[18] M. Zuker, J. A. Jaeger, and D. H. Turner, “A Comparison of Optimal and Suboptimal RNA Secondary Structures Predicted by Free Energy Minimization with Structures Determined by Phylogenetic Comparison,” Nucleic Acids Research, vol. 19, no. 10, pp. 2707-2714, 1991.

[19] Mathews, D., Sabina, J., Zuker, M., and Turner, D. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol , 288(5), 911-940.

[20] Zhu, J. andWartell, R. (1997). The relative stabilities of base pair stacking interactions and single mismatches in long RNA measured by temperature gradient gel electrophoresis. Biochemistry, 36(49), 15326–15335.

[21] J. Holland (1975), Adaptation in Natural and Artificial Systems Michigan Press.

IMPROVED GENETIC ALGORITHMS AND THE APPLICATION OF PREDICTIVE RNA SECONDARY STRUCTURES

Doan Duy Binh, Pham Minh Tuan, Dang Duc Long

ABSTRACTS. *RNA structure is an important area of research. RNA is the center of many stages of protein synthesis, which also plays a role in the structure and function of the cell. Predicting the RNA structure can overcome many problems with determining the physical structure. Currently, physical methods for predicting RNA structures are time consuming and costly, so methods for predicting computations are preferred. Various algorithms have been used to predict RNA structures including dynamic programming and comparative methods. In this pape we introduce genetic algorithm as a method applied to the problem of predicting RNA structure. The genetic algorithm we adopt is that some of the parameters in genetic operators such as selection, crossover and mutation have been improved. Finally, we compared and evaluated with another algorithm for the secondary prediction problem, such as dynamic programming algorithm.*