SỬ DỤNG MẠNG NƠRON HOPFIELD DỰ ĐOÁN CẤU TRÚC THỨ CẤP CỦA RIBONUCLEIC ACID

Đặng Quang Á (*Viện Công nghệ thông tin- Viện KH&CN Việt Nam*)
Nông Thị Hoa (*Khoa Công nghệ thông tin, ĐH Thái Nguyên*)

1. Giới thiệu

RNA (RiboNucleic Acid) là một axít tham gia vào quá trình dịch mã từ DNA (DeoxyriboNucleic Acid) sang Protein. RNA được cấu thành từ các đơn phân tử: AminoAcid, Nucleotide và Monosaccharide (xem [3]). RNA phục vụ hai mục đích sinh học: RNA là phương tiện chuyển thông tin từ DNA vào các sản phẩm của Protein, RNA hoạt động như một thành phần của Ribosom. Việc dự đoán cấu trúc không gian của RNA bằng thực nghiệm rất tốn kém cả về thời gian và chi phí. Vì thế, trong thời gian gần đây người ta đã sử dụng các công cụ toán học và tin học vào mục đích dự đoán cấu trúc của RNA cũng như giải quyết nhiều bài toán khác của công nghệ sinh học.

Trong bài báo này chúng tôi tiếp cận bài toán nhờ mạng nơron nhân tạo Hopfield [7]. Mạng này là một công cụ rất hữu hiệu để giải quyết các bài toán tối ưu hóa và thỏa mãn ràng buộc [1], [2]. Nhờ sử dụng mạng này mà các ràng buộc đặt lên cấu trúc của RNA đã được xử lý. Kết quả mô phỏng trên máy tính cho một số thí dụ đã chứng tỏ tính khả thi của phương pháp dự đoán nhờ mạng nơron.

Bài báo gồm các nội dung sau: Phần 2 mô tả bài toán dự đoán cấu trúc thứ cấp RNA. Phần 3 trình bày lời giải của bài toán nhờ sử dụng mạng nơ ron Hopfield. Phần 4 là kết luận và hướng nghiên cứu tiếp theo.

2. Bài toán dự đoán cấu trúc thứ cấp RNA

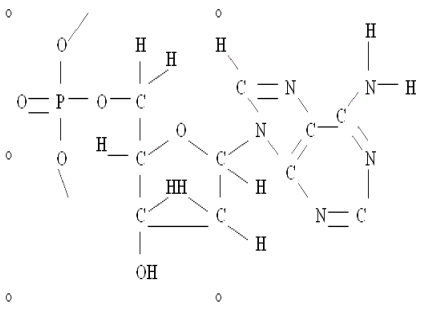
2.1. Cấu trúc của RNA (xem [3], [7])

Axít Nucleic chứa thông tin về cấu trúc và chức năng của một tổ chức sống mà các thông tin này được lưu trữ và di truyền cho lần sinh sản tiếp theo. Có hai kiểu nucleotide trong axít Nucleic: DNA và RNA. Khi các nucleotide hình thành một chuỗi DNA hoặc RNA thì nhóm phosphate của một nucleotide tạo một liên kết hoá học với phân tử oxygen gắn với phân tử cácbon của nucleotide tiếp theo. Đoạn xoắn kép được thiết lập khi hai đầu nút DNA hoặc RNA riêng biệt liên kết với nhau bằng các liên kết hoá học yếu giữa các cặp cơ sở.

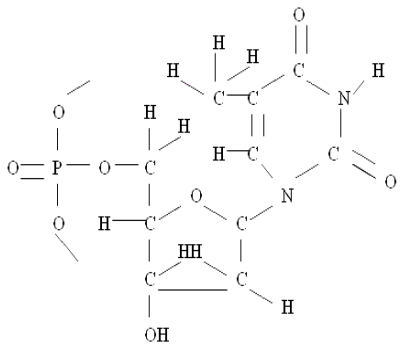
Có ba loại RNA chính tham gia vào quá trình dịch mã sang Protein: mRNA là RNA thông tin, tRNA là RNA vận chuyển, rRNA là RNA risbosome.

RNA là một axit nucleic được cấu tạo từ các chuỗi nucleotide. Mỗi nucleotide của RNA gồm Phosphate, đường Pentoseribose và một trong bốn cơ sở Adenine (A), Guanine (G), Cytosine (C), Uracil (U). Cấu trúc RNA quy định các cặp cơ sở ghép đôi với nhau gồm : **C – G** và **U – A.**

Hình 1 biểu diễn cấu trúc không gian của cơ sở Adenine và Hình 2 biểu diễn cấu trúc không gian của cơ sở Uracil

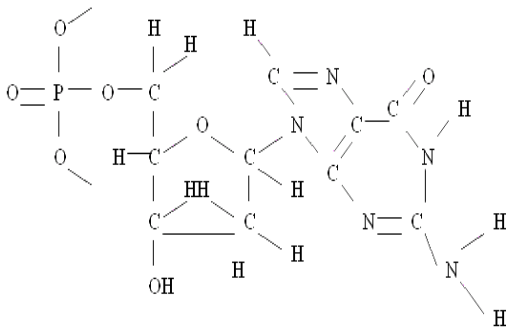


Hình 1

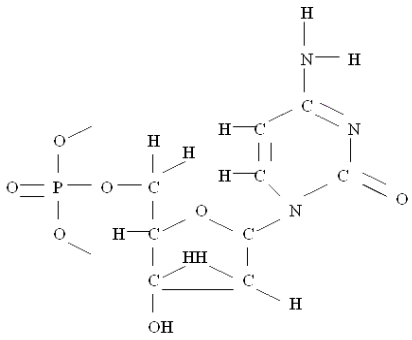


Hình 2

Hình 3 biểu diễn cấu trúc không gian của cơ sở Guanine và Hình 4 biểu diễn cấu trúc không gian của cơ sở Cytosine



Hình 3



Hình 4

2.2. Phát biểu bài toán

***Bài toán**

Cho một đại phân tử RNA gồm một dãy các cơ sở A, G, U, C. Tìm một cấu trúc RNA có số cặp cơ sở ghép đôi lớn nhất (cấu trúc bền vững nhất)**.**

***Cách biểu diễn**

Biểu diễn một phân tử RNA bằng dãy S gồm $S_1,S_2,...,S_n$ với S_i là một trong bốn cơ sở: A, G, U, C. Khi đó cấu trúc thứ cấp P của S được biểu diễn bởi một ma trận tam giác trên bên phải **B** cấp n với.

$\mathbf{B}_{ij}$=1 nếu cơ sở ở vị trí i và vị trí j ghép đôi (với i<j)

$\mathbf{B}_{ij}$=0 nếu cơ sở ở vị trí i và vị trí j không ghép đôi

**** Các ràng buộc chặt của cấu trúc thứ cấp RNA* (xem [6])**

1. Nếu P chứa cặp i.j thì S_i và S_j là G & C hoặc A & U hoặc C & G hoặc U & A.

2. Không có sự xếp chồng các cặp: Nếu P chứa cặp i.j thì không chứa i.k nếu $k \neq j$ và không chứa cặp k.j nếu $k \neq i$.

3. Đối với ∀i, cặp i.i không nằm trong P.

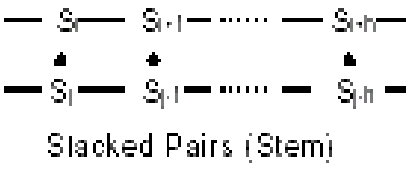
4. Điều kiện không được phép xảy ra:Nếu h < i < j < k thì P không chứa cặp h.j và cặp i.k.

5. Không cho phép có các lặp đột ngột: Nếu P chứa i.j thì cơ sở i và cơ sở j nằm cách xa nhau ít nhất hai cơ sở.

**** Các cấu trúc con của RNA***

Một cấu trúc thứ cấp P của đoạn S gồm sáu cấu trúc con là : lặp kẹp tóc, lặp trong, lặp phức, chỗ phình, đoạn ngăn xếp và vùng thắt nút.

1. Nếu P chứa cặp i.j, (i + 1).(j - 1), …, (i + h).(j - h) thì P chứa đoạn ngăn xếp. (Hai hoặc nhiều cặp ghép đôi liên tiếp gọi là đoạn ngăn xếp) (Hình 5).

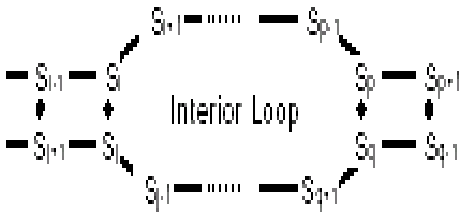


Hình 5

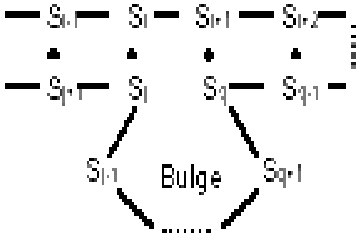
Hình 6

2. Nếu P chứa cặp i.j nhưng không có phần tử bao quanh i + 1…j - 1 nào ghép đôi thì hình thành một lặp kẹp tóc (Hình 6).

3. i + 1 < p < q < j - 1 và P chứa cặp i.j và cặp p.q nhưng các phần tử giữa i và p không ghép đôi và các phần tử giữa q và j không ghép đôi thì 2 vùng không ghép đôi hình thành một lặp trong (Hình 7).



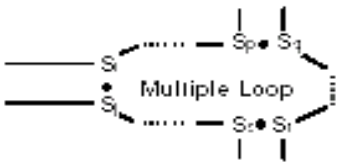
Hình 7



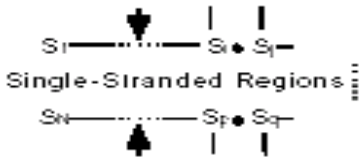
Hình 8

4. Nếu P chứa cặp i.j và (i + 1).q và có vài phần tử không ghép đôi giữa j và q (hoặc P chứa i.j và p.(j - 1) và có các phần tử không ghép đôi giữa i và p) thì những phần tử không ghép đôi hình thành một chỗ phình (Hình 8).

5. Nếu P chứa cặp i.j và i.j bao quanh hai hoặc nhiều hơn các cặp p.q, r.s… mà i.j không bao quanh mỗi nhóm riêng thì một lặp phức được hình thành (Hình 9).



Hình 9



Hình 10

6. Cho r là một dãy con các phần tử trong dãy cơ sở RNA. Nếu r không ghép đôi và không có cặp nào trong P bao quanh r thì r là một vùng thắt nút đơn (Hình 10).

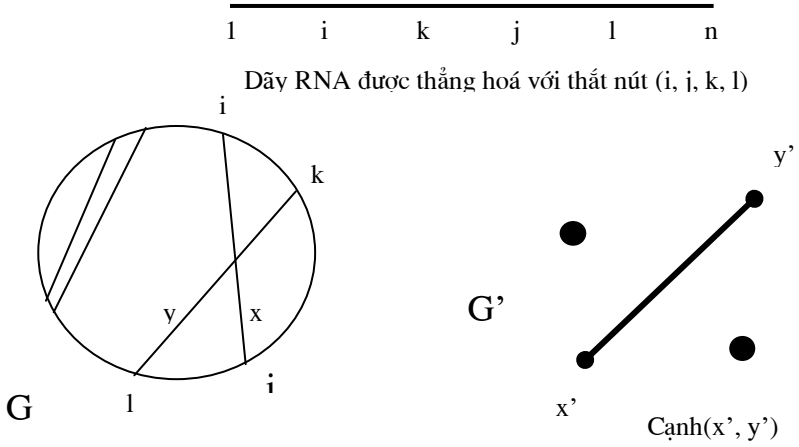
3. Giải bài toán bằng mạng nơron Hopfield

3.1. Đồ thị tròn [7]

Cấu trúc thứ cấp RNA được ánh xạ vào mạng nơron thông qua lý thuyết toán học của đồ thị tròn và đồ thị cạnh liền kề.

Xét một dãy RNA S : S_1S_2. . . S_n trong đó S_i là một trong các cơ sở A, C, G, U. Một cấu trúc thứ cấp của dãy S có thể biểu diễn bởi một đồ thị tròn G với n nút. Có n điểm trên vòng tròn và cạnh (i,j) ứng với cặp cơ sở (i,j) ghép đôi. Các cạnh giao nhau ứng với các thắt nút.

Xây dựng đồ thị G' với số nút bằng số cạnh trong G và mỗi cạnh (x', y') trong G' biểu diễn cho mỗi cặp x=(i,j), y=(k,l) trong G cắt nhau. G' là đồ thị cạnh liền kề cho G.



Cạnh x có Head=i, Tail=j
Cạnh y có Head=k, Tail=l

Các điều kiện trong đồ thị tròn ứng với các ràng buộc chặt của cấu trúc RNA

- Điều kiện tạo ra hai cạnh x và y cắt nhau (tạo ra một thắt nút)

head(x) < head(y) < tail(x) < tail(y)

head(y) < head(x) < tail(y) < tail(x)

- Điều kiện một cơ sở nằm trong 2 cặp ghép đôi

tail(x)=tail(y), head(x)=head(y)

tail(x)=head(y), head(x)=tail(y)

- Điều kiện cặp cơ sở ghép đôi là một lặp kẹp tóc

| head(x)-tail(x) | >=2.

Do đó việc dự đoán cấu trúc thứ cấp RNA giống với vấn đề tìm tập độc lập lớn nhất của đồ thị cạnh liền kề G' của đồ thị tròn G biểu diễn dãy S.

Việc tìm tập độc lập lớn nhất tương ứng với việc tìm ra tập lớn nhất các cặp cơ sở ghép đôi mà không có hai cạnh cắt nhau.

3.2. Thiết kế mạng nơron Hopfield

Cho một dãy S gồm n cơ sở, xây dựng mạng nơron Hopfield một lớp gồm n(n-1)/2 nơron như sau: mỗi nơron biểu diễn một phần tử của phần trên bên phải của ma trận cấp n. Chỉ số i và j để biểu diễn nơron ij trong mạng với i=1, ..., n-1, j=2,..,n. Ký hiệu V_{ij} là thông tin ra, U_{ij} là thông tin vào của nơron ij.

Năng lượng E của mạng được xác định như sau:

$$E = A_1 \sum_{i=1}^{n} \left(\sum_{j \neq i}^{n} V_{ij} - 1 \right)^2 b(i,j) + A_2 \sum_{j=1}^{n} \left(\sum_{i \neq j}^{n} V_{ij} - 1 \right)^2 b(i,j) + C V_{ij} p(i,j,t)$$

$$+ B_1 \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij} \sum_{i<k<j<l} V_{kl} g(i,k,j) g(k,j,l) f(k,l) + B_2 \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij} \sum_{k<i<l<j} V_{kl} g(k,i,l) g(i,l,j) f(k,l)$$

trong đó:

- A_1, A_2, B_1, B_2,C là các tham số mạng
- Hàm b(x,y)=1 nếu x-y là cặp A-U hoặc G-C, ngược lại b(x,y)=0
- Hàm g(x,y,z)=1 nếu x>y>z, ngược lại g(x,y,z)=0
- Hàm f(x,y)=2 nếu x-y là G-C, f(x,y)=1 nếu x-y là A-U còn lại f(x,y)=0
- Hàm p(x,y,z)=1 nếu |x-y|<3, ngược lại g(x,y,z)=0

Từ hệ thức $\frac{dU_{ij}}{dt} = -\frac{\partial E}{\partial V_{ij}}$ ta xác định được phương trình động học của nơron ij :

$$\frac{dU_{ij}}{dt} = -A_1 \left(\sum_{k=1, k \neq i}^{n} V_{ik} - 1 \right) b(i,j) - A_2 \left(\sum_{k=1, k \neq j}^{n} V_{jk} - 1 \right) b(i,j)$$

$$- B_1 \sum_{k<i<l<j} V_{kl} g(k,i,l) g(i,l,j) f(k,l) - B_2 \sum_{i<k<j<l} V_{kl} g(i,k,j) g(k,j,l) f(k,l) - C p(i,j,t)$$

Để tránh hiện tượng mạng bị rơi vào cực tiểu địa phương ta bổ sung thêm một số hạng leo đồi h(x) (*h(x)=1 nếu x=0 ngược lại h(x)=1*) vào phương trình động học của nơron. Khi đó phương trình động của nơ-ron ij có dạng:

$$\frac{dU_{ij}}{dt} = -A_1\left(\sum_{k=1,k\neq i}^{n} V_{ik} - 1\right) b(i,j) - A_2\left(\sum_{k=1,k\neq j}^{n} V_{jk} - 1\right) b(i,j)$$

$$- B_1 \sum_{k<i<l<j} V_{kl} g(k,i,l) g(i,l,j) f(k,l) - B_2 \sum_{i<k<j<l} V_{kl} g(i,k,j) g(k,j,l) f(k,l)$$

$$- Cp(i,j,t) + Dh\left(\sum_{k=1,k\neq i}^{n} V_{ik}\right)$$

với D là tham số mạng. Đối với mạng đã được xây dựng ta sẽ sử dụng hàm kích hoạt **McCuloch-Pitts** với ngưỡng bằng 0.

***** Tìm trạng thái ổn định của mạng

Để tìm trạng thái ổn định của mạng ứng với lời giải của bài toán ta giải phương trình động học bằng phương pháp Euler và xác định đầu ra V_{ij} của mỗi nơron tại mỗi bước thời gian nhờ hàm kích hoạt **McCuloch-Pitts**, tức là V_{ij}=1 nếu U_{ij}>0, ngược lại V_{ij}=0. Giá trị ban đầu $U_{ij}(0)$ được khởi tạo một cách ngẫu nhiên và cho mạng hoạt động đến khi đạt trạng thái ổn định, tức là khi ΔU_{ij}=0.

3.3. Một số đóng góp và kết quả thu được.

Chúng tôi đã thử nghiệm để tìm ra bộ tham số mạng tốt cho bài toán và giá trị khởi tạo ban đầu $U_{ij}(0)$ thích hợp để kết quả thu được sau khi mạng hoạt động và xử lý phù hợp với cấu trúc thực tế.

Chúng tôi đã viết một chương trình thử nghiệm cho bài toán và tốc độ thu được nhanh hơn với các thử nghiệm của Take Fuij do số lần lặp mạng chỉ còn 100 lần và phạm vi dự đoán rộng hơn các thử nghiệm của Take Fuij do có thể dự đoán cho nhiều đoạn RNA có cấu trúc cơ sở khác nhau nhưng kết quả vẫn khá chính xác.

3.4. Một số kết quả mô phỏng

Qua nhiều thử nghiệm chúng tôi đã tìm ra một bộ tham số mạng A_1=1, A_2=1, B_1=0.01, B_2=0.01, C= 1, D=0.1 đảm bảo chương trình cho kết quả nhanh và khá chính xác. Tuy nhiên, để chắc chắn rằng cấu trúc tìm được là hợp lý, sau khi mạng ngừng hoạt động sau một thời gian cho trước chúng tôi thực hiện một số kiểm tra trong chương trình để loại bỏ một số rất ít các cạnh thỏa mãn các ràng buộc.

Nguyên tắc thực hiện loại bỏ cạnh là:

- Đối với 2 cạnh cắt nhau: Kiểm tra 2 cạnh đó có liên quan đến đoạn ngăn xếp và chọn cắt bỏ một cạnh sao cho không làm phá huỷ cấu trúc của đoạn ngăn xếp và số cạnh hợp lệ còn lại là nhiều nhất.

- Đối với một lặp kẹp tóc: nếu 2 cơ sở nối nhau tạo thành lặp kẹp tóc không cách nhau ít nhất 2 cơ sở thì tiến hành cắt bỏ cạnh nối 2 cơ sở đó.

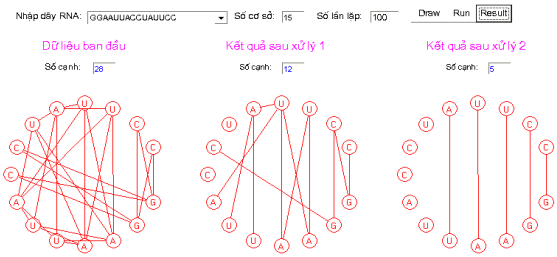
Trạng thái ban đầu của mạng được khởi tạo bằng hàm ngẫu nhiên **Rnd(20)-10** hoặc **Rnd(10)** tuỳ vào dãy RNA thử nghiệm.

Mô hình đã cho kết quả tốt với một số đoạn RNA có độ dài từ 5-30 cơ sở. Thời gian chạy chương trình có thể chấp nhận được, khoảng 3-20 giây.

Dưới đây là kết quả của một số thử nghiệm.

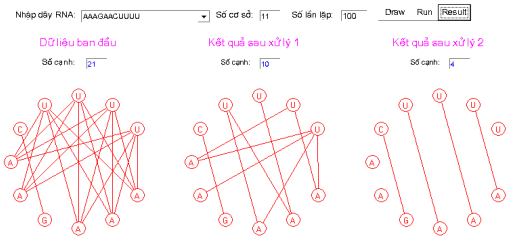
Thử nghiệm 1: Dãy cơ sở: GGAAUUACCUAUUCC.

Kết quả giống với thử nghiệm của Takefuji [7].



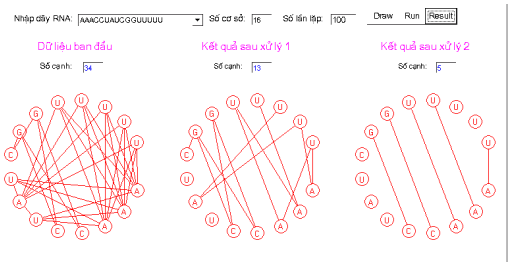
Thử nghiệm 2: Dãy cơ sở: AAAGAaCUUUU

Kết quả giống với thử nghiệm của Zuker [8].



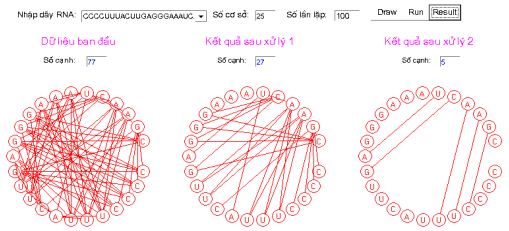
Thử nghiệm 3: Dãy cơ sở: AAACCUAUCGGUUUUU

Kết quả giống với thử nghiệm của Zuker [8].



Thử nghiệm 4: Dãy cơ sở CCCCUUUACUUGAGGGAAAUCAAGC

Cho kết quả giống với thử nghiệm của S.Deogun [5].



4. Kết luận và hướng phát triển

Trong bài báo này chúng tôi đã sử dụng mạng nơron Hopfield để giải quyết bài toán dự đoán cấu trúc thứ cấp RNA. Chúng tôi đã nghiên cứu tìm ra bộ giá trị tốt cho các tham số mạng, cách khởi tạo giá trị ban đầu thích hợp cho mạng và cách cắt tỉa một số ghép đôi không hợp lệ để quá trình dự đoán cấu trúc RNA được tốt hơn, nhanh hơn, chính xác hơn.

Để nâng cao hiệu quả của mạng nơ ron trong việc dự đoán cấu trúc RNA chúng tôi dự kiến sẽ phát triển tiếp các vấn đề sau:

- Thiết kế thêm một thuật toán huấn luyện các tham số mạng.
- Cải tiến thành mô hình mạng nơron Hopfield nhiều lớp.
- Thực hiện chuyển trạng thái ngẫu nhiên để tránh rơi vào cực tiểu địa phương

Tóm tắt

Vấn đề dự đoán cấu trúc thứ cấp RNA là một vấn đề quan trọng trong sinh học phân tử. Cấu trúc thứ cấp có thể được xác định trực tiếp bởi việc phân rã tia X nhưng đó là phương pháp khó khăn, chậm và tốn kém. Vì vậy nhiều mô hình toán học gần đây cho việc dự đoán đã được phát triển và mô phỏng trên máy tính. Trong bài này chúng tôi sử dụng cách tiếp cận mạng nơron, có tên là mạng nơron Hopfiled để xử lý tất cả các ràng buộc của cấu trúc thứ cấp RNA phải thoả mãn. Một chương trình trên máy tính đã được viết để mô phỏng vấn đề trên. Một số thử nghiệm được thực hiện và tất cả chúng đều thể hiện tính hiệu quả của mạng nơron.

Summary

Using Hopfield neural network for RNA secondary structure prediction

The RNA (RiboNucleic Acid) secondary structure prediction problem is a critical one in molecular biology. Secondary structure can be determined directly by x-ray diffraction, but this is difficult, slow, and expensive. Therefore, recently several mathematical models for prediction have been developed and simulated on computers. In this paper we use neural network approach, namely, Hopffield network for treating all constraints that RNA secondary structure must satisfy. A computer program for simulating the problem is written. Some experiments are performed and all they demonstrate the effectiveness of the neural network.

Tài liệu tham khảo

[1] Đặng Quang Á, *Một cách nhìn về việc sử dụng mạng Hopfield giải các bài toán thỏa mãn ràng buộc và tối ưu có ràng buộc*, B/c tại HT quốc gia "*Một số vấn đề chọn lọc của CNTT*", Hải Phòng 6/2001.

[2] Đặng Quang Á (2001) - *Ứng dụng của mạng nơron trong tính toán*, Sách "*Hệ mờ, mạng nơron và ứng dụng*", Chủ biên: Bùi Công Cường, Nguyễn Doãn Phước, Nxb KH và KT, Hà Nội, tr199-211.

[3] Đái Duy Ban (2006), *Công nghệ Gen*, Nxb Khoa học và Kĩ thuật, Hà Nội.

[4] Trần Văn Lăng (2003), *Một số tổng quan về sinh tin học*, Báo cáo nghiên cứu, Phân viện Công nghệ thông tin tại TP Hồ Chí Minh, .

[5] S.Deogun, R.Donis, F.Ma (2004)*RNA Secondary Structure Prediction with Simple Pseudoknots*, The Second Asia Pacific Bioinformatic Conference, New Zealand.

[6] Evan W.Steeg(1993)*, Neural Network, Adaptive Optimization and RNA Secondary Structure Prediction*, In book: *Artificial intelligence and molecular biology*, Lawrence Hunter, ed., American Assoc. Artificial intelligence, tr121-160.

[7] Y.Takefuji (1992), *Neural Network Parallel Computing*, Kluwer Acad. Publ.

[8] M.Zuker (1989), *On finding all suboptionmal folding of an RNA moclecule*, Cambridge University Press, London.