

KẾT HỢP THUẬT GIẢI DI TRUYỀN VỚI GẮN ĐÚNG ĐÁM NGUYÊN TỬ XÁC ĐỊNH VỊ TRÍ GẮN KẾT CO TRÊN DNA

Đến Tòa soạn 2-1-2007

ĐẶNG ỨNG VẬN

Trung tâm ứng dụng tin trong hóa học, Đại học Quốc gia Hà Nội

SUMMARY

The paper deals with combining genetic algorithm and quantum chemistry to determine the docking of small molecules on DNA. The atomic clusters including CO and different atoms of Ademin group of DNA molecule have been chosen to calculate. The results shown that there are five factors affected the solution evolution: the exactness of software using to calculate quantum energy of cluster, the search space, the selection of DNA atoms into cluster, selecting fitness function and mutate probability. The restriction of search space and enlargement of cluster size and selecting appropriate parameters will give the resonable optimal configuration of CO docking on DNA.

I - MỞ ĐẦU

Trong những công trình gần đây [1, 2] chúng ta có thể thấy sự quan tâm của nhiều tác giả đối với quá trình gắn kết những phân tử nhỏ trên các phân tử sinh học DNA, protein .. bằng các thuật giải khác nhau. Việc tìm bến đỗ (docking) cho một phân tử nhỏ trên các phân tử sinh học sẽ cho chúng ta biết được những thông tin về tương tác phân tử, về những vị trí được ưu tiên lựa chọn đối với mỗi phân tử nhỏ khác nhau và trạng thái cân bằng hồi phục (cân bằng động) cũng như năng lượng hồi phục ở mỗi vị trí tương ứng trên DNA. Nhược điểm chung của các phương pháp được trình bày trong [2] là chưa quan tâm đến các tính toán lượng tử cần thiết khi xuất hiện những tiếp cận nguyên tử - nguyên tử ở khoảng cách gần.

Cũng đã có những đề xuất sử dụng mô hình đám (cluster) đối với các hệ phân tử nhỏ - protein tương tự như khi nghiên cứu quá trình hấp phụ phân ly phân tử trên mặt vật rắn [3, 4]. Nhược điểm chung của kỹ thuật đám là tách riêng các nguyên tử được lựa chọn của phân tử sinh học ra khỏi hệ và khảo sát độc lập đối với

các nguyên tử còn lại theo một mô hình định sẵn. Rõ ràng rằng ở đây có vấn đề của bài toán tối ưu toàn cục. Với một phân tử như DNA thì chúng ta sẽ gặp phải vô số các cực tiểu cục bộ có thể giữ phân tử nhỏ cho dù trong một không gian hẹp gần các nhóm đặc trưng.

Theo hướng nghiên cứu sử dụng các gắn đúng lượng tử và bán lượng tử, các kỹ thuật tối ưu, hồi phục (relaxation) và động lực phân tử để nghiên cứu hệ phân tử nhỏ - protein, bài báo này trình bày những kết quả thu được khi sử dụng thuật giải di truyền kết hợp với gắn đúng hóa lượng tử để xác định cấu hình có năng lượng nhỏ nhất của CO khi gắn vào DNA.

II - CƠ SỞ LÝ THUYẾT

Việc áp dụng thuật giải di truyền cho bài toán tối ưu toàn cục đã được nhiều tác giả đề cập đến trong các lĩnh vực ứng dụng khác nhau [5]. Cơ sở lý thuyết của thuật giải di truyền (genetic algorithm - GA) có thể tìm thấy trong các sách giáo khoa [5, 6]. Có thể mô tả tóm tắt như sau:

Thuật giải bắt đầu với một tập các lời giải, mỗi lời giải được gọi là nhiễm sắc thể, tập lời giải được gọi là quần thể. Các lời giải trong một quần thể được dùng để tạo nên một quần thể mới với hy vọng rằng quần thể mới sẽ tốt hơn quần thể cũ. Những lời giải dùng để tạo quần thể mới (offspring) được lựa chọn theo độ thích nghi của chúng (fitness), càng thích nghi thì chúng càng có nhiều cơ hội hơn để tạo sinh. Điều này được lặp lại cho đến khi thỏa mãn một điều kiện nào đó (ví dụ như số quần thể được tạo ra hoặc có được lời giải tốt nhất). Dưới đây là sơ đồ tổng quát thuật giải di truyền cơ bản:

1. [Start] Tạo quần thể hỗn độn có n nhiễm sắc thể (lời giải khả dĩ cho bài toán)

2. [Fitness] Đánh giá độ thích nghi của mỗi nhiễm sắc thể trong quần thể

3. [New population] Tạo một quần thể mới bằng cách lặp lại những bước sau đây cho đến khi quần thể mới này được hoàn thiện:

i. [Selection] Lựa chọn hai nhiễm sắc thể cha mẹ trong quần thể theo độ thích nghi của chúng (càng thích nghi thì cơ hội được chọn càng lớn)

ii. [Crossover] Với xác suất lai ghép (crossover probability) thực hiện lai ghép bố mẹ để tạo một thế hệ con cái mới (offspring). Nếu không có sự lai ghép nào xảy ra thì con cái là bản sao chính xác của bố mẹ.

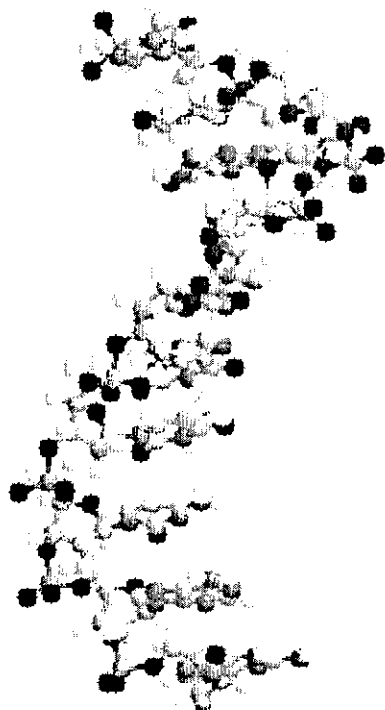
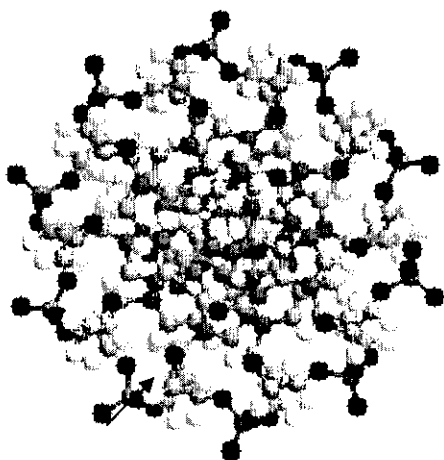
iii. [Mutation] Với xác suất đột biến, thực hiện đột biến thế hệ con cái ở mỗi ổ gen (vị trí trong nhiễm sắc thể).

iv. [Accepting] Xếp thế hệ con cái mới vào quần thể mới.

4. [Replace] Dùng quần thể mới được tạo ra cho việc tiếp tục thực hiện thuật giải.

5. [Test] Nếu điều kiện cuối cùng thỏa mãn thì dừng và cho ra lời giải tốt nhất.

6. [Loop] Quay về bước 2.



Hình 1: Chuỗi DNA (đơn) nhìn từ trục z (a) và mặt xy (b) cùng với phân tử CO (được chỉ bằng mũi tên):

●: O :N ○:P ●:C ○:H

Để có thể thực hiện được một bài toán cụ thể sẽ cần có nhiều tham số và sự sắp đặt cần thiết, ví dụ, phải trả lời các câu hỏi: 1) Làm thế nào để

tạo nhiễm sắc thể và chọn kiểu mã hóa nào. Rồi sau đó là hai toán tử cơ sở của GA là lai ghép và đột biến. 2) Chọn cha mẹ cho lai ghép như thế

nào. Có nhiều cách nhưng ý tưởng chính là chọn cha mẹ khỏe mạnh hơn để có những đứa con khỏe mạnh hơn 3) Sử dụng cá thể trội (elitism) để một lời giải tốt nhất của thế hệ được sao chép nguyên vẹn vào quần thể mới.

Việc áp dụng thuật giải di truyền, vì thế, tùy thuộc vào việc chọn độ thích nghi và khả năng xác định chính xác nó. Trong bài toán tìm bốn đồ cho CO trên DNA thì độ thích nghi chính là tính tối ưu của cấu hình gắn kết CO+DNA được biểu hiện qua năng lượng E.

III - PHƯƠNG PHÁP TÍNH

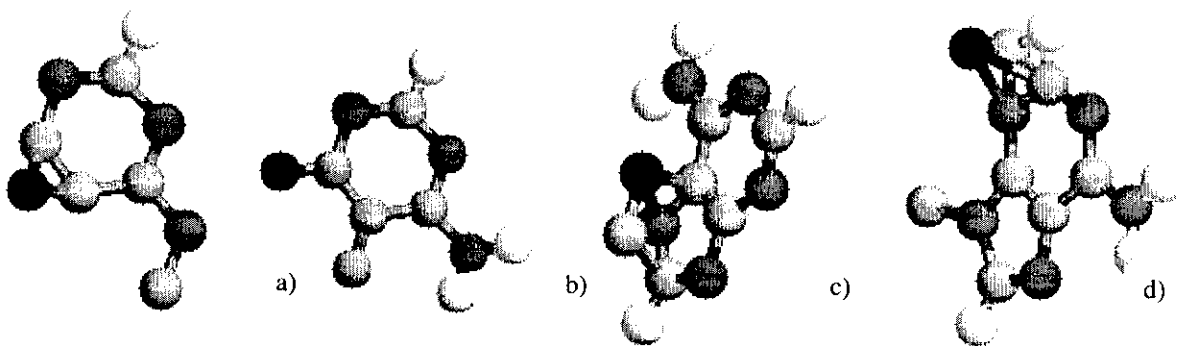
Một chuỗi DNA độc lập được đặt trong hộp mô phỏng có tính chất tuần hoàn. Kích thước hộp (A) là: $x(-9,7600, 9,7600)$, $y(-10,1980, 10,1980)$, $z(-17,6324, 17,6324)$. Chuỗi xoắn DNA được đặt quanh trục Z (xem hình 1). Không gian khảo sát xác lập các tọa độ của 2 nguyên tử O và C trong vùng không gian xung quanh một nhóm đặc trưng nào đó của DNA. Số tham số là 6 tạo nên một nhiễm sắc thể với các giá trị xác định của từng tham số. Số nhiễm sắc thể trong quần thể lời giải là 10.

Một số tham số quan trọng khác của thuật toán gồm xác suất biến dị $\text{pmutate}=0,05$, số thế hệ tối đa $\text{maxgen}=200$, xác suất lai tạp $\text{pcross}=0,8$, số con của mỗi cặp bố mẹ $\text{nchild}=1$, cực trị của các tham số là 0,0 và 1,0 (tức là

khi lập trình tính độ thích nghi chúng ta phải căn chỉnh tọa độ các nguyên tử vào trong giới hạn này) và số bit tối đa (32768) được sử dụng để mã hóa nhị nguyên tập giá trị của 6 tham số. Độ thích nghi của các nhiễm sắc thể được xác định từ năng lượng phân tử -E (dấu trừ để khi năng lượng cực tiểu thì hàm thích nghi cực đại).

Theo gần đúng đám nguyên tử một nhóm xác định các nguyên tử của DNA trong vùng khảo sát được lựa chọn tạo nên một đám phân tử (cluster) trên mặt DNA. Cách này tương tự với nhiều tác giả đã sử dụng (xem ví dụ [5]). Điều khác biệt là đám nguyên tử vẫn định vị trên phân tử sinh học mà không tách biệt khỏi hệ. Tương tác giữa các nguyên tử bên trong đám là tương tác lượng tử, tức là các tính toán lượng tử được áp dụng cho các nguyên tử bên trong đám này, còn tương tác của các phân tử ngoài đám cũng như tương tác giữa các nguyên tử trong đám với các nguyên tử ngoài đám được tính theo mẫu cơ học phân tử và thế kinh nghiệm L-J6-12.

Tùy thuộc vào vị trí tiếp cận của CO trên DNA các nguyên tử khác nhau sẽ được chọn. Số nguyên tử trong đám không được quá lớn và tùy thuộc vào năng lực tính của phần mềm hóa lượng tử và tập cơ sở (basic set) được sử dụng (xem hình 2). Để thích hợp với cách tính đám cần hiệu chỉnh lại subroutine tính lực lượng tử trong phần mềm SQUARED.



Hình 2: Các đám tối ưu thu được từ GA với các kích thước khác nhau

- a) Đám 10 nguyên tử ($\text{CO} + >\text{N}-\text{N}_2\text{HC}_4<$)
- b) Đám 12 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4<$)
- c) Đám 16 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4\text{N}_2\text{CH}$)
- d) Đám 17 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4\text{N}_2\text{C}_2\text{H}$) với cùng không gian khảo sát như c)

IV - KẾT QUẢ VÀ THẢO LUẬN

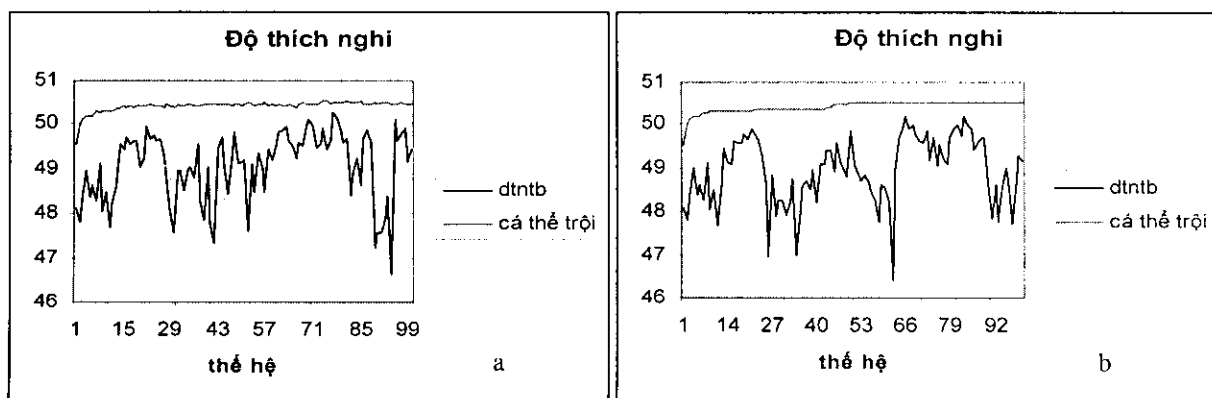
Các yếu tố ảnh hưởng đến sự tiến hóa lời giải

Có một số yếu tố ảnh hưởng đến kết quả tiến hóa lời giải theo thuật giải di truyền. Trước hết là độ ổn định của tính toán năng lượng phân tử của phần mềm hóa lượng tử. Chúng tôi đã sử dụng code của GAMESS. Nhìn chung tính toán trên GAMESS cho kết quả phù hợp với các phần mềm khác. Khi áp dụng cho bài toán "phân tử nhỏ - DNA" có một vướng mắc xảy ra là độ ổn định lời giải không cao. Kết quả tính SCF của hai lần kế tiếp thường là ít trùng nhau. Điều này

thể hiện trên hình 3a. Đường biểu diễn độ thích nghi của cá thể trội qua các thế hệ không nhẵn. Nếu đặt điều kiện để GA chấp nhận giá trị bé nhất của SCF trong các lần tính khác nhau bằng câu lệnh:

```
if (j.eq.irand).and. (funcval.lt.best0)  
funcval=best0
```

trong đó irand là số thứ tự trong quần thể của nhiễm sắc thể được sao chép nguyên xi từ cá thể trội của thế hệ trước đó thì ta sẽ được một đường biểu diễn độ thích nghi của cá thể trội nhẵn và hội tụ nhanh (hình 3b).



Hình 3: Độ thích nghi trung bình (dntnb) và của cá thể trội trong quá trình tiến hóa lời giải GA

Yếu tố thứ hai là không gian khảo sát. Khi mở rộng không gian khảo sát GA có thể tìm được cực tiểu toàn cục trong một không gian rộng hơn. Tuy vậy, nếu mở rộng quá mà không có điều kiện cưỡng bức sẽ rất dễ xảy ra hiện tượng "quá xa" tức là CO bị phân ly hoặc CO gắn kết sang nhóm chức khác của DNA. Đây là đặc điểm cơ bản của cách tìm tối ưu bằng GA so với các phương pháp tối ưu khác. Vì thế, trong quá trình tính toán thoát tiên cần tìm cực đại và cực tiểu tọa độ x, y và z của các nguyên tử thuộc DNA được đưa vào đám nguyên tử để xác định một không gian lõi (core). Không gian khảo sát được mở rộng so với không gian lõi về các hướng một khoảng cách bằng 1,5 Å, lớn hơn một chút so với liên kết C-C và C-O.

Yếu tố thứ ba có liên quan mật thiết đến không gian khảo sát là việc chọn các nguyên tử của DNA tham gia vào đám. Vì rằng tính toán

lượng tử chỉ được áp dụng cho các nguyên tử trong đám còn lực giữa các nguyên tử trong đám với các nguyên tử ngoài đám là lực cơ học phân tử nên dường như là năng lượng của đám được tính độc lập đối với các nguyên tử khác của DNA còn lại. Hệ quả là vị trí tối ưu của C hoặc O sẽ trùng với vị trí của nguyên tử gần đám nhất mà không được đưa vào đám - được hiểu là vị trí cắt đứt liên kết khi tách đám ra khỏi các nguyên tử còn lại của DNA.

Cụ thể là nếu khảo sát sự gắn kết của CO lên nhóm bazơ nitơ Ademin mà chọn đám 10 nguyên tử ($\text{CO} + >\text{N}-\text{N}_2\text{HC}_4<$) thì oxi của CO sẽ ưu tiên gắn kết lên cacbon 4 và 5 của vòng sáu cạnh thay cho hai N trong vòng 5 cạnh bị cắt và C sẽ gắn với N thay cho H của nhóm H_2N (Hình 2a). Hoặc nếu chọn đám 12 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4<$) thì rõ ràng C và O của CO sẽ ưu tiên gắn kết với hai C của vòng 6 cạnh thế

vào chỗ của hai N vòng 5 cạnh bị cắt (Hình 2b). Nếu chúng ta đưa các nguyên tử còn lại của nhóm bazơ nitơ Ademin vào đám - tức là sẽ có đám 16 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4\text{N}^*\text{CH}$) thì hoàn toàn có thể dự báo trước được là CO sẽ gắn vào một trong hai nguyên tử N^* của vòng 5 cạnh chồng lấn với vị trí của nhóm riboz trong DNA. Kết quả tính toán phù hợp với dự báo này (hình 2c). Như vậy, cần một giải pháp hợp lý hơn cho việc lựa chọn đám.

Yếu tố thứ tư nhưng quan trọng nhất là việc chọn hàm thích nghi. Nếu chỉ tính theo E thì về thực chất tính toán của chúng tôi chỉ khác biệt với việc tính toán theo đám phân tử của các tác giả khác về thuật giải tìm tối ưu hình học. Để có thể quan tâm đến cả các tương tác xa (với các nguyên tử ngoài đám đã chọn) chúng tôi đã tính thử với hàm thích nghi -RMS trong đó -RMS là căn bậc hai của tổng bình phương gradien trung bình của tất cả các nguyên tử trong đám. Kết quả thu được dựa trên RMS khi so sánh với cách tính dựa trên -E sẽ khác biệt do chịu ảnh hưởng của cả hai yếu tố: 1) một đẳng là tìm hàm E cực tiểu và một đẳng là tìm RMS cực tiểu; 2) một đẳng chỉ tính đến lực lượng tử trong đám (nếu tính theo E) và một đẳng xét đến toàn bộ tương tác của hệ đến đám này (theo RMS); 3) một đẳng phụ thuộc vào vị trí tương đối của các nguyên tử trong đám (theo E) và một đẳng phụ thuộc vào vị trí tương đối của các nguyên tử trong đám với các nguyên tử còn lại của hệ (theo RMS).

Điều có thể dự báo trước được là vị trí gắn kết của CO sẽ thay đổi về cơ bản so với trường hợp sử dụng hàm thích nghi -E. Chi tiết sẽ được trình bày trong những bài báo sau này. Khó khăn lớn nhất khi chọn RMS làm hàm thích nghi là ở năng lực hội tụ của phần mềm tính lực bán lượng tử.

Yếu tố thứ năm là xác suất đột biến. Khi tăng số nguyên tử trong đám và kèm theo là không gian khảo sát cho tọa độ các nguyên tử C và O của CO cũng rộng ra. Với những tính toán mà vị trí của C và O quá xa hoặc quá gần các nguyên tử trong đám cũng khiến cho việc tính toán năng lượng lượng tử theo thuật toán SCF không hội tụ. Trong trường hợp này E mặc định bằng 0. Với một quần thể mà độ thích nghi của các nhiễm sắc thể đều bằng nhau và bằng 0 thì

rất khó tạo quần thể mới có các nhiễm sắc thể có độ thích nghi lớn hơn nếu xác suất đột biến thấp. Vì thế, thay vì sử dụng các giá trị thông dụng của xác suất đột biến (khoảng 0,02) chúng tôi phải tăng pmutate lên tới 0,05 hoặc lớn hơn.

Hạn chế không gian khảo sát

Sau khi xem xét các yếu tố ảnh hưởng nêu trên chúng tôi đã sử dụng phương pháp hạn chế không gian khảo sát khi sử dụng -E làm hàm thích nghi. Điều này cũng có thể hiểu là những tính toán lượng tử được mở rộng xa hơn một vài nguyên tử so với không gian khảo sát vị trí của CO, tức là tăng thêm số nguyên tử trong đám nhưng không gian khảo sát không thay đổi. Tính toán được thực hiện với đám 17 nguyên tử ($\text{CO} + \text{H}_2\text{N}-\text{N}_2\text{HC}_4\text{N}_2\text{C}^*\text{H}$) trong đó có thêm một nguyên tử C^* gắn với một trong hai nguyên tử N của vòng 5 cạnh trong khi không gian khảo sát vẫn được giữ nguyên như trong trường hợp đám 16 nguyên tử. Kết quả cho thấy nguyên tử C này đã che chắn vị trí ưu tiên có được trong lần tính trước và CO gắn kết không phân li vào N và H của vòng 6 cạnh (hình 2d). Năng lượng đám là 84.05395 a.u. Điều quan trọng là việc tiếp tục tăng thêm số nguyên tử (ví dụ O và H gắn với C nêu trên) mà không mở rộng không gian khảo sát sẽ không ảnh hưởng đến vị trí gắn kết của CO mà chỉ thay đổi năng lượng của đám.

V - KẾT LUẬN

Lần đầu tiên chúng tôi đã áp dụng GA kết hợp với gắn đúng lượng tử khảo sát sự gắn kết phân tử nhỏ lên DNA. Tính toán với gắn đúng đám nguyên tử cho thấy thuật giải GA đáp ứng nhu cầu của bài toán và những điều kiện ảnh hưởng đến kết quả tính toán đã được khảo sát. Trên cơ sở đó tìm được phương pháp hạn chế không gian khảo sát đồng thời với việc mở rộng kích thước đám nguyên tử để xác định cấu hình gắn kết tối ưu và hợp lý của CO trên DNA. Những kết quả thu được về thuật toán và phương pháp tính, cũng như phần mềm tác giả xây dựng được có thể áp dụng cho việc nghiên cứu các hệ phân tử - protein/DNA khác.

Công trình nhận được sự tài trợ từ Bộ Khoa học và Công nghệ trong khuôn khổ đề tài, mã số 5.072.06. Tác giả xin chân thành cảm ơn.

TÀI LIỆU THAM KHẢO

1. A. P. Lyubartsev, A. Laaksonen, J. Biomol. Struc. Dyn., 16, 579 (1998).
2. R. D. Taylor, P. J. Jewsbury, J. W. Essex, J. Computer-Aided Mol. Design, 16, 151 - 166 (2002).
3. H. Luo, M. C. Lin. Chem. Phys. Letters, 343, 219 - 224 (2001).
4. Mai Tuyên, Đặng Đình Bạch, Nguyễn Quang Tùng, Nguyễn Thị Anh Sơn. Báo cáo toàn văn, Hội nghị toàn quốc các đề tài khoa học cơ bản trong lĩnh vực hóa lý - hóa lý thuyết. Hà Nội 2003, Tr. 112 - 119.
5. Hoàng Kiếm và Lê Hoàng Thái. Thuật giải di truyền. Cách giải tự nhiên các bài toán trên máy tính, Nxb. Giáo dục (2001).
6. Đặng Ứng Vận. Giáo trình Hóa tin cơ sở. Khoa Hóa học. Trường đại học Khoa học Tự nhiên, Đại học quốc gia Hà Nội (2006).