

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



VŨ ĐÌNH GIANG

**PHÂN TÍCH TRÌNH TỰ TRONG TIN SINH
HỌC VÀ ỨNG DỤNG TRÊN CƠ SỞ DỮ LIỆU
GENOME TÔM SÚ**

**CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ : 60.48.01**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

**NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. Nguyễn Long Giang**

Thái Nguyên - 2014

MỤC LỤC

MỤC LỤC.....	1
Danh mục các thuật ngữ.....	4
Danh sách bảng.....	5
Danh sách hình vẽ.....	6
MỞ ĐẦU.....	7
MỞ ĐẦU.....	7
Chương 1. CÁC KHÁI NIỆM CƠ BẢN.....	9
1.1. Các khái niệm cơ bản trong sinh học phân tử.....	9
1.2. Các bài toán cơ bản trong tin sinh học	12
1.3. Các ứng dụng của tin sinh học.....	13
1.4. Một số cơ sở dữ liệu sinh học lớn trên thế giới	14
Chương 2. BÀI TOÁN PHÂN TÍCH MỐI QUAN HỆ GIỮA CÁC TRÌNH TỰ	19
2.1. Bài toán phân tích mối quan hệ giữa hai trình tự.....	19
2.1.1. Giới thiệu bài toán	19
2.1.2. Phương pháp giải quyết bài toán	20
2.1.3. Thuật toán Needleman-Wunsch.....	23
2.2. Bài toán phân tích mối quan hệ cục bộ giữa hai trình tự.....	26
2.1.4. Giới thiệu bài toán	26
2.1.5. Thuật toán phân tích mối quan hệ cục bộ giữa hai trình tự	27
2.3. Tìm kiếm trình tự tương đồng bằng BLAST	31
2.3.1. Giới thiệu bài toán	31
2.3.2. Thuật toán BLAST.....	31
2.4. Bài toán phân tích mối quan hệ giữa đa trình tự.....	34
2.4.1. Giới thiệu bài toán	34
2.4.2. Thuật toán quy hoạch động.....	36
2.4.3. Thuật toán ngôi sao.....	39
2.4.4. Thuật toán sắp hàng lũy tiến.....	42
Chương 3. XÂY DỰNG CSDL HỆ GIEN TÔM SÚ VÀ TÍCH HỢP CÔNG CỤ BLAST	48
3.1. Kiến trúc hệ thống	48

3.2. Thiết kế cơ sở dữ liệu	49
3.2.1. Nguồn số liệu đầu vào	49
3.2.2. Thiết kế cơ sở dữ liệu	49
3.3. Thiết kế chức năng hệ thống.....	53
3.3.1. Mô hình phân cấp chức năng.....	53
3.3.2. Mô hình luồng dữ liệu	55
3.3.3. Đặc tả chi tiết một số chức năng cơ bản	57
3.4. Một số giao diện chương trình.....	64
3.4.1. Giao diện trang chủ.....	64
3.4.2. Nạp dữ liệu từ tệp XML.....	64
3.4.3. Nhập dữ liệu các trình tự Protein, Nucleotide, EST	64
3.4.4. Tra cứu thông tin.....	66
3.4.5. Tìm kiếm chuỗi tương đồng bằng BLAST	67
KẾT LUẬN.....	68
Tài liệu tham khảo	69

Danh mục các thuật ngữ

Thuật ngữ tiếng Anh	Thuật ngữ tiếng Việt
Bioinformatics	Tin sinh học
Molecular biology	Sinh học phân tử
Nucleic acid	Axit nuclêic
DNA	AND
RNA	ARN
Nucleotide	Nuclêôtít
Protein	Prôtêin
Amino Acid	Axit amin
Gene	Gien
Genome	Hệ gien
Cromosome	Nhiễm sắc thể
Sequence	Trình tự
Pairwise alignment	Sắp hàng trình tự

Danh sách bảng

Bảng 1.1. Tên đầy đủ, tên viết tắt của 5 loại nuclêôtít:	9
Bảng 1.2. Tên đầy đủ, tên viết tắt của 5 loại nuclêôtít.....	11
Bảng 2.1. Hai trình tự AND X và Y.....	19
Bảng 2.2. Hai trình tự Ξ và Ψ sau khi được sắp hàng.....	20
Bảng 2.3. Các cách sắp hàng khác nhau hai trình tự X và Y.....	21
Bảng 2.4. Ma trận điểm giữa các nuclêôtít.....	22
Bảng 2.5. Các cách sắp hàng khác nhau với tổng điểm khác nhau.....	23
Bảng 2.6. Bảng F của thuật toán quy hoạch động trên hai trình tự ADN.....	25
Bảng 2.7. Sắp hàng hai trình tự X và Y với tổng điểm lớn nhất.....	26
Bảng 2.8. Ma trận quy hoạch động F của bài toán sắp hàng cục bộ hai trình tự AND X và Y.....	30
Bảng 2.9. Sắp hàng cục bộ hai trình tự X và Y.....	30
Bảng 2.10. Minh họa ý tưởng của thuật toán BLAST.....	32
Bảng 2.11. Ba bất cặp XY, XZ, YZ tương thích với nhau có thể kết hợp thành sắp hàng 3 trình tự.....	36
Bảng 2.12. Ba bất cặp XY, XZ, YZ không tương thích với nhau để kết hợp thành sắp hàng 3 trình tự.....	36
Bảng 2.13. Sắp hàng tối ưu ba trình tự X, Y, Z.....	39

Danh sách hình vẽ

Hình 1.1. Minh họa cấu trúc một Axit amin	10
Hình 1.2. Trung tâm thông tin công nghệ sinh học Hoa Kỳ	15
Hình 1.3. Cấu trúc cơ bản của NCBI	16
Hình 2.1. Sắp hàng lũy tiến với 5 trình tự	43
Hình 3.1. Kiến trúc hệ thống CSDL hệ gien tôm Sú.....	49
Hình 3.2. Mô hình CSDL hệ gien tôm Sú.....	50

MỞ ĐẦU

Tin sinh học (bioinformatics) là một lĩnh vực khoa học sử dụng các công nghệ của các ngành tin học, toán học ứng dụng, thống kê và khoa học máy tính để giải quyết các bài toán trong sinh học. Tin sinh học bao gồm việc xây dựng, quản lý và lưu trữ nguồn dữ liệu quy mô toàn cầu liên quan đến sinh học, trên đó xây dựng và hoàn thiện các chương trình máy tính xử lý dữ liệu, là công cụ hỗ trợ hiệu quả cho việc nghiên cứu, khám phá bản chất sinh học của giới tự nhiên và sản xuất ra các sản phẩm sinh học mong muốn phục vụ đời sống con người. Tin sinh học có tính ứng dụng cao trong cuộc sống, đặc biệt là trong lĩnh vực công nghệ sinh học, nông nghiệp và y dược. Các bài toán cơ bản trong tin sinh học bao gồm: *quản lý và lưu trữ dữ liệu, phân tích mối quan hệ giữa các trình tự, dự đoán cấu trúc các trình tự, mô hình hóa, nghiên cứu tiến hóa*. [4]

Một trong những bài toán quan trọng trong tin sinh học là phân tích mối quan hệ giữa các trình tự, gọi tắt là phân tích trình tự. Các bài toán cơ bản trong phân tích trình tự là: tìm kiếm trình tự tương đồng trong cơ sở dữ liệu; sắp hàng trình tự; chuyển đổi trình tự. Mục tiêu của phân tích trình tự là:

- Xác định các gen và các chức năng của từng gen.
- Xác định sự lặp lại của các trình tự.
- Xác định protein dựa trên quy tắc sắp đặt của các biểu thức gen.
- Xác định các vùng chức năng khác nhau của ADN.

Mục tiêu của luận văn là:

1) Nắm bắt được các khái niệm cơ bản trong tin sinh học và các cơ sở dữ liệu sinh học lớn trên thế giới, các phương pháp giải quyết bài toán sắp hàng trình tự, một trong những bài toán cơ bản trong phân tích trình tự.

2) Xây dựng cơ sở dữ liệu cục bộ lưu trữ các chuỗi gen tôm sú (bao gồm các chuỗi nuclêôtit, protein và EST) và tích hợp các công cụ phân tích trình tự nhằm mục đích làm sáng tỏ các vấn đề nghiên cứu lý thuyết. Dữ liệu được thu thập từ Phòng công nghệ AND ứng dụng - Viện Công nghệ sinh học (nay là Viện Genome học) và từ các cơ sở dữ liệu sinh học trên Internet.

Đối tượng nghiên cứu của luận văn là các chuỗi gene tôm Sú được thu thập từ Viện Công nghệ sinh học và các chuỗi gene tôm Sú từ ngân hàng gene thế giới (genbank), bao gồm các chuỗi EST, Nucleotide và Protein.

Phạm vi nghiên cứu lý thuyết là bài toán phân tích trình tự trong tin sinh học, phạm vi nghiên cứu thực nghiệm là xây dựng cơ sở dữ liệu và tích hợp công cụ BLAST tìm kiếm trình tự tương đồng trong cơ sở dữ liệu các trình tự gen tôm Sú (bao gồm các trình tự nucleôtit, protein và EST)

Phương pháp nghiên cứu của luận văn là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Về nghiên cứu lý thuyết: luận văn thực hiện tổng hợp các khái niệm và các kết quả nghiên cứu về sắp hàng trình tự. Về nghiên cứu thực nghiệm: luận văn thực hiện xây dựng cơ sở dữ liệu và tích hợp công cụ tìm kiếm trình tự tương đồng BLAST trên cơ sở dữ liệu được xây dựng nhằm sáng tỏ các vấn đề về lý thuyết.

Bố cục của luận văn gồm phần mở đầu và hai chương nội dung, phần kết luận và danh mục các tài liệu tham khảo.

Chương 1 trình bày các khái niệm cơ bản trong tin sinh học, các bài toán cơ bản trong tin sinh học, các cơ sở dữ liệu sinh học lớn trên thế giới và một số ứng dụng của tin sinh học. Chương 2 trình bày bài toán phân tích mối quan hệ giữa các trình tự và các thuật toán so sánh trình tự. Chương 3 trình bày nội dung về ứng dụng thử nghiệm làm sáng tỏ các vấn đề nghiên cứu lý thuyết, bao gồm thiết kế và xây dựng một cơ sở dữ liệu lưu trữ các trình tự gen của con tôm Sú, tích hợp công cụ tìm kiếm trình tự tương đồng BLAST trên cơ sở dữ liệu cục bộ được xây dựng.

Cuối cùng, phần kết luận nêu những đóng góp của luận văn, hướng phát triển tiếp theo.

Chương 1. CÁC KHÁI NIỆM CƠ BẢN

1.1. Các khái niệm cơ bản trong sinh học phân tử

Tin sinh học (Bioinformatics) là lĩnh vực khoa học mới có tính ứng dụng cao trong cuộc sống, đặc biệt là trong lĩnh vực công nghệ sinh học, nông nghiệp và y-dược. Tin sinh học là lĩnh vực khoa học liên ngành, trong đó sinh học và tin học đóng vai trò chủ đạo. Về cơ bản, tin sinh học tập trung vào nghiên cứu, phát triển và áp dụng các phương pháp và công cụ tin học để giải quyết các bài toán trong sinh học.

Tiếp theo, luận văn giới thiệu một số khái niệm cơ bản trong sinh học phân tử. Sinh học phân tử (molecular biology) là một nhánh của sinh học (biology), tập trung nghiên cứu các sinh vật ở mức độ phân tử. Cụ thể là, sinh học phân tử tập trung giải trình tự (sequencing) và phân tích các trình tự nuclêôtit (trình tự ADN), các trình tự axit amin (trình tự prôtêin). Trong phần này, luận văn tập trung giới thiệu các kiến thức cơ bản trong sinh học phân tử để sử dụng ở các chương sau.

1) Axit nuclêic và nuclêôtit

Axit nuclêic (nucleic acid) là một đại phân tử sinh học (large biological molecule) mang thông tin di truyền mã hóa các chức năng, và đặc điểm của mọi sinh vật sống. Axit nuclêic gồm hai loại: ADN (Axit Deoxyribo Nuclêic) và ARN (Axit Ribo Nuclêic).

Thành phần cơ bản cấu tạo một trình tự axit nuclêic là các phân tử hóa học nuclêôtit (nucleotide). Trình tự ADN chứa bốn loại nuclêôtit khác nhau là: Adenine, Cytosine, Guanine, và Thymine. Trình tự ARN có thành phần tương tự như trình tự ADN, ngoại trừ nuclêôtit Thymine được thay thế bởi nuclêôtit Uracil. Tức là, ARN chứa 4 loại nuclêôtit: Adenine, Cytosine, Guanine, và Uracil. Tên đầy đủ, tên viết tắt của năm loại nuclêôtit được mô tả ở Bảng 1.1.

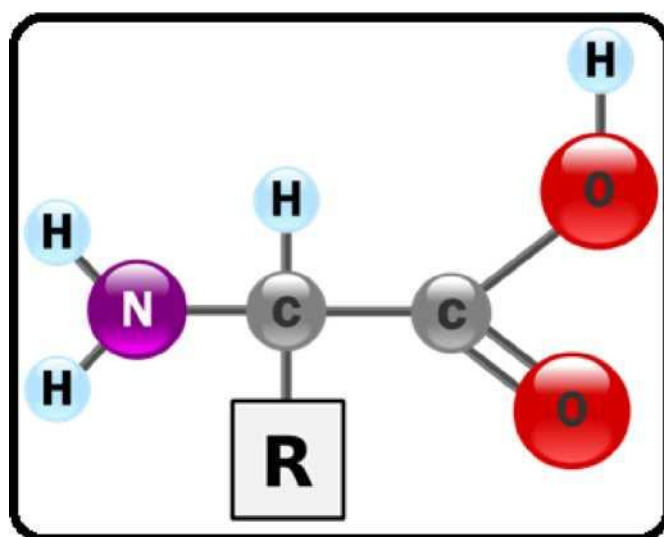
Bảng 1.1. Tên đầy đủ, tên viết tắt của 5 loại nuclêôtit:

STT	Tên đầy đủ	Tên viết tắt
1	Adenine	A
2	Cytosine	C
3	Guanine	G
4	Thymine	T
5	Uracil	U

Do đó, thông tin về một trình tự ADN được biểu diễn bằng một trình tự các nuclêôtít nằm trên một sợi (các nuclêôtít nằm trên sợi còn lại có thể suy luận dựa theo quy tắc trên). Để đơn giản, một trình tự ADN sẽ được biểu diễn bởi một chuỗi ký tự chứa 4 loại ký tự: A, C, G và T (tên viết tắt của 4 loại nuclêôtít). Ví dụ, "CAGTTGACGGCGAACCGTGCGAGCAGACGGTCGTT" là một trình tự ADN. Với cách biểu diễn này, thông tin về các trình tự DN có thể được lưu giữ, tìm kiếm, và trao đổi một cách hiệu quả.

2) Protein và axit amin

Prôtêin/trình tự prôtêin (protein) là loại dữ liệu phổ biến và quan trọng trong sinh học phân tử. Nó quyết định đến chức năng, quá trình phát triển, cũng như các bệnh tật của các sinh vật sống. Prôtêin được cấu tạo bởi một trình tự các axit amin (amino acid), trong đó mỗi axit amin là một hợp chất hữu cơ được tạo bởi ba thành phần chính là: nhóm amin (NH_2), nhóm cacboxyl (COOH) và nhóm R quyết định tính chất của axit amin (xem Hình 1.1)



Hình 1.1. Minh họa cấu trúc một Axit amin

Trong tự nhiên có 20 loại axit amin khác nhau như mô tả ở Bảng 1.2. Mỗi axit amin có tên đầy đủ, tên viết tắt 3 ký tự và tên viết tắt 1 ký tự. Thông thường, chúng ta sử dụng tên viết tắt một ký tự để biểu diễn một axit amin.

Trình tự axit amin có thể được biểu diễn bằng một chuỗi ký tự chứa 20 loại ký tự khác nhau, là tên viết tắt của 20 loại axit amin khác nhau. Ví dụ:

‘ESPQIRRD MGRLCATWPSK DSEDGAGTALRAATPLTANGATTTGLSVTLA
PKQTNWDECWSSPCQNGGTCVDGVAYYNCTCPEGFSGSNCEENVDE’ là
một trình tự axit amin. Với cách biểu diễn này, chúng ta có thể dễ dàng lưu giữ các