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

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	Axít nuclêic
DNA	AND
RNA	ARN
Nucleotide	Nuclêôtít
Protein	Prôtêin
Amino Acid	Axít amin
Gene	Gien
Genome	Hệ gien
Cromosome	Nhiễm sắc thể
Sequence	Trình tự
Pairwise alignment	Sắp hàng trình tự

Danh sách bảng

Danh sách hình vẽ

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 gien và các chức năng của từng gien.
- Xác định sự lặp lại của các trình tự.
- Xác định protêin dựa trên quy tắc sắp đặt của các biểu thức gien.
- 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 gien tôm sú (bao gồm các chuỗi nuclêôtít, protêin 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 gien tôm Sú (bao gồm các trình tự nucleôtít, protêin 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 gien 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 nuclêôtít (trình tự ADN), các trình tự axít 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) Axít nuclêic và nuclêôtít

Axít 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. Axít nuclêic gồm hai loại: ADN (Axít Deoxyribo Nuclêic) và ARN (Axít Ribo Nuclêic).

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

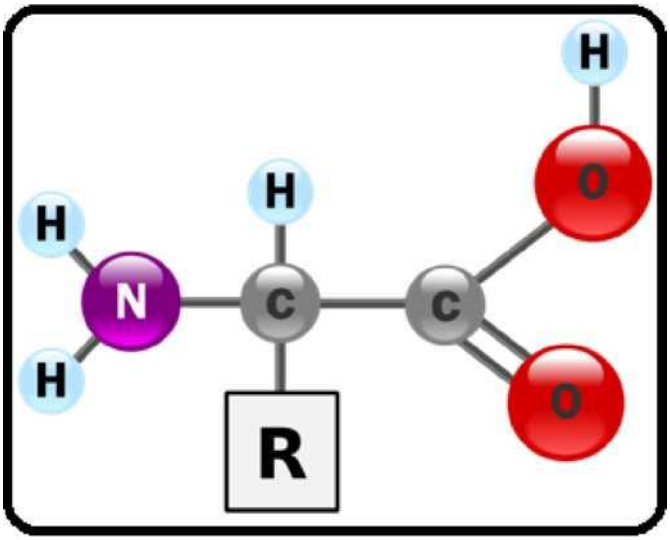
Bảng 1.1. *Tên đầy đủ, tên viết tắt của 5 loại nuclêôtít:*

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 xâu kí tự chứa 4 loại kí 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 axít amin (amino acid), trong đó mỗi axít amin là một hợp chất hữu cơ được tạo bởi ba thành phần chính là: nhóm amin (NH2), nhóm cacboxyl (COOH) và nhóm R quyết định tính chất của axít amin (xem Hình 1.1)



Hình 1.1. *Minh họa cấu trúc một Axit amin*

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

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

'ESPQIRRDMGRLCATWPSKDSEDGAGTALRAATPLTANGATTTGLSVTLA PKQTNWDECWSSPCQNGGTCVDGVAYYNCTCPEGFSGSNCEENVDE' là một trình tự axít amin. Với cách biểu diễn này, chúng ta có thể dễ dàng lưu giữ các