ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN VĂN THÀNH

PHÁT HIỆN CÁC ĐỘT BIẾN ĐẢO ĐOẠN TRONG HỆ GEN GIẢI MÃ TỪ THIẾT BỊ ĐỌC TRÌNH TỰ THẾ HỆ MỚI

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên – 2014

LỜI CAM ĐOAN

Tôi xin cam đoan: Luận văn này là công trình nghiên cứu thực sự của cá nhân, được thực hiện dưới sự hướng dẫn khoa học của ***Tiến sĩ Nguyễn Cường.*** Các số liệu, những kết luận nghiên cứu được trình bày trong luận văn này trung thực và chưa từng được công bố dưới bất cứ hình thức nào.

Tôi xin chịu trách nhiệm về nghiên cứu của mình.

Học viên

Nguyễn Văn Thành

LỜI CÁM ƠN

Lời đầu tiên, tôi xin chân thành cám ơn ***Tiến sĩ Nguyễn Cường*** người đã trực tiếp hướng dẫn tôi hoàn thành luận văn. Với những lời chỉ dẫn, những tài liệu, sự tận tình hướng dẫn và những lời động viên của Thầy đã giúp tôi vượt qua nhiều khó khăn trong quá trình thực hiện luận văn này.

Tôi cũng xin cám ơn quý Thầy (Cô) giảng dạy chương trình cao học "**Khoa học máy tính**" đã truyền dạy những kiến thức quý báu, những kiến thức này rất hữu ích và giúp tôi nhiều khi thực hiện nghiên cứu.

Xin cám ơn các quý Thầy (Cô) công tác tại Trường Đại học Công nghệ thông tin và truyền thông – Đại học Thái Nguyên đã tạo điều kiện cho tôi được tham gia và hoàn thành khóa học.

Tôi xin chân thành cám ơn.

Học viên

Nguyễn Văn Thành

MỤC LỤC

DANH MỤC CÁC HÌNH ẢNH

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC TỪ VIẾT TẮT-THUẬT NGỮ

STT	Từ viết tắt/thuật ngữ	Nghĩa/Mô tả
1	DNA	Deoxyribo Ducleic Acid
2	BP	Base Pair
3	GB	Giga Base Pair
4	NST	Nhiễm sắc thể
5	DNA senquencing	Đọc trình tự DNA
6	HGP	Dự án hệ giải trình tự hệ gen con người
7	Nucleotide	Là các trình tự A,T,G,C
8	SBS	Đọc trình tự bằng sợ tổng hợp
9	SBL	Đọc trình tự gắn nối
10	PCR	Kỹ thuật khuếch đại gen
11	Nanowell	Giếng nano
12	CGIAR	Chuỗi thể hiện số base được ánh xạ/mất/thêm so với tham chiếu
13	SNP	Đa hình đơn điểm/đơn nucleotide
14	CNV	Bản sao số biến thể
15	InDel	Vị trí thể hiện sự chèn hoặc xóa trong gen
16	BWA (hoặc Bowtie)	Công cụ ánh xạ trình tự với dữ liệu tham chiếu
17	TMAP	Chương trình để xây dựng bản đồ di truyền
18	Wgsim	Công cụ mô phỏng các đoạn trình tự ngắn từ dữ liệu hệ gen tham chiếu
19	Single end reads	Phương pháp đọc trình tự theo chiều đơn
20	PPV	Dự đoán dương tính giả
21	Hg19	Trình tự hệ gen người phiên bản 19
22	MAQ	Phần mềm lập bản đồ cho các trình tự ngẫn từ máy đọc trình tự thế hệ mới

PHẦN MỞ ĐẦU

Trong nghiên cứu về sinh học hiện đại có nhiều công nghệ và giải pháp được ứng dụng để phân tích, tổng hợp dữ liệu về cấu trúc và trình tự hệ gen của các loài sinh vật. Việc phân tích và tổng hợp bộ dữ liệu này yêu cầu một hệ thống cấu trúc lưu trữ đáp ứng đủ tính chất về độ phức tạp và độ lớn của bộ dữ liệu kết quả. Các thiết bị đọc trình tự gen được ra đời để giải quyết các vấn đề nêu trên. Các thiết bị đọc trình tự gen là những công cụ xác định thứ tự các nucleotide gắn kết với nhau dọc theo chiều dài của gen và trình tự gắn kết nhau của các nucleotide được gọi là trình tự gen. Trong đó, đọc trình tự thế hệ mới là một bước tiến vượt bậc về công nghệ đọc trình tự, từ khả năng đọc trình tự đoạn ngắn 1500bp (Sanger) hay 100 bp (pyrosequencing) của các thiết bị đọc trình tự trước đó, đọc trình tự thế hệ mới cho phép đọc được từ 8gb đến 600gb, có nghĩa là cho phép đọc trình tự nguyên bộ gen của bất kỳ loài sinh vật nào.

Với mong muốn hiểu chi tiết về cấu trúc gen các nhà nghiêncứu sinh học luôn muốn đọc trình tự hoàn chỉnh các gen của tất cả các loài sinh vât trong tự nhiên, bao gồm cả hệ gen của con người và toàn bộ trình tự gen khác của nhiều động, thực vật, vi sinh vật, đồng thời qua việc nghiên cứu đó có thể phát hiện ra những đột biến cấu trúc trong hệ gen được giải mã. Đặc biệt là dạng đột biến đảo đoạn, loại đột biến này ít gây ảnh hưởng đến sức sống của cá thể, nhưng nógóp phần lớn tăng cường sự sai khác giữa các nhiễm sắc thể (NST) tương đồng điều này dẫn đến tăng sự đa dạng giữa các thứ,các nòi trong cùng một nòi, ít ảnh hưởng tới sức sống của cá thểvà trong đó sự sắp xếp lại hệ gen trên NST do đột biến đảo đoạn góp phần tạo sự đa dạng trong tự nhiên. Đối với con người việc đọc trình tự hệ gen rất quan trọng, nó góp phần trong việc nghiên cứu sinh học cơ bản và trong nhiều lĩnh vực ứng dụng như chẩn đoán bệnh tật, công nghệ sinh học, sinh học pháp y, sinh học hệ thống...Nhận thấy tính thiết thực của vấn đề và với sự

định hướng của giáo viên hướng dẫn, học viên đã chọn đề tài "**Phát hiện các đột biến đảo đoạn trong hệ gen giải mã từ thiết bị đọc trình tự thế hệ mới**" để làm rõ các vấn đề đã nêu trên.

Đối tượng và phạm vi nghiên cứu

- ✓ Kiến trúc về các thành phần và các đột biến cấu trúc trong hệ gen.
- ✓ Ứng dụng thiết bị đọc trình tự thế hệ mới trong công nghệ sinh học.
- ✓ Phương pháp phát hiện các đột biến đảo đoạn khi sử dụng các thiết bị đọc trình tự thế hệ mới để giải mã.

Hướng nghiên cứu của đề tài

- ✓ Nghiên cứu, tìm hiểu mô hình, cách làm việc và giải mã hệ gen từ thiết bị đọc trình tự thế hệ mới.
- ✓ Nghiên cứu cấu trúc dữ liệu, các phương pháp tiền xử lý và lắp ráp hệ gen từ thiết bị đọc trình tự thế hệ mới.
- ✓ Tìm hiểu, tham khảo các tài liệu liên quan đến các đột biến đảo đoạn trong hệ gen, từ đó xây dựng thuật toán phát hiện ra các đột biến gen đảo đoạn trong hệ gen giải mã từ thiết bị đọc trình tự thế hệ mới.

Phương pháp nghiên cứu

- ✓ Nghiên cứu lý thuyết về các thiết bị đọc trình tự thế hệ mới, đột biến gen đảo đoạn và cách phát hiện đột biến đảo đoạn trong hệ gen giải mã từ thiết bị đọc trình tự thế hệ mới.
- ✓ Thiết kế, đặc tả, xây dựng chương trình, phương pháp đọc trình tự gen và phát hiện đột biến đảo đoạn.
- ✓ Qua những phát hiện về đột biến đảo đoạn đưa ra kết luận.

Ý nghĩa khoa học của đề tài

- ✓ Làm cơ sở để phát hiện ra các đột biến đảo đoạn trong hệ gen.
- ✓ Ứng dụng như chẩn đoán bệnh, sinh học pháp y, sinh học hệ thống.