

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 THỂ 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 CẢM ƠN

Lời đầu tiên, tôi xin chân thành cảm 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 cảm 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 cảm 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 cảm ơn.

Học viên

Nguyễn Văn Thành

MỤC LỤC

LỜI CAM ĐOAN.....	1
LỜI CẢM ƠN.....	3
MỤC LỤC	4
DANH MỤC CÁC HÌNH ẢNH	6
DANH MỤC CÁC BẢNG BIỂU	7
DANH MỤC CÁC TỪ VIẾT TẮT-THUẬT NGỮ.....	8
PHẦN MỞ ĐẦU	9
Chương 1.TỔNG QUAN VỀ TIN SINH HỌC VÀ BÀI TOÁN PHÁT HIỆN ĐỘT BIẾN ĐẢO ĐOẠN.....	11
1.1 - Tổng quan về Tin sinh học.....	11
1.2 – Cơ sở lý thuyết bài toán đột biến đảo đoạn	12
1.2.1 - Gen và đột biến cấu trúc hệ gen	12
1.2.2 - Phương pháp phát hiện sự biến đổi cấu trúc trong bản đồ gen.....	15
1.2.3 - Định dạng cơ sở dữ liệu	23
1.2.3 - Bài toán đột biến đảo đoạn	27
1.3 - Các công cụ giải quyết bài toán đảo đoạn.....	33
1.3.1 - Chương trình Wgsim	33
1.3.2 - Chương trình TMAP	33
1.3.3 - Chương trình BWA và Bowtie.	33
Chương 2. MỘT SỐ THUẬT TOÁN PHÁT HIỆN ĐỘT BIẾN	34
2.1 - Thuật toán ma trận điểm.....	35
2.2 - Thuật toán Blast.....	37

2.3 - Thuật toán lai GA-SA.....	42
2.4 - Thuật toán Needleman – Wunsch	45
2.5 - Thuật toán Smith-Waterman	49
Chương 3. CÀI ĐẶT THUẬT TOÁN VÀ ĐÁNH GIÁ KẾT QUẢ	56
3.1 - Ánh xạ các đoạn trình tự.	57
3.2 - Xử lý SAM và khởi tạo điểm dừng khả dĩ.	58
3.3 - Lọc và hoàn thiện điểm dừng.	61
3.4 - Mô phỏng dữ liệu và thống kê kết quả ánh xạ.	64
3.5 - Đánh giá kết quả phân tích.	68
3.6 - So sánh với các phương pháp hiện tại.	74
3.7 - Những hạn chế và cách khắc phục	76
KẾT LUẬN	78
TÀI LIỆU THAM KHẢO	80

DANH MỤC CÁC HÌNH ẢNH

Hình 1.1: Trong mỗi tế bào có một nhân chính giữa.....	13
Hình 1.2: Gen được cấu tạo từ DNA. Mỗi NST có nhiều gen.....	13
Hình 1.3: Cấu trúc một phần của gen.....	13
Hình 1.4: Đột biến đảo đoạn trong hệ gen.	15
Hình 1.5: Các giai đoạn của đọc trình tự thế hệ mới.....	22
Hình 1.6: Định dạng SAM.....	25
Hình 1.7: Bản sao - số biến thể (CNVs).....	28
Hình 1.8: Đồ thị gia tăng CNV và InDel đã thêm vào CSDL từ năm 2002.....	28
Hình 1.9: Đồ thị sự phân phối kích thước CNVs trong cơ sở dữ liệu.....	29
Hình 1.10: Phần lớn sự đảo đoạn đến nay có cỡ 10 đến 100kb.....	30
Hình 2.1: Ma trận thay thế BLOSUM.....	40
Hình 2.2: Ma trận thay thế PAM.....	40
Hình 3.1: Sự giống hệt của quá trình ánh xạ r_1 , r_2 trên vùng đảo ngược...	57
Hình 3.2: Những vùng được lựa chọn dựa vào điểm dừng trái và phải.....	62
Hình 3.4: Số lượng đảo đoạn trong các NST khác nhau.....	65
Hình 3.5: Phân phối kích thước của 90 đảo đoạn.....	65
Hình 3.6: Tổng số trình tự của ánh xạ bởi Map1 và Map2 đọc lý tưởng.....	67
Hình 3.7: Tổng số trình tự của ánh xạ bởi Map1 và Map2 cho trình tự lỗi...	67
Hình 3.8: Những giá trị dương tính giả trong pha thứ 1 và pha thứ 2.....	72
Hình 3.9: Tính nhạy cảm ở pha 1 và pha 2.....	73
Hình 3.10: Dự đoán giá trị dương tính giả ở pha 1 và pha 2.....	73
Hình 3.11: Tính nhạy cảm ở pha 1 và pha 2 cho trình tự có lỗi.....	74
Hình 3.12: PPV ở pha 1 và pha 2 cho trình tự có lỗi.....	74

Hình 3.13: So sánh Inverse Variant với BreakDancer dựa vào điểm dừng.....	76
Hình 3.15: So sánh Inverse Variant với BreakDancer dựa vào tính nhạy cảm, PPV và F-Score.....	76

DANH MỤC CÁC BẢNG BIỂU

Bảng 1.1: Các thẻ định danh trong SAM.....	25
Bảng 1.2: Định nghĩa cờ đảo bit trong SAM.....	25
Bảng 1.3: Mô tả chuỗi CIGAR.....	26
Bảng 1.4 Bảng cho thấy CNVs và đảo đoạn.....	28
Bảng 3.1: Những tham số được đặt mô phỏng cho các đoạn trình tự có lỗi.	66
Bảng 3.2: Kết quả của Inverse Variant ở trình tự lý tưởng có độ dài 100bp.....	69
Bảng 3.3: Kết quả của Inverse Variant ở trình tự lý tưởng có độ dài 200bp.....	69
Bảng 3.4: Kết quả của Inverse Variant ở trình tự lý tưởng có độ dài 400bp.....	69
Bảng 3.5: Kết quả của Inverse Variant ở trình tự lý tưởng với độ bao phủ 10X..	70
Bảng 3.6: Kết quả của Inverse Variant ở trình tự lỗi với độ bao phủ là 10X.....	70
Bảng 3.7: Bảng so sánh Inverse Variant với BreakDancer.....	75

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 sự 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 ngắn 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 nghiên cứu sinh học luôn muốn đọc trình tự hoàn chỉnh các gen của tất cả các loài sinh vật 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 nó góp 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 thể và 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.