

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

NGUYỄN THỊ THANH NGỌC

**MỘT SỐ KỸ THUẬT ỨNG DỤNG ĐỂ LẮP RÁP HỆ
GEN VỚI DỮ LIỆU TRÌNH TỰ NGẮN TRONG TIN
SINH HỌC**

Chuyên ngành: Khoa học máy tính

Mã số chuyên ngành: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN CƯỜNG

Thái Nguyên – 2014

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

NGUYỄN THỊ THANH NGỌC

**MỘT SỐ KỸ THUẬT ỨNG DỤNG ĐỂ LẮP RÁP
HỆ GEN VỚI DỮ LIỆU TRÌNH TỰ NGẮN TRONG
TIN SINH HỌC**

Chuyên ngành: Khoa học máy tính

Mã số chuyên ngành: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN CƯỜNG

Thái Nguyên – 2014

LỜI CAM ĐOAN

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

Học viên

Nguyễn Thị Thanh Ngọc

LỜI CẢM ƠN

Lời đầu tiên, tôi xin chân thành cảm ơn Tiến sĩ Nguyễn Cường, người thầy đã trực tiếp hướng dẫn tôi hoàn thành luận văn này. Thầy đã tận tình hướng dẫn, chỉ bảo và cung cấp những tài liệu liên quan đồng thời động viên tinh thần giúp tôi vượt qua nhiều khó khăn trong quá trình thực hiện luận văn. Tôi cũng xin chân thành cảm ơn quý Thầy (Cô) giảng dạy chương trình cao học chuyên ngành “ Khoa học máy tính” đã truyền đạt những kiến thức hữu ích và giúp tôi 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 khoá học.

Tôi xin chân thành cảm ơn!

Thái Nguyên, ngày tháng năm 2014

Học viên

Nguyễn Thị Thanh Ngọc

MỤC LỤC

LỜI CAM ĐOAN.....	i
LỜI CẢM ƠN	iv
MỤC LỤC.....	v
DANH MỤC CÁC TỪ VIẾT TẮT.....	iv
DANH MỤC BẢNG BIỂU	viii
DANH MỤC HÌNH ẢNH	ix
MỞ ĐẦU.....	1
Chương 1: CƠ SỞ LÝ THUYẾT CỦA BÀI TOÁN LẮP RÁP TRÌNH TỰ GEN	3
1.1 Bài toán lắp ráp trình tự gen.....	3
1.2. Định dạng cơ sở dữ liệu và những sai số thường gặp trong bài toán lắp ráp.....	7
1.3. Ứng dụng của công nghệ đọc trình tự gen	13
Chương 2: MỘT SỐ KỸ THUẬT LẮP RÁP HỆ GEN VỚI DỮ LIỆU TRÌNH TỰ ĐOẠN NGẮN TRONG TIN SINH HỌC.....	15
2.1. Thuật toán Overlap Layout Consensus (OLC)	15
2.2. Thuật toán sử dụng Đồ thị De Bruijn.....	17
2.3. Thuật toán Short Sequence Assembler (SSA)	20
2.3.1. Giới thiệu về thuật toán SSA	21
2.3.2. Sửa lỗi	22
2.3.3. Xây dựng Overlap graph.....	22
2.3.3.1. Bảng băm.....	23
2.3.3.2 Xây dựng cạnh trên Overlap graph.....	23
2.3.3.3. Hạn chế cạnh bắc cầu.....	26
2.3.3.4. Rút gọn các tuyến ghép.....	33
Chương 3: CÀI ĐẶT THỬ NGHIỆM THUẬT TOÁN SSA	36
3.1. Yêu cầu đầu vào và đầu ra của thuật toán:.....	36
3.2. Đánh giá thuật toán và Kết quả thí nghiệm:.....	41
KẾT LUẬN	55
TÀI LIỆU THAM KHẢO.....	56

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt/thuật ngữ	Nghĩa/Mô tả
1.	ADN	(DNA) Deoxyribo Ducleic Acid
2.	BP	Base pair
3.	GB (G base)	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.	DdNTP	Dideoxynucleotide
8.	Nucleotide	Các trình tự A,T,G,C
9.	Sanger (SAGE)	Tên thiết bị đọc trình tự đoạn ngắn (1500bp)
10.	ABI SOLID	Tên thiết bị đọc trình tự
11.	Dntp	Deoxynucleotide
12.	Gdna	DNA thuộc nhiễm sắc thể
13.	SBL	Đọc trình tự gắn nối (sequencing by ligation)
14.	GS20	Tên thiết bị đọc trình tự
15.	Illumina Solexa 1G	Tên thiết bị đọc trình tự
16.	Roche 454 FLX	Tên thiết bị đọc trình tự
17.	Scaffold	(Super cotig) chuỗi các cotig
18.	Tandem Repeat	Các khối nhỏ có kích thước từ vài base đến vài chục base bị lặp đi lặp lại nhiều lần.
19.	Large repeat regions	Chuỗi lặp lớn lên tới vài nghìn base
20.	Fragment	Mảnh DNA
21.	Read	Đoạn trình tự ngắn
22.	Cotig	Đoạn trình tự dài
23.	De Bruijn	Tên một thuật toán lắp ráp hệ gen với dữ liệu
24.	pyrosequencing	Đọc trình tự đoạn ngắn (100bp)
25.	Insert size (fragment	khoảng cách giữa 2 đoạn read xuôi và ngược

	length)	
26.	Coverage	số bản sao chép của hệ gen gốc được giải mã
27.	paired-end short reads	Lắp ráp trình tự sử dụng cặp read ngắn
28.	Ligation error	Lỗi giải trình tự
29.	ALLPAHTS	Tên phương pháp lắp ráp hệ gen với dữ liệu
30.	overlap graph	Đồ thị
31.	Node	Nút trong đồ thị
32.	Tip	một node trong đồ thị mà từ vị trí đó không có cạnh dẫn tới node nào khác
33.	Bubble	Lỗi trong đồ thị, xuất hiện khi tồn tại hai đường dẫn giữa hai điểm node
34.	SSA	(Short Sequence Assembler) thuật toán lắp ráp
35.	Neighbour	Hàng xóm – điểm lân cận

DANH MỤC BẢNG BIỂU

Bảng 3.1. Bảng tóm tắt kết quả lắp ráp giữa thuật toán SSA và Velvet	53
Bảng 3.2: Thống kê tỉ lệ trình tự được sử dụng để lắp ráp	54

DANH MỤC HÌNH ẢNH

Hình 1.1. Quy trình phân tích hệ gen sinh vật từ dữ liệu giải trình tự	4
Hình 1.2. Minh họa phép lắp ráp hệ gen	5
Hình 1.3. Công nghệ giải mã hệ gen	6
Hình 1.4. Nguyên lý lắp ráp trình tự ngắn thành các contig	7
Hình 1.5. Sequencing error	10
Hình 1.6. Ligation error.....	10
Hình 1.7. Sửa lỗi giải trình tự sử dụng nhiều bản sao.....	11
Hình 1.8. Không phải lỗi trong giải trình tự	12
Hình 1.9. Một ví dụ của ‘Tandem repeat’	12
Hình 2.1: Overlap graph.....	16
Hình 2.2. Đồ thị De Bruijn.....	17
Hình 2.3. Đồ thị De Bruijn.....	18
Hình 2.4: Mô tả thuật toán ‘Breadcrumbs’	20
Hình 2.5. Chuỗi băm của Read	24
Hình 2.6. Sử dụng bảng băm để tìm những đoạn giống nhau trong chuỗi	24
Hình 2.7. Đồ thị Overlap graph với 10 read	25
Hình 2.8. Các Read trùng lặp nhau	26
Hình 2.9. Tập hợp các read đầu vào.....	29
Hình 2.10 Các loại Read trùng lặp nhau	33
Hình 2.11: Đồ thị Overlap Graph sau khi hạn chế cạnh bắc cầu	33
Hình 2.12: Đồ thị sau khi rút gọn các tuyến ghép.....	34
Hình 3.1. File config định dạng số liệu về các đoạn trình tự đầu vào	38
Hình 3.2. File H37Rv.scafStatistics thống kê số liệu đã lắp ráp.....	41
Hình 3.3. Chất lượng trung bình của các đoạn trình tự trong bộ dữ liệu	44
Hình 3.4. Chất lượng trình tự theo vị trí base	45
Hình 3.5. Chiều dài các đoạn trình tự trong bộ dữ liệu.....	46
Hình 3.6. Tỷ lệ base chưa xác định trong các trình tự	47
Hình 3.7. Tỷ lệ thành phần base.....	48

Hình 3.8. Tỷ lệ thành phần GC	49
Hình 3.9. Tỷ lệ lặp trình tự.....	50
Hình 3.10. Minh họa hoạt động của SSA	51
Hình 3.11. Minh họa hoạt động của Velvet.....	52