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 LẮPRÁP 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 thày đã trực tiếp hướng dẫn tôi hoàn thành luận văn này. Thày 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

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

DANH MỤC HÌNH ẢNH