Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

QUY TRÌNH LẮP RÁP BỘ GIEN CHLOROPLAST

Huỳnh Phước Hải[1] và Nguyễn Văn Hòa[1]

[1] *Khoa Kỹ thuật Công nghệ Môi trường, Trường Đại học An Giang*

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
An approach to assembly chloroplast genome

Từ khóa:
Mã vạch ADN, chuỗi ADN, xác lập trình tự, ADN, chuỗi ADN ngắn, bộ gien Chloroplast

Keywords:
DNA barcoding, DNA sequencing, genome assembly, chloroplast genome

ABSTRACT

The next generation sequencing (NGS) technologies are capable of producing low-cost data on a giga base-pairs scale in a single run, which usually includes millions of sequencing reads. This revolution allows launching many genome sequencing and re-sequencing projects for various biological applications, such as detection single-nucleotide polymorphism, and assessment of biodiversity. DNA Metabarcoding provides a door to identify the species in a large biological sequence dataset. Chloroplast genome is used as a genetic characteristic to identify species of plants. However, the traditional method to determine chloroplast genome sequence must use a sequence reference. In this paper, we propose a new approach to construct chloroplast genome sequences from raw data without using a reference sequence. To evaluate our approach, we compare the experimental result with four reference chloroplast genome sequences which were determined by biologists. The results show that the chloroplast genome sequences established by our approach are the same as the chloroplast reference sequences.

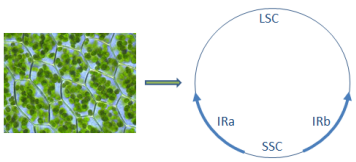
TÓM TẮT

Công nghệ xác lập trình tự gien thế hệ mới có khả năng tạo ra lượng dữ liệu khổng lồ, hàng giga bp trong một lần chạy, với chi phí ngày càng thấp. Bước tiến này cho phép thực hiện nhiều dự án giải trình tự ở các loài chưa được giải hệ gien và cả ở các loài đã giải mã trình tự nhằm thực hiện các ứng dụng sinh học phân tử khác nhau như dữ liệu đa hình đơn SNP, đánh giá sự đa dạng sinh học. Trong đó, Meta barcoding là một dự án cho phép xác định loài thực vật từ những kho dữ liệu trình tự khổng lồ. Trong nghiên cứu sinh học thực vật bộ gien chloroplast (Cp) là cơ sở quan trọng để xây dựng mã vạch sinh học dùng để định danh loài, phân loại và phân tích phát sinh loài. Tuy nhiên, phương pháp xây dựng bộ gien Cp truyền thống có hạn chế phải sử dụng mẫu gien tham chiếu. Phương pháp này không giải quyết được được yêu cầu mẫu dữ liệu đầu vào là dữ liệu thô của dự án Metabarcoding. Trong bài viết này, chúng tôi đề xuất quy trình lắp ráp bộ gien Cp cải tiến để lập trình tự bộ gien Cp từ các dữ liệu thô và không cần sử dụng gien mẫu để tham chiếu. Để đánh giá quy trình, trong thực nghiệm chúng tôi xây dựng bộ gien Cp từ bốn tập dữ liệu gien và so sánh kết quả thực nghiệm với các mẫu gien Cp đã được các nhà sinh học xây dựng.

1 GIỚI THIỆU

Sự ra đời của công nghệ giải trình tự gien thế hệ mới (*Next Generation Sequencing- NGS*) cho phép thu được lượng dữ liệu trình tự ADN khổng lồ với tốc độ nhanh và chi phí thấp hơn so với các phương pháp trước đó (Shendure *et al.,* 2008). Bước tiến này cho phép thực hiện các dự án xác lập trình tự cho các loài đã và chưa giải mã hệ gien. Chẳng hạn dự án giải mã hệ gien người Kinh (Hai DT *et al.*, 2015), dự án 1000 hệ gien (Pagani *et al.*, 2012), dự án chẩn đoán bệnh dựa trên dữ liệu gien của từng cá nhân (Haskin *et al.*, 2011). Thông thường kết quả giải mã hệ gien của loài được công bố và được lưu trữ trong các ngân hàng dữ liệu gien như Genbank. Theo Dennis và các cộng sự thì phiên bản 208 của Genbank có kích thước là khoảng 900Gbp (Dennis *et al.*, 2015). Thống kê cho thấy kích thước của kho dữ liệu này sẽ tăng gấp đôi trong vòng 18 tháng.

Dữ liệu thô (draw data) thu được từ giai đoạn giải mã trình tự thế hệ mới ở dạng các trình tự ADN ngắn (short read) từ 30 đến 100 ký tự và có độ bao phủ (coverage) là khoảng 30 lần. Dữ liệu thô này sẽ được phân tích với nhiều mục tiêu khác nhau như phân tích đa hình đơn nucleotide (Single nucleotide polymorphism) (Li *et al.*, 2009), sửa lỗi trình tự ngắn (Short read correction) (Salmela *et al.*, 2010), lắp ráp trình tự hệ gien (genome assembly) (Li *et al.*, 2010), phân tích nhận dạng tự động dựa trên dữ liệu ADN (ADN metabarcoding) (Coissac *et al.*, 2012). Mã vạch (Barcoding) ADN dựa trên mảnh nhỏ các gien được tìm thấy trong hệ gien của mỗi loài như gien Chloroplast của thực vật. Trình tự ADN của vùng được chọn được xem là đặc tính bổ sung để xác định (loài) và được gọi là barcode. Andersen và các cộng sự đã đưa ra khái niệm Metabarcoding trong trường hợp một kho dữ liệu trình tự ADN được sử dụng để xác định sự hiện diện hệ gien của một loài nào đó (Andersen *et al.*, 2012).



Hình 1: Cấu trúc của gien Chloroplast

Metabarcoding cho các loài thực vật sẽ dựa trên việc xác định các gien chloroplast của từng loài. Chloroplast (Cp) là một dạng lạp thể chỉ có trong các tế bào có chức năng quang hợp diễn ra. Nó cung cấp năng lượng cho các loài thực vật và tảo (Howe *et al.*, 2003). Ngoài ra Cp còn có nhiều chức năng sinh hóa khác trong loài cây (Bausher *et al.*, 2006). Kích thước của gien Cp từ 115 Kbp đến 165 Kbp (Jansen *et al.*, 2005). Cấu trúc của gien Cp có dạng vòng (circuler) bao gồm 2 vùng sao chép IRs (Inverted Repeat regions) (IRa và IRb), LSC (Large Single Copy), SSC (Single copy) như hình 1. Kích thước của vùng IRs từ 10 Kbp đến 30 Kbp (Saski *et al.*, 2007). Gien Cp có nhiều phiên bản sao chép trong một tế bào (khoảng 1000 bản sao chép) (Raubeson *et al.*, 2005).

Trong phiên bản 208 của Genbank hiện có 843 gien Cp (Dennis *et al.*, 2015). Các gien Cp này được xây dựng theo phương pháp truyền thống bởi các nhà sinh học phân tử như trong nghiên cứu xây dựng gien Cp của loài Arabis Alpina (Medolidima *et al.*, 2013). Medolidima và các cộng sự phải chiết tách Cp trước khi tiến hành giải mã trình tự ADN theo công nghệ NGS. Dữ liệu thô được xử lý để loại bỏ các read bị lỗi hoặc bị lặp quá nhiều sau đó tiến hành lắp ráp trình tự, kết quả của giai đoạn này là các contig được lắp ráp từ các read. Sau đó các contig này được ánh xạ vào vị trí của gien (References Genome) để xác định thứ tự các contig giữa các contig vẫn có các vị trí trống gọi là Gap để giảm số lượng Gap và tăng độ dài các contig quy trình tiếp tục mở rộng các contig bằng các chương trình lắp ráp Scaffolding như SSPACE (Boetzer *et al.*, 2011). Kết quả của quy trình này đã xác định gien của Cp từ tập dữ liệu thô ban đầu trong điều kiện phải biết trước được gien mẫu cần xác định.

Quy trình lắp ráp bộ gien Cp theo phương pháp truyền thống bắt buộc phải sử dụng gien Cp mẫu làm cơ sở so sánh cho bước ánh xạ các contig (mapping) để xác định vị trí các contig. Do đó quy trình lắp ráp bộ gien Cp truyền thống không thể sử dụng trong các kho dữ liệu đóng vai trò xác định sự hiện diện của gien Cp của một loài thực vật như là Metabarcoding. Bài viết này giới thiệu quy trình lắp ráp bộ gien Cp mới từ các tập dữ liệu thô chứa các read của gien Cp và các gien khác. Phương pháp đề xuất có thể sử dụng trong Metabarcoding vì không sử dụng gien Cp mẫu.

Phần tiếp theo của bài báo giới thiệu quy trình lắp ráp bộ gien Cp cải tiến do chúng tôi đề xuất trong phần 2. Phần 3 trình bày kết quả đánh giá phân tích quy trình mới dựa trên dữ liệu thực nghiệm. Phần 4 trình bày kết luận và hướng phát triển.

2 QUY TRÌNH LẮP RÁP BỘ GIEN CP

Quy trình đề xuất gồm bốn giai đoạn. Đầu tiên chúng tôi lọc các read có độ phủ tốt từ tập dữ liệu

thô và lắp ráp trình tự để tạo ra các contig. Tiếp theo chúng tôi dựa vào cơ sở dữ liệu của các loài cây để lọc ra các contig thuộc gien Cp. Do cấu trúc của gien Cp là dạng vòng nên các contig sẽ được sắp xếp dựa vào phương pháp đồ thị để tìm được chu trình đi qua các đỉnh, chu trình này chính là bộ gien Cp cần tìm, nếu không tìm được chu trình do các contig đơn độc sẽ tiếp tục giai đoạn mở rộng các contig này sau đó quay lại giai đoạn sắp xếp.

2.1 Chọn read và lắp ráp contig

Do gien Cp được sao chép khoảng 1000 lần trong một tế bào nên để lắp ráp gien Cp trước tiên chúng tôi lọc các read có độ phủ cao từ dữ liệu thô bằng cách dựa vào kết quả thống kê của chương trình phân tích k-mer. Trong bước này chúng tôi sử dụng chương trình DSK (Guillaume Rizk *et al.*, 2012) để thống kê kmer, chương trình này có thể chạy trên các máy tính cá nhân với bộ nhớ tối thiểu là 1GB đồng thời có thể hỗ trợ nhiều giá trị k-mer. Kết quả của chương trình DSK bao gồm tập tin nhị phân chứa kết quả thống kê kmer và đồ thị histogram biểu diễn kết quả này. Do đặc điểm của gien Cp có nhiều vùng trình tự lặp lại nên trong đồ thị biểu diễn sẽ có một khu vực có độ biến thiên đặc biệt so với các khu vực khác, chúng ta dựa vào đồ thị này để xác định ngưỡng độ phủ (threshold) của các k-mer, các k-mer nào có độ phủ lớn hơn ngưỡng có khả năng thuộc gien Cp. Các read chứa các k-mer này có khả năng thuộc gien Cp và chúng tôi xây dựng giải thuật ReadFilter để lọc ra các read này. Kết quả của giai đoạn này chúng tôi được tập tin FASTA chứa các read có độ phủ tốt.

Giải thuật ReadFilter

*Input:*FASTA file, output file of DSK program, coverage threshold.
Output: FASTA file
1: init k-mer hash
2: **while** read each k-mer
3: **do**
4: **if** abundance of k-mer >threshold **then**
5: insert k-mer into k-mer hash
6: **end if**
7: **end while**
8: **for each** read R in FASTA file do
9: **for each** k-mer in R do
10: **if** k-mer exits in hash **then** write R to Output file
11: **end for**
12: **end for**

Tiếp theo chúng tôi sử dụng chương trình Minia (Chikhi *et al.*, 2012) để lắp ráp contig với dữ liệu đầu vào là tập các read có độ phủ cao của chương trình DSK. Chương trình Minia sử dụng phương pháp đồ thị de Bruijn (Idury *et al.*, 1995) để lắp ráp contig. Đây là một chương trình lắp ráp các read ngắn sử dụng bộ nhớ hiệu quả có thể sử dụng được trên các máy tính cá nhân. Kết quả của chương trình Minia là các contig có khả năng thuộc về gien Cp và các gien thuộc vùng trình tự được lặp lại đi lặp lại nhiều lần.

2.2 Lọc contig

Mục tiêu của quy trình mới là xây dựng bộ gien Cp mà không cần sử dụng gien mẫu để tham chiếu. Hiện nay, trên các ngân hàng gien, cơ sở dữ liệu gien của các loài cây (Dennis *et al.*, 2015) gọi là Plastid, chứa các gien Cp. Do đặc tính về di truyền của sinh học nên các gien chung một loài sẽ có trình tự tương đồng nhau. Để có thể xác định gien Cp chúng tôi cần loại bỏ các contig không thuộc gien này. Để thực hiện chúng tôi so khớp các trình tự của các contig với 803 gien Cp bằng chương trình BLAST (Altschul *et al.*, 1990), kết quả của chương trình này chúng tôi được các thông tin của các bắt cặp trình tự. Dựa vào thông tin này chúng tôi xác định được các contig thuộc gien Cp dựa vào đặc điểm sinh học giữa các gien cùng loài phải có độ tương đồng về cấu trúc hơn 80%. Để xác định các contig thuộc gien Cp chúng tôi xây dựng giải thuật ContigFilter để lọc ra các contig dựa tham số vào điều kiện gồm có tỉ lệ chính xác tối thiểu (identity threshold) và tổng độ dài bắt cặp trình tự.

Giải thuật ContigFilter

Input:
- Alignment (alig) file
- Contig file
- Identity threshold
- Align length threshold

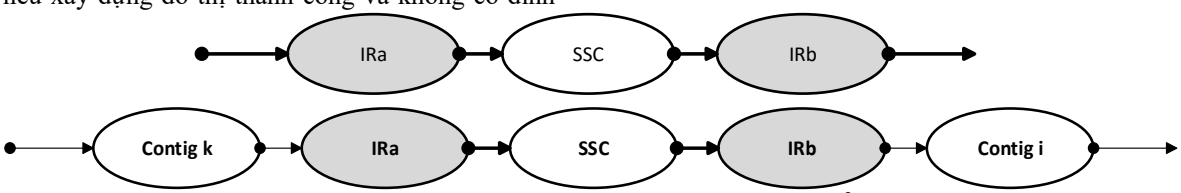
Ouput: a set contig related to chloroplast
1: **for each** align in align file
2: **if** alig identity > identity threshold **then**
3: insert align into listBlast
4: **end for**
5: Sort listBlast by contig's id, plastid seq id
6: **for each** align in listBlast
7: calculate align length and align identity
8: **if** align length and identity ≥ threshold
9: add contig into output file
10: **end for**

2.3 Sắp xếp contig

Chương trình lắp ráp contig Minia sử dụng đồ thị de Bruijn để tạo các contig vì vậy giữa các

contig có một đoạn chồng khớp lên nhau, được gọi là overlap. Chúng tôi dựa vào sự chồng khớp giữa hai contig để sắp xếp lại các contig bằng phương pháp đồ thị (String overlap graph). Chúng tôi tìm trong đồ thị các đỉnh cô lập hoặc các đỉnh treo bằng cách duyệt qua tất cả các đỉnh của đồ thị và tính số bậc của mỗi đỉnh. Nếu có đỉnh cô lập (bậc ra hoặc bậc vào bằng 0) hoặc treo thì chuyển sang giai đoạn tiếp theo để mở rộng các contig đó. Cách duyệt đường đi qua các đỉnh của đồ thị mỗi đỉnh ít nhất một lần. Đường đi qua tất cả các đỉnh và quay trở lại đỉnh đầu tiên chính là cấu trúc của gien Cp cần xác định, nếu không tìm được chu trình hoàn chỉnh thì đường đi dài nhất chính là cấu trúc của gien vì trong trường hợp này dữ liệu đầu vào có khả năng lỗi khi giải trình tự. Sau giai đoạn này nếu xây dựng đồ thị thành công và không có đỉnh treo hoặc đỉnh cô lập thì sẽ xây dựng thành công gien Cp.

Phương pháp lắp ráp bộ gien Cp truyền thống sử dụng gien mẫu để sắp xếp các contig bằng cách ánh xạ chúng lên cấu trúc của gien mẫu. Do đặc điểm cấu trúc của gien Cp dạng vòng nên trong quy trình mới này chúng tôi đi tìm chu trình Hamilton để xác định vị trí của các contig. Mỗi bộ gien Cp bao gồm hai khu vực có trình tự giống nhau nhưng ngược chiều (IRa và IRb) đặc điểm này giúp chúng tôi xác định được sẽ có các contig có chiều ngược lại trong khu vực lặp như Hình 2. Mặt khác, chúng tôi xây dựng chương trình đếm số lần lặp lại các k-mer của các contig trong tập dữ liệu thô để xác định được các contig lặp lại nhiều hơn một lần và các contig không thuộc gien Cp.



Hình 2: Quy trình tìm đường đi Hamilton trong đồ thị

2.4 Mở rộng contig

Trong giai đoạn này, chúng tôi sử dụng các chương trình Scaffolding để mở rộng các contig SSPACE (Boetzer, 2011). Kết quả sẽ tạo các contig được mở rộng từ các tập dữ liệu thô ban đầu. Ở đây, chúng ta cần xác định được mỗi contig cần được mở rộng độ dài L bao nhiêu là vừa đủ. Để xác định cần lấy contig đã mở rộng và so khớp vào tập các contig ban đầu sau đó tìm đoạn giữa khu vực mở rộng (extened regions) có khớp nhau với đoạn bắt đầu hoặc kết thúc của contig khác và chọn độ dài phù hợp. Cuối cùng lấy phần mở rộng thêm vào đầu hoặc cuối contig và lặp lại giai đoạn sắp xếp contig.

3 KẾT QUẢ THỰC NGHIỆM

3.1 Dữ liệu và môi trường thực nghiệm

Để đánh giá quy trình lắp ráp bộ gien Cp, chúng tôi sử dụng bốn tập dữ liệu Arabidopsis Thaliana (SRR616965), Oryzasativa Indica (SRR400297), Sorghum Bicolor (SRR562875) được tải về từ European Nucleotide Archive (http://www.ebi.ac.uk/ ena) và tập dữ liệu Leconten của LECA. Thông tin của các tập dữ liệu được trình bày trong Bảng 1.

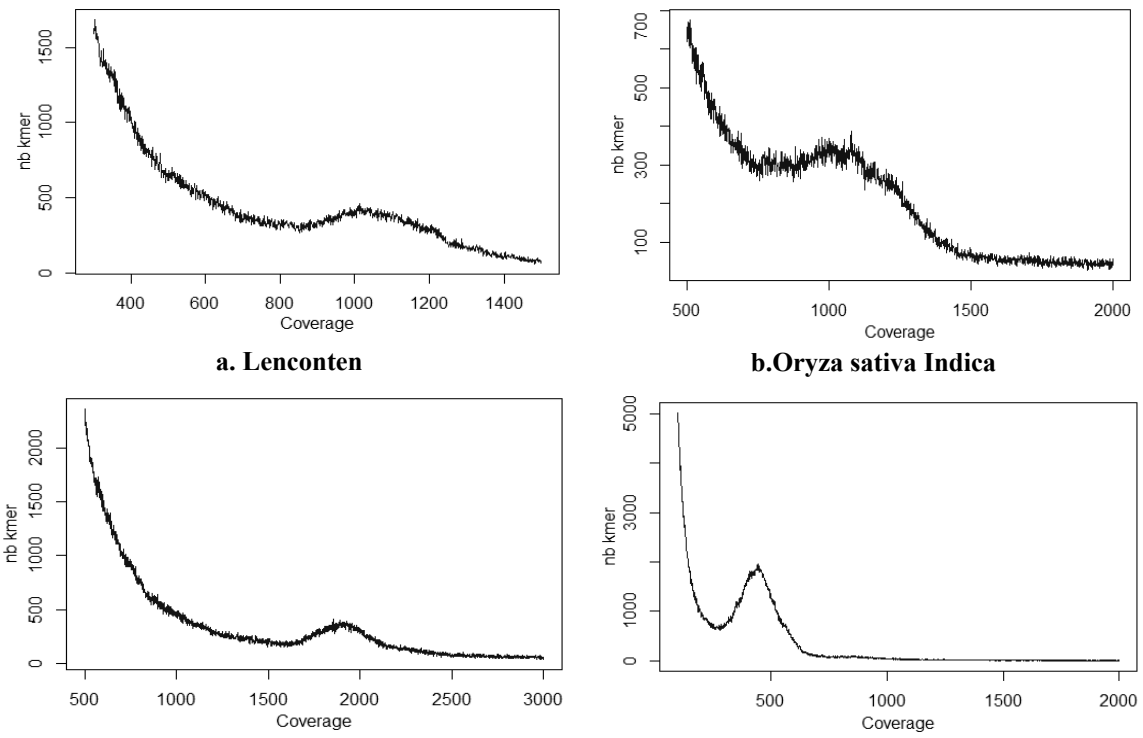
Bảng 1: Thông tin các tập dữ liệu

ID	Scientific name	Number of reads	Read length
SRR120824	Lenconten	38.989.953	100 bp
SRR616965	Arabidopsis Thaliana	53.017.770	100 bp
SRR400297	Oryza sativa Indica	90.317.440	76 bp
SRR562875	Sorghum bicolor	106.358.033	100 bp

Các thực nghiệm được chúng tôi thực hiện trên môi trường máy tính cá nhân với CPU Intel 2.6 GHz, 2MB cache L2, và 4 Gb RAM, hệ điều hành Linux (fedora 18). Chúng tôi xây dựng các chương trình lọc các read (Read Filter), lọc contig (Contig Filter), và sắp xếp contig (Contig Ordering) bằng ngôn ngữ C.

3.2 Kết quả

Chúng tôi sử dụng chương trình DSK (Guillaume Rizk *et al.*, 2012) để đếm và thống kê k-mer của bốn tập dữ liệu, kết quả của chương trình được thể hiện trong bốn đồ thị của Hình 3. Một tập dữ liệu thô có chứa gien Cp khi biểu đồ của k-mer của tập dữ liệu phải có đặc điểm như các đồ thị ở Hình 3.



a. Lenconten **b.Oryza sativa Indica**

c. Sorghum Bicolor **d. Arabidopsis Thaliana**

Hình 3: Đồ thị histogram thống kê k-mer của 4 tập dữ liệu

Tiếp theo chúng tôi sử dụng chương trình Read Filter để lọc ra các read có độ phủ tốt. Kết quả của chương trình được thể hiện trong Bảng 2, cột cuối cùng của bảng cho thấy được tỉ lệ số lượng read được chọn so với số lượng read ban đầu. Cột coverage threshold là tham số ngưỡng độ phủ được sử dụng trong chương trình để lọc các read và cột ba thể hiện số lượng read được chọn từ tập dữ liệu ban đầu.

Bảng 2: Kết quả của chương trình ReadFilter

Tập dữ liệu	Coverage threshold	Số lượng Read được chọn	Tỉ lệ
Lenconten	450	3.462.954 (3.4M)	3.5M (8%)
Arabidopsis Thaliana	550	19.554.206 (19.5M)	18M (36%)
Oryza sativa Indica	300	6.943.242 (6.9M)	6.5M (8%)
Sorghum Bicolor	500	9.610.760 (9.6M)	8.8M (9%)

Tập dữ liệu kết quả của chương trình Read Filter được sử dụng làm dữ liệu đầu vào của chương trình lắp ráp contig là Minia (R. Chikhi *et al.,* 2012). Sau khi thực hiện kết quả của chương trình được trình bày trong Bảng 3. Cột thứ ba của bảng cho thấy số lượng contig thu được sau khi chạy chương trình Minia, trong kết quả này có những contig không thuộc gien Cp bởi vì kích thước của gien Cp từ 115Kbp đến 165 Kbp. Vì vậy, chúng tôi lọc các contig thuộc gien Cp bằng chương trình Filter Contig và được kết quả là số lượng contig và kích thước của các contig ở cột cuối cùng của Bảng 3.

Bảng 3: Kết quả lắp ráp contig và lọc contig

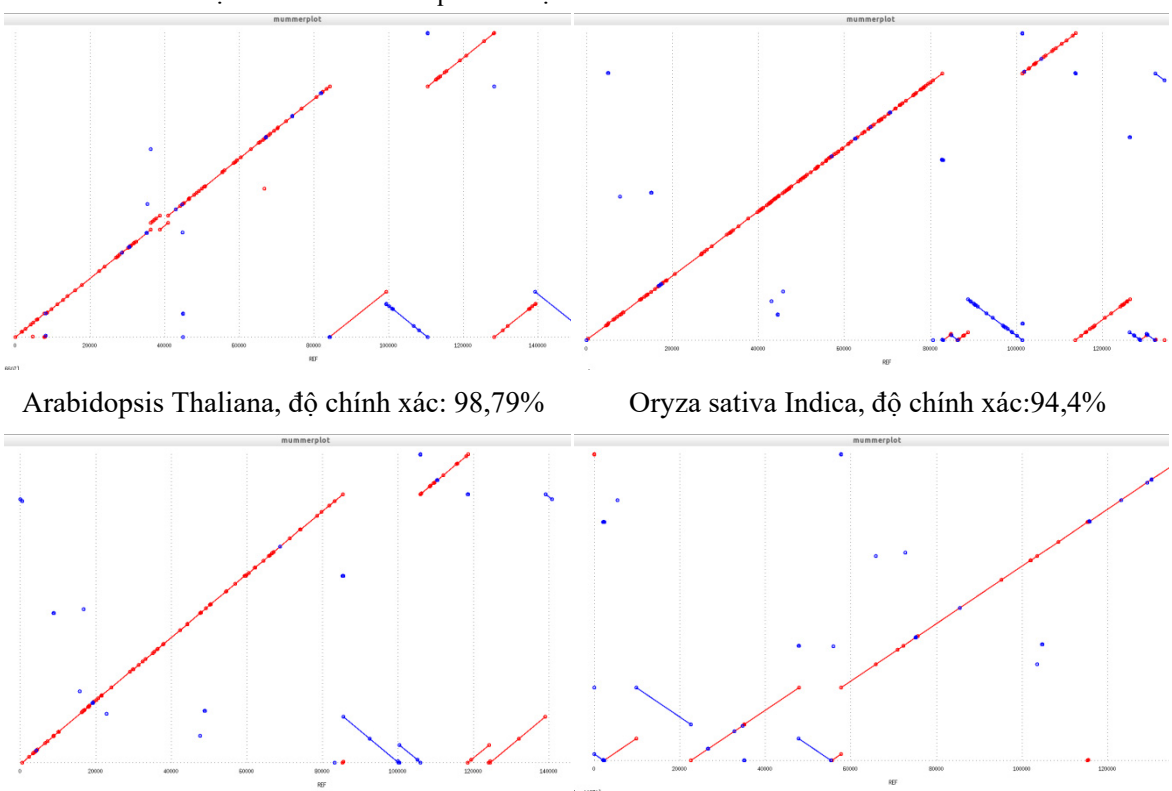
Dataset	Minia		Filtering contigs	
	# contigs	Size of contigs	# contigs	Size of contigs
Lenconten	430	220 Kbp	18	114 Kbp
Arabidopsis Thaliana	180	169 Kbp	17	125 Kpb
Oryza sativa Indica	563	265 Kbp	18	112 Kbp
Sorghum Bicolor	1002	380 Kbp	17	115 Kbp

Tiếp theo chúng tôi sắp xếp lại vị trí của các contig theo các vùng cấu trúc của gien Cp. Trong kết quả thực nghiệm đồ thị contig của tập dữ liệu Arabidopsis và Oryza có chứa các đỉnh cô lập và các contig cô lập sẽ được sử dụng chương trình SPACCE để mở rộng. Trong Bảng 4 cho thấy kết quả sắp xếp các contig của tập dữ liệu Sorghum Bicolor.

Bảng 4: Vị trí các contig trong 4 khu vực của gien (Sorghum Bicolor)

Contig region	Contig name
IRa	r377__len__2158; 378__len__229; r308__len__14547; r483__len__326; r284__len__5576
SSC	316__len__12564
IRb	284__len__5576; ….377__len__2158
LSC	108__len__14408; r67__len__9262; 378__len__229; r644__len__11821; 596__len__15232; 597__len__10844; 260__len__1122; r261__len__2303; 128__len__7918; 947__len__1323; 44__len__4609; 45__len__3793

Để đánh giá hiệu quả của quy trình lắp ráp bộ gien Cp, chúng tôi so sánh kết quả của quy trình mới bằng cách so sánh cấu trúc gien tìm được với các gien mẫu trong các ngân hàng gien được tải về từ cơ sở dữ liệu plastid (Dennis *et al.*, 2015). Để hiển thị kết quả trực quan chúng tôi sử dụng chương trình MUMMER (Stefan Kurtz, 2014) để biểu diễn kết quả khi ánh xạ các contig lên các gien mẫu. Hình 4 thể hiện các biểu đồ kết quả ánh xạ của bốn gien kết quả lên bốn gien mẫu. Trục hoành là thể hiện vị trí các contig còn trục tung là các gien mẫu, kết quả ánh xạ được thể hiện qua đường chéo trên đồ thị. Các đoạn thẳng song song đường chéo phụ thể hiện sự tương đồng giữa các trình tự còn các đoạn thẳng song song đường chéo chính là các trình tự có cấu trúc tương đồng nhưng có chiều ngược lại.



Arabidopsis Thaliana, độ chính xác: 98,79%

Oryza sativa Indica, độ chính xác:94,4%

Sorghum Bicolor, độ chính xác: 98,35%

Lenconten, độ chính xác: 98,39%

Hình 4: So khớp giữa các tập contig và các gien mẫu

Để tính được độ chính xác của quy trình mới, chúng tôi sử dụng chương trình BLAST để so khớp giữa tập các contig của gien tìm được và các gien mẫu. Dựa vào kết quả hiển thị của chương trình BLAST chúng tôi tính được tổng độ dài của các đoạn trình tự tương đồng với gien mẫu. Kết quả độ chính xác được tính bằng thương của tổng độ dài các đoạn trình tự đồng và độ dài của gien mẫu. Độ chính xác của quy trình mới của bốn tập dữ liệu được trình bày trong Bảng 5.

Bảng 5: Độ chính xác của quy trình khi so sánh với gien mẫu

Dataset	Lenconten	Arabidopsis Thaliana	Oryza sativa Indica	Sorghum Bicolor
Độ chính xác	98,39%	98,79%	94,4%	98,35%

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày quy trình lắp ráp bộ gien Chloroplast mới với dữ liệu đầu vào là dữ liệu thô (raw data) chứa nhiều gien khác nhưng không cần sử dụng gien mẫu để ánh xạ. Khác với phương pháp truyền thống chúng tôi đã thành công trong việc sử dụng chu trình Hamilton để sắp xếp vị trí của các contig mà không cần sử dụng gien mẫu để ánh xạ. Kết quả thực nghiệm cho thấy quy trình của chúng tôi có độ chính xác từ 94.4% đến 98.8% so với các gien gốc. Trong tương lai chúng tôi sẽ nghiên cứu hướng tự động hoá trong một số giai đoạn phải tự làm trong giai đoạn mở rộng các contig và mở rộng các giải thuật của các chương trình trong quy trình theo mô hình tính toán song song nhằm tăng tốc các giai đoạn lắp ráp bộ gien Chloroplast.

TÀI LIỆU THAM KHẢO

1. Altschul, S. F., (1990), Basic Local Alignment Search Tool. Journal of Molecular Biology 215, pp. 403–410.
2. Andersen Kenneth and *et al*. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Molecular Ecology, 21 pp. 1966-1979.
3. Bausher MG, Singh ND, Lee S-B, Jansen RK, Daniell H. The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeckvar 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biology 2006: 6:21-29.
4. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W., (2011), Scaffolding pre-assembled contigs using SSPACE, Bioinformatics. Volume 27, Issue 4, pp. 578-579.
5. Coissac E., Riaz T., Puillandre N. (2012), Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology, 21, pp. 1834-1847.
6. Hai DT, Thanh ND, Trang PTM, Quang LS, Hang PTT, Cuong DC, Phuc HK, Duc NH, Dong DD, Minh BQ, Son PB and Vinh LS (2015) Whole genome analysis of a Vietnamese trio. J. Biosci. 40 113–124.
7. Dennis A. Benson, *et al*. (2015). Genbank, Nucleic Acids Research, Vol. 13.
8. Guillaume Rizk, Dominique Lavenier, RayanChikhi, (2012), DSK: k-mer counting with very low memory usage, Bioinformatics journal.
9. Haskin G., Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J. Karczewski, and Russ B. Altman (2011). Bioinformatics challenges for personalized medicine. Bioinformatics 27 (13), pp. 1741-1748.
10. Howe C.J, Barbrook A.C, Koumandou V.L, Nisbet R.E.R, Symington H.A, Wightman T.F 2003 Evolution of the chloroplast genome. Phil. Trans. R. Soc. B. 358, 99–106. doi:10.1098/rstb.2002.1176.
11. Idury, R.M., Waterman, M.S (1995). A new algorithm for DNA sequence assembly. Journal of Computational Biology 2 (2), pp. 291-306.
12. Li R., Zhu H., Ruan J., *et al*., (2010). De novo assembly of human genomes with massively parallel short read sequencing, Genome Research, volume 20, number 2, pp. 265–272.
13. Li R., Li Y., Fang X., Yang H., Wang J., Kristiansen K., Wang J., (2009). SNP detection for massively parallel whole-genome resequencing. Genome Research 19, pp. 1124-1132.
14. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC., (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res; 40(Database issue):D571-9.

15. Pierre Peterlongo and RayanChikhi, (2012), Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. BMC Bioinformatics. Vol 13.
16. Rayan Chikhi and Guillaume Rizk, (2012), Space-efficient and exact de Bruijn graph representation based on a Bloom filter, Algorithms in Bioinformatics, Vol 7534, 2012, pp. 236-248.
17. Salmela L., (2010). Correction of sequencing errors in a maxed set of reads. Bioinformatics 26 (10) pp. 1284-1290.
18. Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol Biol. 2005; 59:309–322.
19. Shendure J. and Ji H., (2008). Next-generation DNA sequencing, Nature biotechnology, volume 26, number 10, pp. 1135-1145.