

TÓM TẮT LUẬN VĂN

So sánh đa trình tự(Multiple Sequence Alignment-MSA) là một trong 10 bài toán lớn của Sinh tin học(Bioinformatics). MSA đóng vai trò quan trọng trong Sinh tin học nói chung và lĩnh vực tìm kiếm gene (Gene Finding) nói riêng. MSA là một bài toán NP, và hoàn toàn chưa có giải pháp trọn vẹn để tìm lời giải tối ưu của bài toán. Nhiều phương pháp sử dụng heuristic đã được đưa ra để giải quyết bài toán khi tập dữ liệu đầu vào lớn, các phương pháp này hướng tới việc tìm 1 lời giải cận tối ưu với thời gian tính toán và bộ nhớ sử dụng chấp nhận được. Progress Algorithm là một phương pháp tốt tiếp cận theo phương thức này.

Đề tài này trình bày một giải thuật mới dựa trên Progressive Algorithm. Sử dụng lời giải của bài toán TSP để mô tả quá trình so sánh(aligned) các sequence. Để cung cấp một Progressive Algorithm có chất lượng, giải thuật đã tối ưu bài toán Pairwise Sequence Alignment(PSA) về độ chính xác và bộ nhớ sử dụng thông qua giải thuật "chia để trị" kết hợp với việc sử dụng 3 ma trận đánh giá BLOSUM. Thông qua quá trình so sánh với CLUSTALW(một chương trình hiện thực Progressive Algorithm được đánh giá là cho kết quả tốt nhất), dựa trên kết quả kiểm thử với tập dữ liệu BALiBASE benchmark và một số nguồn dữ liệu từ NCBI(National Center for Biotechnology Information), chương trình hiện thực giải thuật đã cung cấp một lời giải có độ chính xác khá cao, tiết kiệm bộ nhớ và có thời gian tính toán chấp nhận được.

Từ khoá: Algorithm, Sequence Alignment, Multiple Sequence Alignment, MSA, Pairwise Sequence Alignment, PSA, Progressive Algorithm, Dynamic Programming, Traveling Salesman Problem, TSP, CLUSTALW, BLOSUM, BALiBASE.

MỤC LỤC

LỜI CAM ĐOAN	i
LỜI CẢM ƠN	ii
TÓM TẮT LUẬN VĂN	iii
DANH MỤC HÌNH	vi
DANH MỤC BẢNG	viii
Chương 1. GIỚI THIỆU	1
1.1. Giới thiệu	1
1.2. Kết cấu của luận văn	4
Chương 2. TỔNG QUAN VỀ KHÁI NIỆM SO SÁNH TRÌNH TỰ (SEQUENCE ALIGNMENT)	6
2.1. So sánh trình tự	6
2.1.1. Định nghĩa So sánh trình tự(Sequence Alignment)	6
2.1.2. Phân loại	7
2.1.3. So sánh 2 trình tự (Pairwise Sequence Alignment-PSA).....	8
2.1.4. So sánh nhiều trình tự (Multiple Sequence Alignment-MSA).....	9
2.2. Các khái niệm khác	10
2.2.1. Ma trận đánh giá(Scoring Matrix)	12
2.2.2. Gap.....	14
2.2.3. Phương pháp đánh giá(Scoring Method).....	15
2.3. Các phương pháp giải quyết bài toán so sánh trình tự	18
2.3.1. Phương pháp Quy hoạch động(Dynamic Programming).....	19
2.3.2. Sử dụng các thiết bị phần cứng.....	20
2.3.3. Phương pháp tìm kiếm cục bộ(Local Search).....	21
2.3.4. Sử dụng giải thuật Di truyền(Genetic Algorithm)	21
2.3.5. Sử dụng Mô hình Markov ẩn(Hidden Markov Model-HMM).....	21
Chương 3. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP THỰC HIỆN	24
3.1. Giới thiệu về Dynamic Programming	24
3.2. Bài toán PSA và cách giải quyết bằng kỹ thuật quy hoạch động	24
3.2.1. Giải thuật quy hoạch động cho bài toán PSA	25
3.2.2. Giải thuật Gotoh.....	28
3.2.3. Giải thuật cải tiến không gian nhớ	29
3.3. Giải thuật tính toán phép Alignment tối ưu cho bài toán Multiple Alignment sử dụng kỹ thuật dynamic programming	32
3.3.1. Giải thuật Center Star Alignment Algorithm.....	33
3.3.2. Phương pháp Progressive Algorithm giải quyết bài toán MSA.....	37
3.3.3. Feng-Doolittle Algorithm	38
Chương 4. THIẾT KẾ GIẢI THUẬT VÀ HIỆN THỰC PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN MSA	42
4.1. Giải thuật sử dụng cho bài toán PSA.....	42

4.1.1.	Giải thuật tính toán dựa theo kỹ thuật chia để trị.....	43
4.2.	Giải thuật hiện thực cho bài toán MSA	49
4.2.1.	Bài toán TSP(Travelling Salesman Problem-Bài toán người bán hàng).	50
4.2.2.	Giải thuật 1A.....	51
4.2.3.	Giải thuật 1B(Giải thuật cải tiến gom nhóm nhỏ nhất).....	55
4.3.	Giải thuật di truyền và bài toán TSP.	57
4.3.1.	Đặc điểm giải thuật di truyền.....	57
4.3.2.	Cấu trúc thuật giải di truyền tổng quát.....	59
4.4.	Phần hiện thực giải thuật và chương trình:	60
Chương 5. KẾT QUẢ NHẬN XÉT.....		66
5.1.	Một số kết quả chạy chương trình.	66
5.2.	BAliBASE (Benchmark Alignment Database).....	68
5.3.	So sánh kết quả	69
5.3.1.	Giới thiệu về các chương trình được sử dụng.....	70
5.3.2.	So sánh độ chính xác của kết quả	70
5.3.3.	So sánh về mặt thời gian chạy, bộ nhớ	77
Chương 6. KẾT LUẬN		78
TÀI LIỆU THAM KHẢO.....		80
Phụ lục 1. Bảng đối chiếu Thuật ngữ Anh - Việt.....		83
Phụ lục 2. Từ viết tắt		87
Tham khảo Chỉ mục		88

DANH MỤC HÌNH

Hình 2.1 Ví dụ về PSA.....	7
Hình 2.2 Ví dụ về so sánh trình tự theo hướng toàn cục.....	8
Hình 2.3 Ví dụ về so sánh trình tự theo hướng cục bộ.....	8
Hình 2.4 Cấu trúc 1 PSA.....	8
Hình 2.5 Giới thiệu 1 MSA.....	9
Hình 2.6 Giới thiệu các khái niệm của MSA.....	10
Hình 2.7 Quá trình biến đổi của 2 sequence.....	10
Hình 2.8 Ví dụ về các phép thay thế gap.....	11
Hình 2.9 Ví dụ về Gap.....	15
Hình 2.10 Môi tương quan giữa các chương trình hiện thực cho các phương pháp.....	19
Hình 2.11 Phương pháp tính toán chính xác bằng dynamic programming.....	20
Hình 2.12 Mô hình Markov cho bài toán MSA.....	22
Hình 3.1 Phương pháp quy hoạch động cho bài toán PSA.....	25
Hình 3.2 Các ma trận S, D, I cho 2 chuỗi AGTAC and AAG.....	31
Hình 3.3 Minh họa quá trình tìm 1 MSA tối ưu.....	33
Hình 3.4 Mô hình tiến hoá hình sao.....	34
Hình 3.5 Minh họa Center Star Algorithm.....	35
Hình 3.6 Hình minh họa cho Progressive Algorithm.....	37
Hình 3.7 Minh họa Feng-Doolittle Algorithm.....	39
Hình 3.8 Ví dụ thực thi Feng-Doolittle Algorithm.....	39
Hình 4.1 Mô hình quá trình thực hiện giải thuật PSA.....	43
Hình 4.2 Quá trình xây dựng ma trận của thuật giải cho bài toán PSA.....	48
Hình 4.3 Quá trình align của Center Star Algorithm và phiên bản cải tiến.....	50
Hình 4.4 Bài toán TSP.....	50
Hình 4.5 Kết quả bài toán TSP.....	51
Hình 4.6 Lưu đồ thuật giải 1A.....	52
Hình 4.7 Lưu đồ thuật giải 1B.....	55
Hình 4.8 Cấu trúc chương trình hiện thực.....	61
Hình 4.9 Module PSA.....	61
Hình 4.10 Sơ đồ các khối chức năng của Module MSA.....	62
Hình 4.11 Sơ đồ các khối chức năng của module TSP.....	63
Hình 5.1 Đồ thị tương quan về độ chính xác của MSAPR, CLUSTALW và MULTAL.....	72
Hình 5.2 Đồ thị tương quan về độ chính xác của MSAPR, CLUSTALW và HMMT.....	74
Hình 5.3 Đồ thị tương quan về độ chính xác của MSAPR, CLUSTALW và HMMT.....	75
Hình 5.4 Đồ thị tương quan về độ chính xác của MSAPR, CLUSTALW, SAGA.....	75

Các kỹ thuật toán học cho bài toán so sánh đa trình tự

Hình 5.5 Đồ thị tương quan về độ chính xác của MSAPR, CLUSTALW, SAGA76
Hình 5.6 So sánh thời gian thực thi của MSAPR và CLUSTALW77

DANH MỤC BẢNG

Bảng 2.1 Ma trận BLOSUM62 lưu trữ hàm đánh giá độ tương đồng của tập 23 amino acid.....	12
Bảng 2.2 Một phần ma trận Identity	13
Bảng 3.1 Bảng kết quả giải thuật quy hoạch động cho bài toán PSA	26
Bảng 4.1 Định dạng của file dữ liệu đầu vào	63
Bảng 4.2 Định dạng của file dữ liệu đầu ra.....	64
Bảng 4.3 Định dạng file dữ liệu đầu ra theo chuẩn MSF	64
Bảng 4.4 Bảng tóm tắt các lớp của chương trình.	65
Bảng 5.1 TAT Protein HIV1	66
Bảng 5.2 Kết quả Alignment của MSAPR và CLUSTALW với TAT HIV1	67
Bảng 5.3 Kết quả chạy chương trình với Nhóm 1 có chiều dài nhỏ	71
Bảng 5.4 Kết quả chạy chương trình với Nhóm 1 có chiều dài trung bình.....	71
Bảng 5.5 Kết quả chạy chương trình với Nhóm 1 có chiều dài lớn	72
Bảng 5.6 Kết quả chạy của các chương trình với các sequence của nhóm 2.	73
Bảng 5.7 Kết quả chạy của các chương trình với các sequence của nhóm 3.	74
Bảng 5.8 Kết quả chạy của các chương trình với các sequence của nhóm 4	75
Bảng 5.9 Kết quả chạy của các chương trình với các sequence của nhóm 5	76

Chương 1. GIỚI THIỆU

1.1. Giới thiệu

Cùng với sự phát triển mang tính đột phá của Khoa học kỹ thuật, trong vài thập kỷ qua, sinh học phân tử đã có nhiều bước phát triển mạnh mẽ, một loạt các công cụ ứng dụng sinh học ra đời góp phần thúc đẩy quá trình giải mã một số lượng lớn trình tự bộ gen ở nhiều loài sinh vật. Cho đến nay, nhiều bộ gen vi khuẩn và các sinh vật bậc cao đã được giải mã gần như hoàn toàn. Dự án về bộ gen người được thành lập (1997), và quá trình giải trình tự tất cả 24 cặp nhiễm sắc thể của bộ gen người cũng đã hoàn thành từ cuối năm 2000, cũng như đã giải được khoảng 90% bộ gen người (2001). Lượng thông tin sinh học ngày càng trở nên phong phú và đa dạng. Để có thể xử lý và ứng dụng khối lượng thông tin đồ sộ như vậy, ngành Sinh tin học (hay Bioinformatics) ra đời, đó là sự kết hợp giữa công nghệ thông tin và sinh học, một cách đơn giản sinh tin học giải quyết các vấn đề của sinh học bằng cách sử dụng các kỹ thuật của khoa học máy tính. Các lĩnh vực lớn đang được Sinh tin học giải quyết hiện nay:

- Genomic: nghiên cứu cấu trúc và chức năng của gene.
- Proteinomics: Phân tích một tỉ lệ lớn các protein của một sinh vật
- Pharmacogenomics: phát triển các loại thuốc mới nhắm đến một căn bệnh xác định
- MicroArray: nghiên cứu về DNA chip, protein chip.

Mục tiêu hàng đầu của sinh tin học gắn liền với quá trình phân tích các thông tin sinh học. Điều này được thể hiện qua các nghiên cứu về:

- Tìm kiếm các gene trên các trình tự DNA ở các sinh vật khác nhau.
- Phát triển các phương pháp nhằm dự đoán các trình tự RNA, cấu trúc và chức năng của các protein mới được phát hiện.
- Tập hợp các trình tự có sự tương đồng cao để đưa ra mô hình protein.
- So sánh các trình tự protein tương đồng và thành lập cây phả hệ mô tả mối quan hệ tiến hóa

Trong lĩnh vực nghiên cứu phân tích cấu trúc và chức năng của gene và protein, phân tích trình tự (chuỗi DNA, protein) đóng vai trò quan trọng. Để đơn giản cho việc nghiên cứu, trình tự DNA, protein sẽ được tuần tự hóa và nghiên cứu dưới dạng chuỗi các ký tự. Thông thường khi một gene được phát hiện, một trong những yêu cầu quan trọng là làm thế nào xác định được chức năng của gene này, yêu cầu tương tự cũng được đặt ra khi phát hiện ra protein mới. Một phương pháp tiếp cận phổ biến đó là chúng ta sẽ so sánh, đánh giá sự giống nhau (tương đồng) của chuỗi DNA, protein này với những chuỗi DNA, protein đã biết, từ đó có thể đưa ra dự đoán về chức năng cũng như cấu trúc của những gene mới phát hiện (Sequence Alignment). Quá trình tiến hóa của loài người là một quá trình biến đổi đa dạng, từ một gene (chuỗi DNA) tổ tiên dưới tác động của quá trình tiến hóa đã biến đổi tạo nên những khác biệt so với gene gốc ban đầu. Do đó việc nhận định sự giống nhau của các đoạn gene, trình tự là một vấn đề lớn của sinh tin học. Vấn đề được đặt ra (trong phân tích trình tự) đó là làm thế nào để có được phép so sánh tốt cho các trình tự DNA, khi mà số lượng tế bào trong cơ thể là khoảng 10^{14} và mỗi tế bào mang khoảng $3 \cdot 10^9$ ký tự trong đoạn DNA của chúng. Bài toán so sánh 2 trình tự (Pairwise Sequence Alignment-PSA) đã được giải quyết trọn vẹn bằng nhiều phương pháp khác nhau, đồng thời với việc giải quyết bài toán này, xuất hiện nhu cầu về việc so sánh nhiều trình tự, để so sánh nhiều đoạn gene hoặc tìm ra một phần tử đại diện cho một tập các gene nhằm đáp ứng nhu cầu ngày càng lớn của việc tìm kiếm dự đoán cấu trúc của các gene, protein, khi kho dữ liệu sinh học được tập hợp ngày càng lớn. Bài toán so sánh nhiều trình tự được đặt ra như vấn đề tất yếu. Không như bài toán so sánh 2 trình tự, bài toán so sánh nhiều trình tự (Multiple Sequence Alignment-MSA) là một bài toán NP mở, cho đến hiện tại (2007) vẫn chưa có một giải pháp nào có thể cung cấp một lời giải trọn vẹn cho bài toán, các lời giải thường tập trung vào việc tìm ra phép so sánh “gần” tốt nhất, và mỗi phương pháp tiếp cận sẽ chỉ cho những lời giải thực sự tốt tùy từng yêu cầu tiếp cận và bài toán cụ thể. Progressive Algorithm là một hướng giải quyết tốt cho bài toán so sánh nhiều trình tự. Đây là phương pháp kết hợp Quy hoạch động (Dynamic Programming) với heuristic. Phương pháp này sẽ tăng tốc độ tính toán, giảm độ phức tạp của giải thuật, có thể áp dụng cho các cơ sở dữ liệu gene lớn, phục vụ cho các dự án giải mã gene của các sinh vật bậc cao.

Từ khi được giới thiệu cho đến hiện nay, bài toán MSA đã và vẫn đang là một thách thức cho các nhà khoa học. Nghiên cứu và tìm ra một giải pháp cho bài toán vẫn là động lực thúc đẩy nhiều công trình khoa học về bài toán này.

Xuất phát từ những đặc điểm của bài toán MSA đề tài này cố gắng tập trung vào giải quyết một số vấn đề sau:

- Khảo sát tổng quát các đặc điểm của bài toán MSA, các phương pháp giải quyết bài toán.
- Nghiên cứu về phương pháp dynamic programming, dynamic programming kết hợp với heuristic, Progressive Algorithm.
- Đề xuất một phương pháp giải quyết bài toán dựa trên Progressive Algorithm.
- Xây dựng chương trình hiện thực giải thuật được đề xuất và kiểm thử trên tập dữ liệu thực tế được lấy từ tổ chức NCBI(National Center for Biotechnology Information), và BAliBASE benchmark.

Với những mục tiêu này đề tài đã thu được một số kết quả:

- Cung cấp cái nhìn tổng quan nhất về so sánh trình tự nói chung và bài toán MSA nói riêng.
- Phân loại các phương pháp giải quyết bài toán MSA, phân tích các ưu điểm và nhược điểm của từng phương pháp.
- Xây dựng giải thuật giải quyết bài toán MSA dựa trên việc cải thiện, tối ưu hoá bài toán PSA về độ chính xác cũng như bộ nhớ sử dụng, thông qua việc sử dụng 3 ma trận đánh giá BLOSUM, từ kết quả này của bài toán PSA sử dụng Progressive Algorithm kết hợp với lời giải của bài toán TSP để thực hiện quá trình so sánh nhiều trình tự, tìm ra lời giải cận tối ưu.
- Xây dựng thành công chương trình hiện thực giải thuật, cho phép tìm lời giải cho bài toán MSA với độ chính xác khá cao dựa trên kết quả kiểm thử trên các mẫu dữ liệu thực tế BAliBase benchmark và NCBI. Chương trình

cho phép tiết kiệm bộ nhớ sử dụng, cũng như thời gian tính toán chấp nhận được.

1.2. Kết cấu của luận văn

Luận văn bao gồm 6 chương.

Chương 1. GIỚI THIỆU

Chương này trình bày về bối cảnh, mục tiêu cũng như kết quả thu được của luận văn.

Chương 2. TỔNG QUAN VỀ KHÁI NIỆM SO SÁNH TRÌNH TỰ

Chương này trình bày tổng quát về khái niệm so sánh trình tự, bài toán PSA, MSA, các phương pháp đánh giá chất lượng của MSA, các phương pháp giải quyết bài toán MSA.

Chương 3. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP THỰC HIỆN

Chương này giới thiệu chung về phương pháp quy hoạch động(dynamic programming). Giới thiệu về phương pháp quy hoạch động giải quyết bài toán PSA, giải thuật tính giá trị PSA cải tiến về mặt bộ nhớ sử dụng. Phần tiếp theo của chương này trình bày về cách tiếp cận bài toán MSA hướng đến bài toán chính xác hoàn toàn bằng quy hoạch động thuần túy, những khó khăn khi tiếp cận theo phương pháp này, giới thiệu một cách giải quyết bài toán MSA theo hướng gần đúng dựa trên kỹ thuật quy hoạch động kết hợp heuristic: Center Star Algorithm . Phần cuối chương này trình bày về 3 điểm chính, bao gồm giới thiệu Progressive Algorithm tổng quát, Progressive Algorithm phổ biến nhất, giải thuật Feng-Doolittle(Feng-Doolittle Algorithm) và một số chương trình hiện thực Progressive Algorithm trong thực tế.

Chương 4. THIẾT KẾ GIẢI THUẬT VÀ HIỆN THỰC PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN MSA

Đây là chương dài nhất và cũng là chương giới thiệu những giải pháp mới của đề tài. Chương này trình bày về cách tiếp cận của luận văn để xây dựng giải thuật giải quyết bài toán MSA. Đầu chương giới thiệu về giải thuật tối ưu hoá tìm lời giải bài toán PSA dựa trên việc sử dụng kết hợp giải thuật tính giá trị PSA trình bày ở chương

3 và kỹ thuật chia để trị để tìm lời giải cho bài toán PSA. Phần này giới thiệu thêm việc sử dụng song song 3 ma trận BLOSUM làm hàm đánh giá để cải tiến độ chính xác, phù hợp với thực tế lời giải của bài toán PSA. Tiếp theo chương này đưa ra một giải pháp mới giải quyết bài toán MSA bằng cách kết hợp sử dụng giải thuật cho bài toán PSA vừa thu được, và giải thuật Feng-Doolittle để mô tả cách thức align các nhóm chuỗi trình tự(sequence) với nhau. Sử dụng kết quả bài toán TSP để tìm ra thứ tự align các nhóm sequence, lựa chọn điểm bắt đầu thực hiện quá trình align các sequence thông qua cách thức chọn điểm trung tâm của Center Star Algorithm. Sau nữa, chương này sẽ trình bày 1 cải tiến của giải thuật mới vừa nêu, nhằm nâng cao chất lượng của MSA bằng kỹ thuật gom nhóm theo khoảng cách ngắn nhất dựa trên thứ tự align thu được từ bài toán TSP. Gần cuối chương sẽ giới thiệu về phương pháp giải quyết bài toán TSP bằng giải thuật di truyền(Genetic Algorithm-GA), và cuối cùng sẽ giới thiệu về các module của chương trình hiện thực giải thuật vừa nêu.

Chương 5. KẾT QUẢ, NHẬN XÉT

Chương này giới thiệu về kết quả của chương trình hiện thực. Đánh giá kết quả này, so sánh với một số chương trình giải quyết bài toán MSA.

Chương 6.KẾT LUẬN

Chương này đề cập lại những việc đã thực hiện được của đề tài. Nêu lên hướng mở rộng và phát triển tiếp theo cho đề tài.