

การแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับโดยขั้นตอนวิธีเชิงวิวัฒนาการ



นายพศุภ ธีเหลืองสวัสดิ์

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2547

ISBN 974-17-7100-2

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

MULTIPLE SEQUENCE ALIGNMENT USING EVOLUTIONARY ALGORITHMS



Mr. Pasut Seeluangsawat

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2004

ISBN 974-17-7100-2

หัวข้อวิทยานิพนธ์

การแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับโดยขั้นตอนวิธีเชิง
วิวัฒนาการ

โดย

นายพศุทธิ์ สีเหลืองสวัสดิ์


สาขาวิชา

วิศวกรรมคอมพิวเตอร์

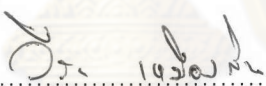
อาจารย์ที่ปรึกษา

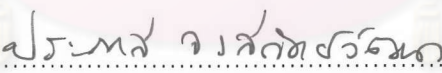
รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา

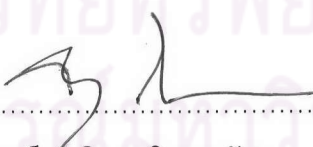
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้
เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

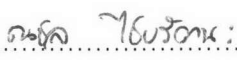

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(อาจารย์ ดร.วีระ เหมืองสิน)


..... อาจารย์ที่ปรึกษา
(รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)


..... กรรมการ
(อาจารย์ ดร.วิชณุ โคตรจรัส)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ณชล ไชยรัตน์)

พศุทธิ์ สีเหลืองสวัสดิ์ : การแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับโดยขั้นตอนวิธีเชิงวิวัฒนาการ.
(MULTIPLE SEQUENCE ALIGNMENT USING EVOLUTIONARY ALGORITHMS) อ. ที่
ปรึกษา : รศ.ดร.ประภาส จงสติกติย์วัฒนา, 64หน้า. ISBN 974-17-7100-2.

ปัญหาการจัดเรียงลำดับเบสหลายลำดับเป็นปัญหาที่สำคัญทางด้านชีวสารสนเทศศาสตร์ ซึ่งปัญหานี้มีการศึกษาค้นคว้าอย่างแพร่หลาย และมีเครื่องมือสำหรับแก้ปัญหามากมาย วิทยานิพนธ์นี้นำเสนอการแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับโดยขั้นตอนวิธีเชิงวิวัฒนาการแบบหลายวัตถุประสงค์ เพื่อพัฒนาคำตอบจากโปรแกรมสำหรับการแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับ โดยผลเฉลยเริ่มต้นของงานวิจัยนี้มาจากโปรแกรม Clustal W, Dialign, MFFT และ T-Coffee

งานวิจัยนี้ทำการทดสอบโดยใช้ชุดข้อมูลจากฐานข้อมูล BALiBASE และผลการทดลองที่ได้จะทำการเปรียบเทียบคำตอบกับโปรแกรมที่มีอยู่ ผลเปรียบเทียบการทดลองแสดงให้เห็นว่าค่าความถูกต้องของคำตอบที่ได้มีการพัฒนาขึ้นอย่างเห็นได้ชัด

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา.... วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต.....*พศุทธิ์ สีเหลืองสวัสดิ์*
สาขาวิชา...วิศวกรรมคอมพิวเตอร์.....ลายมือชื่ออาจารย์ที่ปรึกษา...*ประภาส จงสติกติย์วัฒนา*
ปีการศึกษา2547.....

4670403421 : MAJOR Computer Engineering

KEY WORD: MULTIPLE SEQUENCE ALIGNMENT / EVOLUTIONARY ALGORITHM

PASUT SEELUANGSAWAT : MULTIPLE SEQUENCE ALIGNMENT USING
EVOLUTIONARY ALGORITHMS. THESIS ADVISOR : ASSOC. PROF. PRABHAS
CHONGSTITVATANA, 64 pp. ISBN 974-17-7100-2.

The problem of multiple sequence alignment is important for bioinformatics. This problem is widely studied and there are many popular tools to solve this problem. This thesis introduces a multiple objective evolutionary algorithm to improve solutions obtained from existing tools. An initial solution for the proposed algorithm is derived from Clustal W, Dialign, MFFT and T-Coffee.

The proposed algorithm is tested with the dataset from BALiBASE database. The experiments are conducted to compare the results from the proposed algorithm against the results from existing algorithms. The comparison shows a clear improvement in terms of correctness of the results.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Department.... Computer Engineering.... Student's signature... *Pasut S.*.....

Field of study.... Computer Engineering... Advisor's signature... *P. Chongstitvana*.....

Academic year ...2004.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความช่วยเหลือ และความกรุณาจาก รศ.ดร. ประภาส จงสฤษดิ์วัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ อันเป็นผู้ให้คำปรึกษา และดูแลในการทำวิจัยให้กับข้าพเจ้าเสมอมา

ขอขอบคุณเพื่อน พี่ และ น้องนักศึกษาปริญญาโททุกท่านที่ให้กำลังใจ และร่วมทุกข์ร่วมสุขกันมาตลอด

ท้ายที่สุดขอกราบขอบพระคุณพ่อ คุณแม่ผู้ให้กำเนิด รวมทั้งพี่ชาย ที่คอยสนับสนุน ดูแล และให้กำลังใจข้าพเจ้าจวบจนทุกวันนี้



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญภาพ.....	ญ
สารบัญตาราง.....	ฎ
บทที่	
1 บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์.....	2
1.3 ขอบเขตงานวิจัย.....	2
1.4 ขั้นตอนและวิธีดำเนินงานวิจัย.....	2
1.5 ประโยชน์ที่คาดว่าจะได้รับ.....	3
1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์.....	3
2 ทฤษฎีที่เกี่ยวข้อง.....	4
2.1 ความเป็นมาเกี่ยวกับสายพันธุกรรม.....	4
2.2 ชีวสารสนเทศศาสตร์ (Bioinformatics).....	5
2.2.1 ช่วงเริ่มต้นของชีวสารสนเทศศาสตร์.....	6
2.2.2 ช่วงก่อนโครงการจีโนมมนุษย์เสร็จ.....	6
2.2.3 ช่วงหลังโครงการจีโนมมนุษย์.....	6
2.3 ดีเอ็นเอ อาร์เอ็นเอ และโปรตีน.....	8
2.3.1 องค์ประกอบและโครงสร้างของดีเอ็นเอ.....	8
2.3.1.1 องค์ประกอบทางเคมีของดีเอ็นเอ.....	9
2.3.1.2 โครงสร้างของดีเอ็นเอ.....	10
2.3.1.3 หน้าที่ของดีเอ็นเอ.....	10
2.3.1.4 ความแตกต่างของดีเอ็นเอ.....	10
2.3.2 องค์ประกอบและโครงสร้างของอาร์เอ็นเอ.....	12
2.3.3 องค์ประกอบและโครงสร้างของโปรตีน.....	13

สารบัญ(ต่อ)

	หน้า
2.4 ปัญหาการจัดเรียงลำดับเบสหลายลำดับ	15
2.4.1 รูปแบบข้อมูลที่ใช้กับปัญหา	15
2.4.2 คะแนนที่ใช้คำนวณความถูกต้อง	16
2.4.2.1 สมการผลรวมคู่เบส และตารางการแทน	16
2.4.2.2 วิธีการคำนวณหาเอ็นโทรปีต่ำสุด	17
2.4.2.3 การหักคะแนนเมื่อมีแก๊ป	18
2.5 การแก้ปัญหาแบบหลายวัตถุประสงค์ (Multiple objective optimization)	19
2.5.1 ทฤษฎีเบื้องต้นของการแก้ปัญหาแบบหลายวัตถุประสงค์	19
2.6 ขั้นตอนวิธีเชิงวิวัฒนาการ (Evolutionary Algorithm)	20
2.6.1 ขั้นตอนวิธีพันธุกรรม (Genetic Algorithms)	20
2.6.2 รายละเอียดของขั้นตอนวิธีเชิงพันธุกรรม	21
2.6.2.1 การสร้างกลุ่มประชากรของผลเฉลยตั้งต้น	21
2.6.2.2 การตรวจสอบค่าความแข็งแรงของผลเฉลย	22
2.6.2.3 การสร้างกลามประชากรของผลเฉลยรุ่นใหม่	22
2.6.2.4 การค้นหาคำตอบ	23
3 งานวิจัยที่เกี่ยวข้อง	24
3.1 งานวิจัยเกี่ยวกับขั้นตอนวิธีเชิงวิวัฒนาการแบบหลายวัตถุประสงค์	24
3.1.1 แบบไม่ใช้พารามิเตอร์	24
3.1.2 แบบใช้พารามิเตอร์	25
3.2 งานวิจัยเกี่ยวกับการจัดเรียงลำดับเบสหลายลำดับ	25
3.2.1 การใช้ขั้นตอนวิธีเชิงวิวัฒนาการในการแก้ปัญหา	25
3.2.2 การใช้ขั้นตอนวิธีต่างๆ	26
3.3 ตัวอย่างขั้นตอนวิธีในการแก้ไขปัญหาการจัดเรียงลำดับเบสหลายลำดับ	27
3.3.1 การจัดเรียงองค์รวม (Global alignment)	27
3.3.2 Clustal W	29
3.3.3 MSA-EA2002	31
3.3.4 MSA-EA2003	35
3.4 ฐานข้อมูล BALiBASE	38
4 รายละเอียดงานวิจัย	41

สารบัญ(ต่อ)

หน้า

4.1	ขั้นตอนวิธีโปรแกรม MOMSA.....	41
4.2	การเข้ารหัสปัญหา	42
4.3	การสร้างประชากรเริ่มต้น	43
4.4	การกำหนดฟังก์ชันวัตถุประสงค์.....	43
4.5	การกำหนดลำดับที่	43
4.6	การคัดเลือกประชากรรุ่นถัดไป.....	44
4.7	การคัดเลือกผลเฉลยเพื่อปรับปรุง.....	45
4.8	การไขว้เปลี่ยน และการกลายพันธุ์	45
4.8.1	การย้ายเบสแบบสุ่ม	45
4.8.2	การย้ายเบสแบบย้ายฝั่ง.....	46
4.8.3	การเลื่อนแถว	46
4.8.4	การไขว้เปลี่ยนแบบสองจุด.....	47
4.9	การหาผลเฉลยสุดท้าย	47
4.10	พารามิเตอร์ที่ใช้.....	48
5	การทดลอง	50
5.1	ข้อมูลที่ใช้ในการทดลอง	50
5.2	การคำนวณความถูกต้องของผลเฉลย	50
5.3	ผลการทดลอง	51
5.3.1	ผลการเปรียบเทียบระหว่างการใช้ฟังก์ชันวัตถุประสงค์เดียวกับการใช้ สองฟังก์ชันวัตถุประสงค์	51
5.3.2	ผลการเปรียบเทียบการใช้โปรแกรมต่างๆในการสร้างกลุ่มประชากร เริ่มต้นให้กับโปรแกรม MOMSA	53
6	สรุปผลการวิจัยและข้อเสนอแนะ	60
6.1	สรุปผลการวิจัย.....	60
6.2	ข้อเสนอแนะ	60
	รายการอ้างอิง.....	61
	ประวัติผู้เขียนวิทยานิพนธ์	64

สารบัญภาพ

หน้า

รูปที่ 2.1 ตัวอย่างแสดง SNP	7
รูปที่ 2.2 องค์ประกอบของดีเอ็นเอ	8
รูปที่ 2.3 สายนิวคลีโอไทด์ที่เชื่อมด้วยพันธะฟอสโฟไดเอสเตอร์จากปลาย 5' ไปปลาย 3'	9
รูปที่ 2.4 โครงสร้างของดีเอ็นเอ	10
รูปที่ 2.5 การกลายพันธุ์บนสายดีเอ็นเอ	11
รูปที่ 2.6 การขาดหาย หรือการขยายจำนวนของดีเอ็นเอในกระบวนการไขว้เปลี่ยน	12
รูปที่ 2.7 การขยายจำนวนของดีเอ็นเอที่เกิดจากการเลื่อนของเบส	12
รูปที่ 2.8 หน้าทีของอาร์เอ็นเอ	13
รูปที่ 2.9 โครงสร้างกรดอะมิโนทั้ง 20 ชนิด	14
รูปที่ 2.10 โครงสร้างโปรตีน	15
รูปที่ 2.11 ตัวอย่างข้อมูลรูปแบบเอ็มเอสเอฟของรหัสกรดอะมิโน	15
รูปที่ 2.12 ตัวอย่างลำดับเบสหลายลำดับสำหรับคำนวณเอ็นโทรปี	17
รูปที่ 2.13 ตัวอย่างการคำนวณการหักคะแนนเมื่อมีแก๊ป	18
รูปที่ 2.14 กราฟแสดงปัญหาแบบสองวัตถุประสงค์	19
รูปที่ 2.15 รหัสเทียบของขั้นตอนวิธีเชิงพันธุกรรม	21
รูปที่ 2.16 แสดงตัวอย่างการกลายพันธุ์จากผลเฉลยต้นแบบ (ก) ไปเป็นผลเฉลยใหม่ (ข)	23
รูปที่ 2.17 ตัวอย่างการไขว้เปลี่ยน จากผลเฉลยต้นแบบ 2 ผลเฉลย	23
รูปที่ 3.1 รหัสเทียบของขั้นตอน Needleman-Wunsch	27
รูปที่ 3.2 การแก้ปัญหาโดยขั้นตอน Needleman-Wunsch	28
รูปที่ 3.3 แสดงตารางค่าระยะทาง	29
รูปที่ 3.4 แสดง Neighbor-joining tree	30
รูปที่ 3.5 ยกตัวอย่างลำดับแรกที่มีการจัดเรียง	30
รูปที่ 3.6 แสดงการสร้างประชากรของผลเฉลยตั้งต้น	32
รูปที่ 3.7 แสดงวิธี LocalShuffle	32
รูปที่ 3.8 แสดงวิธี GrowMatchedColumn	33
รูปที่ 3.9 แสดงวิธี RecombineMatchedColumn	34
รูปที่ 3.10 รหัสเทียบขั้นตอนวิธีของ MSA-EA2002	35
รูปที่ 3.11 รหัสเทียบขั้นตอนวิธีของ MSA-EA2003	37

สารบัญภาพ(ต่อ)

	หน้า
รูปที่3.12 สารบบในฐานข้อมูล BALIBASE.....	38
รูปที่4.1 รหัสเทียบชั้นตอนวิธีของโปรแกรม MOMSA.....	42
รูปที่4.2 รูปการแก้ปัญหาที่มีฟังก์ชันวัตถุประสงค์สองค่า.....	44
รูปที่4.3 การสร้างประชากร และหน่วยเก็บถาวรรุ่นต่อไป.....	44
รูปที่4.4 แสดงการย้ายเบสแบบสุ่ม	45
รูปที่4.5 แสดงการย้ายเบสแบบย้ายฝั่ง.....	46
รูปที่4.6 แสดงการเลื่อนแถว	46
รูปที่4.7 แสดงการไขว้เปลี่ยนแบบสองจุด.....	47
รูปที่4.8 รูปการหาผลเฉลยสุดท้าย.....	48
รูปที่5.1 รหัสเทียบชั้นตอนวิธีของ SOMSA.....	51



 ศูนย์วิทยทรัพยากร
 จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

	หน้า
ตารางที่ 3.1 แสดงลักษณะความยาวของกลุ่มลำดับอ้างอิงในแต่ละสารบบ.....	39
ตารางที่ 5.1 รายละเอียด 9 ชุดข้อมูลที่ใช้ในการทดลอง.....	50
ตารางที่ 5.2 เปรียบเทียบค่า SPS/CS ที่ได้จากโปรแกรม MOMSA, SOMSA1,..... SOMSA2.....	52
ตารางที่ 5.3 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม Clustal W กับโปรแกรม MOMSA	53
ตารางที่ 5.4 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม Dialign กับโปรแกรม MOMSA	54
ตารางที่ 5.5 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม MAFFT version fft-n-1 กับโปรแกรม MOMSA	55
ตารางที่ 5.6 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม MAFFT version fft-n-2 กับโปรแกรม MOMSA.....	56
ตารางที่ 5.7 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม T-Coffee..... กับโปรแกรม MOMSA	57
ตารางที่ 5.8 แสดงผลที่ได้จากโปรแกรม MOMSA เมื่อ + คือคำตอบมีค่าดีขึ้น และ..... - คือคำตอบมีค่าแย่ลง.....	58
ตารางที่ 5.9 แสดงค่า SPS/CS ที่ได้จากโปรแกรม MOMSA	59

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย