



ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เรื่อง การประกอบชิ้นส่วนดีเอ็นเอแบบอัตโนมัติและเป็นลำดับโดยใช้กฎเมตาฮิวริสติก
โดย นายกัมพล สุกใส

ได้รับอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า

คณบดีบัณฑิตวิทยาลัย

(อาจารย์ ดร.มงคล หวังสถิตย์วงศ์)

21 พฤษภาคม 2550

คณะกรรมการสอบวิทยานิพนธ์

๑๖๕๘ ๗๖๖๓๖: ประธานกรรมการ

(รองศาสตราจารย์ ดร.ณชด ไชยรัตน์)

กรรมการ

(รองศาสตราจารย์ ดร.วรา วราวิทย์)

กรรมการ

(ดร.ศิษณุศ ทองสีมา)

การประกอบชิ้นส่วนดีเอ็นเอแบบอัตโนมัติและเป็นลำดับโดยใช้กฎเมตาอีวริสติก

นายกัมพล สุกใส

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปีการศึกษา 2549

ลิขสิทธิ์ของสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ชื่อ : นายกัมพล สุกใส
ชื่อวิทยานิพนธ์ : การประกอบชิ้นส่วนดีเอ็นเอแบบอัตโนมัติและเป็นลำดับโดยใช้กฎ
เมตาฮิวริสติก
สาขาวิชา : วิศวกรรมไฟฟ้า
: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ที่ปรึกษาวิทยานิพนธ์ : รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
ปีการศึกษา : 2549

บทคัดย่อ

การประกอบชิ้นส่วนดีเอ็นเอกลายเป็นปัญหา เนื่องจากการหาลำดับเบสของสายดีเอ็นเอที่มีความยาวมาก ๆ นั้น จำเป็นจะต้องทำการหาลำดับแบบซ้ำซ้อนกัน โดยการทำให้ดีเอ็นเอมีชิ้นส่วนที่เล็กพอที่จะนำมาหาลำดับเบสได้ ซึ่งหลังจากนั้นนำชิ้นส่วนดีเอ็นเอที่ได้ลำดับเบสแล้วมาทำการประกอบกลับเข้าเป็นสายยาวอีกครั้ง ปัญหาในการประกอบชิ้นส่วนดีเอ็นเอนี้ อยู่ที่การหาอันดับและทิศทางที่ถูกต้องของชิ้นส่วนดีเอ็นเอแต่ละชิ้น และเมื่อจัดเรียงแล้ว จะต้องเป็นลำดับที่สอดคล้องกับสายสายดีเอ็นเอดั้งเดิม วิธีการประกอบชิ้นส่วนดีเอ็นเอที่น่าเสนอนี้จะใช้กฎเมตาฮิวริสติก ซึ่งประกอบด้วย 3 ขั้นตอนคือการหาคะแนนค่าซ้อนเหลื่อม ซึ่งได้จากวิธีสมิท-วอเตอร์แมน จากนั้นจะเป็นขั้นตอนการหาอันดับและทิศทางของชิ้นส่วนดีเอ็นเอ ซึ่งได้จากกฎฮิวริสติกเพื่อนบ้านที่ใกล้ที่สุดร่วมกับวิธีเอชไอเอส และนำค่าที่ได้มาจัดแถวลำดับชิ้นส่วนดีเอ็นเอให้เป็นคอนทิกด้วยโปรแกรมที่เขียนขึ้น โดยผลลัพธ์คอนทิกแต่ละเส้นที่ได้นี้ สามารถพิจารณาเป็นชิ้นส่วนดีเอ็นเอย่อยเพื่อนำกลับไปคำนวณด้วยวิธีข้างต้นอีกครั้งได้ และผลลัพธ์ที่ได้จากการทดลองนี้ จะนำมาเทียบเคียงด้วยโปรแกรมแปลกับสายดีเอ็นเอต้นแบบ

(วิทยานิพนธ์มีจำนวนทั้งสิ้น 38 หน้า)

คำสำคัญ : กฎเมตาฮิวริสติก, การจัดแถวลำดับ, การประกอบชิ้นส่วนดีเอ็นเอ

อาจารย์ที่ปรึกษาวิทยานิพนธ์

Name : Mr.Kampol Suksai
Thesis Title : Hierarchical Automated DNA Fragment Assembly by Metaheuristics
Major Field : Electrical EngineeringKing Mongkut's Institute of Technology North
Bangkok
Thesis Advisor : Associate Professor Dr.Nachol Chaiyaratana
Academic Year : 2006

Abstract

DNA fragment assembly is an important step in genome sequencing projects. The assembly problem generally arises during the sequencing of large strands of DNA where the strands are needed to be shotgun-replicated and broken into fragments that are small enough for sequencing. The assembly problem can thus be classified as a combinatorial optimization problem where the aim is to find the right order and orientation of each fragment in the ordering layout that leads to the formation of a consensus sequence that truly reflects the original DNA strands. This thesis presents a metaheuristic procedure for the task where the procedure contains three main steps: fragment overlap score calculation, fragment layout determination and contiguous sequence (contig) construction. The overlap score is obtained from a Smith-Waterman algorithm while the score-based fragment layout including the order and orientation of each fragment is generated by a nearest neighbor heuristic rule and an ant colony system algorithm. With the availability of the fragment layout, contigs can be constructed where each contig can also be viewed as a new DNA fragment. As a result, the metaheuristic procedure can be recursively applied to the resulting contig sets. The experimental results indicate that the performance of the metaheuristic procedure is comparable to that of a CAP3 assembly program.

(Total 38 pages)

Keywords : Metaheuristics, Sequence Alignment, DNA Fragment Assembly

Advisor

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ได้ด้วยความช่วยเหลืออย่างดียิ่งของรองศาสตราจารย์ ดร.ณชด ไชยรัตน์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำปรึกษา แนะนำ ข้อคิดเห็น และแนวทางแก้ไขต่างๆ อันเป็นประโยชน์อย่างยิ่งในการทำวิทยานิพนธ์นี้มาโดยตลอด และขอขอบคุณเพื่อนๆ ทุกคนที่ภาควิชาวิศวกรรมไฟฟ้า เพื่อนกลุ่มบ้านเซเวียร์ทุกคนที่ให้กำลังใจและคอยช่วยเหลือสิ่งต่างๆ จนกระทั่งวิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ที่คอยช่วยเหลือสนับสนุนทั้งด้านทุนทรัพย์ และให้กำลังใจรวมทั้งเป็นแรงผลักดันเสมอมาจนสำเร็จการศึกษา

กัมพล สุกใส

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ข
บทคัดย่อภาษาอังกฤษ	ค
กิตติกรรมประกาศ	ง
สารบัญตาราง	ฉ
สารบัญภาพ	ช
บทที่ 1 บทนำ	1
บทที่ 2 ปัญหาการประกอบชิ้นส่วนดีเอ็นเอ	5
2.1 ขั้นตอนวิธีระบบอาณานิคมมด	6
2.2 กฎการเปลี่ยนสถานะในขั้นตอนวิธีเอซีเอส	7
2.3 กฎการปรับฟีโรโมนเฉพาะที่ในขั้นตอนวิธีเอซีเอส	8
2.4 บัญชีคู่แข่ง	8
2.5 การประยุกต์ใช้ขั้นตอนวิธีเอซีเอสในการแก้ปัญหาการประกอบชิ้นส่วนดีเอ็นเอ	10
บทที่ 3 วิธีดำเนินการวิจัย	17
3.1 ขั้นตอนการเตรียมและปรับปรุงข้อมูล	17
3.2 กำหนดค่าช้อนเหลื่อมของแต่ละข้อมูล	18
3.3 การหาค่าลำดับสอดคล้องจากกฎฮิวริสติกเพื่อนบ้านและขั้นตอนวิธีอาณานิคมมด	19
3.4 ขั้นตอนการประกอบชิ้นส่วนดีเอ็นเอ	22
3.5 ขั้นตอนการเปรียบเทียบผลคอนทิกกับสายต้นแบบดีเอ็นเอ	26
บทที่ 4 ผลการทดลอง	28
4.1 จำนวนคอนทิกจากวิธีเอ็นเอ็นเอชกับวิธีเอซีเอส เทียบกับโปรแกรมแคปทรี	28
4.2 คุณภาพของผลเฉลยจากการประกอบชิ้นส่วนดีเอ็นเอ	30
4.3 วิเคราะห์และสรุปผลทดลอง	33
บทที่ 5 สรุปผลการวิจัย	35
บรรณานุกรม	36
ประวัติผู้วิจัย	38

สารบัญตาราง

ตารางที่	หน้า
2-1	แสดงการจัดวาง 4 รูปแบบที่เป็นไปได้ระหว่างชั้นส่วนสองชั้น
3-1	รายละเอียดชุดข้อมูล
3-2	การกำหนดค่าพารามิเตอร์สำหรับการคำนวณหาค่าชั้นเสื่อมในไฟล์ประกอบ พีเอชอาร์เอพี(PHRAP)
3-3	การกำหนดค่าพารามิเตอร์สำหรับกฎฮิวริสติกเพื่อนบ้านที่ใกล้ที่สุด
3-4	การกำหนดค่าพารามิเตอร์สำหรับขั้นตอนวิธีระบบอาณานิคม
4-1	จำนวนคอนทิกจากผลเฉลยที่ได้จากวิธีเอ็นเอ็นเอชกับวิธีเอชไอเอสและโปรแกรมแคปทรี
4-2	ค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกเทียบกับชุดข้อมูลกับ จำนวนรอบในการทดลองและโปรแกรมแคปทรี
4-3	ค่าความผิดพลาดในการครอบคลุมเทียบกับชุดข้อมูลกับจำนวนรอบในการทดลอง และโปรแกรมแคปทรี

สารบัญภาพ

ภาพที่	หน้า
2-1 แผนผังแสดงกระบวนการประกอบชิ้นส่วน	5
2-2 สรุปลขั้นตอนวิธีระบบอาณานิคมมด	9
2-3 แสดงการตัดสินใจที่เป็นไปได้ในการจัดวางชิ้นส่วนสองชิ้น	12
2-4 อาร์เรย์ขนาด $2 \times 2n \times cl$ ที่ใช้เก็บข้อมูลบัญชีคู่แข่ง	14
3-1 ลักษณะไฟล์โครมาโตแกรม	18
3-2 แสดงชิ้นส่วนสายปกติ และชิ้นส่วนสายประกอบแบบอ่านจากหลังมาหน้า	18
3-3 แสดงลักษณะไฟล์ Score_seq.txt และ Frag_seq.txt	22
3-4 การจัดเรียงชิ้นส่วนดีเอ็นเอ คู่แรก	23
3-5 การจัดเรียงชิ้นส่วนดีเอ็นเอคู่ 2	23
3-6 การจัดเรียงชิ้นส่วนดีเอ็นเอ ภายในอาร์เรย์ผลลัพธ์	23
3-7 การจัดเรียงชิ้นส่วนดีเอ็นเอคู่ 3	24
3-8 การจัดเรียงชิ้นส่วนดีเอ็นเอ ในอาร์เรย์ผลลัพธ์	24
3-9 การจัดเรียงชิ้นส่วนดีเอ็นเอที่เหลือในอาร์เรย์ผลลัพธ์	25
3-10 คอนทิกที่ได้จากการใช้กฎการเลือกตามเสียงข้างมาก	25
3-11 จำนวนคอนทิกของแต่ละข้อมูลตามจำนวนรอบของการทดลอง	25
3-12 ตัวอย่างความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออก	26
3-13 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากโปรแกรม SIM.EXE	27
4-1 แสดงจำนวนคอนทิกของแต่ละข้อมูลตามจำนวนรอบของการทดลอง	30
4-2 ตารางเปรียบเทียบค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกต่อจำนวนรอบ	31
4-3 กราฟเปรียบเทียบค่าความผิดพลาดในการครอบคลุมต่อจำนวนรอบ	33

บทที่ 1

บทนำ

ปัญหาการประกอบชิ้นส่วนดีเอ็นเอเกี่ยวข้องกับการค้นหาอันดับ (Ordering) และทิศทาง (Orientation) ที่ถูกต้องของชิ้นส่วนดีเอ็นเอแต่ละชิ้นในลำดับการจัดเรียงตัว ซึ่งนำไปสู่การสร้างลำดับสอดคล้องที่แสดงถึงสายดีเอ็นเอต้นแบบ ทั้งนี้ปัญหาการประกอบชิ้นส่วนดีเอ็นเอสามารถพิจารณาให้เป็นปัญหาการหาค่าเหมาะที่สุดเชิงการจัด (Combinatorial Optimisation Problem) ได้ ซึ่งความเหมาะสมในการจัดเรียงตัวของชิ้นส่วนสามารถกำหนดโดยผลรวมของค่าซ้อนเหลื่อมจะคำนวณมาจากชิ้นส่วนแต่ละคู่ที่เรียงต่อกัน วิธีการค้นหาทั้งแบบเชิงกำหนดและแบบเฟ้นสุ่มต่างๆ (Deterministic and Stochastic Search Techniques) ได้ถูกนำมาใช้ในการแก้ปัญหการประกอบชิ้นส่วนดีเอ็นเอ (Pop et al., 2002) เช่น Staden (1980), Huang (1992), Sutton et al. (1995), Huang (1996), Huang และ Madan (1999) และ Green (2004) ได้ใช้ขั้นตอนวิธีละโมบ (Greedy Search Algorithm) เข้ามาแก้ปัญหายังไรก็ดีผลลัพธ์ที่ได้จากคอมพิวเตอร์จำเป็นต้องมีการนำมาปรับ (Manual Manipulation) เพื่อให้สอดคล้องกับผลลัพธ์ในเชิงชีววิทยา และยังมีการนำวิธีการค้นหาเชิงกำหนดแบบอื่นๆมาใช้ ได้แก่ ขั้นตอนวิธีขยายและตัด (Branch-and-Cut Algorithm) (Ferreira et al., 2002) ขั้นตอนวิธีเชิงทฤษฎีกราฟ (Graph-Theoretic Algorithm) ซึ่งชิ้นส่วนดีเอ็นเอจะถูกแทนให้เป็นปมกราฟ (Graph Nodes) (Kececiloglu and Myers, 1995) หรือขอบกราฟ (Graph Edges) (Pevzner et al., 2001) นอกจากนี้ยังมีการศึกษาถึงการนำขั้นตอนวิธีการค้นหาแบบเฟ้นสุ่ม อาทิเช่น ขั้นตอนวิธีจำลองการอบอ่อน (Simulated Annealing Algorithm) (Churchill et al., 1993) ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) และเทคนิคการทำนายซึ่งอาศัยข่ายงานระบบประสาท (Neural Network Based Prediction Technique) (Angeleri et al., 1999) มาใช้แก้ปัญหาด้วย ผลการประกอบชิ้นส่วนดีเอ็นเอที่ดีที่สุดที่ได้จากวิธีการค้นหาแบบเฟ้นสุ่ม ถูกตีพิมพ์ไว้ในบทความของ Parsons and Johnson (1997) และ Kim and Mohan (2003) ซึ่งแสดงให้เห็นว่าการใช้ขั้นตอนวิธีเชิงพันธุกรรมจะให้ประสิทธิภาพที่ดีกว่าการใช้ขั้นตอนวิธีละโมบ โดยไม่จำเป็นต้องนำผลลัพธ์มาปรับเปลี่ยน (Manual Manipulation) แต่อย่างใด ถึงแม้ว่าผลลัพธ์ที่ได้จะดีกว่าการใช้ขั้นตอนวิธีละโมบประสิทธิภาพในการค้นหายังสามารถปรับปรุงได้อีก ถ้าสามารถกำจัดความซ้ำซ้อน (Redundancy) ของการแสดงผลเฉลย (Solution Representation) ในขั้นตอนวิธีการค้นหา (Parsons et al., 1995)

ในทำนองเดียวกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมในการหาค่าเหมาะที่สุดเชิงการจัดผลเฉลยในการจัดเรียงชิ้นส่วนดีเอ็นเอจะถูกแสดงในลักษณะมีการสับเปลี่ยนลำดับ (Permutation Representation) ซึ่งการแสดงผลลัพธ์ในลักษณะนี้ทำให้ผลเฉลยที่มีการเรียงลำดับแตกต่างกันสามารถสร้างลำดับสอดคล้องเดียวกันได้ เนื่องจากธรรมชาติของขั้นตอนวิธีเชิงพันธุกรรมที่มีลักษณะการค้นหาเป็นแบบขนาน (Parallel search) (Holland, 1975; Goldberg, 1989) ความซ้ำซ้อนของการแสดงผล (Representation Redundancy) จะทำให้ประสิทธิภาพในการค้นหาลดลง แนวทางหนึ่งที่สามารถแก้ปัญหานี้ก็คือ การกำจัดผลของความซ้ำซ้อนออกจากกลไกการค้นหาที่จะนำมาใช้ ทั้งนี้ขั้นตอนวิธีการค้นหาแบบเฟ้นสุ่มที่ไม่มีผลของความซ้ำซ้อนวิธีหนึ่ง คือ ขั้นตอนวิธีระบบอาณานิคมมด (Dorigo and Gambardella, 1997)

ขั้นตอนวิธีระบบอาณานิคมมดพัฒนามาจากลักษณะทางธรรมชาติของมดที่อาศัยอยู่ร่วมกันในอาณานิคม โดยมดเหล่านี้สามารถค้นหาเส้นทางการเดินทางที่สั้นที่สุดระหว่างแหล่งอาหารกับรังของมัน (Beckers et al., 1992) โดยไม่มีการทิ้งร่องรอยให้เห็น หากแต่อาศัยฟีโรโมน (Pheromone) โดยในระหว่างการเดินมดจะทิ้งฟีโรโมนลงบนพื้น และมีแนวโน้มที่จะเดินตามฟีโรโมนที่ถูกทิ้งโดยมดตัวอื่นก่อนหน้านี้ ซึ่งเราสามารถอธิบายการใช้ฟีโรโมนในการหาระยะทางที่สั้นที่สุดระหว่างจุดสองจุดได้ดังนี้ โดยพิจารณาสถานการณ์ที่กลุ่มมดจำเป็นต้องตัดสินใจว่าจะเดินไปตามเส้นทางใดระหว่างเส้นทางสองเส้นทางที่เป็นไปได้ในการเดินไปหาเป้าหมายและเดินย้อนกลับมา เนื่องจากว่าไม่มีร่องรอยใดที่แสดงให้เห็นว่าเส้นทางไหนดีกว่ากัน พวกมดจึงจำเป็นต้องสุ่มเลือกเส้นทางใดเส้นทางหนึ่ง ซึ่งสามารถสันนิษฐานได้ว่าครั้งหนึ่งของมดจะเลือกไปทางเส้นทางหนึ่ง ส่วนอีกครึ่งหนึ่งจะเดินไปในเส้นทางที่เหลือ สมมติว่ามดทุกตัวเดินด้วยความเร็วที่ใกล้เคียงกัน ฟีโรโมนที่ถูกปล่อยจะมีการสะสมอย่างรวดเร็วในเส้นทางที่สั้นที่สุด หลังจากนั้นไม่นานความแตกต่างระหว่างปริมาณฟีโรโมนบนเส้นทางทั้งสองจะมีมากพอที่จะส่งผลต่อการตัดสินใจของมดตัวอื่นที่เดินมาถึงจุดที่ต้องทำการตัดสินใจนี้ ดังนั้นมดที่เดินตามหลังมาจะเลือกเส้นทางที่สั้นกว่า เนื่องจากมันสามารถรับรู้ได้ถึงปริมาณฟีโรโมนที่มีมากกว่า ทำให้อุณหภูมิของมดทุกตัวจะใช้เส้นทางที่สั้นกว่า ดังนั้นถ้าพวกมดต้องเดินไปยังเป้าหมายที่แตกต่างกัน n เป้าหมายจนครบรอบโดยไม่คำนึงถึงลำดับของเป้าหมายที่จะเดินไป เส้นทางที่สั้นที่สุดที่เกิดขึ้นในการเดินนี้ จะเป็นผลเฉลยของปัญหาการเดินทางของพนักงานขาย (Travelling Salesman Problem, TSP) ที่ต้องเดินทางไปยังเมือง n เมืองจากลักษณะดังกล่าวข้างต้น Dorigo และ Gambardella (1997) ได้พัฒนาขั้นตอนวิธีระบบอาณานิคมมดสำหรับการหาค่าเหมาะที่สุดเชิงการจัด วิธีการนี้ถูกนำไปใช้ได้อย่างดีในปัญหาหลากหลาย อาทิเช่น ปัญหาการจัดเรียงลำดับอย่างต่อเนื่อง (Gambardella and Dorigo, 2000) ปัญหาการหาเส้นทางคมนาคม (Vehicle Routing Problem) ปัญหาการหาเส้นทางในเครือข่ายของระบบสื่อสาร

(Telecommunication Network Routing Problem) (Di Caro and Dorigo, 1998) ปัญหาการจัดวางตัวของลำดับหลายลำดับเข้าด้วยกัน (Multiple Sequence Alignment Problem) (Moss and Johnson, 2003) แม้ว่าขั้นตอนวิธีเอซีเอส (ACS) จะใช้กลไกการค้นหาแบบขนานในเชิงสุ่ม (Stochastic Parallel Search Mechanism) แต่ประสิทธิภาพของขั้นตอนวิธีนี้จะไม่ขึ้นกับรูปแบบการแสดงผลเฉลย (Solution Representation) ทั้งนี้เนื่องจากว่าคำตอบที่เหมาะสมที่สุดที่ได้มาจากขั้นตอนวิธีเอซีเอสจะเกิดขึ้นในลักษณะที่เป็นเส้นทางที่สั้นที่สุดเพียงคำตอบเดียว ดังนั้นจึงกล่าวได้ว่าวิธีการนี้มีความเหมาะสมอย่างมากในการนำมาแก้ปัญหาการประกอบชิ้นส่วน ดีเอ็นเอ เพราะว่าจะไม่เกิดปัญหาในเรื่องของความซ้ำซ้อนของการแสดงผลเฉลย (Redundancy in Solution Representation)

จากที่ได้กล่าวมางานวิจัยนี้จึงมุ่งเน้นที่จะนำเสนอแนวความคิดในการประกอบชิ้นส่วน ดีเอ็นเอ โดยนำวิธีวิวัฒนาการที่ได้กล่าวมาข้างต้นมาใช้ประกอบในการหาลำดับชิ้นคอนทิก และทำการประกอบชิ้นส่วนดีเอ็นเอเข้าด้วยกัน โดยใช้โปรแกรมเป็นตัวแทนแบบอัตโนมัติ แพลตฟอร์มที่ได้เป็นคอนทิกเหล่านี้ยังสามารถพิจารณาเป็นชิ้นส่วนดีเอ็นเอและนำกลับมาทดลองซ้ำอีกครั้ง เพื่อเทียบเคียงผลคำตอบกับโปรแกรมแคปทรี

1.1 วัตถุประสงค์

- 1.1.1 นำกฎวิวัฒนาการหลายรูปแบบมาช่วยในการหาลำดับชิ้น ของคอนทิกแต่ละคอนทิก
- 1.1.2 สร้างโปรแกรมในการประกอบชิ้นส่วนดีเอ็นเอ และแสดงผลออกมาโดยอัตโนมัติ

1.2 ขอบเขตของการวิจัย

- 1.2.1 ทำการประกอบชิ้นส่วนดีเอ็นเอจากข้อมูลที่มี โดยวิธีการสร้างลำดับสอดคล้องจาก ฟังก์ชันกฎวิวัฒนาการเพื่อนบ้านที่ใกล้ที่สุด และระบบอาณานิคมมด
- 1.2.2 สร้างโปรแกรมในการประกอบชิ้นส่วนดีเอ็นเอและนำผลที่ได้มาเปรียบเทียบกับ โปรแกรมแคปทรี

1.3 วิธีการวิจัย

- 1.3.1 ศึกษาวิธีการประกอบชิ้นส่วนดีเอ็นเอเบื้องต้น และการใช้งาน โปรแกรม และขั้นตอนวิธี สมิท-วอเตอร์แมน
- 1.3.2 ศึกษาแนวความคิดและวิธีการประกอบชิ้นส่วนดีเอ็นเอ แบบอัตโนมัติโดยใช้โปรแกรม ภาษาไพธอน (Python Language Programming)

1.3.3 ศึกษาวิธีการระบบอาณานิคมมด และกฎฮิวริสติกเพื่อนบ้านที่ใกล้ที่สุด เพื่อนำมาใช้กับ
ตัวงาน

1.3.4 จัดเตรียมข้อมูลที่จะนำมาใช้ จากเว็บไซต์ <http://www.ncbi.nlm.nih.gov>

1.3.5 นำข้อมูลที่ได้ มาใช้กับฟังก์ชันกฎฮิวริสติกเพื่อนบ้านที่ใกล้ที่สุด และระบบอาณานิคม
มด เพื่อหาคำลำดับ และทิศทาง

1.3.6 เขียนโปรแกรม และทดสอบการประกอบชิ้นส่วนดีเอ็นเอ

1.3.7 เก็บข้อมูลและประเมินผล

1.3.8 สรุปทำรายงานการวิจัย และนำเสนอผลงานที่ได้จากงานวิจัย

1.4 เครื่องมือที่ใช้

1.4.1 เครื่องคอมพิวเตอร์

1.4.2 โปรแกรม Python และ โปรแกรม Borland C++ เวอร์ชัน 5.02

1.4.3 โปรแกรม Precap และ โปรแกรมแคปทีรี (CAP3)

1.5 ประโยชน์ของการวิจัย

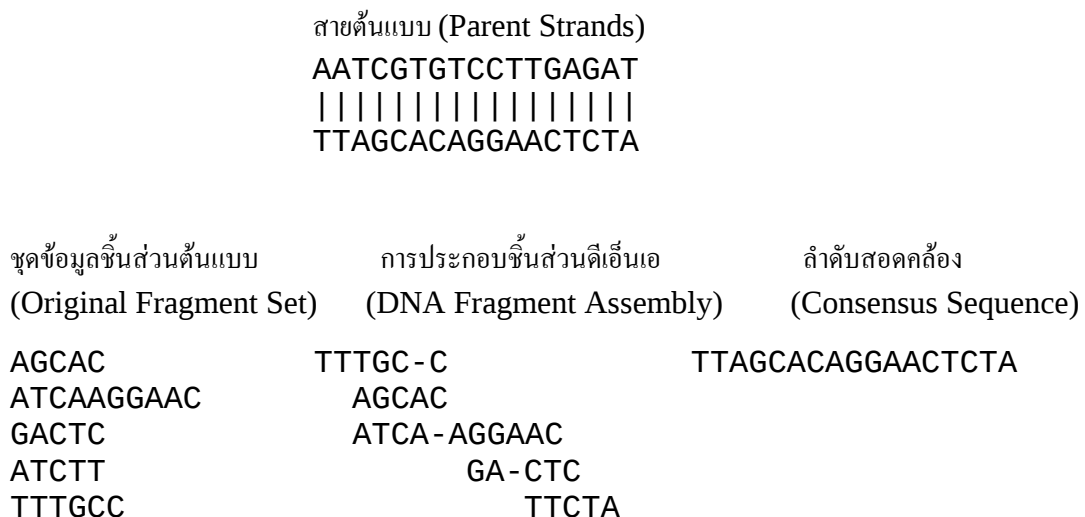
1.5.1 สามารถทำการประกอบชิ้นส่วนดีเอ็นเอโดยใช้วิธีการฮิวริสติกหลายแบบ มาใช้ และทำ
การประกอบชิ้นส่วนดีเอ็นเอ ได้แบบอัตโนมัติ และมีประสิทธิภาพที่ดีกว่าเมื่อเทียบกับโปรแกรม
แคปทีรี (Huang and Madan, 1999)

1.5.2 ได้โปรแกรมฟรีแวร์ สามารถนำไปพัฒนาต่อ เพื่อเพิ่มประสิทธิภาพให้มากขึ้น

บทที่ 2

ปัญหาการประกอบชิ้นส่วนดีเอ็นเอ

ในการสร้างแผนที่ดีเอ็นเอจะเริ่มจากการนำสายดีเอ็นเอต้นแบบมาจำลองหลายๆ ครั้ง จากนั้นแต่ละสายของโพลีนิวคลีโอไทด์ จะถูกส้อมแบ่งออกเป็นชิ้นๆ จากนั้นจะทำการอ่านการจัดเรียงเบส (Base Calling Procedure) ในแต่ละชิ้นส่วนซึ่งจะอ่านจากฟลูออเรสเซนซ์ (Fluorescent Trace-Data) (Allex et al., 1996) เมื่อเบสของแต่ละชิ้นส่วนได้จัดเรียงแล้ว ชิ้นส่วนเหล่านี้จะถูกนำมาจัดวางเข้าด้วยกันเพื่อสร้างลำดับสอดคล้อง ซึ่งจะแสดงถึงสายดีเอ็นเอต้นแบบหรือสายดีเอ็นเอดั้งเดิม ทั้งนี้ในการวางชิ้นส่วนสองชิ้นเข้าด้วยกันจะสามารถทำได้เมื่อมีบางช่วงของชิ้นส่วนแต่ละชิ้นที่สามารถเข้าคู่กันได้ โดยอาจเป็นการเข้าคู่กันของลำดับการจัดเรียงเบสในแบบที่เรียงต่อเป็นสายเดียวกัน (Same-Base Ordering Sequence) หรือเป็นแบบที่ประกอบเข้าด้วยกัน (Complementary-Base Ordering Sequence) ทั้งนี้เพราะชิ้นส่วนเหล่านี้อาจมาจากสายโพลีนิวคลีโอไทด์เดียวกันหรือจากสายประกอบ (Complementary Strands) ก็ได้ การจัดเรียงตัวของชิ้นส่วนสามารถแสดงได้ดังภาพที่ 2-1



ภาพที่ 2-1 แผนผังแสดงกระบวนการการประกอบชิ้นส่วน

จากภาพที่ 2-1 ลำดับสอดคล้องที่ได้จะเป็นสายข้างหนึ่งของสายดีเอ็นเอต้นแบบ เนื่องจากในแต่ละคู่ของชิ้นส่วนที่นำมาจัดวางจะมีการซ้อนเหลื่อม (Overlap) กัน โดยจำนวนเบสที่ตรงกัน

(Matching) และการทำโทษ (Penalty) ในกรณีที่เบสไม่ตรงกัน (Mismatch) หรือเกิดช่องว่าง (Gap) จะถูกนำมากำหนดเป็น ค่าซ้อนเหลื่อม (Overlap Score) ของชิ้นส่วนสองชิ้นที่นำมาจัดวางเข้าด้วยกัน โดยก่อนที่จะได้ค่าซ้อนเหลื่อม ชิ้นส่วนทั้งสองจะถูกจัดวางในทิศทางต่างๆ ที่เป็นไปได้ ทั้งนี้การซ้อนเหลื่อม ทิศทางและการวางตัวจะถูกกำหนดในลักษณะที่จะให้จำนวนเบสที่ตรงกันมากที่สุด และให้มีเบสที่ไม่ตรงกันและมีช่องว่างน้อยที่สุด ถ้าการค้นหาวางตัวที่เป็นไปได้ของชิ้นส่วนที่สนใจกับชิ้นส่วนอื่นๆ ให้ค่าซ้อนเหลื่อมเป็นศูนย์ จะส่งผลให้มีช่องว่างในลำดับสอดคล้องเกิดขึ้น ซึ่งในกรณีนี้ลำดับสอดคล้องที่ได้จะเกิดเป็นลำดับที่มีความไม่ต่อเนื่อง (Multiple Disjoint Sequences) ซึ่งเรียกว่า หลายลำดับติดกัน (Contiguous sequences) หรือหลายคอนทิก (Contig) กล่าวอีกนัยหนึ่งก็คือ เราต้องการให้มีเพียงหนึ่งคอนทิกในลำดับสอดคล้อง ทั้งนี้แต่ละเบสของลำดับสอดคล้องจะถูกกำหนดขึ้นจากการใช้กฎการเลือกตามเสียงข้างมาก (Majority Vote Rule) ในแต่ละหลัก (Column) ของเบสในชิ้นส่วนที่นำมาจัดเรียงกัน

ในการทดลองนี้ได้ใช้ฟังก์ชัน 2 แบบ มาใช้เพื่อหาทิศทางและลำดับในการประกอบชิ้นส่วนดีเอ็นเอ และใช้โปรแกรมที่เขียนขึ้นมา เพื่อทำการประกอบชิ้นส่วนดีเอ็นเอ ให้ทำการจัดเรียงคู่เบส และในกรณีที่มิช่องว่างขึ้น ก็ให้ทำการเปิดช่องว่างโดยอัตโนมัติ เพราะตามปกติแล้วนักชีววิทยา มักจะทำการจัดเรียงด้วยมือเอง (Manual)

2.1 ขั้นตอนวิธีระบบอาณานิคม

เอชีเอตเป็นขั้นตอนวิธีการค้นหามีรากฐานมาจากการศึกษาพฤติกรรมกลุ่มของแมลง (Dorigo and Gambardella, 1997) ขั้นตอนวิธีนี้เหมาะกับการแก้ปัญหาการหาค่าเหมาะที่สุดเชิงการจัด ซึ่งเป็นปัญหาที่มีลักษณะเดียวกับปัญหาการเดินทางของพนักงานขาย หรือ ทีเอสพี โดยมดจะเลือกเดินทางจากเมืองหนึ่งไปยังเมืองถัดไป โดยใช้ข้อมูลด้านระยะทางและปริมาณของฟีโรโมนที่ถูกทิ้งไว้บนเส้นทางระหว่างเมืองต่างๆ ฟีโรโมนเป็นสารชนิดหนึ่งที่ถูกทิ้งไว้บนเส้นทางโดยมดตัวอื่นซึ่งเคยเดินผ่านเส้นทางนั้นมาก่อน ระดับของฟีโรโมนจะสัมพันธ์โดยตรงกับจำนวนมดที่เดินทางระหว่างเมือง และเป็นสัดส่วนผกผันกับระยะทางระหว่างเมือง ในเส้นทางที่มีปริมาณฟีโรโมนมาก จำนวนมดที่เลือกเดินเส้นทางนั้นก็จะมากตามไปด้วย หลังจากที่มีมดทุกตัวได้เดินทางครบรอบโดยเดินผ่านทุกเมืองแล้ว เส้นทางที่ใช้เดินทางได้สั้นที่สุดจะได้รับปริมาณฟีโรโมนเพิ่มเข้าไปอีก ซึ่งจะมีส่วนช่วยเสริมให้มดมาใช้เส้นทางเหล่านั้นมากขึ้น และก่อให้เกิดเส้นทางที่สั้นที่สุดจากการเดินทางทั้งหมด โดยวิธีการนี้ผลเฉลยที่เหมาะสมที่สุดของปัญหาทีเอสพีจะเกิดขึ้นจากการร่วมมือกันของมดทุกตัวในอาณานิคม

จากคำบรรยายข้างต้น ขั้นตอนวิธีนี้ประกอบไปด้วยสามส่วนหลักด้วยกันคือ กฎการเปลี่ยนสถานะ (State Transition Rule) กฎการปรับฟีโรโมนเฉพาะที่ (Local Pheromone-Updating Rule) และกฎการปรับฟีโรโมนวงกว้าง (Global Pheromone-Updating Rule) นอกจากนี้ Dorigo และ Gambardella (1997) ได้เสนอการใช้ชุดข้อมูลที่เรียกว่า บัญชีคู่แข่ง (Candidate List) เพื่อสร้างลำดับของเมืองที่จะถูกเลือกให้แก่มด กฎทั้งสามข้อและบัญชีคู่แข่งสามารถอธิบายได้ดังต่อไปนี้

2.2 กฎการเปลี่ยนสถานะในขั้นตอนวิธีเอชีเอส

มดจะมีการเดินทางหรือเปลี่ยนสถานะจากเมืองหนึ่งไปอีกเมืองหนึ่งโดยอาศัยกฎการเปลี่ยนสถานะ กฎนี้ประกอบไปด้วยสองทางเลือกที่เป็นไปได้ นั่นคือ การเลือกสำรวจเพื่อหาผลเฉลยที่เป็นไปได้ใหม่ (Exploration) และการเลือกผลเฉลยโดยอาศัยความรู้เดิมที่มีอยู่ก่อนแล้ว (Exploitation) ให้มด k อยู่ที่เมือง r มันจะเลือกที่จะเปลี่ยนสถานะไปอยู่ที่เมือง s โดย

$$s = \begin{cases} \arg \max_{u \in J_k(r)} \{ [\tau(r, u)] [\eta(r, u)]^\beta \} & \text{if } q \leq q_0 \text{ (exploitation)} \\ S & \text{if } q > q_0 \text{ (exploration)} \end{cases} \quad (2-1)$$

เมื่อ q เป็นตัวเลขสุ่มที่มีการกระจายแบบสม่ำเสมอในช่วงศูนย์ถึงหนึ่ง และ q_0 เป็นความน่าจะเป็นที่จะเลือกจากความรู้เดิมที่มีอยู่ ($0 \leq q_0 \leq 1$) β เป็นเลขยกกำลังสำหรับคำนวณผลคูณระหว่างฟีโรโมนกับความใกล้ชิด (pheromone-closeness product) $\tau(r, u)$ แสดงถึงปริมาณฟีโรโมนบนเส้นทางเดินระหว่างเมือง r และ u พจน์ $\eta(r, u) = 1/\delta(r, u)$ เป็นส่วนกลับของระยะทางระหว่างเมือง r และ u ส่วน $J_k(r)$ เป็นเซตของเมืองที่มด k ยังไม่เคยเดินทางไปมาก่อน โดยปัจจุบันมดอยู่ที่เมือง r และ S คือเมืองที่จะถูกเลือกถ้ามด k ตัดสินใจที่จะทำการสำรวจ ความน่าจะเป็นที่มด k ซึ่งอยู่ที่เมือง r จะเลือกเดินทางไปยังเมือง S ($p_k(r, S)$) ถูกกำหนดโดย

$$p_k(r, S) = \begin{cases} \frac{[\tau(r, S)] [\eta(r, S)]^\beta}{\sum_{u \in J_k(r)} [\tau(r, u)] [\eta(r, u)]^\beta} & \text{if } S \in J_k(r) \\ 0 & \text{otherwise} \end{cases} \quad (2-2)$$

จากสมการที่ (1) ถ้ามด k เลือกที่จะใช้ความรู้เดิมที่ได้รับมาจากมดตัวอื่นๆ มด k จะเลือกเดินทางจากเมือง r ไปเมือง s ซึ่งกำหนดได้จากเส้นทางที่ให้ผลคูณระหว่างฟีโรโมนกับความใกล้ชิดที่มากที่สุดในการที่มด k เลือกที่จะสำรวจเส้นทางใหม่ จากสมการที่ (2) เส้นทางที่มีผลคูณระหว่างฟีโรโมนกับความใกล้ชิดมากเท่าไร ก็มีโอกาที่จะถูกเลือกมากขึ้นเท่านั้น หลังจากที่มีมดทำการเปลี่ยนสถานะ มดจะทิ้งฟีโรโมนจำนวนหนึ่งบนเส้นทางที่เดินผ่าน กฎการเปลี่ยนแปลงระดับฟีโรโมนบนเส้นทางจะได้อธิบายในหัวข้อถัดไป

2.3 กฎการปรับฟีโรโมนเฉพาะที่ในขั้นตอนวิธีเอชีเอส

หลังจากที่มดได้เดินทางจากเมือง r ไปยังเมือง s ระดับฟีโรโมนบนเส้นทางระหว่างเมือง r และ s จะถูกปรับโดยสมการ

$$\tau(r, s) \leftarrow (1 - \rho)\tau(r, s) + \rho\Delta\tau(r, s) \quad (2-3)$$

เมื่อ ρ คือ อัตราการปรับฟีโรโมนเฉพาะที่ (Local Pheromone-Updating Rate) ($0 < \rho < 1$) และ $\Delta\tau(r, s) = \tau_0$ เป็นแฟกเตอร์การปรับฟีโรโมน ซึ่งกำหนดโดย $\tau_0 = (n \cdot L_{\text{opt}})^{-1}$ เมื่อ n คือจำนวนเมืองทั้งหมดในปัญหานี้ และ L_{opt} เป็นความยาวของเส้นทางที่เหมาะสมที่สุด (Optimal Tour Length) ที่กำหนดจากการใช้กฎฮิวริสติกเพื่อนบ้านใกล้ที่สุด เห็นได้ว่าพจน์แรกของสมการทางด้านขวาแสดงถึงปริมาณฟีโรโมนที่มีอยู่บนเส้นทางได้ระเหยไปจำนวนหนึ่ง ส่วนพจน์ที่สองคือปริมาณฟีโรโมนที่มดทิ้งไว้บนเส้นทาง นอกเหนือจากการใช้กฎการปรับฟีโรโมนเฉพาะที่แล้ว ปริมาณฟีโรโมนบนเส้นทางยังสามารถเปลี่ยนแปลงโดยกฎการปรับฟีโรโมนวงกว้าง ซึ่งจะได้อธิบายในหัวข้อถัดไป

2.4 บัญชีคู่แข่ง

จากกฎการเปลี่ยนสถานะในขั้นตอนวิธีเอชีเอสที่กล่าวข้างต้น ในแต่ละครั้งที่มด k ซึ่งอยู่ที่เมือง r ตัดสินใจที่จะเดินทางไปยังเมือง s ซึ่งเมือง s นี้ถูกเลือกจากเซตของเมืองทั้งหมดที่มดยังไม่ได้ไป หรือเซต $J_k(r)$ ในกฎการเปลี่ยนสถานะ เมื่อทำการเลือกเมืองที่อยู่ใกล้กับเมือง r มากที่สุดจำนวน cl เมืองจากเซต $J_k(r)$ แล้วนำเมืองเหล่านี้มาจัดเรียงลำดับ จากเมืองที่ใกล้ที่สุดถึงเมืองที่ไกลที่สุด เซตของเมืองจำนวน cl เมืองที่ได้จัดเรียงลำดับแล้วจะถูกเรียกว่า บัญชีคู่แข่ง ทั้งนี้เมื่อมดได้ตัดสินใจที่จะเดินทางหรือเปลี่ยนสถานะ เมืองที่อยู่ในบัญชีคู่แข่งจะมีระดับความสำคัญสูงกว่าเมืองที่อยู่นอกบัญชี หลังจากที่มีเมืองในบัญชีได้ถูกเลือกหมดแล้ว มดจึงจะสามารถเลือกเมืองที่อยู่นอกบัญชีได้ Dorigo และ Gambardella (1997) ได้อธิบายไว้ว่าการใช้บัญชีคู่แข่งจะช่วยเพิ่มประสิทธิภาพในการค้นหาของขั้นตอนวิธีเอชีเอส

ภาพรวมของขั้นตอนวิธีเอชีเอสสามารถสรุปเป็นรหัสเทียม (Pseudo-Code) ได้ดังภาพที่ 2-2 ซึ่งประกอบด้วยกฎการเปลี่ยนสถานะ กฎการปรับฟีโรโมนและการใช้บัญชีคู่แข่ง (Bonabeau et al., 1999)

```

/* กำหนดค่าเริ่มต้น (Initialisation) */
For ทุกเส้นทางระหว่างเมือง  $r$  และ  $s$  do
     $\tau(r, s) = \tau_0$  โดยที่  $\tau_0 = (n \cdot L_{nn})^{-1}$ 
End For
กำหนดให้  $T^+$  เป็นเส้นทางที่สั้นที่สุดขณะเริ่มต้นที่ได้จากกฎฮิวริสติกเพื่อนบ้านใกล้ที่สุดและ  $L^+$  คือความยาวของเส้นทางนั้น
/* รอบหลัก (Main Loop) */
For  $t = 1$  ถึง  $t_{\max}$  do
    For  $k = 1$  ถึง  $m$  do
        วางมด  $k$  ในเมืองที่สุ่มเลือกมา
    End For
    For  $k = 1$  ถึง  $m$  do
        สร้างเส้นทาง  $T^k(t)$  โดยทำตามขั้นตอนต่อไปนี้เป็นจำนวน  $n-1$  ครั้ง
        If มีเมือง  $s$  อย่างน้อยหนึ่งเมือง  $\in$  บัญชีคู่แข่ง then
            เลือกเมือง  $s$  ถัดไป,  $s \in J_k(r)$ , ในบัญชีคู่แข่งซึ่งมีจำนวน  $cl$  เมืองดังนี้
            
$$s = \begin{cases} \arg \max_{u \in J_k(r)} \{ \tau(r, u) [\eta(r, u)]^\beta \} & \text{if } q \leq q_0; \\ S & \text{if } q > q_0, \end{cases}$$

            โดย  $S \in J_k(r)$  จะถูกเลือกจากความน่าจะเป็น :
            
$$p_k(r, S) = \frac{[\tau(r, S)][\eta(r, S)]^\beta}{\sum_{u \in J_k(r)} [\tau(r, u)][\eta(r, u)]^\beta}$$

            และ  $r$  เป็นเมืองปัจจุบันที่สนใจ
        Else
            เลือกเมือง  $s$  ที่ใกล้ที่สุด  $\in J_k(r)$ 
        End If
        หลังการเปลี่ยนสถานะมด  $k$  จะใช้กฎการปรับฟีโรโมนเฉพาะที่:
         $\tau(r, s) \leftarrow (1 - \rho)\tau(r, s) + \rho\tau_0$ 
    End For /* วนรอบ  $k$  */
    For  $k = 1$  ถึง  $m$  do
        หาความยาว  $L^k(t)$  ของเส้นทาง  $T^k(t)$  ที่มด  $k$  ทำได้
    End For
    If มีเส้นทางที่ดีกว่าเดิม then
        ปรับค่า (update)  $T^+$  และ  $L^+$ 
    End If
    For ทุกเส้นทางระหว่างเมือง  $r$  และ  $s \in T^+$  do
        ปรับปริมาณฟีโรโมนด้วยกฎ:
         $\tau(r, s) \leftarrow (1 - \alpha)\tau(r, s) + \alpha\Delta\tau(r, s)$  โดย  $\Delta\tau(r, s) = 1/L^+$ 
    End For
End For /* วนรอบ  $t$  */
Print เส้นทางที่สั้นที่สุด  $T^+$  และความยาวของมัน  $L^+$ 
Stop

```

ภาพที่ 2-2 สรุปขั้นตอนวิธีระบบอาณานิคมมด

ในภาพที่ 2-2 ขั้นตอนวิธีนี้เริ่มจากกำหนดค่าเริ่มต้นโดยมดแต่ละตัวจะถูกวางในเมืองแบบสุ่มและฟีโรโมนของทุกๆ เส้นทางระหว่างเมืองต่างๆ ถูกกำหนดให้มีระดับเท่ากัน นอกจากนี้ค่าเริ่มต้นของเส้นทางที่ดีที่สุดและความยาวของเส้นทางเดินนั้นจะกำหนดจากกฎฮิวริสติกเพื่อนบ้านใกล้ที่สุด มดแต่ละตัวจะใช้กฎการเปลี่ยนสถานะและกฎการปรับฟีโรโมนเฉพาะที่ หลังจากที่มีมดทุกตัวสิ้นสุดการเดินทางก็จะได้เส้นทางที่ดีที่สุดมา จากนั้นก็ใช้กฎการปรับฟีโรโมนวงกว้าง ถ้าเส้นทางเดินที่ดีที่สุดที่ได้มานี้ดีกว่าเส้นทางเดิมก็จะถูกบันทึกไว้แทนที่ แล้วทำการสุ่มวางมดที่เมืองต่างๆ และทำการวนซ้ำทั้งหมด จนกว่าจะได้เงื่อนไขตามเกณฑ์การหยุด (Termination Criterion)

แม้ว่าที่กล่าวมาข้างต้นจะเป็นการอธิบายการใช้ขั้นตอนวิธีระบบอาณานิคมมดในปัญหาทีเอสพี แต่วิธีการนี้สามารถนำมาประยุกต์ใช้กับปัญหาการประกอบชิ้นส่วนดีเอ็นเอได้อย่างดี โดยค่าซ้อนเหลื่อมซึ่งให้ข้อมูลว่าชิ้นส่วนสองชิ้นจัดวางกันได้ดีเพียงใด จะเทียบเคียงได้กับค่าส่วนกลับของระยะทางระหว่างเมืองสองเมือง ซึ่งรายละเอียดของการนำขั้นตอนวิธีเอซีเอสมาแก้ปัญหการประกอบชิ้นส่วนดีเอ็นเอจะได้นำเสนอในหัวข้อถัดไป

2.5 การประยุกต์ใช้ขั้นตอนวิธีเอซีเอสในการแก้ปัญหการประกอบชิ้นส่วนดีเอ็นเอ

ปัญหาการประกอบชิ้นส่วนดีเอ็นเอสามารถพิจารณาให้เป็นปัญหาการหาค่าเหมาะที่สุดเชิงการจัดซึ่งมีความใกล้เคียงกับปัญหาการเดินทางของพนักงานขายอย่างมาก ในหัวข้อถัดไปจะแสดงการเปรียบเทียบระหว่างสองปัญหานี้ และแสดงให้เห็นว่าจะนำขั้นตอนวิธีเอซีเอสมาใช้แก้ปัญหการประกอบชิ้นส่วนดีเอ็นเอได้อย่างไร

2.5.1 การเปรียบเทียบระหว่างปัญหาการประกอบชิ้นส่วนดีเอ็นเอกับทีเอสพี เมื่อเปรียบเทียบสองปัญหาเข้าด้วยกัน จะเห็นว่าเราสามารถพิจารณาให้ชิ้นส่วนต่างๆ ที่จะนำมาประกอบแทนเมืองในปัญหาทีเอสพี และค่าซ้อนเหลื่อมจะสามารถนำมาใช้แทนค่าส่วนกลับของระยะทางระหว่างเมือง ทั้งนี้เนื่องจากเป้าหมายของทีเอสพี คือการหาเส้นทางเดินแบบวงรอบที่สั้นที่สุดที่สามารถเดินเชื่อมต่อไปทุกๆ เมืองได้ ในขณะที่เป้าหมายของปัญหาการประกอบชิ้นส่วนดีเอ็นเอนั้น ต้องการที่จะหาการจัดวางชิ้นส่วนที่ให้ค่าผลรวมของค่าซ้อนเหลื่อมระหว่างชิ้นส่วนที่วางต่อเนื่องกันในลำดับการจัดเรียงมีค่าสูงที่สุด

อย่างไรก็ตามปัญหาการประกอบชิ้นส่วนดีเอ็นเอ จัดเป็นปัญหาทีเอสพีแบบสมมาตร (Symmetric TSP) ที่มีลักษณะพิเศษทั้งนี้ในปัญหาทีเอสพีแบบสมมาตรจะให้ระยะทางระหว่างเมือง r และ s ในการเดินทางไปและกลับเท่ากัน ปัจจัยสองตัวที่ทำให้ปัญหาการประกอบชิ้นส่วนดีเอ็นเอเป็นปัญหาทีเอสพีแบบสมมาตรที่มีลักษณะพิเศษ ได้แก่ ทิศทางสัมพัทธ์ (Relative Orientation) ของ

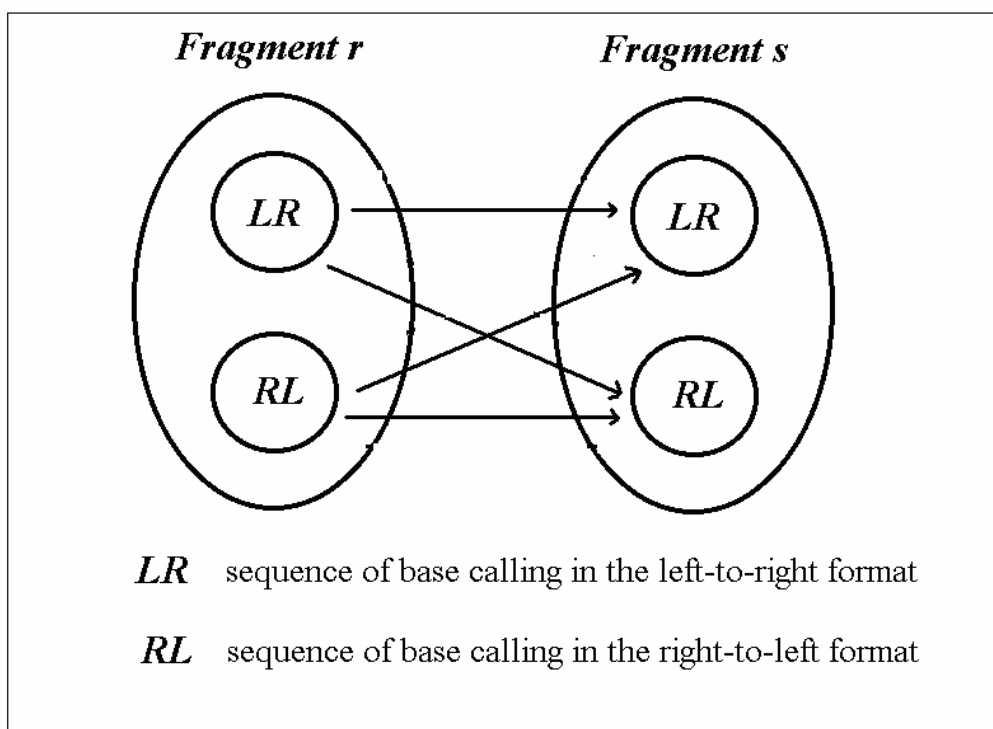
ชิ้นส่วนเมื่อเทียบกับชิ้นส่วนอื่นๆ ในลำดับการเรียงตัว และการพิจารณาถึงสายดีเอ็นเอต้นแบบที่ชิ้นส่วนถูกแบ่งมา ทั้งนี้เมื่อมีการเปลี่ยนแปลงทิศทางสัมพัทธ์ของชิ้นส่วน ลำดับการอ่านเบสบนชิ้นส่วนจะถูกเรียงกลับด้านและทำให้ค่าซ้อนเหลื่อมของชิ้นส่วนนี้เปลี่ยนแปลงตาม และเมื่อพิจารณาถึงที่มาของแต่ละชิ้นส่วน ซึ่งมีความเป็นไปได้เท่าๆกันว่าอาจมาจากสายใดสายหนึ่งในสองสาย ถ้าชิ้นส่วนนั้นมาจากสายต่างกันจะให้ค่าซ้อนเหลื่อมที่ไม่เท่ากัน ด้วยเหตุนี้เมื่อพิจารณาปัจจัยทั้งสองเข้าด้วยกัน ในการหาค่าซ้อนเหลื่อมระหว่างชิ้นส่วนสองชิ้นจะสามารถสร้างแบบการจัดวางชิ้นส่วนที่เป็นไปได้ 8 แบบ ดังแสดงได้ในตารางที่ 2-1

ตารางที่ 2-1 แสดงการจัดวาง 4 รูปแบบที่เป็นไปได้ระหว่างชิ้นส่วนสองชิ้น

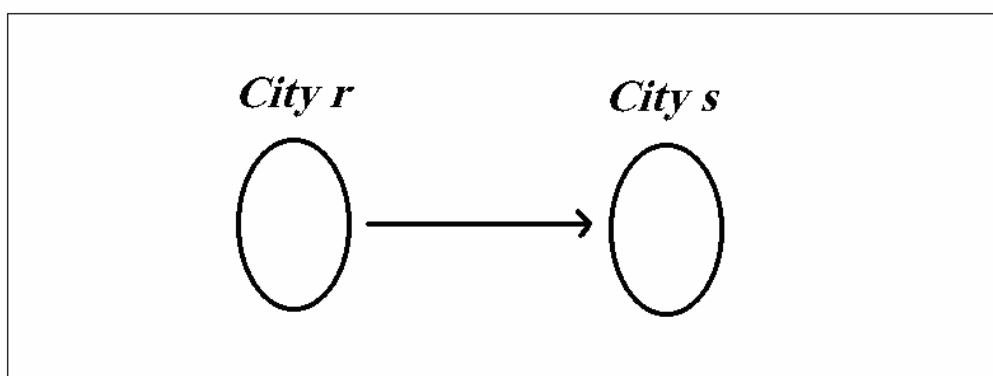
แบบการจัดวาง (Configuration)	ทิศทางสัมพัทธ์ระหว่างชิ้นส่วนสองชิ้น (Relative Orientation between Two Fragments)	
	ลำดับการอ่านเบสบนชิ้นส่วน r (Sequence of Base Calling on r)	ลำดับการอ่านเบสบนชิ้นส่วน s (Sequence of Base Calling on s)
1	ซ้ายไปขวา	ซ้ายไปขวา
2	ซ้ายไปขวา	ขวาไปซ้าย
3	ขวาไปซ้าย	ขวาไปซ้าย
8	ขวาไปซ้าย	ซ้ายไปขวา

2.5.2 การประเมินค่าซ้อนเหลื่อม จากหัวข้อย่อยก่อนหน้านี้แสดงให้เห็นว่ามีการจัดวางตัวที่เป็นไปได้ 8 รูปแบบ สำหรับการจัดวางชิ้นส่วนสองชิ้นเข้าด้วยกัน ค่าซ้อนเหลื่อมในการจัดวางสองชิ้นส่วนนี้ คำนวณจากขั้นตอนวิธีสมิท-วอเตอร์แมน (Smith and Waterman, 1981) ซึ่งวิธีการนี้จะคำนวณการจัดวางที่ดีที่สุดเฉพาะที่ (Locally Best Alignment) ของชิ้นส่วนสองชิ้นโดยใช้เทคนิคไดนามิกโปรแกรมมิ่ง (Dynamic Programming) ในขั้นตอนวิธีสมิท-วอเตอร์แมนต้องมีการกำหนดพารามิเตอร์คือ เมตริกซ์ค่าความเหมือน (Similarity Score Matrix) และค่าการทำโทษเมื่อเกิดช่องว่าง (Gap Penalty) ในบทความนี้เมตริกซ์ค่าความเหมือนและค่าการทำโทษเมื่อเกิดช่องว่างจะกำหนดขึ้นมาโดยอ้างอิงจากค่าที่กำหนดไว้อยู่แล้ว (Default) ในโปรแกรมสำหรับ

DNA Fragment Assembly Problem



Travelling Salesman Problem



ภาพที่ 2-3 แสดงการตัดสินใจที่เป็นไปได้ในการจัดวางชิ้นส่วนสองชิ้น

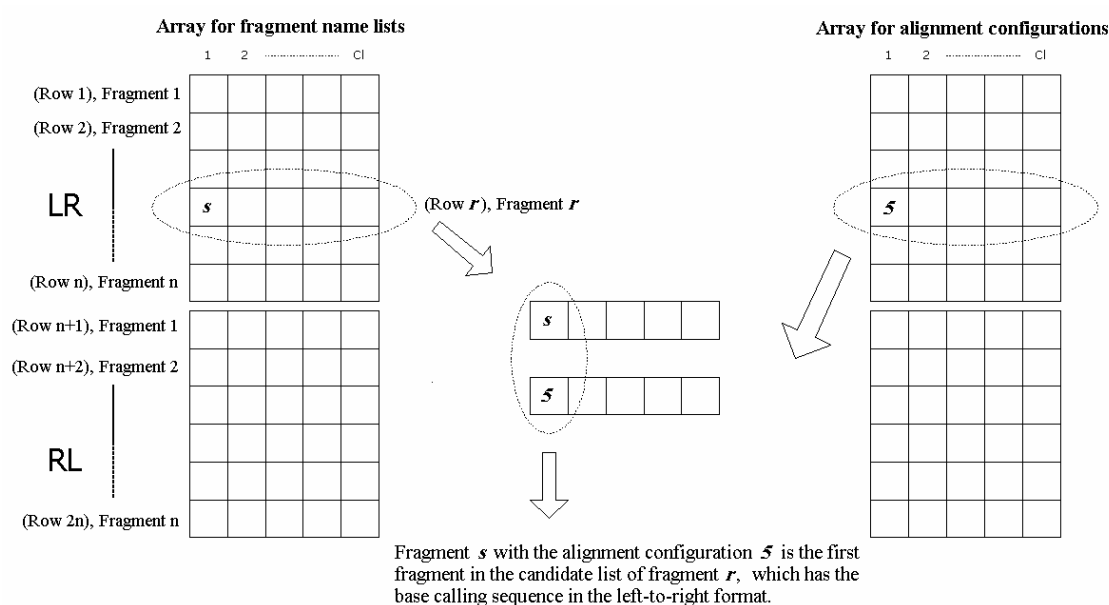
การประกอบพีเอชอาร์เอฟ (PHRAP) (Green, 2004) โดยเราจะให้คะแนน +1 และ -9 สำหรับกรณีที่เบสเอ ซี จี หรือ ที ตรงกัน (Match) และไม่ตรงกัน (Mismatch) ตามลำดับ และให้คะแนน 0 สำหรับกรณีที่เบสอื่น (N) ตรงกันและไม่ตรงกัน ซึ่งเบสอื่นจะเป็นเบสที่ไม่สามารถระบุได้ชัดเจนว่าเป็นเบสใด คะแนน -11 จะให้สำหรับกรณีที่มิช่องว่างเกิดขึ้นใหม่ (โดยจะให้กับส่วนแรกของการเกิดช่องว่าง) และคะแนน -10 จะให้สำหรับการขยายช่องว่างที่มีอยู่เดิมแล้ว ในบทความนี้โปรแกรม

ซิม (SIM.EXE) (Huang and Miller, 1991) ซึ่งเป็นการนำขั้นตอนวิธีสมิท-วอเตอร์แมนมาใช้ โดยมีการพัฒนาด้วยภาษาซีจะถูกนำมาหาค่าซ้อนเหลื่อมระหว่างชิ้นส่วนสองชิ้น โดยการใช้ขั้นตอนวิธีสมิท-วอเตอร์แมน ทิศทางสัมพัทธ์ระหว่างชิ้นส่วน และการพิจารณาสายดีเอ็นเอดั้งเดิมที่ชิ้นส่วนถูกแบ่งออกมา จะทำให้ได้ค่าซ้อนเหลื่อมที่เป็นไปได้ 2 ค่า ดังสรุปไว้ในตารางที่ 2-1 สังเกตว่าในตารางที่ 2-1 เมื่อคำนวณค่าซ้อนเหลื่อมในแบบที่ 1 หรือ 2 ลำดับการอ่านเบสของชิ้นส่วนทั้งสองจะมีรูปแบบแตกต่างกัน (ชิ้นส่วนชิ้นหนึ่งจะเป็นแบบซ้ายไปขวา ขณะที่อีกชิ้นส่วนจะเป็นแบบขวาไปซ้าย) ภายหลังจากการคำนวณค่าซ้อนเหลื่อมในลักษณะนี้ ลำดับการจัดเรียงชิ้นส่วนจะถูกสร้างขึ้นโดยการเพิ่มชิ้นส่วนเข้าไปในทีละชิ้น โดยทำการต่อเข้ากับลำดับการจัดเรียงที่มีอยู่เดิมจนเกิดเป็นลำดับสอดคล้อง โครงร่างของลำดับการจัดเรียงชิ้นส่วนที่ได้จะถูกถ่ายขึ้นบันได ซึ่งเป็นผลมาจากการที่ตำแหน่งสัมพัทธ์ของชิ้นส่วนในการจัดวางตัวถูกกำหนดด้วยค่าซ้อนเหลื่อม

2.5.3 จุดประสงค์ในการค้นหา ในบทความนี้ได้ศึกษาฟังก์ชันจุดประสงค์สองรูปแบบ โดยฟังก์ชันจุดประสงค์แรกจะเป็นการทำให้ผลรวมของค่าซ้อนเหลื่อมระหว่างชิ้นส่วนที่ต่อเนื่องกันในลำดับการจัดเรียงตัวมีค่าสูงที่สุด ซึ่งมีข้อพิสูจน์แล้วว่าลำดับการจัดเรียงชิ้นส่วนที่ให้ผลรวมค่าซ้อนเหลื่อมมากที่สุด สามารถให้ลำดับสอดคล้องที่ใกล้เคียงกับสายดีเอ็นเอต้นแบบ (Parsons et al., 1995; Parsons and Johnson, 1997) ส่วนฟังก์ชันจุดประสงค์ที่สองเป็นการทำให้ทั้งจำนวนคอนทิกและความแตกต่างระหว่างความยาวของคอนทิกที่ยาวที่สุดกับคอนทิกที่สั้นที่สุดมีค่าน้อยที่สุด จากฟังก์ชันจุดประสงค์นี้ผลเฉลยที่ให้จำนวนคอนทิกน้อยกว่าจะถูกพิจารณาให้เป็นผลเฉลยที่ดีกว่า ในการทดลองนี้ตำแหน่งเริ่มและสิ้นสุดของแต่ละคอนทิกในลำดับการจัดเรียงชิ้นส่วน จะเป็นตำแหน่งที่ค่าซ้อนเหลื่อมระหว่างชิ้นส่วนสองชิ้นที่ต่อเนื่องกันมีค่าน้อยกว่าค่าขีดเริ่มเปลี่ยน (Threshold Value) อย่างไรก็ตามในกรณีที่มีผลเฉลยที่ให้จำนวนคอนทิกเท่ากันมากกว่าหนึ่งผลเฉลยนั้น เมื่อพิจารณาผลต่างระหว่างความยาวของคอนทิกที่ยาวที่สุดกับคอนทิกที่สั้นที่สุด ผลเฉลยที่มีค่าผลต่างน้อยที่สุดจะเป็นผลเฉลยที่ดีกว่า ฟังก์ชันจุดประสงค์นี้ได้มาจากการต้องการที่จะได้ผลเฉลยที่เป็นเป้าหมายสูงสุด (Ultimate Goal Solution) ซึ่งเป็นผลเฉลยที่มีเพียงหนึ่งคอนทิกหรือมีจำนวนคอนทิกน้อยที่สุดเท่าที่เป็นไปได้ โดยที่แต่ละคอนทิกมีความยาวพอเหมาะ

2.5.4 บัญชีคู่แข่งสำหรับการประกอบชิ้นส่วนดีเอ็นเอ ในหัวข้อย่อยนี้ได้เสนอแนวคิดของบัญชีคู่แข่ง ซึ่งภายในบัญชีได้บรรจุชิ้นส่วนคู่แข่ง (Candidate Fragment) ที่เป็นไปได้ที่สามารถต่อเข้ากับลำดับการจัดเรียงชิ้นส่วนที่มีอยู่ในระหว่างการใช้กฎการเปลี่ยนสถานะในขั้นตอนวิธีเอชเอส ส่วนกรณีของทีเอสที่บัญชีคู่แข่งจะมีเพียงเมืองต่างๆ แต่ในการประกอบชิ้นส่วนดีเอ็นเอบัญชีคู่แข่งจำเป็นต้องมีข้อมูลเพิ่มเติมเกี่ยวกับแบบของการอ่านเบสของชิ้นส่วนสุดท้ายในลำดับการจัดเรียงชิ้นส่วนปัจจุบัน และข้อมูลเกี่ยวกับรูปแบบการจัดวางตัวที่เป็นไปได้ สมมติว่ามีชิ้นส่วนในบัญชี

คู่แข่งเป็นจำนวน cl ชิ้น สำหรับปัญหาการประกอบชิ้นส่วนดีเอ็นเอที่มี n ชิ้นส่วนจะต้องใช้อาร์เรย์ (Array) ขนาด $2 \times 2n \times cl$ ในการเก็บข้อมูลบัญชีคู่แข่ง โดยเซลล์ขนาด $2n \times cl$ ชุดแรกใช้เก็บบัญชีคู่แข่งของชื่อชิ้นส่วน (Fragment Names) สำหรับชิ้นส่วนทั้งหมด n ชิ้น โดยเก็บลำดับการอ่านเบส (Base Calling Sequence) ทั้งในแบบซ้ายไปขวาและขวาไปซ้าย ส่วนเซลล์ขนาด $2n \times cl$ ชุดที่สองใช้เก็บรูปแบบการจัดวางตัวที่สอดคล้องกับชิ้นส่วนที่อยู่ในเซลล์ขนาด $2n \times cl$ ชุดแรก แผนผังของอาร์เรย์สำหรับเก็บข้อมูลบัญชีคู่แข่งสามารถแสดงได้ดังภาพที่ 2-4 เซลล์ขนาด $2n \times cl$ ชุดแรกจะแบ่งออกเป็นสองส่วน โดยส่วนบนจะใช้เก็บรายการชื่อชิ้นส่วน (List of Fragment Names) สำหรับกรณีที่ชิ้นส่วนสุดท้ายของลำดับการจัดเรียงชิ้นส่วนที่มีอยู่มีลำดับการอ่านเบสแบบซ้ายไปขวา ในขณะที่ส่วนล่างของเซลล์จะใช้เก็บรายการชื่อชิ้นส่วน (List of Fragment Names) สำหรับกรณีที่ชิ้นส่วนสุดท้ายของลำดับการจัดเรียงชิ้นส่วนที่มีอยู่มีลำดับการอ่านเบสแบบขวาไปซ้าย ในเซลล์ขนาด $2n \times cl$ ชุดที่สองก็สามารถแบ่งได้เป็นสองส่วนเช่นเดียวกัน จากตารางที่ 2-1 รูปแบบการวางตัวที่เป็นไปได้ที่จะถูกเก็บไว้ในส่วนบนของเซลล์ขนาด $2n \times cl$ ชุดที่สอง ส่วนรูปแบบการวางตัวที่เป็นไปได้ที่จะถูกเก็บไว้ในส่วนล่างของเซลล์ขนาด $2n \times cl$ ชุดที่สอง



ภาพที่ 2-4 อาร์เรย์ขนาด $2 \times 2n \times cl$ ที่ใช้เก็บข้อมูลบัญชีคู่แข่ง

2.5.5 กฎอภิศตติกเพื่อนบ้านใกล้ที่สุด จากกฎการปรับพีโรโมนเฉพาะที่ในขั้นตอนเอชีเอสที่ได้อธิบายไว้ในหัวข้อย่อย 2.2 แสดงให้เห็นว่า แฟลคเตอร์การปรับพีโรโมนคือส่วนกลับของความยาวของการเดินทางในทีเอสพีซึ่งหาได้จากกฎอภิศตติกเพื่อนบ้านใกล้ที่สุดการใช้กฎอภิศตติกในการหาผลเฉลยของปัญหาการประกอบชิ้นส่วนดีเอ็นเอมีขั้นตอนดังนี้ ขั้นแรกทำการสุ่มเลือก

ชิ้นส่วนหนึ่งชิ้นซึ่งเรียกว่าชิ้นส่วน r จากเซตของชิ้นส่วน โดยรูปแบบการอ่านเบส (ซ้ายไปขวาหรือขวาไปซ้าย) ของชิ้นส่วน r จะถูกสุ่มเลือกด้วยเช่นกัน ชิ้นส่วนคู่แข่ง s ที่สามารถจัดวางเข้ากับชิ้นส่วน r ได้ดีที่สุดจะถูกค้นหาจากชิ้นส่วนที่เหลืออยู่ ซึ่งค่าซ้อนเหลื่อมและรูปแบบการวางตัวที่ได้จะเป็นแบบการอ่านเบสที่เหมาะสมที่สุดสำหรับชิ้นส่วน s จากนั้นกระบวนการหาชิ้นส่วนคู่แข่งของการจัดวางกับชิ้นส่วน s จะถูกดำเนินต่อไปในลักษณะเดียวกันนี้ จนกระทั่งชิ้นส่วนทั้งหมดถูกจัดวางเข้าด้วยกัน จนเป็นลำดับการจัดเรียงชิ้นส่วน ชิ้นส่วนที่ถูกใช้ในลำดับการจัดเรียงชิ้นส่วนแล้ว จะถูกลบออกจากรายการชิ้นส่วนที่เหลือ (Remaining Fragment List) ขั้นตอนสุดท้ายของการประกอบ ค่าซ้อนเหลื่อมระหว่างชิ้นส่วนสุดท้ายกับชิ้นส่วนแรกในลำดับการจัดเรียงตัวจะเป็นค่าที่ได้จากการจัดวางตัวที่กำหนดจากรูปแบบของลำดับการอ่านเบสของชิ้นส่วนทั้งสอง เนื่องจากลำดับสอดคล้องจะไม่มีลักษณะเป็นวงกลม ดังนั้นลำดับการจัดเรียงชิ้นส่วนที่ได้เป็นวงกลมนั้น จะถูกแบ่งที่ตำแหน่งที่ให้ค่าซ้อนเหลื่อมระหว่างสองชิ้นส่วนในลำดับการจัดเรียงที่น้อยที่สุด จากขั้นตอนข้างต้นการเลือกชิ้นส่วนแรกในลำดับการจัดเรียงจะเป็นตัวกำหนดว่าลำดับสอดคล้องจะเป็นอย่างไร ในปัญหาการประกอบชิ้นส่วน n ชิ้น ลำดับสอดคล้องทั้งหมด n ลำดับจะถูกสร้างขึ้นโดยใช้กฎเอ็นเอ็นเอช จะสังเกตได้ว่ารูปแบบของลำดับการอ่านเบสของชิ้นส่วนชิ้นแรกจะไม่ส่งผลต่อจำนวนผลเฉลยที่เป็นไปได้สูงสุดที่สามารถสร้างขึ้นมาได้ ทั้งนี้เนื่องจากการเปลี่ยนรูปแบบของลำดับการอ่านเบสของชิ้นส่วนชิ้นแรกจะเป็นแค่การกลับทิศทางสัมพัทธ์ของทุกชิ้นส่วนในลำดับการจัดเรียง หรือกล่าวอีกนัยหนึ่งว่า ลำดับสอดคล้องที่สร้างขึ้นจะมีลำดับการอ่านเบสที่กลับทิศกับลำดับสอดคล้องที่ได้จากการใช้รูปแบบเดิมในการอ่านลำดับเบสของชิ้นส่วนแรก

2.5.6 กฎการเปลี่ยนสถานะและกฎการปรับฟีโรโมนในขั้นตอนวิธีเอชีเอส ในการใช้ขั้นตอนวิธีเอชีเอสในการประกอบชิ้นส่วนดีเอ็นเอ จำเป็นต้องปรับเปลี่ยนพจน์บางพจน์ที่ใช้ในขั้นตอนวิธีเอชีเอส ได้แก่ พจน์ส่วนกลับของระยะทางและแฟคเตอร์การปรับฟีโรโมน ในกฎการปรับสถานะของขั้นตอนวิธีเอชีเอส การเลือกเปลี่ยนสถานะจะอาศัยผลคูณระหว่างฟีโรโมนกับความใกล้เคียง พจน์ส่วนกลับของระยะทาง ($\eta = 1/\delta$) ในสมการที่ (1) และ (2) จะถูกแทนด้วยค่าซ้อนเหลื่อมระหว่างชิ้นส่วนสองชิ้น เช่นเดียวกับในกฎการปรับฟีโรโมนความยาวของเส้นทางที่ใช้ในแฟคเตอร์การปรับฟีโรโมน ($\Delta \tau$) จะถูกแทนด้วยผลรวมของส่วนกลับของค่าซ้อนเหลื่อมที่ได้จากลำดับการจัดเรียงชิ้นส่วนที่ถูกสร้างขึ้น ในกฎการปรับฟีโรโมนเฉพาะที่ซึ่งถูกแสดงไว้ในสมการที่ (3) พจน์ $\Delta \tau(r, s) = \tau_0 = (n \cdot L_m)^{-1}$ จะถูกเปลี่ยนเป็น $\Delta \tau(r, s) = \tau_0 = (n \cdot IS_m)^{-1}$ โดยที่ IS_m เป็นผลรวมของส่วนกลับของค่าซ้อนเหลื่อมที่ได้จากกฎเอ็นเอ็นเอช และพจน์ $\Delta \tau(r, s) = (L_{gb})^{-1}$ โดย $(r, s) \in$ เส้นทางเดินทางที่ดีที่สุดจากทั้งหมดในกฎการปรับฟีโรโมนวงกว้างที่แสดงในสมการที่ (5) จะเปลี่ยนเป็น $\Delta \tau(r, s) = (IS_{gb})^{-1}$ ซึ่ง $(r, s) \in$ ผลเฉลยที่ดีที่สุดจากทั้งหมด โดยที่ IS_{gb} เป็นผลรวมของ

ส่วนกลับค่าซ้อนเหลื่อมที่ได้จากผลเฉลี่ยที่ดีที่สุดจากทั้งหมด ผลเฉลี่ยที่ดีที่สุดจากทั้งหมดสามารถระบุได้ตามฟังก์ชันจุดประสงค์ชนิดใดชนิดหนึ่งตามที่ได้กล่าวไว้ทั้งนี้ลำดับสอดคล้องสุดท้ายที่ได้จากขั้นตอนวิธีเอชีเอสจะได้จากการแบ่งลำดับการจัดเรียงขึ้นส่วนที่เป็นวงกลม ณ ตำแหน่งที่ให้ค่าซ้อนเหลื่อมระหว่างขึ้นส่วนสองขึ้นน้อยที่สุด ดังที่ได้อธิบายไว้ในหัวข้อการใช้กฎเอ็นเอ็นเอชสำหรับการค้นหา

บทที่ 3

วิธีดำเนินการวิจัย

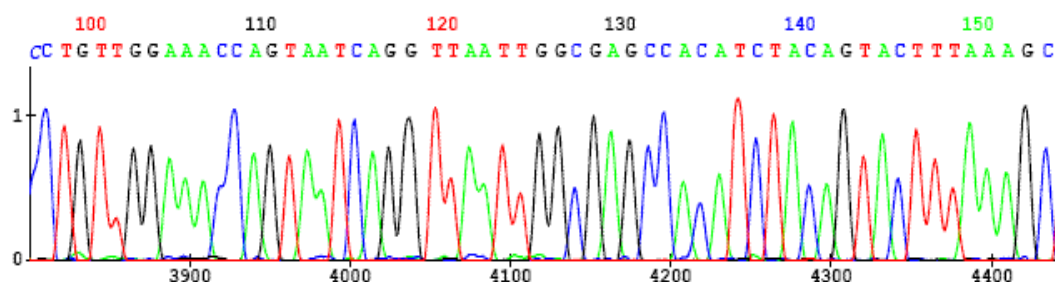
3.1 ขั้นตอนการเตรียมและปรับปรุงข้อมูล

ข้อมูลที่ใช้ในการวิเคราะห์นำมาจากชุดข้อมูลของฐานข้อมูลเจนแบงก์ (Genbank Database) ที่ได้จากศูนย์ข้อมูลเทคโนโลยีชีวภาพแห่งชาติ (The National Center for Biotechnology Information) หรือ NCBI (<http://www.ncbi.nlm.nih.gov>) โดยสายดีเอ็นเอต้นแบบในกรณีนี้มาจากแบคทีเรีย บาซิลลัส แอนทราซิส (Bacillus Anthracis) ที่มีรายละเอียดดังตารางที่ 3-1

ตารางที่ 3-1 รายละเอียดชุดข้อมูล

Contig ID (หมายเลขคอนทิก)	Accession no. (เลขทะเบียนข้อมูล)	Size (bp) (ความยาว - คู่เบส)
842	AAEP01000012	9,936
873	AAEP01000043	19,679
836	AAEP01000006	29,178
851	AAEP01000021	43,268
838	AAEP01000008	51,179

ในชุดฐานข้อมูลนี้จะมีไฟล์ฟาสเอ (FASTA), ข้อมูลเชิงคุณภาพ (Quality score), และ ไฟล์โครมาโตแกรม (Chromatogram File) ที่มีนามสกุล .scf โดยไฟล์ฟาสเอนี้จะประกอบไปด้วยคู่เบส A G C T และช่องว่าง (Gap) และไม่มีเบส N ซึ่งเป็นเบสที่ไม่สามารถระบุได้ชัดเจนว่าเป็นเบสใด อยู่ในข้อมูลไฟล์ฟาสเอ



ภาพที่ 3-1 ลักษณะไฟล์โครมาโตแกรม

หลังจากได้ข้อมูลมาครบทั้งหมดแล้ว จะใช้โปรแกรมพรีแกป (Pregap) เปิดไฟล์โครมาโตแกรมทั้งหมดของแต่ละชุดเพื่อทำการตัดชิ้นส่วนดีเอ็นเอที่มีความผิดพลาด และไฟล์ผลลัพธ์ที่ได้จากโปรแกรมจะเป็นไฟล์นามสกุล EXP ภายในจะแสดงตำแหน่งเริ่มต้น และตำแหน่งสิ้นสุดของกลุ่มเบสที่ใช้ได้ในแต่ละชิ้นส่วนดีเอ็นเอ และใช้โปรแกรมที่เขียนขึ้นมา ทำการสร้างข้อมูลไฟล์ฟาสเอ (Fasta) และค่าข้อมูลเชิงคุณภาพ(Quality Score)ขึ้นมาใหม่ ข้อมูลใหม่นี้จะมีทั้งสายปกติและสายประกอบแบบอ่านจากหลังมาหน้า (Reverse Complementary Strand) โดยการเปลี่ยนข้อมูลจากชิ้นส่วนสายปกติ เป็นชิ้นส่วนสายประกอบแบบอ่านจากหลังมาหน้านี้ จะทำการสลับตัวอักษรคู่เบส A $\Leftrightarrow$ G และ C $\Leftrightarrow$ T และเริ่มอ่านไฟล์จากคู่เบสสุดท้ายสุดมาด้านหน้า และทำการเก็บไฟล์ฟาสเอแบบสายปกติในชื่อแฟ้มข้อมูล data_normal และเก็บไฟล์ฟาสเอในแบบสายประกอบแบบอ่านจากหลังมาหน้าลงในชื่อแฟ้มข้อมูล data_recomp

ชิ้นส่วนสายปกติ

AGTCGTCCAT

ชิ้นส่วนสายประกอบแบบอ่านจากหลังมาหน้า

CGTTCATCAG

ภาพที่ 3-2 แสดงชิ้นส่วนสายปกติ และชิ้นส่วนสายประกอบแบบอ่านจากหลังมาหน้า

3.2 คำนวณค่าซ้อนเหลื่อมของแต่ละข้อมูล

จากหัวข้อบทที่ 2 ก่อนหน้านี้แสดงให้เห็นว่าการหาค่าซ้อนเหลื่อมในการจัดวางสองชิ้นส่วนนี้คำนวณได้มาจากโปรแกรม SIM.EXE ซึ่งใช้ขั้นตอน วิธีสมิท-วอเตอร์แมน (Smith and Waterman, 1981) ในขั้นตอนวิธีสมิท-วอเตอร์แมนต้องมีการกำหนดพารามิเตอร์คือ เมตริกซ์ค่าความเหมือน (Similarity Score Matrix) และค่าการทำโทษเมื่อเกิดช่องว่าง (Gap Penalty) ในการทดลองนี้เมตริกซ์ค่าความเหมือนและค่าการทำโทษเมื่อเกิดช่องว่างจะกำหนดขึ้นมาโดยอ้างอิงจากค่าที่กำหนดไว้อยู่แล้ว (Default) ในโปรแกรมสำหรับการประกอบ ฟิเชอราเอฟี (PHRAP) ตามตารางที่ 3-2

ตารางที่ 3-2 การกำหนดค่าพารามิเตอร์สำหรับการคำนวณหาค่าซ้อนเหลื่อมในไฟล์
ประกอบพีเอชเอเอพี (PHRAP)

พารามิเตอร์ (Parameter)	ค่า (Value)
คู่เบสตรงกัน (match)	+1
คู่เบสไม่ตรงกัน (mismatch)	-9
มีช่องว่างเกิดขึ้นครั้งแรก	-11
มีช่องว่างเพิ่มขึ้นมาอีก	-10

การหาค่าซ้อนเหลื่อมนี้จะคำนวณหาทั้งสองแบบจากไฟล์ฟาสเอในแต่ละชุด คือขึ้นส่วนปกติ กับขึ้นส่วนปกติ และขึ้นส่วนปกติ กับ ขึ้นส่วนประกอบที่อ่านจากหลังมาหน้า ผลลัพธ์ที่ได้จะเป็นลักษณะของเมทริกสมมาตรอย่างละ 2 ไฟล์ต่อข้อมูล 1 ชุด

3.3 การหาลำดับสอดคล้องจากกฎฮิวริสติกเพื่อนบ้าน และขั้นตอนวิธีอาณานิคมมด

หลังจากที่ได้ค่าซ้อนเหลื่อมมาแล้ว ขั้นตอนต่อไปจะทำการหาค่าผลเฉลยที่เป็นไปได้โดยใช้กฎฮิวริสติกเพื่อนบ้านใกล้เคียงที่สุดในที่นี้เป็นโปรแกรม NNH2.EXE ซึ่งผลลัพธ์ที่ได้จะเป็นจำนวนผลเฉลยที่เป็นไปได้ทั้งหมด n ผลเฉลย โดย n เป็นจำนวนชิ้นส่วน ผลเฉลยถูกสร้างขึ้นด้วยจุดประสงค์ที่จะให้ผลรวมของค่าซ้อนเหลื่อมสูงที่สุด ผลเฉลยที่ดีที่สุด 10 อันดับแรกจะถูกเลือกมา ซึ่งจะได้อันดับเริ่มแรกของชิ้นส่วนดีเอ็นเอ 10 ชิ้นส่วนเช่นเดียวกัน เมื่อได้ชิ้นส่วนดีเอ็นเอ 10 ชิ้นส่วนนี้แล้วก็จะเข้าขั้นตอนต่อไปเพื่อหาค่าผลเฉลยที่เป็นไปได้ โดยใช้กฎฮิวริสติกเพื่อนบ้านที่ใกล้เคียงที่สุด โดยในขั้นตอนนี้จะเป็นโปรแกรม NNH ซึ่งผลลัพธ์ที่ได้ จะมีทั้งลำดับขั้นของการจัดเรียงชิ้นส่วนดีเอ็นเอ (Order) และทิศทางของชิ้นส่วนแต่ละลำดับขั้น (Orientation) และค่าซ้อนเหลื่อมของแต่ละลำดับขั้น (Score) โดยการพิจารณาผลลัพธ์ในส่วนนี้ จะพิจารณาว่า ชุดคำตอบใดที่มีผลรวมของค่าซ้อนเหลื่อมมากที่สุด (Maximum Score) ก็จะใช้ชุดคำตอบนั้น เป็นค่าเริ่มต้นเพื่อเข้าสู่ขั้นตอนวิธีอาณานิคมมดต่อไป

ตารางที่ 3-3 การกำหนดค่าพารามิเตอร์สำหรับกฎฮิวริสติกเพื่อนบ้านที่ใกล้ที่สุด

พารามิเตอร์ (Parameter)	ค่า (Value)
กฎฮิวริสติกเพื่อนบ้านใกล้ที่สุด (Nearest Neighbour Heuristic Rule)	
ค่าขีดเริ่มเปลี่ยนของค่าซ้อนเหลื่อมที่ต่ำที่สุดสำหรับนำมาใช้แบ่งลำดับการจัดเรียงออกเป็นหลายคอนทิก (Minimum Overlap Score Threshold for Grouping an Ordering Sequence into Contigs)	95

สำหรับขั้นตอนวิธีอาณานิคมมด จะใช้เพื่อหาอันดับ(Order) ทิศทาง (Orientation) และค่าซ้อนเหลื่อมของ แต่ละลำดับ (Score) โดยการนำผลลัพธ์ที่ได้จากโปรแกรม เอ็นเอ็นเอช มาเป็นค่าเริ่มต้นของโปรแกรมเอชีเอส เมื่อได้ผลลัพธ์จากโปรแกรมเอชีเอส ก็จะเริ่มนำมาใช้ในโปรแกรมที่เขียนขึ้นมา

ตารางที่ 3-4 การกำหนดค่าพารามิเตอร์สำหรับขั้นตอนวิธีระบบอาณานิคมมด

พารามิเตอร์ (Parameter)	ค่า (Value)
ขั้นตอนวิธีระบบอาณานิคมมด (Ant Colony System Algorithm)	
กฎการเปลี่ยนสถานะ (State Transition Rule)	
จำนวนมด (Number of Ants), m	10
จำนวนชิ้นส่วนในบัญชีคู่แข่ง (Number of Fragments in a Candidate List), cl	15
ความน่าจะเป็นในการเลือกใช้ความรู้เดิม (Exploitation Probability), q_0	0.9
เลขยกกำลังสำหรับการคำนวณผลคูณระหว่างฟีโรโมนกับความใกล้เคียง (Exponent Coefficient for the Calculation of a Pheromone-closeness Product), β	2
กฎการปรับฟีโรโมนเฉพาะที่ (Local Pheromone-Updating Rule)	
อัตราการปรับฟีโรโมน (Pheromone-updating Rate Parameter), ρ	0.1
แฟกเตอร์การปรับฟีโรโมน (Pheromone-updating Factor), $\Delta\tau(r, s)$ หรือ τ_0	$(n \cdot IS_{nn})^{-1}$
กฎการปรับฟีโรโมนวงกว้าง (Global Pheromone-Updating Rule)	
อัตราการปรับฟีโรโมน (Pheromone-updating Rate Parameter), α	0.1
จำนวนการวนรอบสูงสุดของรอบหลัก (Maximum Number of Algorithm Repetitions (main loops)), t_{max}	100
ค่าขีดเริ่มเปลี่ยนของค่าซ้อนเหลื่อมที่ต่ำที่สุดสำหรับนำมาใช้แบ่งลำดับการจัดเรียงออกเป็นหลายคอนทิก	95

3.4 ขั้นตอนการประกอบชิ้นส่วนดีเอ็นเอ

ขั้นตอนนี้จะทำการอ่านชื่อไฟล์ข้อมูลฟาสทั้งหมด (Data_list.txt), คะแนนค่าซ้อนเหลื่อมของลำดับชิ้นส่วนดีเอ็นเอ (Score_seq.txt), ลำดับการจัดเรียงของชิ้นส่วนดีเอ็นเอ (Frag_seq) และทิศทางการจัดเรียงของชิ้นส่วนดีเอ็นเอ (Orient_seq.txt) ซึ่งผลเหล่านี้ได้มาจากขั้นตอนวิธีระบบอาณานิคมมด ขั้นตอนการจัดเรียงนี้ จะทำการจัดเรียงทีละคอนทิก โดยเริ่มพิจารณาจากคะแนนค่าซ้อนเหลื่อมในไฟล์ Score_seq.txt ในภาพที่ 3-3 โดยคะแนนของลำดับแรก จะไม่นำมาคิด เพราะเป็นคะแนนของชิ้นส่วนแรกกับชิ้นส่วนสุดท้าย

score_seq.txt	frag_seq.txt	orient_seq.txt
1 9	1 21	1 R
2 155	2 168	2 R
3 801	3 85	3 R
4 867	4 22	4 R
5 781	5 0	5 L
6 867	6 157	6 L
7 594	7 138	7 R
8 801	8 24	8 L
9 571	9 56	9 L
10 853	10 5	10 R
11 674	11 37	11 L
12 875	12 150	12 L
13 192	13 171	13 R
14 803	14 47	14 R
15 793	15 144	15 L
16 812	16 16	16 L
17 861	17 142	17 R
18 551	18 159	18 R
19 749	19 64	19 L
20 910	20 88	20 L
21 447	21 41	21 R
22 853	22 35	22 R
23 774	23 67	23 L
24 753	24 78	24 L
25 669	25 154	25 R
26 669	26 44	26 R
27 884	27 129	27 R
28 232	28 49	28 R
29 13	29 1	29 R
30 725	30 163	30 L

ภาพที่ 3-3 แสดงลักษณะไฟล์ Score_seq.txt และ Frag_seq.txt

3.4.1 กรณีที่คะแนนค่าซ้อนเหลื่อมของชิ้นส่วนดีเอ็นเอมีคะแนนมากกว่า 95 (Threshold Score) จากภาพที่ 3-3 จะพบว่าคะแนนของลำดับที่ 2 จากไฟล์ Score_seq.txt มีคะแนน 155 และไปพิจารณาลำดับที่ 1 และ 2 จากไฟล์ Frag_seq.txt จะพบว่าป็นตัวเลข 21 และ 168 หมายความว่า

ลำดับไฟล์ฟาสเอที่ 22 และ 169 (i+1) เทียบกับลำดับที่และพิจารณาลำดับที่ 1 และ 2 ของไฟล์ orient_seq.txt จะพบว่าเป็นอักษร R ทั้งคู่ หมายความว่าใช้ไฟล์ฟาสเอลำดับที่ 22 และ 169 ในแฟ้มข้อมูล data_recomp และนำไฟล์ฟาสเอลำดับที่ 22 และ 169 มาผ่านโปรแกรม SIM.EXE เพื่อหาว่าชิ้นส่วน 2 ชิ้นนี้ มีการซ้อนเหลื่อมกันที่ตำแหน่งคู่เบสไหน และทำการเก็บค่าที่ได้ลงในอาร์เรย์

3.4.2 กรณีที่คะแนนค่าซ้อนเหลื่อมของชิ้นส่วนดีเอ็นเอมีคะแนนน้อยกว่า 95 จากภาพที่ 3-3 พิจารณาลำดับที่ 29 คะแนนของไฟล์ Score_seq.txt มีคะแนน 13 และพิจารณาลำดับเดียวกันของไฟล์ Frag_seq.txt จะเป็นตัวเลข 1 และลำดับเดียวกันของไฟล์ Orient_seq.txt เป็น R หมายความว่า ลำดับไฟล์ฟาสเอที่ 1 จากแฟ้มข้อมูล data_recomp และลำดับไฟล์ฟาสเอที่ 50 จากแฟ้มข้อมูล data_recomp ไม่ซ้อนเหลื่อมกัน และถือว่าให้เริ่มเป็นคอนทิกใหม่

3.4.3 การจัดเรียงชิ้นส่วนดีเอ็นเอจากผลของโปรแกรม SIM.EXE ที่เรานำไปใช้คือส่วนที่แสดงถึงจุดเริ่มต้นคู่เบสของการซ้อนเหลื่อม จากไฟล์ฟาสเอแต่ละคู่ และทำการจัดเรียงชิ้นส่วนดีเอ็นเอโดยผลของการซ้อนเหลื่อมนี้ โดยจะยกตัวอย่างจากลำดับไฟล์ฟาสเอ ตามภาพที่ 3-3

Upper Sequence: ลำดับที่ 22	AGTCTACGTTCAAGTGCTT
Lower Sequence: ลำดับที่ 169	ACGACGTT - AAGTGCTACGC

ภาพที่ 3-4 การจัดเรียงชิ้นส่วนดีเอ็นเอคู่แรก

ไฟล์ฟาสเอลำดับที่ 22 และ ลำดับที่ 169 เมื่อทำการหาค่าซ้อนเหลื่อมและทำการจัดเรียงชิ้นส่วนให้ตรงตามผล จะได้ตามภาพที่ 8 และจัดเก็บลงอาร์เรย์ผลลัพธ์ และนำไฟล์ลำดับถัดไปมาหาค่าซ้อนเหลื่อมต่อ ก็คือไฟล์ลำดับที่ 169 และลำดับที่ 86

Upper Sequence: ลำดับที่ 169	ACGACGTT - AAGTGCTACGC
Lower Sequence: ลำดับที่ 86	ACGTTCAAGTGCT - CGCTAC

ภาพที่ 3-5 การจัดเรียงชิ้นส่วนดีเอ็นเอคู่ 2

ชิ้นส่วนลำดับที่ 22	AGTCTACGTTCAAGTGCTT
ชิ้นส่วนลำดับที่ 169	ACGACGTT - AAGTGCTTTAGC
ชิ้นส่วนลำดับที่ 86	ACGTTCAAGTGCT - CGCTAC

ภาพที่ 3-6 การจัดเรียงชิ้นส่วนดีเอ็นเอภายในอาร์เรย์ผลลัพธ์

จากภาพที่ 3-5 จะเป็นการจัดเรียงชิ้นส่วนดีเอ็นเอคู่ที่ 2 โดยสิ่งที่ต้องพิจารณาก็คือลำดับด้านบน (Upper Sequence) ของคู่ 2 นี้ จะเป็นไฟล์เดียวกันกับลำดับสุดท้าย ภายในอาร์เรย์ผลลัพธ์ ซึ่งหากลำดับด้านบนของภาพที่ 3-5 มีการเปลี่ยนแปลงไปจากลำดับสุดท้ายภายในอาร์เรย์ผลลัพธ์ของภาพที่ 3-4 คือมีช่องว่างเพิ่มขึ้น จะทำให้การจัดเรียงไฟล์ในอาร์เรย์ผลลัพธ์ มีการเปลี่ยนแปลง แต่ในกรณีภาพที่ 3-5 นี้ไฟล์ฟาสต์ลำดับด้านบนไม่มีการเปลี่ยนแปลงเมื่อเทียบกับภายในอาร์เรย์ผลลัพธ์ ก็ให้ทำการใส่ผลลัพธ์และจัดเรียงกันตามปกติภายในอาร์เรย์ผลลัพธ์ และเริ่มทำการหาค่าซ้อนเหลื่อมต่อไป

Upper Sequence: ลำดับที่ 86 ACGTTCAAG - TGCT - TAGCTAC
Lower Sequence: ลำดับที่ 23 CTTCAAGATGCTTTAG - TACCG

ภาพที่ 3-7 การจัดเรียงชิ้นส่วนดีเอ็นเอคู่ 3

จากภาพที่ 3-7 จะพบว่าลำดับด้านบนของภาพนี้ มีการเปลี่ยนแปลงเมื่อเทียบกับชิ้นส่วนสุดท้าย ภายในอาร์เรย์ผลลัพธ์จากภาพที่ 3-6 คือมีการเปิดช่องว่างเพิ่มขึ้นจากเดิม ทำให้เมื่อประกอบชิ้นส่วนลงภายในอาร์เรย์ผลลัพธ์ จะต้องมีการเปิดช่องว่าง ณ ตำแหน่งที่เพิ่มเข้ามาใหม่ในแต่ละหลัก (Column) ของเบสในทุกลำดับก่อนนี้ จะได้ดังภาพที่ 3-8

ชิ้นส่วนลำดับที่ 22 AGTCTACGTTCAAG - TGCTT
ชิ้นส่วนลำดับที่ 169 ACGACGTT - AAG - TGCTTTAGC
ชิ้นส่วนลำดับที่ 86 ACGTTCAAG - TGCT - CGCTAC
ชิ้นส่วนลำดับที่ 23 CTTCAAGATGCTTTAG - TACCG

ภาพที่ 3-8 การจัดเรียงชิ้นส่วนดีเอ็นเอในอาร์เรย์ผลลัพธ์

ให้ทำการประกอบชิ้นส่วนดีเอ็นเอ จนหมดชุดคอนทิก ตามคะแนนค่าซ้อนเหลื่อมจากไฟล์ score_seq.txt แล้วผลลัพธ์ทั้งหมดจะอยู่ในอาร์เรย์ผลลัพธ์ ดังภาพที่ 3-9 หลังจากนั้นจะมาหาคอนทิกจากอาร์เรย์ผลลัพธ์ใช้กฎการเลือกตามเสียงข้างมาก (Majority Vote Rule) โดยให้พิจารณาในแต่ละหลัก (Column) ของเบสในชิ้นส่วนที่นำมาจัดเรียง ว่ามีอักษรไหนมากที่สุด (A,G,T,C, -) และหากหลักใดมีอักษรที่เท่ากัน จะใช้การสุ่มเลือกอักษรที่มีจำนวนเท่ากันออกมา และจากภาพที่ 3-9 เมื่อใช้กฎการเลือกตามเสียงข้างมาก จะได้ดังภาพที่ 3-10

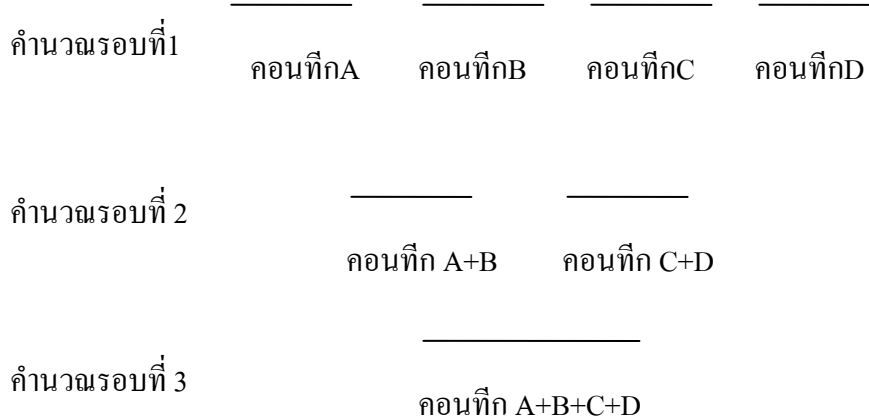
ชิ้นส่วนลำดับที่ 22	AGTCTACGTTCAAG - TGCTT
ชิ้นส่วนลำดับที่ 169	ACGACGTT - AAG - TGCTTTAGC
ชิ้นส่วนลำดับที่ 86	ACGTTCAAG - TGCT - CGCTAC
ชิ้นส่วนลำดับที่ 23	CTTCAAGATGCTTTAG - TACCG
ชิ้นส่วนลำดับที่ 1	ACAAGATGCTTTAGCTACATT
ชิ้นส่วนลำดับที่ 158	AAGATGCTTTAGCTACATAG
ชิ้นส่วนลำดับที่ 139	GCAAGATGCTTTAGCTACATAG

ภาพที่ 3-9 การจัดเรียงชิ้นส่วนดีเอ็นเอที่เหลือในอาร์เรย์ผลลัพธ์

AGTCTACGTTCAAGATGCTTTAGCTACCTAG

ภาพที่ 3-10 แสดงคอนทิกที่ได้จากการใช้กฎการเลือกตามเสียงข้างมาก

เมื่อได้ผลลัพธ์ของคอนทิกนี้แล้วให้ทำตามขั้นตอนที่ 3.4 ต่อไปจนหมดลำดับไฟล์ตามภาพที่ 3-3 ซึ่งจะได้จำนวนคอนทิกมาจำนวนหนึ่ง แต่ในการทดลองนี้สามารถนำค่าคอนทิกที่ได้กลับไปทดลองซ้ำตามขั้นตอนที่ 3.2 โดยจะพิจารณาคอนทิกแต่ละเส้น เป็นเส้นดีเอ็นเอย่อย (Fragment) เพื่อทำการประกอบชิ้นส่วนดีเอ็นเออีกครั้งดังภาพที่ 3-11 เพื่อให้ได้ค่าที่ใกล้เคียงกับเส้นดีเอ็นเอต้นแบบได้มากที่สุด



ภาพที่ 3-11 การนำผลลัพธ์คอนทิกที่ได้กลับมามีจำนวนซ้ำอีกครั้ง

ดีเอ็นเอเส้นหลัก (Coverage) โดยการหาค่าความผิดพลาดนี้จะทำการหาทุกจำนวนรอบที่มีการทดลองซ้ำเช่นกัน

Number : Local Alignment
 Similarity Score : 83
 Match Percentage : 97%
 Number of Matches : 122
 Number of Mismatches : 2
 Total Length of Gaps : 1
 Begins at (4825,1) and Ends at (4948,125)

ภาพที่ 3-13 ตัวอย่างจากไฟล์ผลลัพธ์ที่ได้จากโปรแกรม SIM.EXE

จากภาพที่ 3-13 ในตำแหน่งบรรทัดสุดท้ายหมายความว่าคอนทิกที่ได้จากการทดลองนี้ มีการครอบคลุมตั้งแต่ตำแหน่งที่ 4825 ถึง 4948 ของสายดีเอ็นเอหลัก โดยตัวมันเองครอบคลุมตั้งแต่ตำแหน่งที่ 1 ถึง 125 และให้ทำการเก็บค่าความผิดพลาดนี้ทุกคอนทิกที่มี และทุกจำนวนที่ทำการทดลอง เพื่อนำมาวิเคราะห์และสรุปหาสาเหตุของผลที่ได้ในบทต่อไป

บทที่ 4

ผลการทดลอง

ขั้นตอนการหาประกอบชิ้นส่วนดีเอ็นเอ สิ่งที่ต้องรู้เป็นอันดับแรก ก็คือลำดับของชิ้นส่วน ซึ่งการทดลองนี้ ได้ใช้กฎฮิวริสติกเพื่อนบ้านใกล้ที่สุด และวิธีระบบอาณานิคมมด (เอชีเอส) มาใช้หาลำดับชิ้นส่วนร่วมกัน และนำมาประกอบเข้าด้วยกัน เพื่อนำมาเปรียบเทียบกับสายต้นแบบ กับโปรแกรมแคปทรี ในกรณีของกฎฮิวริสติกเพื่อนบ้านใกล้ที่สุดนั้น ได้พยายามทำการประกอบคอนทิกของดีเอ็นเอ โดยค่าซ้อนเหลื่อมระหว่างคอนทิกแต่ละคอนทิกสามารถหาได้จากขั้นตอนวิธีสมิท-วอเตอร์แมน จำนวนผลเฉลยที่เป็นไปได้ทั้งหมด n ผลเฉลยจะถูกสร้างขึ้น โดย n เป็นจำนวนชิ้นส่วน ผลเฉลยถูกสร้างขึ้นด้วยจุดประสงค์ที่จะให้ผลรวมของค่าซ้อนเหลื่อมสูงที่สุด ผลเฉลยที่ดีที่สุดจะถูกเลือกมา จากนั้นจะใช้กฎเอ็นเอ็นเอชในขั้นตอนต่อไป เปรียบเทียบกับกระบวนการประกอบก่อนหน้านี้ ผลเฉลยที่เป็นไปได้ทั้งหมด m ผลเฉลยจะถูกสร้างขึ้นโดย m เป็นจำนวนของคอนทิกที่ได้จากการประกอบในขั้นแรก และผลเฉลยที่ดีที่สุดจะถูกเลือกให้เป็นผลเฉลยสุดท้าย ภายหลังจากการทดลองซ้ำในเอชีเอสสิ้นสุดลง ผลเฉลยที่ดีที่สุดจะถูกเลือกและคอนทิกจะได้รับการประกอบชิ้นส่วนเข้าด้วยกัน จากนั้นการประกอบคอนทิกจะเริ่มค้นหาลำดับขึ้น โดยใช้โปรแกรมและวิธีการตามหัวข้อย่อยบทที่ 3.4

จากผลการประกอบเราได้ทำการวิเคราะห์ใน 2 เรื่อง ซึ่งได้แก่ การเปรียบเทียบประสิทธิภาพระหว่างการใช้กฎเอ็นเอ็นเอชร่วมกับขั้นตอนวิธีเอชีเอสในขั้นตอนแรกของการประกอบจำนวนคอนทิกในผลเฉลยการประกอบสุดท้าย คุณภาพของผลเฉลยสุดท้าย

4.1 จำนวนคอนทิกจากวิธีเอ็นเอ็นเอชกับวิธีเอชีเอส เทียบกับโปรแกรมแคปทรี

จากตารางที่ 7 จะพบว่าการใช้วิธีเอ็นเอ็นเอชกับวิธีเอชีเอสร่วมกัน จำนวนคอนทิกที่หาได้จะมีจำนวนที่สูง ผลลัพธ์ยังได้แสดงให้เห็นว่าเมื่อขนาดของปัญหาใหญ่ขึ้น จำนวนคอนทิกที่ได้จากวิธีเอชีเอสกับเอ็นเอ็นเอชจะเพิ่มขึ้นด้วย แต่เมื่อนำผลคอนทิกที่ได้จากรอบแรกเข้าสู่กระบวนการและทำการทดลองซ้ำอีกครั้ง จำนวน คอนทิกที่ได้จะลดลงไป ทั้งนี้เพราะชิ้นส่วนคอนทิกบางชิ้นส่วน สามารถแทรกหรือซ้อนเข้าไปในคอนทิกชิ้นที่มีขนาดใหญ่กว่าได้ นอกจากนี้จะเห็นว่าขนาดของปัญหาไม่ได้สัมพันธ์กับจำนวนคอนทิกสำหรับกรณีที่ใช้โปรแกรมแคปทรีจาก

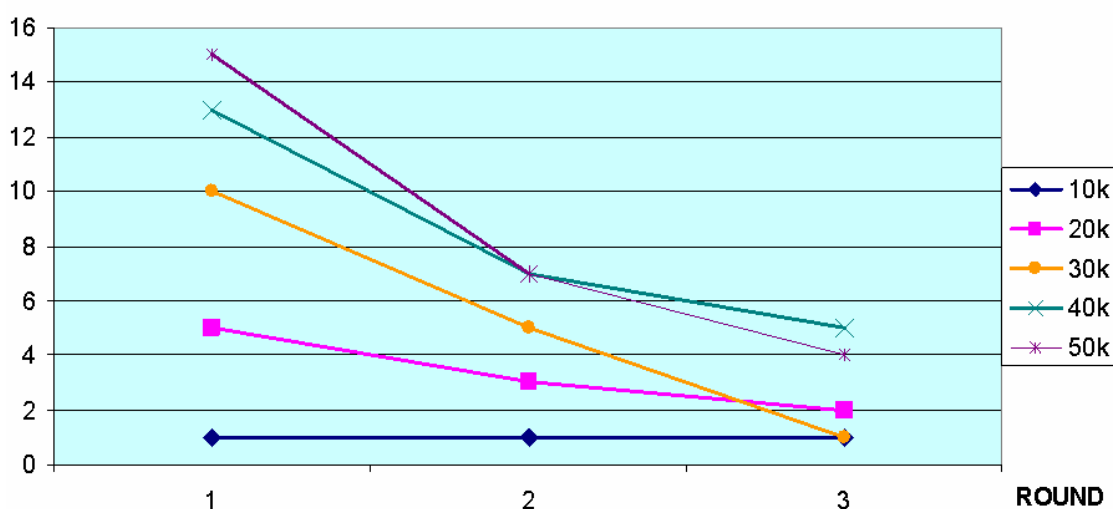
การสังเกตข้างต้นสามารถกล่าวได้ว่าประสิทธิภาพโดยรวมของวิธีเอซีเอสกับเอ็นเอ็นเอชจะใกล้เคียงกับโปรแกรมแคปทรี เมื่อทำการทดลองซ้ำมากกว่า 1 ครั้ง

ตารางที่ 4-1 จำนวนคอนทิกจากผลเฉลยที่ได้จากวิธีเอ็นเอ็นเอชกับวิธีเอซีเอส และจากโปรแกรมแคปทรี

CONTIC ID (เลขทะเบียนข้อมูล)	จำนวนคอนทิก (Number of Contigs)			
	เอ็นเอ็นเอชกับ เอซีเอส รอบที่ 1 (NNH+ACS)	เอ็นเอ็นเอชกับ เอซีเอส รอบที่ 2 (NNH+ACS)	เอ็นเอ็นเอชกับ เอซีเอส รอบที่ 3 (NNH+ACS)	แคปทรี (CAP3)
AAEP01000012 (10K เบส)	1	1	1	1
AAEP01000043 (20Kเบส)	5	3	2	2
AAEP01000006 (30Kเบส)	10	5	1	1
AAEP01000021 (40Kเบส)	13	6	4	1
AAEP01000008 (50Kเบส)	15	7	3	1

จากตารางที่ 4-1 แสดงให้เห็นว่าจำนวนคอนทิกที่ได้จากเอ็นเอ็นเอชและเอซีเอสในรอบแรกของข้อมูล AAEP01000012 จะได้เท่ากับผลที่ได้จากแคปทรีและข้อมูล AAEP01000043 เมื่อดูผลคอนทิกในรอบที่ 3 จะพบว่ามีค่าเท่ากับผลที่ได้จากแคปทรีเช่นกัน และในส่วนของผลลัพธ์คอนทิกของข้อมูลชุดอื่นๆ จะพบว่ายิ่งทำการคำนวณหลายรอบจำนวนคอนทิกจะลดลงมากตามจำนวนรอบ ซึ่งเป็นการพิสูจน์ว่าการทดลองนี้สามารถนำผลคอนทิกกลับไปคำนวณอีกครั้งได้ และแนวโน้มของการลดลงของจำนวนคอนทิกตามจำนวนรอบ พิจารณาได้ตามภาพที่ 4-1

CONTIG



ภาพที่ 4-1 แสดงจำนวนคอนทิกของแต่ละข้อมูลตามจำนวนรอบของการทดลอง

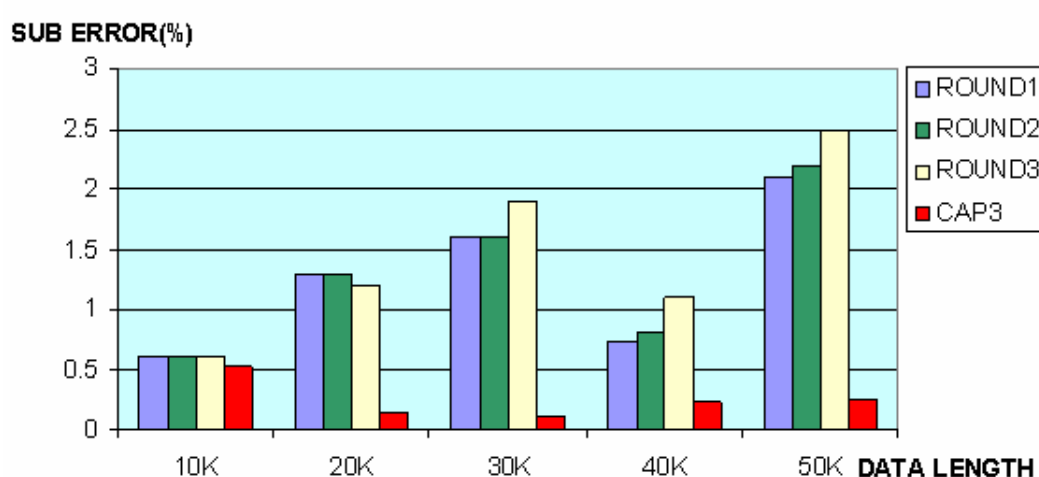
4.2 คุณภาพของผลเฉลยจากการประกอบชิ้นส่วนดีเอ็นเอ

หัวข้อนี้จะได้ศึกษาคุณภาพของคอนทิกเหล่านี้ คุณภาพของคอนทิกจะถูกวัดจากความแตกต่างของเบสระหว่างลำดับดีเอ็นเอต้นแบบและคอนทิกที่สนใจ โดยอาศัยขั้นตอนวิธี สมิท-วอเตอร์แมนพร้อมการกำหนดค่าดังที่ได้อธิบายในหัวข้อ 2.2 ความแตกต่างนี้สามารถอธิบายด้วยความผิดพลาดในการประกอบ ซึ่งสามารถแบ่งออกเป็นสามส่วนด้วยกันคือ ความผิดพลาดจากการแทนที่ (Substitution Error) ความผิดพลาดจากการใส่เพิ่มและดึงออก (Insertion/Deletion (Indel) Error) และความผิดพลาดของการครอบคลุม (Coverage Error) โดยที่ความผิดพลาดจากการแทนที่ที่เกิดขึ้นเมื่อเบสของหนึ่งในสองลำดับที่นำมาจัดเรียง กล่าวคือเป็นลำดับดีเอ็นเอต้นแบบหรือคอนทิกที่สนใจ ไม่ตรงกับเบสที่ตำแหน่งเดียวกันของอีกลำดับหนึ่ง เมื่อเบสในลำดับที่นำมาจัดเรียงเสมือนว่ามีเบสถูกลบออกไปเมื่อเปรียบเทียบกับเบสของอีกลำดับที่นำมาเรียงเข้ากัน การหายไปของเบสในลำดับนั้นจะเรียกว่าเกิดความผิดพลาดจากการดึงออก อีกนัยหนึ่ง เมื่อเบสเสมือนว่าถูกใส่เพิ่มเข้าไปจนเป็นลำดับที่ยาวขึ้น ลำดับที่ถูกเสริมเข้าไปนั้นจะมีความผิดพลาดจากการใส่เพิ่มเกิดขึ้น การดึงออกจากลำดับหนึ่งสามารถที่จะพิจารณาให้เป็นการใส่เพิ่มในอีกลำดับหนึ่ง ด้วยเหตุนี้ความผิดพลาดทั้งสองกรณีโดยทั่วไปจะสามารถพิจารณาเป็นคู่ควบกัน โดยจะเรียกเป็นความผิดพลาดจากการใส่เพิ่มและดึงออกหรือความผิดพลาดแบบอินเดล และสำหรับความผิดพลาดของการครอบคลุมจะเกิดขึ้นเมื่อมีเบสในลำดับดีเอ็นเอต้นแบบที่อยู่ภายนอกคอนทิกที่ไม่ให้ความสอดคล้องกับลำดับดีเอ็นเอต้นแบบนั้นที่สุดและไม่ได้ถูกครอบคลุมด้วยคอนทิกใด

ตารางที่ 4-2 ค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกเทียบกับชุดข้อมูลกับจำนวนรอบในการทดลองและโปรแกรมแคปทรี

ความยาว คอนทิก	ความผิดพลาดของการแทนที่กับการใส่เพิ่ม และดึงออก (%) (Substitution & Indel Errors (%))			
	เอ็นเอ็นเอชกับเอชีเอส รอบที่ 1 (NNH+ACS)	เอ็นเอ็นเอชกับเอชีเอส รอบที่ 2 (NNH+ACS)	เอ็นเอ็นเอชกับเอชีเอส รอบที่ 3 (NNH+ACS)	แคปทรี Cap3
10Kเบส	0.6	0.6	0.6	0.53
20Kเบส	1.3	1.3	1.2	0.14
30Kเบส	1.6	1.6	1.9	0.12
40Kเบส	0.74	0.8	1.1	0.23
50Kเบส	2.1	2.2	2.5	0.25

จากตารางที่ 4-2 จะพบว่าค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกของผลเฉลี่ยที่ได้จากการหาลำดับด้วยวิธีการเอ็นเอ็นเอช และ เอชีเอส มีจำนวนสูงกว่าผลที่ได้จากโปรแกรมแคปทรี โดยยิ่งข้อมูลมีความยาวของกลุ่มเบสที่มาก ก็ยิ่งมีความผิดพลาดมากขึ้นตามไป แต่ในทางกลับกันกับโปรแกรมแคปทรี ที่ค่าความผิดพลาดแทบไม่แตกต่างกัน แม้ว่าความยาวของกลุ่มเบสจะมีมากขึ้นก็ตาม



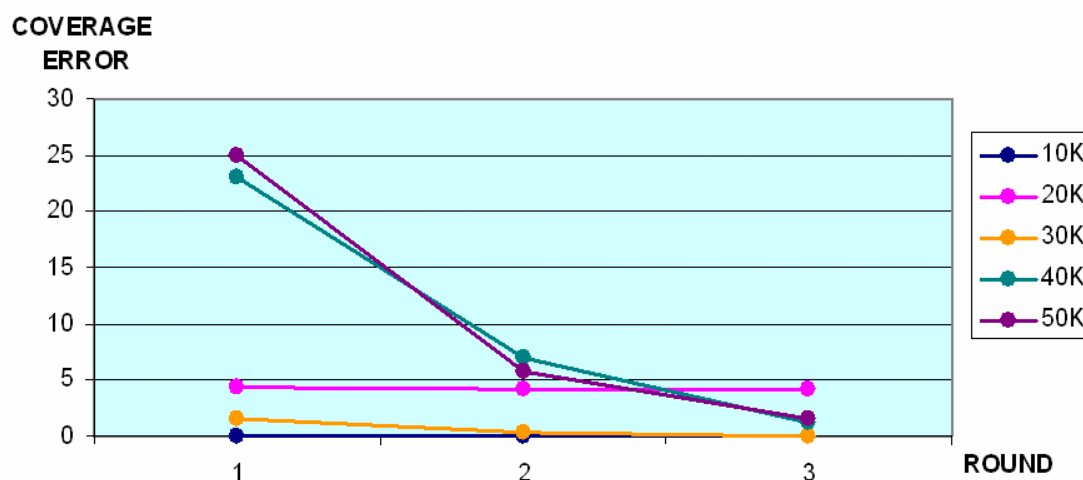
ภาพที่ 4-2 ตารางเปรียบเทียบค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกต่อจำนวนรอบ

จากภาพที่ 4-2 แสดงให้เห็นว่าค่าความผิดพลาดของการแทนที่กับการใส่เพิ่มและดึงออกจะมีแนวโน้มมากขึ้นตามจำนวนรอบที่ทำการทดลอง สาเหตุส่วนหนึ่งเป็นเพราะวิธีการคัดเลือกกลุ่มเบสในขั้นตอนประกอบชิ้นส่วนดีเอ็นเอ ในการทดลองนี้ใช้วิธีคัดเลือกตามส่วนมาก (Majority Vote Rule) และทำการสุ่มอักษรหากมีจำนวนอักษรที่เท่ากัน ทำให้มีกลุ่มเบสที่ผิดไปตั้งแต่รอบแรก และยังทำการทดลองซ้ำในรอบต่อไป โอกาสที่จะผิดก็มีมากขึ้น

ตารางที่ 4-3 ค่าความผิดพลาดในการครอบคลุมเทียบกับชุดข้อมูลกับจำนวนรอบในการทดลอง และโปรแกรมแคปทรี

ความยาว กลุ่มเบส	ค่าความผิดพลาดในการครอบคลุม (%) (Coverage Error (%))			
	เอ็นเอ็นเอชกับเอซีเอส รอบที่ 1 (NNH+ACS)	เอ็นเอ็นเอชกับเอซีเอส รอบที่ 2 (NNH+ACS)	เอ็นเอ็นเอชกับเอซีเอส รอบที่ 3 (NNH+ACS)	แคปทรี Cap3
10Kเบส	0	0	0	0.2
20Kเบส	4.3	4.2	4.1	4.5
30Kเบส	1.6	0.3	0	0.12
40Kเบส	23	7	1.3	0.1
50Kเบส	25	5.7	1.6	0.2

สำหรับค่าความผิดพลาดในการครอบคลุม ข้อมูลที่ 10K เบสเมื่อทำการทดลองรอบเดียวก็พบว่าครอบคลุมได้หมดเมื่อเทียบกับเส้นดีเอ็นเอต้นแบบ และเมื่อเทียบกับแคปทรีจะพบว่า ผลที่ได้จากโปรแกรมเอ็นเอ็นเอชและเอซีเอสให้ค่าได้ดีกว่าและสำหรับข้อมูลที่มีขนาดความยาวกลุ่มเบส 20K ที่ไม่มีการครอบคลุมครบหมด แต่เมื่อพิจารณารูปแบบตามเว็บไซต์ต้นฉบับ จะพบว่าข้อมูลต้นฉบับก็มีการครอบคลุมที่ไม่ครบเช่นกัน ส่วนข้อมูลชุดอื่นๆนั้น ผลลัพธ์ที่ได้จากหาลำดับด้วยวิธีการเอ็นเอ็นเอชและเอซีเอส จะดีกว่าผลลัพธ์จากโปรแกรมแคปทรี แม้ว่าขนาดความยาวของกลุ่มเบสจะเพิ่มขึ้นด้วยก็ตาม โดยข้อมูลที่มีขนาดความยาวกลุ่มเบส 10K ถึง 30K จะมีการครอบคลุมกับเส้นดีเอ็นเอต้นแบบที่ดีกว่าโปรแกรมแคปทรี ส่วนข้อมูลที่มีความยาวกลุ่มเบส 40K ถึง 50K จะพบว่าค่าครอบคลุมที่ได้มีค่ามาก ถ้าเทียบกับค่าจากโปรแกรมแคปทรี แต่มีแนวโน้มลดลงมากเมื่อทำการทดลองในรอบที่มากขึ้น



ภาพที่ 4-3 กราฟเปรียบเทียบค่าความผิดพลาดในการครอบคลุมต่อจำนวนรอบ

จากภาพที่ 4-3 แสดงให้เห็นความสัมพันธ์ของจำนวนรอบและค่าความผิดพลาดในการครอบคลุม โดยผลลัพธ์จากกราฟบอกว่าจำนวนรอบยิ่งมากขึ้น ค่าความผิดพลาดของการครอบคลุมก็ยิ่งน้อยลงเรื่อยๆ เมื่อพิจารณา รอบแรกและรอบที่ 2 จะพบว่าค่าที่ได้นั้น มีการลดลงเกินครึ่งหนึ่งจากค่าเดิม สาเหตุส่วนหนึ่งที่ทำให้ค่าความผิดพลาดในการครอบคลุมดีขึ้น เมื่อทำการทดลองซ้ำ มาจากการใช้กฎเอ็นเอ็นเอชและกฎเอชีเอส เข้ามาช่วย ทำให้มีการจัดเรียงลำดับและทิศทางได้ดีกว่าค่าที่ได้จากโปรแกรมแคปทรี

4.3 วิเคราะห์และสรุปผลการทดลอง

จากผลการทดลองในวิทยานิพนธ์นี้ ผลลัพธ์ที่สนใจคือจำนวนคอนทิกที่ได้จากการหาลำดับและการประกอบชิ้นส่วนดีเอ็นเอ มาเปรียบเทียบกับโปรแกรมแคปทรี โดยผลลัพธ์ที่ได้จะพบว่า สำหรับข้อมูลชุดที่มีความยาวไม่เกิน 30K จะให้จำนวนที่เท่ากับกับโปรแกรมแคปทรี โดยทำการทดลองซ้ำไม่เกิน 3 รอบ แต่สำหรับข้อมูลชุดที่มีความยาวเกิน 40K จะพบว่ามีความแตกต่างกันมากกับผลลัพธ์ที่ได้จากโปรแกรมแคปทรี แต่ก็สามารถปรับปรุงคุณภาพของจำนวนคอนทิกได้ ด้วยการนำคอนทิกที่ได้มาทำการหาค่าซ้อนเหลื่อม และหาลำดับ และทำการประกอบชิ้นส่วนเข้าด้วยกันอีกครั้ง ซึ่งจากตารางที่ 4-1 จะพบว่า เมื่อทำซ้ำรอบที่สอง จำนวนของคอนทิกจะลดลงเป็นอย่างมาก เมื่อเทียบกับผลรอบแรก และหากทำซ้ำอีกหลายรอบ อาจจะมีแนวโน้มลดลงเรื่อยๆจนใกล้เคียงกับผลที่ได้จากการคำนวณด้วยโปรแกรมแคปทรี

ผลลัพธ์อีกส่วนที่สนใจก็คือ คุณภาพของคอนทิกที่ได้ โดยในการทดลองนี้ จะพิจารณาจากค่าความผิดพลาดจากการใส่เพิ่มและดึงออก และความผิดพลาดของการครอบคลุม ซึ่งผลลัพธ์ที่ได้

จากการประกอบขึ้นส่วนดีเอ็นเอจะพบว่ายีนข้อมูลมีความยาวมากเท่าไร ก็ยังมีความผิดพลาดทั้งจากการแทนที่ และการดึงออก มากขึ้นตามกัน สาเหตุส่วนหนึ่งเป็นเพราะวิธีการคัดเลือกคู่เบสในขั้นตอนประกอบขึ้นส่วนดีเอ็นเอ ในการทดลองนี้ใช้วิธีคัดเลือกตามส่วนมาก (Majority Vote Rule) และทำการสุ่มอักษรหากมีจำนวนอักษรที่เท่ากัน ซึ่งทำให้ผลลัพธ์ที่ได้มีค่าสูงขึ้นเมื่อทำการทดลองซ้ำหลายรอบ การปรับปรุงคุณภาพของผลเฉลยในส่วนนี้ อาจจะทำให้ได้ด้วยการเปลี่ยนวิธีการคัดเลือกคู่เบสในขั้นตอนประกอบขึ้นส่วนดีเอ็นเอ จากวิธีคัดเลือกตามส่วนมาก เป็นการนำไฟล์คะแนนคุณภาพ (Quality Score) มาใช้เพื่อทำการคัดเลือกตัวอักษรคู่เบส

สำหรับค่าความผิดพลาดในการครอบคลุมจะพบว่า เมื่อเทียบกับโปรแกรมแคปทรีแล้ว ในชุดข้อมูลที่มีความยาวไม่เกิน 30K จะให้ค่าความผิดพลาดในการครอบคลุมที่ดีกว่าผลลัพธ์ที่ได้จากโปรแกรมแคปทรี ส่วนชุดข้อมูลที่มีความยาว 40K-50K จะพบว่าค่าความผิดพลาดในการครอบคลุมมีค่ามากกว่าค่าจากโปรแกรมแคปทรี แต่เมื่อพิจารณาจากภาพที่ 4-3 พบว่าค่าความผิดพลาดได้ลดลงมาก เมื่อทำการทดลองซ้ำ ซึ่งเป็นผลมาจากการหาลำดับและทิศทางจากขั้นตอนเอ็นเอ็นเอ และขั้นตอนเอซีเอส

นอกจากการนำผลลัพธ์คอนทิกมาทำซ้ำหลายๆรอบเพื่อลดจำนวนคอนทิกแล้ว การปรับปรุงโปรแกรมเอ็นเอ็นเอและเอซีเอส ให้สามารถรองรับการคำนวณไฟล์ที่มีขนาดความยาวมากๆ รวมทั้งการนำไฟล์คะแนนคุณภาพ (Quality Score) มาใช้ในการคัดเลือกคู่เบสในช่วงขั้นตอนการประกอบขึ้นส่วนดีเอ็นเอ ก็ควรจะถูกนำมาศึกษาในโอกาสต่อไป

บทที่ 5

สรุปผลการวิจัย

วิทยานิพนธ์นี้นำเสนอวิธีการประกอบชิ้นส่วนดีเอ็นเอแบบอัตโนมัติ ซึ่งปัญหาของการประกอบชิ้นส่วนดีเอ็นเอ จะเกิดขึ้นเนื่องจากในขั้นตอนการหาลำดับเบสของสายดีเอ็นเอที่มีความยาวมากกว่า 600 คู่เบส สายดีเอ็นเอจะถูกจำลองแล้วแบ่งออกเป็นชิ้นส่วนเล็กๆที่สั้นพอที่จะสามารถหาลำดับเบสได้โดยเครื่องมืออัตโนมัติ ทั้งนี้สายดีเอ็นเอจะถูกแบ่งแบบสุ่มทำให้ไม่ทราบว่ชิ้นส่วนนั้นเป็นอันดับใดในสายต้นแบบและไม่ทราบว่ชิ้นส่วนนั้นมาจากสายใด โดยการหาลำดับของชิ้นส่วนนี้ จะใช้วิธีสมิท-วอเตอร์แมน (Smith-Waterman algorithm) เพื่อหาค่าซ้อนเหลื่อม จากนั้นจะใช้ขั้นตอนกฏฮิวริสติกเพื่อนบ้านใกล้ที่สุด (NNH) ร่วมกับวิธีเอซีเอส (ACS) ในการหาลำดับส่วนการจัดเรียงชิ้นส่วนดีเอ็นเอให้กลายเป็นคอนทิก และจะใช้โปรแกรมที่เขียนขึ้นเป็นตัวจัดเรียงชิ้นส่วนดีเอ็นเอเหล่านี้ให้โดยอัตโนมัติ โดยผลลัพธ์ที่ได้จะถูกนำไปเปรียบเทียบกับโปรแกรมแคปทรีซึ่งเป็นโปรแกรมมาตรฐานที่มีอยู่ในปัจจุบันสำหรับการประกอบชิ้นส่วนดีเอ็นเอ และผลลัพธ์คอนทิกที่ได้นี้ยังสามารถพิจารณาเป็นเสมือนชิ้นส่วนดีเอ็นเอแล้วนำมาทดลองซ้ำ จากผลลัพธ์ที่ได้แสดงให้เห็นว่า จำนวนคอนทิกที่ได้แม้จะมีจำนวนมากกว่าผลลัพธ์ที่ได้จากโปรแกรมแคปทรี แต่เมื่อทำการวนทดลองซ้ำด้วยข้อมูลคอนทิกที่ได้มาในรอบแรก พบว่าจำนวนคอนทิกที่ลดลงและเทียบเคียงกันกับผลของโปรแกรมแคปทรี ในส่วนคุณภาพของผลเฉลยที่ได้จากโปรแกรมแคปทรีจะดีกว่าที่ได้จากเทคนิคที่นำเสนอในแง่ของความผิดพลาดในการแทนที่และความผิดพลาดของการใส่เพิ่มและดึงออก แต่ในส่วนของความผิดพลาดในการครอบคลุม ผลลัพธ์ที่ได้จากการทดลอง มีผลที่ดีกว่าที่ได้จากโปรแกรมแคปทรี ในชุดข้อมูลที่มีความยาวของคู่เบสไม่เกิน 30 K แสดงได้ว่าการนำขั้นตอนกฏฮิวริสติกเพื่อนบ้านใกล้ที่สุดร่วมกับวิธีเอซีเอส สามารถหาลำดับและทิศทางเพื่อนำมาประกอบชิ้นส่วนดีเอ็นเอ และได้ผลลัพธ์ที่เทียบเคียงได้ตามโปรแกรมสำเร็จรูป เช่น โปรแกรมแคปทรี

บรรณานุกรม

ภาษาอังกฤษ

- Allex, C. F., Baldwin, S. F., Shavlik, J. W. and Blattner, F. R. "Improving the Quality of Automatic DNA Sequence Assembly Using Fluorescent Trace-Data Classifications." Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. 3-14. Edited by D. J. States, P. Agarwal, T. Gaasterland, L. Hunter and R. F. Smith. Menlo Park, CA : AAAI Press, 1996.
- Angeleri, E., Apolloni, B., de Falco, D. and Grandi, L. "DNA Fragment Assembly Using Neural Prediction Techniques." International Journal of Neural Systems. 9 (1999) : 523-544.
- Beckers, R., Deneubourg, J. L. and Goss, S. "Trails and U-turns in the Selection of the Shortest Path by the Ant *Lasius Niger*." Journal of Theoretical Biology. 159 (1992) : 397-415.
- Bonabeau, E., Dorigo, M. and Theraulaz, G. Swarm Intelligence. New York, NY : Oxford University Press, c1999.
- Churchill, G., Burks, C., Eggert, M., Engle, M. and Waterman, M. Assembling DNA Sequence Fragments by Shuffling and Simulated Annealing. Technical Report LAUR 2287, Los Alamos National Lab, 1993.
- Di Caro, G. and Dorigo, M. "AntNet: Distributed Stigmergetic Control for Communications Networks." Journal of Artificial Intelligence Research. 9 (1998) : 317-365.
- Dorigo, M. and Gambardella, L. M. "Ant Colony System : A Cooperative Learning Approach to the Traveling Salesman Problem." IEEE Transactions on Evolutionary Computation. 1 (1997) : 53-66.
- Gambardella, L. M. and Dorigo, M. "An Ant Colony System Hybridized with a New Local Search for the Sequential Ordering Problem." INFORMS Journal on Computing. 12 (2000) : 237-255.
- Goldberg, D. E. Genetic Algorithms : In Search, Optimization and Machine Learning. Reading, MA : Addison-Wesley, c1989.

- Green, P. 2004. Phrap Documentation. Phred, Phrap, and Consed. Available online at <http://www.phrap.org>
- Holland, J. H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI : University of Michigan Press, c1975.
- Huang, X. and Miller, W. "A Time-Efficient, Linear-Space Local Similarity Algorithm." Advances in Applied Mathematics. 12 (1991) : 337-357.
- Huang, X. "A Contig Assembly Program Based on Sensitive Detection of Fragment Overlaps." Genomics. 14 (1992) : 18-25.
- Huang, X. "An Improved Sequence Assembly Program." Genomics. 33 (1996) : 21-31.
- Huang, X. and Madan, A. "CAP3 : A DNA Sequence Assembly Program." Genome Research. 9 (1999) : 868-877.
- Kececioglu, J. D. and Myers, E. W. "Combinatorial Algorithms for DNA Sequence Assembly." Algorithmica. 13 (1995) : 7-51.
- Kim, K. and Mohan, C. K. "Parallel Hierarchical Adaptive Genetic Algorithm for Fragment Assembly." Proceedings of the 2003 Congress on Evolutionary Computation. 600-607. Edited by R. Saker, R. Reynolds, H. Abbass, K. C. Tan, B. McKay, D. Essam, T. Gedeon. Piscataway, NJ : IEEE, 2003.
- Moss, J. D. and Johnson, C. G. "An Ant Colony Algorithm for Multiple Sequence Alignment in Bioinformatics." Artificial Neural Nets and Genetic Algorithms. 182-186. Edited by D. W. Pearson, N. C. Steele and R. F. Albrecht. New York, NY : Springer, 2003.
- Parsons, R. J. and Johnson, M. E. "DNA Sequence Assembly and Genetic Algorithms New Results and Puzzling Insights." Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology. 277-284. Edited by C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak. Menlo Park, CA : AAAI Press, 1995.
- Parsons, R. J. and Johnson, M. E. "A Case Study in Experimental Design Applied to Genetic Algorithms with Applications to DNA Sequence Assembly." American Journal of Mathematical and Management Sciences. 17 (1997) : 369-396.

- Pevzner, P. A., Tang, H. and Waterman, M. S. "An Eulerian Path Approach to DNA Fragment Assembly." Proceedings of the National Academy of Sciences of the United States of America. 98 (2001) : 9748-9753.
- Smith, T. F. and Waterman, M. S. "Identification of Common Molecular Subsequences." Journal of Molecular Biology. 147 (1981) : 195-197.
- Staden, R. "A New Computer Method for the Storage and Manipulation of DNA Gel Reading Data." Nucleic Acids Research. 8 (1980) : 3673-3694.
- Sutton G., White, O., Adams, M., and Kerlavage, A. "TIGR Assembler : A New Tool for Assembling Large Shotgun Sequencing Projects." Genome Science and Technology. 1 (1995) 9-19.

ประวัติผู้วิจัย

ชื่อ : นายกัมพล สุกใส
 ชื่อวิทยานิพนธ์ : การประกอบชิ้นส่วนดีเอ็นเอแบบอัตโนมัติและเป็นลำดับโดยใช้กฎเมตาฮิวริสติก
 สาขาวิชา : วิศวกรรมไฟฟ้า

ประวัติ

จบการศึกษาระดับปริญญาตรี ที่สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
 คณะวิศวกรรมศาสตร์ สาขาวิชาวิศวกรรมไฟฟ้า ปีการศึกษา 2545

สถานที่ติดต่อ 363/569 ซ.พหลโยธิน52 ถ.พหลโยธิน แขวงคลองถนน เขตสายไหม กรุงเทพฯ