การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธีเชิงวิวัฒนาการ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

น.ส.สุภาวดี ศรีคำดี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

PREDICTION OF RNA SECONDARY STRUCTURE USING EVOLUTIONARY ALGORITHM

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

Miss Supawadee Srikamdee

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Computer Engineering) in Computer Engineering

Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2018

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธีเชิงวิวัฒนาการ
โดย	น.ส.สุภาวดี ศรีคำดี
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

.... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

.... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

.... กรรมการ
(ดร.ดวงดาว วิชาดากุล)

.... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล)

.... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.พันธุ์ปิติ เปี่ยมสง่า)

สุภาวดี ศรีคำดี : การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธีเชิงวิวัฒนาการ. (PREDICTION OF RNA SECONDARY STRUCTURE USING EVOLUTIONARY ALGORITHM) อ.ที่ปรึกษาหลัก : ศ. ดร.ประภาส จงสถิตย์วัฒนา

วิทยานิพนธ์นี้นำเสนอขั้นตอนวิธีแบบใหม่ชื่อว่า Hybrid-EDAfold ซึ่งเป็นขั้นตอนวิธีเชิงวิวัฒนาการที่อยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจงแบบผสมสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ ขั้นตอนวิธีที่นำเสนอประกอบด้วย 2 ขั้นตอนวิธีประมาณการแจกแจงและดำเนินการอยู่บนเทคนิคการทำนายโครงสร้างที่มีค่าพลังงานต่ำสุด ขั้นตอนวิธีที่นำเสนอใช้ทั้งกลุ่มคำตอบดีและกลุ่มคำตอบด้อยร่วมกันในการปรับปรุงแบบจำลองความน่าจะเป็นเพื่อส่งเสริมให้ขั้นตอนวิธีสามารถค้นหาได้ทั่วทั้งปริภูมิค้นหา ใช้ข้อมูลจากคำตอบด้อยเพื่อบ่งบอกว่าบริเวณไหนไม่น่าสนใจที่จะเข้าไปสำรวจเมื่อต้องดำเนินการกับข้อมูลที่มีจำนวนมิติที่ค่อนข้างสูง วิธีการที่นำเสนอมีการเพิ่มเติมตัวดำเนินการกลายพันธุ์ในขั้นตอนวิธีประมาณการแจกแจงหนึ่งเพื่อสนับสนุนการค้นหาแบบท้องถิ่น ช่วยเพิ่มความหลากหลายของคำตอบและบรรเทาการลู่เข้าก่อนกำหนด นอกจากนี้ วิธีการที่นำเสนอยังรองรับการทำนายหลายโครงสร้างทั้งโครงสร้างที่มีค่าพลังงานต่ำสุดและโครงสร้างที่มีค่าพลังงานต่ำรองเพื่อเพิ่มโอกาสที่จะพบโครงสร้างที่ใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากยิ่งขึ้น การประเมินประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เมื่อเปรียบเทียบกับขั้นตอนวิธีในกลุ่มของกำหนดการพลวัตที่เป็นที่รู้จักกันดี ได้แก่ Mfold, RNAfold และ RNAstructure บนข้อมูลอาร์เอ็นเอจาก 15 ชนิด จำนวน 760 สายลำดับ พบว่า ขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายเฉลี่ยดีกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบในทุกตัวชี้วัด และ เปรียบเทียบกับขั้นตอนวิธีในกลุ่มเมตาฮิวริสติกด้วยอาร์เอ็นเอ 20 สายลำดับ ผลลัพธ์แสดงให้เห็นว่าวิธีการที่นำเสนอมีค่า F-measure เฉลี่ยดีกว่า RnaPredict และ SARNA-Predict และ มีผลลัพธ์เทียบเคียงได้กับ TL-PSOfold

สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต	....
ปีการศึกษา	2561	ลายมือชื่อ อ.ที่ปรึกษาหลัก	....

2430366845

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

5771474521 : MAJOR COMPUTER ENGINEERING
KEYWORD: RNA Secondary Structure / Evolutionary Algorithm / Estimation of Distribution Algorithm

Supawadee Srikamdee : PREDICTION OF RNA SECONDARY STRUCTURE USING EVOLUTIONARY ALGORITHM . Advisor: Prof. Prabhas Chongstitvatana, Ph.D.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

This thesis proposed a new method namely Hybrid-EDAfold which is an evolutionary algorithm (EA) based on a hybrid estimation of distribution algorithms (EDAs) for RNA secondary structure prediction. The proposed method consists of two EDAs and using minimum free energy technique. The Hybrid-EDAfold uses both good and poor solutions enabling the algorithm to search throughout the search space. Using information from poor solutions can indicate which area is unappealing to explore when conducting a search with high-dimensional data. In addition, one of the EDA uses a mutation operator to support local search which increases the diversity and moderately avoid early convergence. Moreover, the proposed method returns the answer as a set of structures consisting of optimal structure and suboptimal structures to increase the chance of finding a predicted structure closer to the real structure. Comparison of the Hybrid-EDAfold was evaluated with well-known web servers namely Mfold, RNAfold, and RNAstructure on 15 RNA types with 760 RNA sequences total. The Hybrid-EDAfold yields better results than other methods in every metrics. The proposed method was also compared with metaheuristic methods on 20 RNA sequences collected from their literature. The results showed that the Hybrid-EDAfold yields better results than RnaPredict and SARNA-Predict and is comparable to TL-PSOfold.

Field of Study: Computer Engineering		Student's Signature ....
Academic Year: 2018		Advisor's Signature ....

กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จลุล่วงลงได้เพราะได้รับการดูแลจากท่านอาจารย์ ศ.ดร. ประภาส จงสถิตย์วัฒนา ตลอดระยะเวลาของการเรียนปริญญาเอก อาจารย์ได้ให้คำแนะนำทั้งทางด้านวิชาการและแนวทางในการใช้ชีวิตของป.เอก ช่วงเวลาที่มีปัญหา หรือ ติดขัดสิ่งใด อาจารย์ก็จะให้การช่วยเหลือจนผ่านพ้นอุปสรรคต่าง ๆ ไปได้ด้วยดี

ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ ศ. ดร. บุญเสริม กิจศิริกุล, รศ.ดร. พันธุ์ปิติ เปี่ยมสง่า, ดร. ดวงดาว วิชาดากุล, และ ผศ. ดร. พีรพล เวทีกูล ที่ให้ความอนุเคราะห์ในการเป็นกรรมการสอบ รวมทั้งให้คำแนะนำ ข้อคิดเห็นอันเป็นประโยชน์สำหรับปรับปรุงงานวิจัย ประสบการณ์ที่ได้รับในครั้งนี้จะช่วยพัฒนาทักษะทางด้านการวิจัยของผู้เรียนให้ดียิ่งขึ้น

ขอขอบคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย สำหรับความรู้ ความเมตตา ตลอดระยะเวลาของการศึกษาที่นี่ และ คณาจารย์จากคณะวิทยาการสารสนเทศ ม.บูรพา สำหรับโอกาส และ ทุนการศึกษาตลอดระยะเวลา 3 ปี

ขอขอบคุณ ผศ. ดร. สุนิสา ริมเจริญ สำหรับการดูแล และ ความช่วยเหลือในทุก ๆ ด้านตั้งแต่การศึกษาในระดับปริญญาตรี ปริญญาโท จนกระทั่งระดับปริญญาเอกอาจารย์ก็ยังคงห่วงใยและให้การสนับสนุนในด้านต่าง ๆ แก่ผู้เรียนเสมอมา

ขอขอบคุณพี่ ๆ น้อง ๆ สมาชิก Intelligent Systems Laboratory (ISL) รวมทั้งเพื่อน ๆ ป.โท ป.เอก ในภาควิชาวิศวกรรมคอมพิวเตอร์ ตลอดช่วงเวลาที่ผู้เรียนศึกษาอยู่ที่นี่ มิตรภาพและความทรงจำดี ๆ ได้เกิดขึ้นมากมาย ขอบคุณสำหรับทุก ๆ การช่วยเหลือ และน้ำใจที่ทุกคนมอบให้

สุดท้ายขอขอบคุณ “ครอบครัว” ที่คอยให้การช่วยเหลือ ให้กำลังใจ และสนับสนุนผู้เรียนในทุก ๆ ด้าน ถือเป็นกำลังใจที่สำคัญที่ทำให้ผู้เรียนก้าวข้ามอุปสรรคต่าง ๆ และไม่ละความพยายามที่จะทำเป้าหมายนี้ให้สำเร็จ

สุภาวดี ศรีคำดี

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญรูปภาพ ..... ฎ

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 6

1.3 ขอบเขตการวิจัย 6

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย 6

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย 7

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 8

2.1 ทฤษฎีที่เกี่ยวข้อง 8

2.1.1 กรดไรโบนิวคลีอิก หรือ อาร์เอ็นเอ 8

2.1.2 ชนิดของอาร์เอ็นเอ 9

2.1.3 โครงสร้างทุติยภูมิของอาร์เอ็นเอ 12

2.1.4 การกำหนดโครงสร้างทุติยภูมิของอาร์เอ็นเอ 14

2.1.5 ขั้นตอนวิธีประมาณการแจกแจง 22

2.2 งานวิจัยที่เกี่ยวข้อง 33

2.2.1 งานวิจัยเกี่ยวกับการทำนายโครงสร้างด้วย 1 สายลำดับ 33

2430366845

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.2.2 งานวิจัยเกี่ยวกับการทำนายโครงสร้างโดยใช้หลายสายลำดับ 37
บทที่ 3 วิธีดำเนินการวิจัย..... 42
3.1 ระบุฮีลิกที่เป็นไปได้ทั้งหมดใน 1 สายลำดับที่เป็นข้อมูลนำเข้า..... 43
3.1.1 การระบุฮีลิกโดยใช้ข้อมูลความน่าจะเป็นของคู่เบส 44
3.1.2 ประเมินประสิทธิภาพของขั้นตอนการจัดเตรียมฮีลิก 47
3.1.3 การปรับปรุงเซตของฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิก..... 49
3.2 การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold..... 54
3.2.1 กำหนดค่าเริ่มต้นให้กับโมเดล 59
3.2.2 การสร้างประชากร 61
3.2.3 การประเมินคุณภาพของประชากร..... 67
3.2.4 การปรับปรุงค่าในเวกเตอร์ความน่าจะเป็น 70
3.2.5 การปรับปรุงข้อมูลในอาไคว์ 76
3.3 การประเมินค่าความถูกต้องของโครงสร้างที่ทำนายได้..... 77
บทที่ 4 ผลการวิจัย..... 79
4.1 ประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เมื่อกำหนดค่าพารามิเตอร์แตกต่างกัน 80
4.2 เปรียบเทียบประสิทธิภาพของวิธีการทำนายหลายโครงสร้างที่งานวิจัยนี้นำเสนอกับวิธีการที่ใช้ในโปรแกรมอื่น ๆ 85
4.3 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตบนข้อมูล pre-miRNA ของมนุษย์จำนวน 10 รายการ 91
4.4 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มเมตาฮิวริสติกด้วยข้อมูลอาร์เอ็นเอจำนวน 20 รายการ 97
4.5 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตด้วยข้อมูลอาร์เอ็นเอจำนวน 750 รายการจากฐานข้อมูล RNA STARND v2.0 100
4.5.1 การเปรียบเทียบค่าความถูกต้องโดยเฉลี่ย..... 101
4.5.2 ผลการจัดอันดับ F-measure สำหรับอาร์เอ็นเอแต่ละชนิด..... 106

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

4.6 สรุปสิ่งที่ได้จากการทดสอบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold 117

บทที่ 5 สรุปผล.. 120

5.1 สรุปผลการวิจัย... 120

5.2 งานวิจัยในอนาคต 126

บรรณานุกรม... 127

ประวัติผู้เขียน.. 139

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

สารบัญตาราง

หน้า

ตารางที่ 3.1 ตัวอย่างความน่าจะเป็นของคู่เบสที่ได้จากโปรแกรม RNAfold 45

ตารางที่ 3.2 การประเมินประสิทธิภาพของขั้นตอนวิธีการระบุฮีลิก 47

ตารางที่ 3.3 การเปรียบเทียบเซตของฮีลิกที่พบในโครงสร้างคำตอบกับเซตของฮีลิกที่สร้างได้......... 49

ตารางที่ 3.4 ตัวอย่างการแชร์ตำแหน่งเบสร่วมกันของสองฮีลิกที่ตรงกับกรณีที่ 1............................ 51

ตารางที่ 3.5 ตัวอย่างการแชร์ตำแหน่งเบสร่วมกันของสองฮีลิกที่ตรงกับกรณีที่ 2............................ 52

ตารางที่ 3.6 การพิจารณาตำแหน่งของคู่เบสที่มีการทับซ้อนกันของ $helix_1$ และ $helix_6$ 53

ตารางที่ 3.7 ตัวอย่างการกำหนดค่าเริ่มต้นให้เวกเตอร์ความน่าจะเป็น 59

ตารางที่ 3.8 ตัวอย่างการกำหนดค่าในเมทริกซ์ตรวจสอบความเข้ากันได้ของฮีลิก 60

ตารางที่ 3.9 การประเมินค่าความเหมาะสมสำหรับประชากร 69

ตารางที่ 3.10 ความสอดคล้องกันระหว่างค่าความเหมาะสมกับคุณภาพคำตอบ 69

ตารางที่ 3.11 การจำแนกโครโมโซมดีและโครโมโซมด้อยในกลุ่มประชากร 71

ตารางที่ 3.12 ความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมดี 72

ตารางที่ 3.13 ความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมด้อย.. 72

ตารางที่ 3.14 การปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ EDA-G 73

ตารางที่ 3.15 การเปรียบเทียบค่าความเหมาะสมของบรรพบุรุษกับลูกหลานที่สร้างได้................... 74

ตารางที่ 3.16 ความถี่ของหมายเลขฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษ 75

ตารางที่ 3.17 ความถี่ของหมายเลขฮีลิกที่ถูกสุ่มเพิ่มเติมในการสร้างโครโมโซมลูก........................... 75

ตารางที่ 3.18 การปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นสำหรับ EDA-L 76

ตารางที่ 4.1 คุณลักษณะของ 20 สายลำดับอาร์เอ็นเอ... 81

ตารางที่ 4.2 การทดสอบพารามิเตอร์ของขั้นตอนวิธี Hybrid-EDAfold บน 20 อาร์เอ็นเอ 82

ตารางที่ 4.3 เปรียบเทียบค่า free energy ของโครงสร้างที่ทำนายได้จากโปรแกรม Mfold, RNAstructure และ Hybrid-EDAfold กับโครงสร้างคำตอบบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ ... 86

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.4 การเปรียบเทียบโครงสร้างที่มีค่า free energy ต่ำสุดที่ทำนายด้วยโปรแกรม Mfold, RNAstructure และ Hybrid-EDA กับโครงสร้างคำตอบบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ 87

ตารางที่ 4.5 ประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Mfold, RNAstructure และ Hybrid-EDAfold เมื่อแต่ละขั้นตอนวิธีรองรับการทำนายหลายโครงสร้างบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ 89

ตารางที่ 4.6 คุณลักษณะของสายลำดับ pre-miRNA ของมนุษย์ 91

ตารางที่ 4.7 การเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold กับวิธีในกลุ่มกำหนดการพลวัตเมื่อทดสอบกับข้อมูล pre-miRNA ของมนุษย์ 92

ตารางที่ 4.8 การเปรียบเทียบผลการทำนายโครงสร้างของ Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มเมตาฮิวริสติก 98

ตารางที่ 4.9 ข้อมูลสรุปของอาร์เอ็นเอ 14 ชนิดจากฐานข้อมูล RNA STRAND v2.0 100

ตารางที่ 4.10 การเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตบนข้อมูลสายลำดับอาร์เอ็นเอ 14 ชนิด 102

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

สารบัญรูปภาพ

หน้า

รูปที่ 2.1 ตัวอย่างโครงสร้างทุติยภูมิของอาร์เอ็นเอส่งถ่าย 9

รูปที่ 2.2 การแทนโครงสร้างทุติยภูมิในรูปแบบสัญลักษณ์จุดและวงเล็บ 12

รูปที่ 2.3 รูปร่างพื้นฐานที่สามารถพบได้ในโครงสร้างทุติยภูมิของอาร์เอ็นเอ [50] 13

รูปที่ 2.4 ตัวอย่างซูโดนอท 19

รูปที่ 2.5 รหัสเทียมของขั้นตอนวิธีประมาณการแจกแจง 23

รูปที่ 2.6 ขั้นตอนการทำงานของขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ 27

รูปที่ 2.7 ขั้นตอนการทำงานของขั้นตอนวิธีคอยน์ 28

รูปที่ 2.8 การกำหนดค่าเริ่มต้นให้เมทริกซ์ 29

รูปที่ 2.9 การประเมินค่าความเหมาะสมของแต่ละโครโมโซม 30

รูปที่ 2.10 ตัวอย่างการจำแนกกลุ่มของประชากร 30

รูปที่ 2.11 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับสมาชิก [1,3] 31

รูปที่ 2.12 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับโครโมโซมที่เป็นคำตอบดี 32

รูปที่ 2.13 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับโครโมโซมที่เป็นคำตอบด้อย 32

รูปที่ 3.1 ภาพรวมของขั้นตอนวิธีการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ 42

รูปที่ 3.2 ตัวอย่างการระบุบริเวณที่เป็นฮีลิกในสายลำดับอาร์เอ็นเอยาว 20 นิวคลีโอไทด์ 43

รูปที่ 3.3 ตัวอย่างการสร้าง และการเข้ารหัสฮีลิก 46

รูปที่ 3.4 ตัวอย่างการสร้างโครงสร้างอาร์เอ็นเอด้วยการเลือกฮีลิกครั้งละชิ้น 54

รูปที่ 3.5 ภาพรวมของขั้นตอนวิธี Hybrid-EDAfold 57

รูปที่ 3.6 ขั้นตอนวิธีการสร้างประชากรสำหรับ EDA-G 62

รูปที่ 3.7 ขั้นตอนวิธีการสร้างประชากรสำหรับ EDA-L 65

รูปที่ 3.8 ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ EDA-G 72

รูปที่ 3.9 ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ EDA-L 74

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

รูปที่ 3.10 การเปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบ 78

รูปที่ 4.1 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-16-1..................... 94

รูปที่ 4.2 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-let-7f-2............... 96

รูปที่ 4.3 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-29a 96

รูปที่ 4.4 การจัดอันดับ F-measure เมื่อทดสอบกับ Transfer Messenger RNA........................ 106

รูปที่ 4.5 การจัดอันดับ F-measure เมื่อทดสอบกับ 16S Ribosomal RNA 107

รูปที่ 4.6 การจัดอันดับ F-measure เมื่อทดสอบกับ Transfer RNA 108

รูปที่ 4.7 การจัดอันดับ F-measure เมื่อทดสอบกับ Ribonuclease P RNA 108

รูปที่ 4.8 การจัดอันดับ F-measure เมื่อทดสอบกับ Synthetic RNA... 109

รูปที่ 4.9 การจัดอันดับ F-measure เมื่อทดสอบกับ Signal Recognition Particle RNA 110

รูปที่ 4.10 การจัดอันดับ F-measure เมื่อทดสอบกับ 23S Ribosomal RNA............................. 111

รูปที่ 4.11 การจัดอันดับ F-measure เมื่อทดสอบกับ 5S Ribosomal RNA............................... 111

รูปที่ 4.12 การจัดอันดับ F-measure เมื่อทดสอบกับ Group I Intron 112

รูปที่ 4.13 การจัดอันดับ F-measure เมื่อทดสอบกับ Hammerhead Ribozyme 113

รูปที่ 4.14 การจัดอันดับ F-measure เมื่อทดสอบกับ Other Ribosomal RNA.......................... 113

รูปที่ 4.15 การจัดอันดับ F-measure เมื่อทดสอบกับ Other Ribozyme.................................... 114

รูปที่ 4.16 การจัดอันดับ F-measure เมื่อทดสอบกับ Group II Intron 115

รูปที่ 4.17 การจัดอันดับ F-measure เมื่อทดสอบกับ Cis-regulatory element........................ 116

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

กรดไรโบนิวคลีอิก (Ribonucleic, RNA) ทำหน้าที่หลักในการเป็นตัวกลางเพื่อถอดรหัสข้อมูลทางพันธุกรรมจากดีเอ็นเอไปเป็นโปรตีน กล่าวคือ ในกระบวนการสังเคราะห์โปรตีนจะใช้ดีเอ็นเอเป็นแม่แบบ โดยโครงสร้างโมเลกุลที่เป็นเกลียวคู่ของดีเอ็นเอจะเกิดการคลายตัวออก จากนั้นสายด้านหนึ่งของดีเอ็นเอจะถูกถอดรหัสโดยเอนไซม์อาร์เอ็นเอพอลิเมอเรส (RNA polymerase) ได้เป็นอาร์เอ็นเอนำรหัส (messenger RNA, mRNA) จากนั้นรหัสในอาร์เอ็นเอนำรหัสจะถูกแปลงเป็นกรดอะมิโนด้วยอาร์เอ็นเอส่งถ่าย (transfer RNA, tRNA) โดยมีอาร์เอ็นเอไรโบโซม (ribosomal RNA, rRNA) ทำหน้าที่เป็นส่วนประกอบของไรโบโซม เมื่อกรดอะมิโนเหล่านั้นเรียงต่อกันจะได้เป็นโปรตีน

อาร์เอ็นเอมีบทบาทสำคัญในทางชีววิทยา อาทิ การเร่งปฏิกิริยาเคมี (catalysis) การควบคุมยีน และ ช่วยในกระบวนการสังเคราะห์โปรตีน การเข้าใจโครงสร้างของอาร์เอ็นเอสำคัญต่อความรู้พื้นฐานทางด้านพันธุศาสตร์ ปัจจุบันงานวิจัยทางการแพทย์มีการค้นพบว่าอาร์เอ็นเอบางชนิด เช่น microRNA สามารถถูกใช้เป็นดัชนีชี้วัดทางชีวภาพ (biomarker) เพื่อบ่งชี้การเป็นมะเร็งในร่างกายมนุษย์ โดยโครงสร้างอาร์เอ็นเอจะมีลักษณะเป็นลำดับชั้นไล่เรียงไปตั้งแต่โครงสร้างปฐมภูมิ (primary structure) ซึ่งประกอบด้วยตัวอักษร ‘A’, ‘C’, ‘G’, ‘U’ ที่แทนนิวคลีโอไทด์เรียงต่อกันเป็นสายพอลิเมอร์ โครงสร้างทุติยภูมิ (secondary structure) เป็นกลุ่มของเบสที่เข้าคู่กัน และโครงสร้างตติยภูมิ (tertiary structure) เป็นโครงสร้างสามมิติ

วิธีการพิจารณาโครงสร้างทุติยภูมิของอาร์เอ็นเอแบ่งได้เป็น 2 กลุ่มหลัก ๆ [1] ได้แก่ วิธีเชิงการทดลอง (experimental approaches) เช่น chemical probing, x-ray crystallography และนิวเคลียร์แมกเนติกเรโซแนนซ์สเปกโทรสโคปี (nuclear magnetic resonance spectroscopy, NMR) ซึ่งให้ค่าความแม่นยำในการกำหนดโครงสร้างที่ค่อนข้างสูง แต่มีข้อเสียคือมีความอ่อนไหวต่อสภาพแวดล้อม แพง และใช้ระยะเวลานาน ในขณะที่วิธีการอีกกลุ่มหนึ่งคือ วิธีเชิงการคำนวณ (computational approaches) ซึ่งได้รับความนิยมในปัจจุบัน เนื่องจากให้ค่าความถูกต้องในการทำนายโครงสร้างที่ดีเพียงพอและปริมาณงานที่ทำได้ในหนึ่งหน่วยเวลาสูง (high throughput)

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยวิธีเชิงการคำนวณเริ่มตั้งแต่ปี ค.ศ. 1978 โดย Nussinov และคณะได้นำเสนอวิธีการหาคู่เบสมากสุด (maximal base-pairing) ที่ริเริ่มแปลงปัญหาการทำนายโครงสร้างอาร์เอ็นเอไปเป็นปัญหาการตัดสินใจเหมาะสมสุด (optimal decision problem) และแก้ปัญหาด้วยกำหนดการพลวัต (dynamic programming) งานวิจัยนี้ค่อนข้างมีความสำคัญอย่างมากแม้ว่าค่าความถูกต้องในการทำนายจะยังไม่สูงมากแต่วิธีการที่ใช้มีความเรียบง่ายและตรงไปตรงมา [2] ต่อมา Zuker ได้พัฒนาวิธีค่าพลังงานต่ำสุด (minimum free energy, MFE) [3] ซึ่งวิธีนี้ได้ถูกปรับปรุงและพัฒนาเป็นเครื่องมือที่ได้รับความนิยมใช้งานอยู่ในปัจจุบัน อาทิ Mfold [4] และ RNAfold ซึ่งอยู่ใน ViennaRNA package [5, 6] โปรแกรมดังกล่าวรับข้อมูลนำเข้าเป็นสายลำดับอาร์เอ็นเอที่ต้องการทำนายโครงสร้างและให้ผลการทำนายโครงสร้างทุติยภูมิที่อยู่ในรูปแบบสัญลักษณ์จุดและวงเล็บ (dot-bracket notation) ซึ่งจุดแทนบริเวณนิวคลีโอไทด์อิสระและคู่ของวงเล็บแทนตำแหน่งของนิวคลีโอไทด์ที่มีการจับคู่กันภายในโครงสร้าง

นอกเหนือจากวิธีการในกลุ่มกำหนดการพลวัตมีกลุ่มงานวิจัยที่นำเสนอวิธีการเชิงสุ่ม (Stochastic approach) สำหรับกำหนดโครงสร้างอาร์เอ็นเอ เช่น stochastic context-free grammar (SCFG) [7-9] Bayesian statistical [10] และ partition function [11] และวิธีการในกลุ่มที่อาศัยวิธีเมตาฮิวริสติก (metaheuristic) ได้แก่ RNAPredict [12] ที่ทำการหาโครงสร้างอาร์เอ็นเอที่มีค่าพลังงานต่ำสุดด้วยขั้นตอนวิธีทางพันธุกรรม (Genetic Algorithm, GA) SARNA-Predict [13] อาศัยหลักการของแบบจำลองการอบเหนียว (annealing schedule) ด้วยการกลายพันธุ์แบบต่าง ๆ และ TL-PSOfold [14] อาศัยหลักการของการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (particle swarm optimization) โดยขั้นตอนวิธีนี้ได้แบ่งการทำงานเป็น 2 ระดับ แต่ละระดับใช้ฟังก์ชันวัตถุประสงค์ (objective function) ที่แตกต่างกัน

งานวิจัยในช่วงหลังเริ่มเพิ่มความซับซ้อนของขั้นตอนวิธีสำหรับทำนายโครงสร้างมากขึ้นโดยใช้หลายเทคนิคร่วมกัน เช่น งานวิจัย [15] ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับวิธีการเปรียบเทียบ (comparative approach) หรือเพิ่มจำนวนฟังก์ชันวัตถุประสงค์มากขึ้น เช่น TL-PSOfold ที่ใช้ 2 ฟังก์ชันวัตถุประสงค์ ได้แก่ หาจำนวนคู่เบสที่ทำให้เกิดพันธะไฮโดรเจนมากสุดร่วมกับการคำนวณค่าพลังงานต่ำสุด และงานวิจัย [16] ใช้การคำนวณค่าพลังงานต่ำสุดและการคำนวณค่าความถูกต้องที่คาดหวังสูงสุด (maximum expected accuracy, MEA) เป็นฟังก์ชันวัตถุประสงค์และทำนายโครงสร้างด้วยการโปรแกรมเชิงจำนวนเต็ม (integer programming) แสดงให้เห็นว่าวิธีการพื้นฐาน

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

อาจไม่เพียงพอที่จะทำให้ได้ค่าความถูกต้องของการทำนายโครงสร้างในระดับที่น่าพึงพอใจโดยเฉพาะเมื่อดำเนินการกับกลุ่มข้อมูลสายลำดับอาร์เอ็นเอที่ค่อนข้างยาว เช่น 16S Ribosomal RNA หรือ 23S Ribosomal RNA

จากโปรแกรมสำหรับทำนายโครงสร้างทุติยภูมิที่มีการใช้งานในปัจจุบันทำให้วิทยานิพนธ์นี้เกิดแรงจูงใจที่จะนำเสนอขั้นตอนวิธีสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่มีค่าความถูกต้องมากยิ่งขึ้น วิธีการที่นำเสนออยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจง (Estimation of distribution algorithms, EDA) ซึ่งถูกนำเสนอเป็นครั้งแรกโดย Mühlenbein และ Paass ในปี 1996 [17] โดยขั้นตอนวิธีประมาณการแจกแจงจะแตกต่างไปจากขั้นตอนวิธีเชิงวิวัฒนาการ (Evolutionary Algorithms, EAs) แบบดั้งเดิม [18] เนื่องจากขั้นตอนวิธีนี้จะใช้แบบจำลองความน่าจะเป็นเพื่อสร้างประชากร โดยแบบจำลองความน่าจะเป็นนี้ถูกสร้างโดยการเรียนรู้จากเซตคำตอบที่ดีจากประชากรในรุ่นก่อนหน้าและถูกใช้สำหรับสร้างประชากรรุ่นถัดไป ข้อดีหลักของขั้นตอนวิธีประมาณการแจกแจงที่เหนือกว่าขั้นตอนวิธีเชิงพันธุกรรม คือ ไม่มีพารามิเตอร์ที่จะต้องถูกปรับให้เหมาะ เช่น ความน่าจะเป็นในการไขว้เปลี่ยน (mutation probability) และความน่าจะเป็นในการกลายพันธุ์ (crossover probability) และไม่ใช้ตัวดำเนินการทางพันธุกรรมแต่ใช้แบบจำลองความน่าจะเป็นซึ่งแทนข้อมูลเชิงสถิติที่มีความหมาย [19] เป็นที่รู้กันว่าขั้นตอนวิธีประมาณการแจกแจงประสบความสำเร็จในการแก้ปัญหาการหาค่าเหมาะสุดเชิงการจัด [20] นอกจากนี้ ขั้นตอนวิธีประมาณการแจกแจงถูกพิสูจน์ว่ามีประสิทธิภาพและประสิทธิผลในการแก้ปัญหาจีโนมที่เป็น NP-hard ต่าง ๆ [21] โดยขั้นตอนวิธีประมาณการแจกแจงถูกประยุกต์ใช้ในงานทางด้านชีวสารสนเทศมาตั้งแต่ปี 2000 เช่น การใช้ขั้นตอนวิธีประมาณการแจกแจงเพื่อวิเคราะห์โครงสร้างของยีน (Gene structure analysis) เนื่องจากยีนอาจประกอบด้วยหลายส่วนที่แตกต่างกัน ปัญหาการทำนายโครงสร้างของยีนสามารถถูกมองเป็นปัญหาการแบ่งส่วน (segmentation) และใช้เทคนิคการเลือกคุณลักษณะย่อย (feature subset selection, FSS) เพื่อหาคุณลักษณะที่เกี่ยวข้องสำหรับการจดจำหน่วยทางโครงสร้างของยีน ตัวอย่างงานวิจัยได้แก่ [21-23] ข้อดีของการใช้ขั้นตอนวิธีประมาณการแจกแจงคือได้รายละเอียดเชิงลึกที่มากขึ้นที่เกี่ยวข้องกับแต่ละคุณลักษณะว่าคุณลักษณะใดที่มีความเกี่ยวข้องมาก เกี่ยวข้องน้อย หรือไม่เกี่ยวข้อง นอกจากนี้ มีงานวิจัยที่ใช้ขั้นตอนวิธีประมาณการแจกแจงในการออกแบบโปรตีน (Protein design) และการทำนายโครงสร้างของโปรตีน (Protein structure prediction) ตัวอย่างงานวิจัยได้แก่ [24-27] อย่างไรก็ตาม ยังไม่พบงานวิจัยที่ประยุกต์ใช้ขั้นตอนวิธีประมาณการแจกแจงสำหรับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ในแต่ละรุ่นประชากรของขั้นตอนวิธีประมาณการแจกแจงถูกสร้างจากแบบจำลองความน่าจะเป็นทำให้ขั้นตอนวิธีนี้อาจสูญเสียความหลากหลายของประชากรคำตอบและมีแนวโน้มลู่เข้าก่อนกำหนดหลังจากผ่านกระบวนการวิวัฒนาการไปเพียงไม่กี่รุ่น [28] งานวิจัยส่วนหนึ่งจึงนำเสนอกรอบการทำงานในลักษณะของขั้นตอนวิธีประมาณการแจกแจงแบบผสม (Hybrid EDAs) โดยการรวมขั้นตอนวิธีประมาณการแจกแจงเข้ากับวิธีการทางเมตาฮิวริสติก อื่น ๆ เพื่อแก้ไขข้อจำกัดนี้ เช่น ในปัญหาทางด้านการจัดตารางเวลาการไหลเวียนงาน (permutation scheduling flowshop) มีงานวิจัยจำนวนหนึ่งนำเสนอขั้นตอนวิธีประมาณการแจกแจงแบบผสม เช่น TSSB-HEDA [29] ใช้ทั้งแบบจำลองความน่าจะเป็นของขั้นตอนวิธีประมาณการแจกแจงและตัวดำเนินการทางพันธุกรรมของขั้นตอนวิธีเชิงพันธุกรรม งานวิจัย [30] ใช้ขั้นตอนวิธีประมาณการแจกแจงร่วมกับการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค และ งานวิจัย [31] รวมแนวคิดของระบบอาณาจักรมด (ant colony system, ACS) เข้ากับขั้นตอนวิธีประมาณการแจกแจง จากผลการทดลองพบว่าการทำงานร่วมกันดังกล่าวช่วยปรับปรุงประสิทธิภาพของขั้นตอนวิธีที่นำเสนอและให้ผลลัพธ์ที่ดีขึ้น

จากประสิทธิภาพที่ดีขึ้นเมื่อใช้ขั้นตอนวิธีแบบผสม งานวิจัยนี้จึงนำเสนอขั้นตอนวิธีประมาณการแจกแจงแบบผสมสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชื่อว่า Hybrid-EDAfold ซึ่งแตกต่างจากงานวิจัยที่ได้นำเสนอไป เนื่องจากขั้นตอนวิธีที่นำมาทำงานร่วมกันเป็นขั้นตอนวิธีประมาณการแจกแจงทั้งคู่ แต่ละขั้นตอนวิธีอยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจงในกลุ่มที่ตัวแปรไม่ขึ้นต่อกัน (univariate) ซึ่งอยู่บนสมมุติฐานที่ว่าทุกตัวแปรอิสระ สมมุติฐานนี้อาจไม่เป็นจริงในบริบทของโครงสร้างทุติยภูมิอาร์เอ็นเอแต่เนื่องจากความง่ายและใช้ต้นทุนในการคำนวณต่ำ ผู้วิจัยจึงเลือกใช้กรอบการทำงานนี้ อ้างอิงจากหลาย ๆ งานวิจัย พบว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพของขั้นตอนวิธีประมาณการแจกแจง คือ การคัดเลือกกลุ่มประชากรย่อย และการใช้ความรู้จากกลุ่มประชากรย่อยนั้นในการสร้างและปรับปรุงแบบจำลองความน่าจะเป็น งานวิจัยนี้จึงออกแบบให้แต่ละขั้นตอนวิธีประมาณการแจกแจงมีพฤติกรรมของทั้งสองกระบวนการนี้แตกต่างกันเพื่อรองรับการค้นหาทั้งในระดับโกลบอล (global) และ ระดับโลคอล (local) กล่าวคือ ในขั้นตอนการคัดเลือกงานวิจัยนี้จะใช้ทั้งกลุ่มคำตอบย่อยที่มีค่าความเหมาะสมดีสุดและด้อยสุด n ลำดับแรก ซึ่งแตกต่างจากขั้นตอนวิธีประมาณการแจกแจงมาตรฐานทั่วไปที่มักจะใช้แค่กลุ่มคำตอบย่อยที่มีค่าความเหมาะสมดีเป็นลำดับต้น ๆ เท่านั้น และใช้ความรู้จากทั้งสองกลุ่มคำตอบนี้ในการปรับปรุงแบบจำลองความน่าจะเป็นเพื่อนำทางการค้นหาไปในทิศทางของกลุ่มคำตอบดีและออกห่างจากกลุ่มคำตอบด้อย

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

และในกระบวนการสร้างประชากรรุ่นถัด ๆ ไปของแต่ละขั้นตอนวิธีก็มีพฤติกรรมที่แตกต่างกัน โดยขั้นตอนวิธีประมาณการแจกแจงตัวแรกจะดำเนินการแบบขั้นตอนวิธีประมาณการแจกแจงมาตรฐาน คือสุ่มสร้างประชากรจากแบบจำลองความน่าจะเป็นโดยตรง แต่สำหรับขั้นตอนวิธีประมาณการแจกแจงอีกตัวหนึ่งจะมีการเพิ่มเติมตัวดำเนินการไขว้เปลี่ยนเพื่อเพิ่มความสามารถในการค้นหาแบบโลคอลโดยการไขว้เปลี่ยนจะเป็นแบบหลายตำแหน่งและพึ่งพาอยู่บนแบบจำลองความน่าจะเป็นของขั้นตอนวิธีประมาณการแจกแจงแทนการใช้ความน่าจะเป็นของการไขว้เปลี่ยน

นอกจากนี้ ขั้นตอนวิธี Hybrid-EDAfold ที่งานวิจัยนี้นำเสนอยังรองรับการทำนายหลายโครงสร้าง กล่าวคือรายงานผลการทำนายทั้งโครงสร้างที่มีค่าพลังงานต่ำสุดและต่ำรองลงมา อ้างอิงจากหลาย ๆ งานวิจัยที่อยู่บนพื้นฐานของการทำนายโครงสร้างที่มีค่าพลังงานต่ำสุด พบว่า ในบางอาร์เอ็นเอโครงสร้างที่พบในทางธรรมชาติอาจไม่ใช่โครงสร้างที่มีค่าพลังงานต่ำสุด [32] ดังนั้นการรองรับการทำนายหลายโครงสร้างจะช่วยลดข้อจำกัดที่เกิดจากความไม่สมบูรณ์ของพารามิเตอร์ที่ใช้ในการคำนวณค่าพลังงานและส่งเสริมให้โปรแกรมทำนายโครงสร้างสามารถกำหนดโครงสร้างได้ใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากยิ่งขึ้น และเพื่อประเมินประสิทธิภาพของขั้นตอนวิธีที่นำเสนอในแง่ของความถูกต้องและการรองรับการทำนายโครงสร้างของข้อมูลอาร์เอ็นเอที่หลากหลาย ขั้นตอนวิธี Hybrid-EDAfold ถูกทดสอบด้วยอาร์เอ็นเอ 14 ชนิดจากฐานข้อมูล RNA STARND v2.0 [33] และข้อมูล pre-miRNA ของมนุษย์ที่รวบรวมจากงานวิจัย [34] โดยทำการเปรียบเทียบกับทั้งโปรแกรมที่ได้รับความนิยมใช้งานในปัจจุบันที่อยู่บนหลักการของกำหนดการพลวัต และ ขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ เพื่อมุ่งหวังว่าขั้นตอนวิธีที่นำเสนอจะสามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอได้ความถูกต้องที่มากขึ้น หรือ เทียบเคียงได้กับประสิทธิภาพที่พบในขั้นตอนวิธีอื่น ๆ และเนื่องจากวิธีการที่นำเสนออยู่บนพื้นฐานขั้นตอนวิธีประมาณการแจกแจง ข้อดีที่เหนือกว่าวิธีการเมตาฮิวริสติกอื่น ๆ คือความยืดหยุ่นของแบบจำลองความน่าจะเป็น กล่าวคือ นักชีวสารสนเทศสามารถจัดการกับแบบจำลองความน่าจะเป็นเพื่อให้ได้คำตอบที่น่าพึงพอใจมากขึ้นด้วยการกำหนดค่าบางส่วนไว้ล่วงหน้า (pre-established partial configurations) นอกจากนี้ แบบจำลองความน่าจะเป็นที่ถูกสร้างในระหว่างกระบวนการค้นหาสามารถถูกสำรวจเพื่อเปิดเผยข้อมูลที่เป็นประโยชน์เกี่ยวกับปัญหานั้น [35, 36] ถือเป็นอีกทางเลือกหนึ่งที่น่าสนใจในการใช้กรอบการทำงานนี้เพื่อช่วยในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ หรือเป็นแนวทางในการประยุกต์ใช้สำหรับวิเคราะห์ข้อมูลในงานทางด้านชีวสารสนเทศอื่น ๆ ต่อไปในอนาคต

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 นำเสนอขั้นตอนวิธีเชิงวิวัฒนาการสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ

1.2.2 วิเคราะห์ประสิทธิภาพการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชนิดต่าง ๆ ของขั้นตอนวิธีที่นำเสนอ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

1.3 ขอบเขตการวิจัย

1.3.1 นำเสนอขั้นตอนวิธีการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจาก 1 สายลำดับอาร์เอ็นเอ (RNA sequence)

1.3.2 ศึกษาการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชนิดต่าง ๆ จาก 2 ฐานข้อมูล

- microRNA (miRNA) จากฐานข้อมูล miRBase [37]
- transfer RNA (tRNA) ribosomal RNA (rRNA) จากฐานข้อมูล RNA STRAND 2.0 [33]

1.3.3 แสดงผลการทำนายโครงสร้างในรูปแบบสัญลักษณ์จุดและวงเล็บ (Dot-Bracket Notation, DBN)

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย

1.4.1 ศึกษาข้อมูลเกี่ยวกับโครงสร้างของอาร์เอ็นเอ

1.4.2 ศึกษางานวิจัยเกี่ยวกับการกำหนดโครงสร้างทุติยภูมิของอาร์เอ็นเอ

1.4.3 วิเคราะห์และออกแบบขั้นตอนวิธีสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ

1.4.4 พัฒนาขั้นตอนวิธีสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ

1.4.5 ทดสอบและประเมินผลขั้นตอนวิธีที่นำเสนอ

1.4.6 ปรับปรุงประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ

1.4.7 สรุปผลและจัดทำวิทยานิพนธ์

1.4.8 เผยแพร่ผลงานตีพิมพ์

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1.5.1 ได้ขั้นตอนวิธีแบบใหม่สำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชื่อว่า Hybrid-EDAfold ซึ่งเป็นขั้นตอนวิธีที่อยู่บนพื้นฐานของขั้นตอนวิธีประมาณแจกแจง

1.5.2 ขั้นตอนวิธี Hybrid-EDAfold สามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชนิดต่าง ๆ ได้หลากหลาย และให้ค่าความถูกต้องในการทำนายที่ดีกว่าหรือเทียบเคียงได้กับผลการทำนายโครงสร้างด้วยกำหนดการพลวัต และ วิธีการเมตาฮิวริสติกอื่น ๆ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้ประกอบด้วย 2 ส่วน ส่วนแรกอธิบายทฤษฎีพื้นฐานที่เกี่ยวข้องกับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ ประกอบด้วย ทฤษฎีเกี่ยวกับอาร์เอ็นเอนำเสนอในหัวข้อ 2.1.1 ชนิดของอาร์เอ็นเอนำเสนอในหัวข้อ 2.1.2 โครงสร้างทุติยภูมิของอาร์เอ็นเอนำเสนอในหัวข้อ 2.1.3 การกำหนดโครงสร้างทุติยภูมิของอาร์เอ็นเอนำเสนอในหัวข้อ 2.1.4 และ ขั้นตอนวิธีประมาณการแจกแจงนำเสนอใน 2.1.5 ส่วนที่สองนำเสนองานวิจัยที่เกี่ยวข้องกับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอประกอบด้วยกลุ่มงานวิจัยที่ทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจาก 1 สายลำดับ และกลุ่มงานวิจัยที่ทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจากหลายสายลำดับ รายละเอียดเป็น ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 กรดไรโบนิวคลีอิก หรือ อาร์เอ็นเอ

อาร์เอ็นเอเป็นสารประกอบที่มีความสำคัญต่อสิ่งมีชีวิต หน้าที่หลักของอาร์เอ็นเอคือการถอดรหัสข้อมูลพันธุกรรมจากดีเอ็นเอและเปลี่ยนรหัสพันธุกรรมนั้นไปเป็นโปรตีน อาร์เอ็นเอส่วนใหญ่พบอยู่ในไซโทพลาสซึม มีโครงสร้างเป็นสายเดี่ยวแต่อาจเกิดการพับกลับเข้าหาตัวเองและเกิดการจับคู่กันของเบสภายในสายนั้นได้

โมเลกุลอาร์เอ็นเอประกอบด้วยนิวคลีโอไทด์มากมายเรียงต่อกันเป็นสายพอลิเมอร์และเชื่อมต่อกันด้วยพันธะฟอสโฟไดเอสเธอร์ (phosphodiester bond) โดยแต่ละนิวคลีโอไทด์ประกอบด้วย น้ำตาลไรโบส หมู่ฟอตเฟต และ เบส โดยอาร์เอ็นเอมีเบสที่แตกต่างกัน 4 ชนิด ได้แก่ อะดินีน (Adenine) กวานีน (Guanine) ไซโตซีน (Cytosine) และ ยูราซิล (Uracil)

โดยปกติมีการจับคู่กันของเบสอะดินีนกับเบสยูราซิล (A-U) และ เบสกวานีนกับเบสไซโตซีน (G-C) เรียกว่าการจับคู่เบสแบบวัทสัน-คริก (Watson-Crick base pairs) และการจับคู่กันของเบสกวานีนกับเบสยูราซิล (G-U) ก็สามารถเกิดขึ้นได้แต่ได้รับความพึงพอใจในแง่ของพลังงานน้อยกว่า ดังนั้น คู่เบสนี้จึงถูกเรียกว่าคู่เบสวอบเบิล (wobble base pair) และการจับคู่กันของเบสทั้ง 2 กลุ่มถูกเรียกรวมว่าคู่เบสคาโนนิคอล (canonical base pairs)

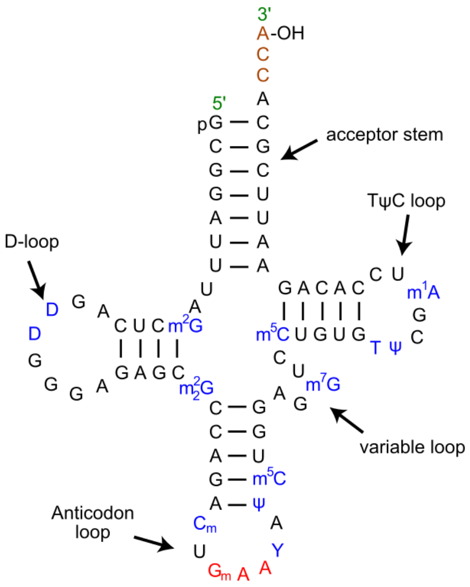
โครงสร้างของอาร์เอ็นเอมีทิศทางจาก 5’ ไป 3’ กล่าวคือ ปลายของนิวคลีโอไทด์ตัวเริ่มต้นจะมีฟอตเฟตอยู่ตำแหน่งที่ 5 ของน้ำตาลตัวแรกปลายนี้จึงเรียกว่าปลาย 5’ ส่วนอีกปลายหนึ่งจะมี 3-OH ของน้ำตาลตัวสุดท้ายซึ่งไม่เชื่อมกับหมู่ฟอสเฟตปลายนี้เรียกว่าปลาย 3’

2.1.2 ชนิดของอาร์เอ็นเอ

ชนิดของอาร์เอ็นเอสามารถแบ่งออกเป็น 3 กลุ่มใหญ่ตามการทำงานและโครงสร้าง ได้แก่ อาร์เอ็นเอนำรหัส อาร์เอ็นเอส่งถ่าย และ อาร์เอ็นเอไรโบโซม

โดยอาร์เอ็นเอนำรหัสทำหน้าที่ในการขนส่งข้อมูลพันธุกรรมจากดีเอ็นเอไปที่ไรโบโซม และเป็นโมเลกุลอาร์เอ็นเอที่ใหญ่ที่สุด โดยอาร์เอ็นเอนำรหัสถูกค้นพบโดยนักวิทยาศาสตร์ 2 คน คือ Elliot Volkin และ Lazarus Astachan ในปี 1956 โดยดีเอ็นเอถูกคัดลอกไปเป็นอาร์เอ็นเอนำรหัส จากนั้นถูกถอดรหัสไปเป็นโปรตีน ซึ่ง 1 โมเลกุลของอาร์เอ็นเอนำรหัสถูกใช้เพื่อเข้ารหัสสารสนเทศของ 1 โปรตีน แต่สำหรับแบคทีเรียหลายสารสนเทศของโปรตีนสามารถถูกเข้ารหัสด้วย 1 โมเลกุลของอาร์เอ็นเอนำรหัส [38]

อาร์เอ็นเอส่งถ่ายมีบทบาทในลักษณะเป็นการเชื่อมต่อทางกายภาพระหว่างดีเอ็นเอและสายลำดับอาร์เอ็นเอของกรดนิวคลีอิกและสายโปรตีนของกรดอะมิโน อาร์เอ็นเอส่งถ่ายช่วยถอดรหัสของอาร์เอ็นเอนำรหัส โดยอาร์เอ็นเอส่งถ่ายเป็นองค์ประกอบที่สำคัญของการแปลโปรตีนและมีจำนวน 75-95 นิวคลีโอไทด์ [39] ตัวอย่างโครงสร้างทุติยภูมิของอาร์เอ็นเอส่งถ่ายเป็นดังรูปที่ 2.1 (https://en.wikipedia.org/wiki/Transfer_RNA)



รูปที่ 2.1 ตัวอย่างโครงสร้างทุติยภูมิของอาร์เอ็นเอส่งถ่าย

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ในอาร์เอ็นเอส่งถ่ายแต่ละโมเลกุลจะมีเบสส่วนหนึ่งที่สามารถจับคู่กับเบสในรหัสของอาร์เอ็นเอนำรหัส เบสส่วนนี้ถูกเรียกว่าแอนติโคดอนประกอบด้วยเบส 3 ตัวที่เข้าคู่กับ 1 โคดอนบนอาร์เอ็นเอนำรหัส นอกจากนี้ อาร์เอ็นเอส่งถ่ายจะมีเบสส่วนหนึ่งทำหน้าที่พ่วงกรดอะมิโนที่มีความสัมพันธ์กับแอนติโคดอนนั้น เช่น อาร์เอ็นเอส่งถ่ายที่มีแอนติโคดอนเป็น UAG จะจับคู่กับโคดอน AUC ของอาร์เอ็นเอนำรหัสและพ่วงกรดอะมิโนไอโซลิวซีน

อาร์เอ็นเอไรโบโซมเป็นหนึ่งในองค์ประกอบของไรโบโซมซึ่งจำเป็นต่อการสังเคราะห์โปรตีน เนื่องจากฟังก์ชันที่สำคัญของอาร์เอ็นเอไรโบโซมการสังเคราะห์โปรตีนจึงเกิดขึ้น ไรโบโซมมีอยู่ในทุกสิ่งมีชีวิตและช่วยแปลสารสนเทศในอาร์เอ็นเอนำรหัสไปเป็นโปรตีน ไรโบโซมมีอาร์เอ็นเอไรโบโซมประมาณ 60% [40]

โปรคารีโอต สามารถแบ่งโมเลกุลของอาร์เอ็นเอไรโบโซมได้เป็น 3 ชนิด คือ

- 23S rRNA มีขนาดประมาณ 2,904 นิวคลีโอไทด์
- 16S rRNA มีขนาดประมาณ 1,541 นิวคลีโอไทด์
- 5S rRNA มีขนาดประมาณ 120 นิวคลีโอไทด์

ยูคารีโอต สามารถแบ่งโมเลกุลของอาร์เอ็นเอไรโบโซมได้เป็น 4 ชนิด คือ

- 28S rRNA มีขนาดประมาณ 4,718 นิวคลีโอไทด์
- 18S rRNA มีขนาดประมาณ 1,874 นิวคลีโอไทด์
- 5.8S rRNA มีขนาดประมาณ 160 นิวคลีโอไทด์
- 5S rRNA มีขนาดประมาณ 120 นิวคลีโอไทด์

นอกเหนือจากอาร์เอ็นเอ 3 ชนิดหลักที่นำเสนอไปในข้างต้น ยังมีอาร์เอ็นเอชนิดอื่น ๆ ที่มีบทบาทสำคัญ [41] รายละเอียดเป็น ดังนี้

- Non-coding RNA (ncRNA) เป็นโมเลกุลของอาร์เอ็นเอที่ไม่ถูกเข้ารหัสไปเป็นโปรตีนแต่ non-coding RNA ยังคงประกอบด้วยสารสนเทศที่สำคัญและมีหลายฟังก์ชัน ฟังก์ชันหนึ่งของ ncRNA คือควบคุมการแสดงออกของยีนที่ระดับการคัดลอก (transcription level) [42]

- Transfer-Messenger RNA (tmRNA) เป็นโมเลกุลอาร์เอ็นเอที่มีคุณลักษณะเหมือนกับทั้งอาร์เอ็นเอส่งถ่ายและอาร์เอ็นเอนำรหัส โดย tmRNA ทำงานเหมือนเป็นระบบควบคุมคุณภาพที่คอยตรวจสอบการสังเคราะห์โปรตีน [43]

- Small nuclear RNA (snRNA) มีโมเลกุลอาร์เอ็นเอเล็ก ๆ และเกี่ยวข้องในการจับคู่หรือมีปฏิกิริยาอื่น ๆ ของอาร์เอ็นเอ ความยาวโดยประมาณของ snRNA คือ 150 นิวคลีโอไทด์ snRNA ที่ซับซ้อนถูกเรียกว่า snRNP ซึ่งเก็บรักษา RNA-protein complexes โดยทั่วไปเรียกว่า snurps [44]

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

- microRNA (miRNA) มักมาจากกลุ่มของ non-coding RNA มีบทบาทสำคัญในการควบคุมการแสดงออกของยีนโดยเข้าไปจับกับอาร์เอ็นเอนำรหัสบริเวณตำแหน่งที่เป็นคู่สมกันซึ่งทำให้กระบวนการอ่านรหัสและถอดรหัสจากอาร์เอ็นเอนำรหัสนั้น ๆ ถูกยับยั้ง ปัจจุบันนักวิจัยหลายกลุ่มค้นพบว่า microRNA สามารถถูกใช้เป็นตัวบ่งชี้ทางชีวภาพในการเกิดโรคต่าง ๆ ของมนุษย์โดยเฉพาะโรคมะเร็ง โดย miRNA เป็นโมเลกุลอาร์เอ็นเอสายเดี่ยวและมีความยาวค่อนข้างสั้นประมาณ 22 นิวคลีโอไทด์ ทั้งสัตว์และพืชมี microRNA [45]

- Small interfering RNA (siRNA) มักถูกเรียกว่า Silencing RNA หรือ short interfering RNA อาร์เอ็นเอชนิดนี้ทำหน้าที่ปิดการทำงานของยีนในช่วงเวลาสั้น ๆ siRNA เป็น synthetic RNA ที่ถูกสร้างจากโมเลกุลอาร์เอ็นเอสายคู่ที่มีความยาว 20-25 คู่เบส ถูกใช้สำหรับหยุดยีนที่เป็น non-protein coding [46]

- Small nucleolar RNA (snoRNA) เป็นกลุ่มพิเศษของ small RNA molecule ที่มีความสำคัญต่อการเปลี่ยนแปลงทางเคมีของอาร์เอ็นเอชนิดอื่น เช่น อาร์เอ็นเอในกลุ่มของอาร์เอ็นเอไรโบโซม อาร์เอ็นเอส่งถ่าย และ snoRNA มีการนำเสนอว่าอาร์เอ็นเอเหล่านี้มาจากการวิวัฒนาการในการทำสำเนายีนของอาร์เอ็นเอส่งถ่าย [47]

- Antisense RNA (asRNA) เป็นอาร์เอ็นเอสายเดี่ยวที่สัมพันธ์กับอาร์เอ็นเอนำรหัส บางครั้งอาร์เอ็นเอชนิดนี้ถูกเรียกว่า mRNA-interfering complementary (micRNA) แต่ micRNA ไม่ได้รับความนิยมและไม่ถูกนำไปใช้อย่างแพร่หลาย โดย asRNA ถูกใช้เพื่อหยุดการควบคุมยีนโดยการยับยั้งกระบวนการการแสดงออกของยีน [48]

- Signal recognition particle RNA (SRP RNA) เป็นสารประกอบไรโบนิวคลีโอโปรตีน (RNP complex) ที่พบได้ทั่วไปซึ่งมีผลต่อการหลั่งโปรตีนและเยื่อหุ้มเซลล์และจำเป็นต่อการร่วมแปล (co-translational) โปรตีนที่เป็นเป้าหมาย [49]

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

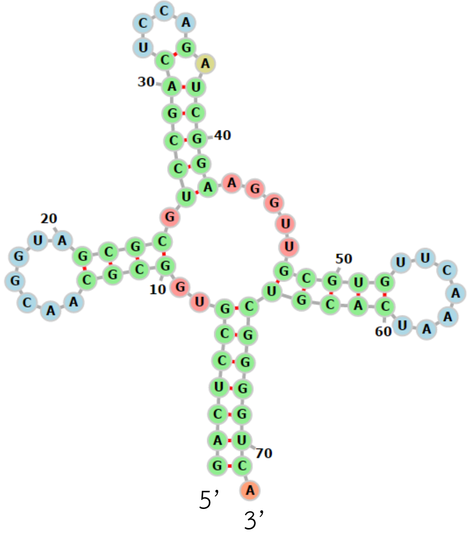
2.1.3 โครงสร้างทุติยภูมิของอาร์เอ็นเอ

โครงสร้างของอาร์เอ็นเอเป็นลำดับชั้น กล่าวคือ โครงสร้างปฐมภูมิเป็นลำดับเบสเรียงต่อกันเป็นสาย เบสแต่ละตัวถูกแทนด้วยอักษร A, C, G และ U โครงสร้างทุติยภูมิจะประกอบด้วยการจับคู่ของเบสต่าง ๆ อ้างอิงตามคู่เบสคาโนนิคอล โครงสร้างตติยภูมิคือโครงสร้างสามมิติของโมเลกุลอาร์เอ็นเอ

โมเลกุลอาร์เอ็นเอมีแนวโน้มชื่นชอบโครงสร้างทุติยภูมิที่มีจำนวนคู่เบสมากสุด และเบสต้องเข้าคู่กันอย่างเป็นระเบียบในลักษณะที่ไม่ทับซ้อนกับคู่เบสอื่น สัญลักษณ์ที่ใช้แทนโครงสร้างทุติยภูมิประกอบด้วย 3 ตัวอักษร “(“ , “)” และ “.” โดยวงเล็บเปิดและวงเล็บปิดแทนนิวคลีโอไทด์ที่มีการจับกันเป็นคู่เบส ในขณะที่ “จุด” แทนนิวคลีโอไทด์ที่ไม่ได้ถูกจับคู่ ด้วยวิธีการเช่นนี้ ทุก ๆ โครงสร้างทุติยภูมิที่ถูกต้องสามารถถูกแทนอย่างมีลักษณะเฉพาะด้วยสัญลักษณ์จุดและวงเล็บด้วยความยาวที่เท่ากับจำนวนนิวคลีโอไทด์ที่ปรากฏในสายลำดับอาร์เอ็นเอดังรูปที่ 2.2

GACUCCGUGGCGCAACGGUAGCGCGUCCGACUCCAGAUCGGAAGGUUGCGUGUUCAAAUCACGUCGGGGUCA

(((((((..((((.......)))).((((((....).))))).....(((((.......)))))))))))).

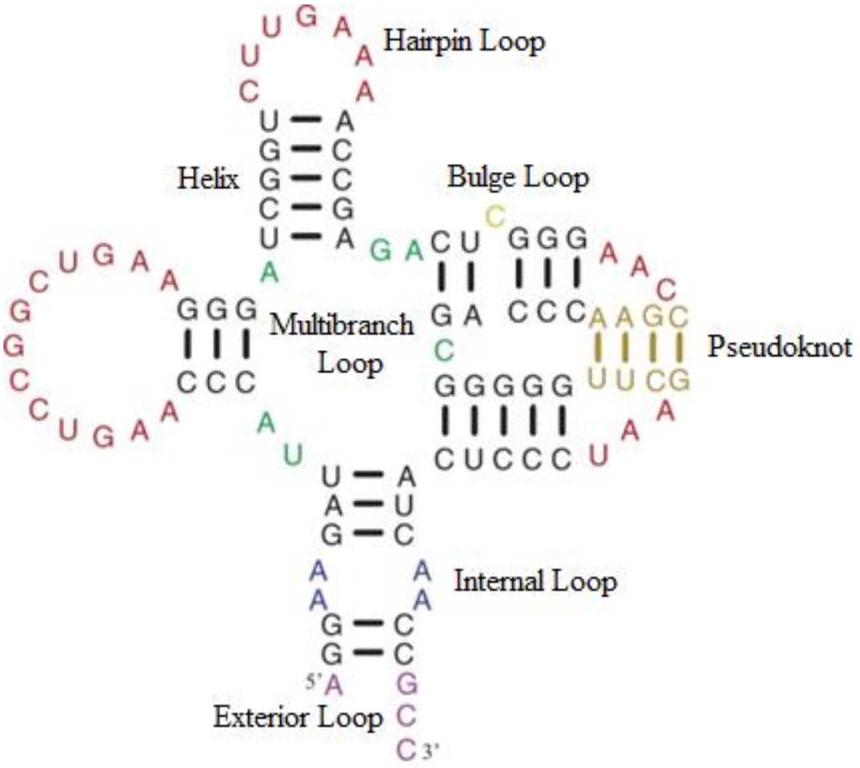


รูปที่ 2.2 การแทนโครงสร้างทุติยภูมิในรูปแบบสัญลักษณ์จุดและวงเล็บ

รูปร่างพื้นฐานของโครงสร้างทุติยภูมิของอาร์เอ็นเอเป็นดังรูปที่ 2.3 [50] จากรูปบริเวณที่เรียกว่าฮีลิก (helix) คือ บริเวณที่มีการจับคู่ของเบสอ้างอิงตามคู่เบสคาโนนิคอลเรียงซ้อนกันไป ส่วนบริเวณที่เรียกว่าลูป (loop) คือ ส่วนของนิวคลีโอไทด์ที่ไม่ได้จับคู่กัน โดยลูปชนิดแฮร์พิน (hairpin loop) จะมีเพียง 1 ฮีลิก ในขณะที่ลูปชนิดอินเทอร์นอล (internal loop) และ ลูปชนิดบัลจ์ (bulge loop) จะมี 2 ฮีลิก ซึ่งลูปชนิดอินเทอร์นอลจะมีนิวคลีโอไทด์ที่ไม่ได้จับคู่อยู่ทั้งสองด้านของลูป ส่วนลูปชนิดบัลจ์นั้นจะมีนิวคลีโอไทด์ที่ไม่ได้จับคู่อยู่เพียงด้านใดด้านหนึ่งของลูป และ ลูปชนิดมัลติบรานซ์

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

(multibranch loop) หรือบางครั้งเรียกว่าทางเชื่อมฮีลิก (helical junction) จะมีตั้งแต่ 3 ฮีลิกขึ้นไป และ ลูปชนิดเอ็กเทียเรียร์ (exterior loop) จะเป็นส่วนปลายของสายลำดับเบสและมี 1 ฮีลิกหรือมากกว่า บริเวณที่เรียกว่าซูโดนอท (pseudoknots) คือ คู่เบสคาโนนิคอลที่มีการเชื่อมบริเวณที่เป็นลูปหนึ่งมาติดกับลูปอื่น โดยหลักการซูโดนอทเกิดขึ้นเมื่อมีอย่างน้อย 2 คู่เบส แทนด้วย i - j และ i'-j' ที่เป็นไปตามเงื่อนไข คือ ตำแหน่งของนิวคลีโอไทด์เป็น ดังนี้ $i < i' < j < j'$



รูปที่ 2.3 รูปร่างพื้นฐานที่สามารถพบได้ในโครงสร้างทุติยภูมิของอาร์เอ็นเอ [50]

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.1.4 การกำหนดโครงสร้างทุติยภูมิของอาร์เอ็นเอ

จากความสำคัญของโครงสร้างทุติยภูมิและฟังก์ชันการทำงานของอาร์เอ็นเอชนิดต่าง ๆ วิธีการกำหนดโครงสร้างทุติยภูมิของอาร์เอ็นเอสามารถแบ่งออกเป็น 2 กลุ่มหลัก ๆ คือ วิธีเชิงการทดลอง และวิธีเชิงการคำนวณที่ดำเนินการกับสายลำดับของอาร์เอ็นเอ ในหัวข้อนี้จะนำเสนอวิธีการดำเนินการคร่าว ๆ ของแต่ละวิธี รายละเอียดดังนี้

2.1.4.1 วิธีเชิงการทดลอง

วิธีเชิงการทดลองสำหรับศึกษาโครงสร้างทุติยภูมิของอาร์เอ็นเอประกอบด้วย วิธีทางชีวเคมี เช่น RNase footprinting และ chemical probing และ วิธีการทางฟิสิกส์ เช่น X-ray crystallography และ นิวเคลียร์แมกเนติกเรโซแนนซ์สเปกโทรสโคปี [1] แม้ว่าวิธีเหล่านี้จะมีความแตกต่างกันในแง่ของกลไกและการดำเนินการ แต่ผลลัพธ์สุดท้ายเป็นไปในทำนองเดียวกันคือมีคุณภาพของการทำนายที่สูงแต่ก็แลกมาด้วยความพยายามที่มาก (low throughput) แต่ละวิธีมีการดำเนินการ ดังนี้

1. X-ray crystallography

วิธีนี้ดำเนินการโดยนำอาร์เอ็นเอบริสุทธิ์ที่ต้องการศึกษามาทำการตกผลึกเพื่อให้ได้คริสตัล จากนั้นฉายรังสีเอกซ์ (x-ray) ไปยังคริสตัลนั้นแล้วทำการวิเคราะห์รูปแบบการหักเหของแสงจากความหนาแน่นของอิเล็กตรอน ซึ่งแสงที่เกิดการรวมตัวกันจนมีความเข้มสูง ๆ สามารถนำเครื่องตรวจจับ (detector) มาวัดสังเกตเห็นเป็นจุดสีดำ โดยจุดดำเหล่านี้ทำหน้าที่เปรียบเหมือนเป็นแผนที่ลายแทงที่ช่วยบอกว่าอิเล็กตรอนใดบนโมเลกุลที่ก่อให้เกิดการหักเหของแสงบนเครื่องตรวจจับ จากนั้นทำการวิเคราะห์โดยผู้เชี่ยวชาญเพื่อกำหนดโครงสร้างของอาร์เอ็นเอต่อไป วิธีการนี้ใช้เวลาและความพยายามมากเนื่องจากจำนวนของการตกผลึกที่จำเป็นต้องทดสอบเพื่อผลิตคริสตัลที่จะนำไปสร้างข้อมูลการหักเหของแสงมีจำนวนค่อนข้างมาก [51]

2. นิวเคลียร์แมกเนติกเรโซแนนซ์สเปกโทรสโคปี

วิธีการนี้เป็นเทคนิคที่เกี่ยวข้องกับการวัดระดับพลังงานที่แตกต่างกันของนิวเคลียสที่อยู่ภายใต้อิทธิพลของสนามแม่เหล็ก ประกอบด้วยหลายวิธีแต่หลักการคือชนิดที่แตกต่างกันของนิวเคลียสจะคายคุณลักษณะทางเคมีที่แตกต่างกันส่งผลให้เกิดการเคลื่อนย้ายความถี่เมื่อถูกฉายด้วยสนามแม่เหล็ก ข้อมูลที่มีการเคลื่อนย้ายเหล่านี้สามารถถูกใช้เพื่อศึกษาพลังงานและการเคลื่อนที่ของอาร์เอ็นเอ [52] วิธีการนี้มีประสิทธิภาพมากแต่ก็ให้ปริมาณงานในหนึ่งหน่วยเวลาที่ต่ำ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3. Chemical probing

สารเคมีต่าง ๆ สามารถเปลี่ยนแปลงอาร์เอ็นเอและการเปลี่ยนแปลงนี้สามารถถูกอธิบายผ่านการประเมินคุณลักษณะของโครงสร้าง เช่น พันธะไฮโดรเจน การเข้าถึงตัวทำละลาย และ การเข้าถึงตำแหน่งนิวคลีโอไทด์ การวิเคราะห์ความไวปฏิกิริยาต่อเบสร่วมกับเทคนิคที่อยู่บนพื้นฐานของค่าพลังงาน (free energy-based modeling) สามารถถูกใช้เพื่ออนุมานโครงสร้างทุติยภูมิ [53]

คุณสมบัติทางเคมีของ SHAPE (Selective 2′ Hydroxyl Acylation analyzed by Primer Extension, SHAPE) ถูกใช้อย่างกว้างขวางเพื่อศึกษาโครงสร้างทุติยภูมิของหลาย ๆ อาร์เอ็นเอ ความถูกต้องของการทำนายโครงสร้างทุติยภูมิด้วยวิธีการนี้สูงมากและมีประสิทธิภาพเทียบเคียงได้กับวิธีการกำหนดโครงสร้างเชิงการคำนวณที่ดีที่สุดที่สามารถพบได้ขณะนี้ [1]

4. RNase footprinting

เป็นวิธีการทางชีวเคมีเพื่อพิจารณาโครงสร้างทุติยภูมิของอาร์เอ็นเอที่ใช้ประโยชน์จากไรโบนิวเคลียส (RNases) [54] ที่สามารถแยกบริเวณที่เป็นน้ำตาลทั้งในส่วนของคู่เบส เช่น RNase V1 หรือ ส่วนที่ไม่ใช่คู่เบส เช่น RNases ONE, T1 และ A โดยบริเวณที่มีการแยกจะถูกทำให้เห็นโดยการบันทึกภาพรังสีหรือกระบวนการย้อนการถอดรหัส (reverse transcription)

2.1.4.2 วิธีเชิงการคำนวณสำหรับทำนายโครงสร้างทุติยภูมิด้วย 1 สายลำดับ

ในกรณีที่ไม่มีข้อมูลเกี่ยวกับรูปร่างลักษณะที่คล้ายกัน (homolog) ของสายลำดับ วิธีการที่ได้รับความนิยมมากสุด คือ การทำนายโครงสร้างทุติยภูมิจาก 1 สายลำดับ [55]

1. เทคนิคการค่าพลังงานต่ำสุด (Free Energy Minimization, MFE)

เป็นวิธีการทำนายโครงสร้างทุติยภูมิที่ได้รับความนิยม โดยค่าพลังงาน (free energy) สามารถถูกประเมินโดยใช้แบบจำลองเพื่อนบ้านที่ใกล้ที่สุด (nearest neighbor model) ซึ่งแบบจำลองนี้อยู่บนสมมุติฐานที่ว่าการเปลี่ยนแปลงค่าพลังงานสำหรับ 1 คู่เบสขึ้นอยู่กับคู่เบสนั้นกับคู่เบสเพื่อนบ้านที่ติดกัน และ ค่าพลังงานที่สัมพันธ์กับบริเวณที่เป็นลูปและ รูปร่างเชิงโครงสร้าง (motif) อื่น ๆ ไม่ได้ขึ้นกับปัจจัยภายนอกลูปนั้น ดังนั้นค่าพลังงานสำหรับ 1 โครงสร้างสามารถคำนวณได้ง่าย ๆ โดยการหาผลรวมของทุกค่าพลังงานที่สัมพันธ์กับบริเวณที่เป็นฮีลิกและรูปร่างเชิงโครงสร้างอื่น ๆ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

พารามิเตอร์ที่ใช้ในการคำนวณถูกกำหนดจากการทดลองหลอมละลายด้วยแสง (optical melting experiments) โดยโครงสร้างที่ถูกทำนายว่ามีค่าพลังงานต่ำสุดสามารถคำนวณด้วยกำหนดการพลวัต โดยพิจารณาทุกโครงสร้างที่เป็นไปได้และรับประกันคำตอบเหมาะสมสุด

กำหนดการพลวัตที่หาโครงสร้างที่มีค่าพลังงานต่ำสุดใช้เวลา $O(N^3)$ เมื่อ N คือความยาวของสายลำดับอาร์เอ็นเอหรือจำนวนนิวคลีโอไทด์ในสายลำดับอาร์เอ็นเอ

โครงสร้างที่ถูกทำนายสำหรับสายลำดับที่ยาวไม่เกิน 800 นิวคลีโอไทด์ด้วยเทคนิคการหาค่าพลังงานต่ำสุดมีค่าเฉลี่ยของความอ่อนไหว (sensitivity) เท่ากับ 74% ซึ่งคำนวณจากสัดส่วนจำนวนคู่เบสที่ทำนายถูกต้องตรงกับจำนวนคู่เบสที่พบในโครงสร้างที่เป็นคำตอบ และค่าความจำเพาะ (specificity) เท่ากับ 66% ซึ่งคำนวณจากสัดส่วนจำนวนคู่เบสที่ทำนายได้ถูกต้องเทียบกับจำนวนคู่เบสทั้งหมดที่พบในโครงสร้างที่ทำนายได้

ในอาร์เอ็นเอชนิดต่าง ๆ เมื่อสายลำดับยาวขึ้นค่าความถูกต้องที่ได้จะต่ำลง เช่น การทำนายโครงสร้างของ rRNA มีค่าความอ่อนไหวเป็น 47.1% และ ค่าความจำเพาะเป็น 56.2% [56]

2. การทำนายความน่าจะเป็นของคู่เบส (Predicting base pair probabilities)

การทำนายโครงสร้างด้วยการคำนวณค่าพลังงานต่ำสุดอยู่ภายใต้สมมุติฐาน 3 ข้อ

1) อาร์เอ็นเออยู่ในสภาวะสมดุล

2) สายลำดับอาร์เอ็นเอนั้นสร้างได้เพียง 1 โครงสร้าง และ

3) พารามิเตอร์ของแบบจำลองเพื่อนบ้านใกล้เคียงสุดไม่มีความผิดพลาด

สมมุติฐานข้อที่ 3 ไม่เป็นความจริงเพราะจากการสังเกตพบผลกระทบของค่าพลังงานที่ไม่ใช่เพื่อนบ้านใกล้เคียงสุด (non-nearest neighbor) [57] นอกจากนี้ สมมุติฐาน 2 ข้อแรกก็อาจไม่เป็นความจริงสำหรับทุกสายลำดับอาร์เอ็นเอ [58]

โดยเฉลี่ยโครงสร้างทุติยภูมิที่ถูกทำนายจะมีคู่เบสที่ทำนายอย่างถูกต้องและคู่เบสที่ทำนายผิด เทคนิคการหาค่าพลังงานต่ำสุดสามารถถูกเติมเต็มให้มีความสมบูรณ์มากขึ้นด้วยการคำนวณฟังก์ชันพาร์ทิชัน (partition function) ซึ่งมีส่วนช่วยแนะนำว่าคู่เบสเหล่านั้นมีแนวโน้มที่จะจับคู่กันมากแค่ไหนเพื่อส่งผลให้สามารถทำนายตำแหน่งที่เบสจับคู่กันได้ถูกต้องมากยิ่งขึ้น โดยฟังก์ชันพาร์ทิชัน Q แทนผลรวมของค่าคงที่สมดุล (equilibrium constants) K_i ของทุกโครงสร้างที่เป็นไปได้ คำนวณได้ดังสมการที่ 2.1

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

$$Q = \sum_{all\ structure} K_i = \sum_{all\ structure} e^{-\Delta G_i^{\circ}/RT} \quad (2.1)$$

ดังนั้น ความน่าจะเป็นของโครงสร้าง *i* ที่จะถูกพบในโครงสร้างคำตอบสามารถคำนวณได้ดังสมการที่ 2.2

$$P_{\mathrm{i}} = \frac{e^{-\Delta G_i^{\circ}/RT}}{Q} \quad (2.2)$$

และความน่าจะเป็นของเบสตำแหน่งที่ *i* และ *j* จะจับคู่กัน สามารถคำนวณโดยการรวมค่าคงที่สมดุลของโครงสร้างที่มีคู่เบสนั้นแล้วหารด้วยฟังก์ชันพาร์ทิชันดังสมการที่ 2.3

$$P(i-j) = \sum_k \frac{e^{-\Delta G_i^{\circ}/RT}}{Q} \quad (2.3)$$

เมื่อ *k* คือจำนวนโครงสร้างที่มีคู่เบส *i* จับคู่กับ *j*

คู่เบสที่มีความน่าจะเป็นสูงอ้างอิงตามฟังก์ชันพาร์ทิชันเป็นคู่เบสที่มีความเป็นไปได้สูงที่จะถูกทำนายได้อย่างถูกต้อง การคำนวณฟังก์ชันพาร์ทิชันใช้เวลา $O(N^3)$ และการคำนวณฟังก์ชันพาร์ทิชันสามารถถูกใช้ร่วมกับเทคนิคการหาค่าพลังงานต่ำสุดเพื่อบ่งชี้บริเวณที่มีโอกาสทำนายได้อย่างถูกต้องและบริเวณที่มีโอกาสทำนายผิด รวมทั้งสามารถพิจารณาโครงสร้างในภาพรวมว่าโครงสร้างที่ทำนายได้มีความเป็นไปได้มากน้อยเพียงใด ดังนั้นข้อแนะนำคือในการคำนวณค่าพลังงานต่ำสุดสำหรับสายลำดับใด ๆ ควรเสริมด้วยการคำนวณฟังก์ชันพาร์ทิชันเพื่อให้คำอธิบายเพิ่มเติมสำหรับโครงสร้างที่ทำนายได้ด้วยความน่าจะเป็นของคู่เบส วิธีการนี้พบในทั้งโปรแกรม RNAstructure [59, 60] และโปรแกรมทำนายโครงสร้างที่อยู่ใน Vienna package

3. การทำนายโครงสร้างที่มีค่าความถูกต้องที่คาดหวังสูงสุด (Maximum Expected Accuracy Structures, MEA)

วิธีการนี้เป็นอีกทางเลือกหนึ่งของการทำนายโครงสร้าง คือการสร้างโครงสร้างด้วยคู่เบสที่มีความน่าจะเป็นสูงสุด วิธีการนี้ถูกริเริ่มโดยใช้ฐานความรู้ที่มีศักยภาพเพื่อทำนายความน่าจะเป็นของการจับคู่เบส โดยค่าความถูกต้องที่คาดหวัง (Expected accuracy) คำนวณได้ดังสมการที่ 2.4 [61]

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

$$Expected\ accuracy\ (S) = \gamma \sum_{(i,j)\epsilon BP} 2P_{BP}(i,j) + \sum_{k\epsilon SS} P_{ss}\ (k) \quad (2.4)$$

โดยที่ $P_{BP}(i,j)$ คือ ความน่าจะเป็นที่เบสตำแหน่งที่ *i* และ เบสตำแหน่งที่ *j* จะจับคู่กัน

$P_{ss}(k)$ คือความน่าจะเป็นที่เบสตำแหน่งที่ *k* จะอยู่เดี่ยว ๆ (ไม่จับคู่กับเบสอื่น)

γ คือ ค่าน้ำหนัก (weight factor) ของความน่าจะเป็นทั้ง 2 ส่วนที่ถูกนำมารวมกัน

ผลรวมของ 2 ค่านี้ดำเนินการกับทุกคู่เบสและทุกเบสเดี่ยว ๆ ที่พบในโครงสร้าง โดยโครงสร้างที่ทำนายด้วยเทคนิคนี้สามารถถูกประกอบจากความน่าจะเป็นของคู่เบสที่ถูกคำนวณโดยฟังก์ชันพาร์ทิชันโดยใช้โปรแกรมที่ชื่อว่า MaxExpect [62] ใช้เวลา O(N^3) เมื่อ *N* คือความยาวของสายลำดับอาร์เอ็นเอ

ท่ามกลางอาร์เอ็นเอชนิดต่าง ๆ โปรแกรม MaxExpect มีค่าความอ่อนไหวเท่ากับวิธีทำนายโครงสร้างที่มีค่าพลังงานต่ำสุดคือประมาณ 73% แต่มีค่าความจำเพาะที่ดีขึ้น กล่าวคือเทคนิคการทำนายโครงสร้างที่มีค่าพลังงานต่ำสุดมีค่าความจำเพาะ 66% ในขณะที่เทคนิคการทำนายโครงสร้างที่มีค่าความถูกต้องที่คาดหวังสูงสุดมีค่าความจำเพาะประมาณ 66-68%

4. การทำนายโครงสร้างที่มีความเหมาะสมรอง (Suboptimal structure prediction)

โครงสร้างเหมาะสมรอง (Suboptimal structure) คือโครงสร้างที่มีคะแนนเท่ากับหรือใกล้เคียงกับโครงสร้างที่ถูกทำนายว่ามีคะแนนดีสุด เช่น ในเทคนิคทำนายโครงสร้างที่มีค่าพลังงานต่ำสุดนั้นโครงสร้างเหมาะสมรองคือโครงสร้างที่มีค่าพลังงานต่ำรองลงมา

เนื่องจากโครงสร้างที่มีค่าพลังงานต่ำสุดหรือมีค่าความถูกต้องที่ความหวังสูงสุดอาจไม่ใช่โครงสร้างที่ตรงกับโครงสร้างที่เป็นคำตอบเสมอไป โครงสร้างเหมาะสมรองอาจใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากกว่า นอกจากนี้ บางสายลำดับอาร์เอ็นเออาจมีหลายโครงสร้างทุติยภูมิ เช่น riboswitches ซึ่งโครงสร้างทุติยภูมิเปลี่ยนตาม ligand binding

โครงสร้างเหมาะสมสุด (Optimal structure) เพียงอย่างเดียวไม่สามารถเก็บสารสนเทศเชิงโครงสร้างที่สำคัญได้หมด มีหลายวิธีที่สามารถใช้สร้างโครงสร้างเหมาะสมรองที่มีค่าพลังงานต่ำสุด [63] แนวทางหนึ่งคือ วิธีฮิวริสติกที่คำนวณโครงสร้างทางเลือกที่เป็นตัวแทน (representative alternative structure)

แม้ว่าค่าความอ่อนไหวเฉลี่ยของโครงสร้างที่มีค่าพลังงานต่ำสุดจะมีค่าประมาณ 73% แต่โครงสร้างเหมาะสมรองที่ถูกต้องตรงกับโครงสร้างที่เป็นคำตอบมากที่สุดในกลุ่มมีค่าความอ่อนไหวเฉลี่ยเป็น 87% และเมื่อพิจารณาทุกคู่เบสในโครงสร้างเหมาะสมรองใด ๆ ค่าความอ่อนไหวในทั้งเซตของโครงสร้างเหมาะสมรองมีค่าเฉลี่ยเป็น 97% [64]

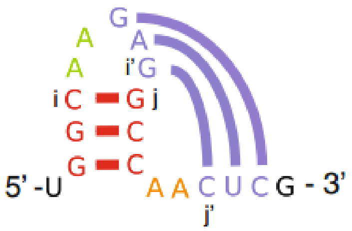
2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

มีความเป็นไปได้ที่จะสร้างโครงสร้างเหมาะสมรองโดยสร้างทุกโครงสร้างที่เป็นไปได้ภายในช่วงค่าพลังงานที่กำหนดเทียบกับโครงสร้างที่มีค่าพลังงานต่ำสุด [65] แต่การทำเช่นนี้ใช้ต้นทุนการคำนวณที่สูงเพราะจำนวนโครงสร้างเหมาะสมรองที่เป็นไปได้เพิ่มขึ้นเป็นเอกโพเนลเชียลเมื่อค่าพลังงานเพิ่มขึ้น การแจงนับอย่างครบถ้วนของโครงสร้างเหมาะสมรองมีประโยชน์ในกรณีที่ข้อมูลจากการทดลองสามารถถูกใช้เพื่อเลือกโครงสร้างที่สมเหตุสมผลจากชุดของโครงสร้างที่ทำนายได้ อีกแนวทางหนึ่งในการสร้างโครงสร้างเหมาะสมรองคือสุ่มโครงสร้างตามความน่าจะเป็นในโบลทซ์มันน์ (Boltzmann ensemble)

วิธีการที่ใช้ในโปรแกรม Sfold [66] RNAstructure และ Vienna Package อาจเกี่ยวข้องกับ 1 กลุ่ม (cluster) หรือหลายกลุ่มที่ใกล้เคียงกับโครงสร้างที่กำลังพิจารณา กล่าวคือโครงสร้างตัวแทนที่เรียกว่าโครงสร้างเซนทรอยด์ (centroid structure) อาจเป็นโครงสร้างที่ให้ค่าความถูกต้องมากกว่าโครงสร้างที่มีค่าพลังงานต่ำสุด

5. ซูโดนอทและการทำนายซูโดนอท

ซูโดนอทแทนปัญหาที่เฉพาะเจาะจงสำหรับขั้นตอนวิธีในการทำนายโครงสร้างอาร์เอ็นเอ ซูโดนอทถูกสร้างโดยคู่เบสที่ไม่ได้เรียงกันอย่างเป็นระเบียบ (non-nested base pair) กล่าวคือ 1 ซูโดนอทถูกกำหนดโดยอย่างน้อย 2 คู่เบส *i-j* และ *i'-j'* ซึ่งตำแหน่งของนิวคลีโอไทด์ *i* อยู่ก่อน *i'* ตำแหน่งของนิวคลีโอไทด์ *i'* อยู่ก่อน *j* และ ตำแหน่งของนิวคลีโอไทด์ *j* อยู่ก่อน *j'* ตัวอย่างซูโดนอทแสดงดังรูปที่ 2.4 [55]



รูปที่ 2.4 ตัวอย่างซูโดนอท

การทำนายโครงสร้างที่มีค่าพลังงานต่ำสุดที่มีซูโดนอทถูกพิสูจน์ว่าเป็นปัญหา NP-hard [67] และความท้าทายที่เพิ่มขึ้นคือค่าพารามิเตอร์สำหรับคำนวณค่าพลังงานสำหรับซูโดนอทไม่ได้ถูกกำหนดโดยการทดลอง อย่างไรก็ตาม มีชุดพารามิเตอร์ที่แยกออกมา [68] ซึ่งกำหนดโดยใช้แบบจำลองพอลิเมอร์ (polymer model) แบบจำลองแลตทิซ (lattice model) และวิธีการเชิงประจักษ์ (empirical approach) หลาย ๆ อัลกอริทึมรวมทั้ง Mfold

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

และ RNAfold ยังไม่รองรับการทำนายในส่วนของซูโดนอท อย่างไรก็ตาม อัลกอริทึมที่พยายามทำนายซูโดนอทมีการดำเนินการดังนี้

วิธีการแรกคือใช้กำหนดการพลวัตที่สามารถทำนายซูโดนอทโดยจำกัดอยู่แค่บางโครงสร้าง (topology) [69] วิธีการที่สองใช้การวนทำซ้ำ (iterative approach) เพื่อประกอบโครงสร้างจากขั้นตอนวิธีที่ไม่สามารถทำนายซูโดนอทได้ภายในรอบการทำงานเดียว [70] นอกจากนี้ วิธีการประกอบโครงสร้างที่มีค่าความถูกต้องที่คาดหวังมากสุดที่สามารถทำนายซูโดนอทของโครงสร้างใด ๆ ถูกนำเสนอใน [56] อย่างไรก็ตาม ในภาพรวมค่าความถูกต้องของขั้นตอนวิธีในการทำนายคู่เบสที่เป็นซูโดนอทยังต่ำอยู่และยังคงเป็นงานวิจัยที่ต้องมีการพัฒนาต่อไป

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.1.4.3 วิธีเชิงการคำนวณสำหรับทำนายโครงสร้างทุติยภูมิจากหลายสายลำดับที่ถูกจัดตำแหน่ง (multiple-aligned sequence)

เนื่องจากความยากของการทำนายโครงสร้างทุติยภูมิด้วย 1 สายลำดับจากที่นำเสนอไปในหัวข้อก่อนหน้า ขั้นตอนวิธีในกลุ่มดังกล่าวประสบปัญหาจากความไม่สมบูรณ์ของพารามิเตอร์ค่าพลังงานทำให้ค่าความถูกต้องของการทำนายอยู่ระหว่าง 45 – 70% [71] ต่อมาจึงเริ่มมีการใช้สารสนเทศอื่น ๆ เพิ่มเติม ภายใต้หลักการที่ว่าโมเลกุลอาร์เอ็นเอที่ฟังก์ชันมีความสัมพันธ์กันมักมีโครงสร้างที่สัมพันธ์กัน ทำให้สามารถหาโครงสร้างที่ดีที่สุดจากชุดของโมเลกุลที่มีความสัมพันธ์กันได้ เช่น อาร์เอ็นเอส่งถ่ายจำนวนหนึ่งจะไม่แสดงโครงสร้างที่เป็นรูปร่างคล้ายใบถั่ว (clover-leaf shape) เมื่อถูกพับโดยใช้แค่ 1 สายลำดับแต่หากทำนายโครงสร้างที่สอดคล้องกัน (consensus structure) จาก 2 – 3 สายลำดับก็เพียงพอที่จะระบุโครงสร้างที่มีรูปร่างคล้ายใบถั่วได้อย่างถูกต้อง

โครงสร้างที่สอดคล้องกันหมายความว่าแม้ข้อมูลในสายลำดับจะเปลี่ยนแปลงแต่ความสามารถในการสร้างคู่เบสยังคงเดิม เช่น ในสายลำดับหนึ่งอาจจะมีคู่เบสเป็น AU ในขณะที่อีกสายลำดับหนึ่งมีคู่เบสเป็น GC เรียกลักษณะนี้ว่าการกลายพันธุ์ชดเชย (compensatory mutation) แต่หากมีการเปลี่ยนแปลงแค่ด้านหนึ่งของคู่เบส เช่นจาก GU เป็น GC เรียกว่าการกลายพันธุ์ที่สม่ำเสมอ (consistent mutation) ดังนั้น สามารถใช้ข้อมูลของหลาย ๆ สายลำดับอาร์เอ็นเอที่สอดคล้องกันเพื่อช่วยเพิ่มประสิทธิภาพในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอได้

วิธีการทำนายโครงสร้างทุติยภูมิจากหลายสายลำดับที่มีการจัดตำแหน่ง (alignment) [72] ให้ค่าความถูกต้องที่ดีกว่าวิธีการทำนายโครงสร้างที่ใช้เพียง 1 สายลำดับ แต่ก็ใช้ต้นทุนการคำนวณที่สูงกว่าทั้งในแง่ของเวลาและหน่วยความจำ วิธีการในกลุ่มนี้สามารถแบ่งได้เป็น 3 กลุ่ม หลัก ๆ [73] ได้แก่ กลุ่มขั้นตอนวิธีที่จัดตำแหน่งก่อนแล้วจึงพับ (align and then fold) กลุ่มขั้นตอนวิธีที่พับและจัดตำแหน่งไปพร้อมกัน และ กลุ่มขั้นตอนวิธีที่พับก่อนแล้วจึงจัดตำแหน่ง (fold and then align)

1. กลุ่มขั้นตอนวิธีที่จัดตำแหน่งก่อนแล้วจึงพับโครงสร้าง

เป็นวิธีการฮิวริสติกที่เป็นไปได้ในทางปฏิบัติ กล่าวคือใช้เครื่องมือที่มีความสามารถในการจัดตำแหน่งเพื่อหาตำแหน่งที่สอดคล้องกันในสายลำดับเหล่านั้น จากนั้นจึงทำการพับโครงสร้างได้เป็นโครงสร้างทุติยภูมิของอาร์เอ็นเอที่ต้องการ ประสิทธิภาพของขั้นตอนวิธีในกลุ่มนี้ถูกกำหนดโดยคุณภาพของการจัดตำแหน่ง ในขั้นตอนของการพับโครงสร้างก็สามารถใช้วิธีการต่าง ๆ ในทำนองเดียวกับการทำนายโครงสร้างที่ใช้เพียง 1 สายลำดับ ตัวอย่างขั้นตอนวิธีในกลุ่มนี้ เช่น RNAalifold [74] ซึ่งใช้กำหนดการพลวัตมาตรฐานที่นำเสนอโดย Zuker ในขั้นตอนของการพับโครงสร้างจากสายลำดับที่มีการจัดตำแหน่งแล้ว อีกตัวอย่างคือ Pfold [8] ซึ่งใช้ SCFGs ในการทำนายโครงสร้าง และ ILM [70] ซึ่งใช้การจับคู่ลูปแบบวนทำซ้ำ (iterated loop maching) กล่าวคือมีการคำนวณคะแนนของคู่เบส จากนั้นเมื่อหาฮีลิกที่ดีที่สุดในโครงสร้างได้ก็ตัดออกจากส่วนที่มีการจัดตำแหน่ง จากนั้นพิจารณาโครงสร้างส่วนที่เหลือไปเรื่อย ๆ ซึ่งขั้นตอนวิธีนี้สามารถทำนายโครงสร้างในส่วนของซูโดนอทได้ด้วย

2. กลุ่มขั้นตอนวิธีที่พับโครงสร้างและจัดตำแหน่งไปพร้อมกัน

ในทางปฏิบัติ การจัดตำแหน่งไม่เหมาะสมและไม่สามารถช่วยให้ค่าความถูกต้องในการทำนายโครงสร้างดีขึ้นเมื่อ 2 สายลำดับอาร์เอ็นเอใด ๆ มีความคล้ายคลึงต่ำกว่า 50% [75] วิธีการแรกในการพับหลาย ๆ สายลำดับอาร์เอ็นเอที่มีลักษณะคล้ายกันถูกนำเสนอโดย David Sankoff [76] ซึ่งพิจารณาการจัดตำแหน่งและการพับโครงสร้างของสายลำดับอาร์เอ็นเอจำนวนเท่าไหร่ก็ได้ไปพร้อมกันใน 1 การคำนวณ ขั้นตอนวิธีนี้ใช้เวลาในการคำนวณเป็น $O(N^{3s})$ และ ใช้หน่วยความจำเป็น $O(N^{2s})$ เมื่อ s คือจำนวนสายลำดับอาร์เอ็นเอที่ยาว N นิวคลีโอไทด์ ขั้นตอนวิธีนี้ไม่สามารถทำนายซูโดนอทได้และใช้ต้นทุนการคำนวณที่สูงโดยเฉพาะเมื่อดำเนินการกับสายลำดับอาร์เอ็นเอที่มากกว่า 2 สายขึ้นไป

จากข้อจำกัดในแง่ของการนำไปใช้งานจริงของขั้นตอนวิธีที่นำเสนอโดย Sankoff ต่อมาจึงมีการนำเสนอ FOLDALIGN [77] ซึ่งใช้เทคนิคจำนวนคู่เบสมากสุด (base pair maximization) แทนการหาค่าพลังงานต่ำสุดและมีการตัดการพิจารณาในส่วนของโครงสร้างที่เป็นทางแยกทิ้งไป (branced structure) ผลก็คือสามารถลดเวลาลงเหลือ $O(N^4)$ จากนั้นเวอร์ชันต่อมาของ FOLDALIGN [78] มีการเพิ่มเติมให้สามารถสนับสนุนการทำนายโครงสร้างที่เป็นทางแยกได้ สามารถใช้แบบจำลองค่าพลังงานและใช้ฮิวริสติกเพื่อทำการตัดเล็มข้อมูลบางส่วนทิ้งไปเพื่อเพิ่มความเร็วในการคำนวณ ผลก็คือขั้นตอนวิธีนี้มีค่าความถูกต้องเพิ่มขึ้นอย่างมีนัยสำคัญ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3. กลุ่มขั้นตอนวิธีที่พับโครงสร้างก่อนแล้วจึงจัดตำแหน่ง

ขั้นตอนวิธีในกลุ่มนี้มีการใช้งานแพร่หลายน้อยสุด กล่าวคือ ทำการพับสายลำดับด้วยวิธีการทำนายโครงสร้างจาก 1 สายลำดับ จากนั้นทำการจัดตำแหน่งโครงสร้างที่ทำนายได้โดยใช้ตัวชี้วัดที่อยู่บนพื้นฐานของต้นไม้ (tree-based metrics) [79] ข้อเสียหลักของวิธีการในกลุ่มนี้คือการทำนายโดยใช้เพียง 1 สายลำดับมักมีความผิดพลาดอยู่แล้ว เมื่อนำผลที่ได้มาทำการวิเคราะห์ต่อก็ยิ่งจะได้รับผลกระทบจากค่าความผิดพลาดในขั้นตอนของการทำนายโครงสร้าง ตัวอย่างขั้นตอนวิธีในกลุ่มนี้ เช่น RNAshapes [80] ซึ่งทำการแจกแจงรูปร่าง (abstract shape) ที่เป็นไปได้ของแต่ละสายลำดับอย่างอิสระและคำนวณความน่าจะเป็นของแต่ละรูปร่าง จากนั้นระบุโครงสร้างที่เหมาะสมสุดเชิงค่าพลังงาน ซึ่งขั้นตอนวิธี RNAshapes เองไม่ได้มีการดำเนินการในส่วนของการจัดตำแหน่งแต่สามารถทำได้ในภายหลังโดยใช้ RNAforester [81]

2.1.5 ขั้นตอนวิธีประมาณการแจกแจง

ขั้นตอนวิธีประมาณการแจกแจง [17, 82] เป็นขั้นตอนวิธีหนึ่งที่อยู่ในกลุ่มของขั้นตอนวิธีเชิงวิวัฒนาการ แนวคิดหลักของขั้นตอนวิธีในกลุ่มนี้คือการรักษาแบบจำลองความน่าจะเป็นที่ชัดเจนเพื่อแทนการกระจายตัวของคำตอบที่เป็นไปได้และปรับปรุงแบบจำลองนั้นโดยอาศัยผลลัพธ์ของการประเมินค่าความเหมาะสมของคำตอบเหล่านี้ ดังนั้น อัลกอริทึมมีแนวโน้มว่าจะสร้างคำตอบที่ดีขึ้นในอนาคต สังเกตว่าการใช้แบบจำลองความน่าจะเป็นที่ชัดเจนทำให้ขั้นตอนวิธีประมาณการแจกแจงค่อนข้างแตกต่างจากเมตาฮิวริสติกอื่น ๆ เช่น ขั้นตอนวิธีเชิงพันธุกรรม [83] หรือ แบบจำลองการอบเหนียว [84] ในแง่ที่การแจกแจงความน่าจะเป็นถูกใช้เพื่อสร้างคำตอบใหม่มักถูกกำหนดโดยอ้อมผ่านตัวดำเนินการค้นหาต่าง ๆ

ขั้นตอนวิธีประมาณการแจกแจงเริ่มต้นด้วยการสุ่มสร้างประชากรเริ่มต้น จากนั้นประชากรเหล่านี้ถูกประเมินโดยใช้ฟังก์ชันวัตถุประสงค์ซึ่งจะประเมินว่าแต่ละคำตอบมีคุณภาพดีแค่ไหนในปัญหานั้น ๆ และทำซ้ำกระบวนการเหล่านี้ไปจนกระทั่งได้คำตอบที่ดีที่สุดหรือเป็นไปตามเงื่อนไขหยุดการทำงาน ได้แก่ การคัดเลือกกลุ่มประชากรย่อย โดยโครโมโซมที่มีค่าความเหมาะสมดีกว่าก็มีโอกาสถูกเลือกมากกว่า จากนั้นแบบจำลองความน่าจะเป็นจะถูกสร้างจากกลุ่มประชากรย่อยที่ถูกเลือก และประชากรรุ่นถัดไปจะถูกสุ่มจากแบบจำลองความน่าจะเป็นนี้ รหัสเทียมของขั้นตอนวิธีประมาณการแจกแจง แสดงดังรูป 2.5

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

```
g <- 0 // เริ่มต้นการทำงาน
1. กำหนดค่าเริ่มต้นให้กับแบบจำลองความน่าจะเป็น M(0)
2. ทำซ้ำขั้นตอนต่อไปนี้ตราบใดที่ยังไม่เป็นไปตามเงื่อนไขจบการทำงาน
        3.1 สุ่มสร้างประชากร P(g) อ้างอิงตาม M(g)
        3.2 ประเมินประชากร P(g) ด้วยฟังก์ชันวัตถุประสงค์
        3.3 เลือกกลุ่มประชากรย่อยแทนด้วย S(g)
        3.4 ปรับปรุงแบบจำลองความน่าจะเป็น M(g) ด้วย S(g)
        3.5 g <- g + 1
```

รูปที่ 2.5 รหัสเทียมของขั้นตอนวิธีประมาณการแจกแจง

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.1.5.1 ประเภทของขั้นตอนวิธีประมาณการแจกแจงแบ่งตามลักษณะการขึ้นแก่กันของตัวแปร

ขั้นตอนวิธีประมาณการแจกแจงสามารถแบ่งได้เป็น 3 กลุ่มหลัก ๆ ตามลักษณะการขึ้นต่อกันของตัวแปร รายละเอียด เป็นดังนี้

1. ขั้นตอนวิธีประมาณการแจกแจงแบบที่ตัวแปรไม่ขึ้นต่อกัน (Univariate estimation of distribution algorithm)

ขั้นตอนวิธีในกลุ่มนี้อยู่บนสมมุติฐานที่ว่าทุกตัวแปรเป็นอิสระจากตัวแปรอื่น ๆ นั่นคือ การแจกแจงความน่าจะเป็น $P(X_1, X_2, . . . , X_n)$ ของเวกเตอร์ $(X_1, X_2, ... , X_n)$ ของ n ตัวแปร คือ ผลคูณการแจกแจงของแต่ละตัวแปร ดังสมการ 2.5

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i) \tag{2.5}$$

ตัวอย่างขั้นตอนวิธีประมาณการแจกแจงที่อยู่บนพื้นฐานแบบจำลองที่ตัวแปรไม่ขึ้นต่อกัน เช่น ขั้นตอนวิธีเชิงพันธุกรรมแบบสมดุล (equilibrium genetic algorithm, EGA) [85] การเรียนรู้เพิ่มขึ้นแบบอาศัยประชากร (population-based incremental learning, PBIL) [86] ขั้นตอนวิธีแจกแจงตามขอบหนึ่งตัวแปร (univariate marginal distribution algorithm, UMDA) [17] ขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ (compact genetic algorithm, cGA) [87] เป็นต้น

2.ขั้นตอนวิธีประมาณการแจกแจงแบบที่ตัวแปรขึ้นต่อกันเป็นคู่ (Bivariate estimation of distribution algorithm)

แม้ว่าขั้นตอนวิธีประมาณการแจกแจงแบบที่ตัวแปรไม่ขึ้นต่อกันจะสามารถทำงานได้อย่างมีประสิทธิภาพ แต่ในหลาย ๆ กรณีขั้นตอนวิธีในกลุ่มนี้ก็ไม่สามารถแก้ปัญหาที่ให้ประสิทธิภาพที่ดีกว่าการใช้ขั้นตอนวิธีเชิงพันธุกรรมมาตรฐาน เพื่อเอาชนะข้อจำกัดนั้นขั้นตอนวิธีประมาณการแจกแจงแบบที่ตัวแปรขึ้นต่อกันเป็นคู่จึงถูกนำเสนอ

แบบจำลองความน่าจะเป็นในกลุ่มนี้ความสัมพันธ์ระหว่างตัวแปรแทนด้วยต้นไม้ 1 ต้นหรือกราฟที่เป็นป่า (forest tree) การแทนแบบจำลองด้วยต้นไม้แต่ละตัวแปรยกเว้นรากของต้นไม้ถูกเงื่อนไขด้วยทุกตัวแปรที่เป็นโหนดพ่อแม่ของมัน ในทางตรงกันข้าม การแทนแบบจำลองด้วยกราฟที่เป็นป่าคือกลุ่มของต้นไม้ที่ไม่ต่อเนื่องกัน และป่านั้นก็จะประกอบด้วยทุกตัวแปรของปัญหา หากกำหนดให้ $X = (X_1, X_2, . . . , X_n)$ เป็นตัวแปรที่ถูกเก็บในเวกเตอร์ การแจกแจงของขั้นตอนวิธีในกลุ่มนี้สามารถแสดงได้ดังสมการ 2.6

$$P(X_1, X_2, \dots, X_n) = \prod_{X_i \epsilon R} P(X_i) \prod_{X_i \epsilon X \setminus R} P(X_i | parent(X_i)) \qquad (2.6)$$

ตัวอย่างขั้นตอนวิธีประมาณการแจกแจงที่อยู่บนพื้นฐานของแบบจำลองที่ตัวแปรมีการขึ้นแก่กันแบบคู่ เช่น จัดกลุ่มข้อมูลนำเข้าที่อยู่ร่วมกันสูงสุด (mutual information maximizing input clustering: MIMIC) [88] เป็นขั้นตอนวิธีประมาณการแจกแจงที่อยู่บนพื้นฐานของต้นไม้พึ่งพา (dependency trees) [89] และ ขั้นตอนวิธีแจงแจงตามขอบสองตัวแปร (bivariate marginal distribution algorithm: BMDA) [90]

3. ขั้นตอนวิธีประมาณการแจกแจงแบบตัวแปรหลายตัวขึ้นต่อกัน (Multivariate estimation of distribution algorithm)

แบบจำลองหลายตัวแปรแทนความสัมพันธ์โดยใช้ทั้งกราฟมีทิศทางแบบไม่มีวัฏจักร (directed acyclic graphs) หรือ กราฟแบบไม่มีทิศทาง (undirected graphs) รูปแบบการแทนแบบจำลองที่ได้รับความนิยมในขั้นตอนวิธีประมาณการแจกแจง ได้แก่ (1) เครือข่ายแบบเบย์ (Bayesian networks) และ (2) เครือข่ายมาร์คอฟ (Markov networks) โดยเครือข่ายแบบเบย์ถูกแทนด้วยกราฟมีทิศทางแบบไม่มีวัฏจักรซึ่งแต่ละโหนดแทนแต่ละตัวแปร และแต่ละเส้นเชื่อมแทนการขึ้นต่อกันอย่างมีเงื่อนไขแบบมีทิศทาง การแจกแจงความน่าจะเป็นที่ถูกเข้ารหัสโดยเครือข่ายแบบเบย์เขียนได้ดังสมการ 2.7

$$P(X_1, X_2, \dots, X_2) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \qquad (2.7)$$

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

เครือข่ายแบบเบย์แทนการแบ่งปัญหาบนสมมุติฐานการไม่ขึ้นต่อกันอย่างมีเงื่อนไข นั่นคือ ความสัมพันธ์ของแต่ละโหนดจะไม่วนมาหาโหนดเดิม และ แต่ละโหนดจะมีความสัมพันธ์กันตามทิศทางที่แสดงในเครือข่าย ถ้ามีลูกศรจากโหนด X_1 ชี้ไปหาโหนด X_2 จะเรียกว่า โหนด X_1 เป็นโหนดพ่อแม่ของ X_2 และแต่ละโหนด X_i จะมีเงื่อนไขการแจกแจงความน่าจะเป็น $P(X_i|\text{parents}(X_i))$ ซึ่งส่งผลต่อโหนดพ่อแม่ของแต่ละโหนด

ในเครือข่ายมาร์คอฟ 2 ตัวแปรถูกสมมุติว่าอิสระจากกันภายในสับเซตของตัวแปรที่กำหนดเงื่อนไขก็ต่อเมื่อทุกเส้นเชื่อมระหว่างตัวแปรเหล่านั้นถูกแยกด้วย 1 หรือ หลายตัวแปรในเงื่อนไขนั้น

ตัวอย่างขั้นตอนวิธีประมาณการแจกแจงที่อยู่บนพื้นฐานของแบบจำลองที่มีหลายตัวแปรขึ้นต่อกัน เช่น factorized distribution algorithm (FDA), learning FDA (LFDA) [91], estimation of Bayesian network algorithm (EBNA) [92], Bayesian optimization algorithm (BOA) [93] และ extended compact genetic algorithm (ecGA) [94]

2.1.5.2 ตัวอย่างการประยุกต์ใช้ขั้นตอนวิธีประมาณการแจกแจงในการแก้ปัญหา

เนื่องจากขั้นตอนวิธีที่งานวิจัยนี้นำเสนอถูกพัฒนาขึ้นโดยได้รับแรงบันดาลใจจากขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ [87] และ ขั้นตอนวิธีคอยน์ (Coincidence algorithm, COIN) [95] กล่าวคือ ขั้นตอนวิธี Hybrid-EDAfold ใช้แบบจำลองความน่าจะเป็นในลักษณะเดียวกับที่ใช้ในขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ และ มีกระบวนการคัดเลือกกลุ่มประชากรย่อยและการปรับปรุงแบบจำลองความน่าจะเป็นที่คล้ายกับขั้นตอนวิธีคอยน์ ในหัวข้อนี้ จึงนำเสนอการทำงานคร่าว ๆ ของทั้ง 2 ขั้นตอนวิธีดังกล่าว โดยยกตัวอย่างการประยุกต์ใช้งานกับปัญหาอย่างง่าย รายละเอียดเป็นดังนี้

1. ตัวอย่างการแก้ปัญหา Onemax ด้วยขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ

ขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับเป็นขั้นตอนวิธีการประมาณการแจกแจงในกลุ่มที่ตัวแปรไม่ขึ้นต่อกัน ขั้นตอนวิธีนี้ใช้แบบจำลองความน่าจะเป็นแทนการใช้กลุ่มประชากรแบบในขั้นตอนวิธีเชิงพันธุกรรมมาตรฐานทำให้ใช้พื้นที่หน่วยความจำลดลง และไม่มีตัวดำเนินการเชิงพันธุกรรมอย่างการไขว้เปลี่ยน และ การกลายพันธุ์ แต่ยังคงความสามารถในการค้นหาคำตอบที่เทียบเท่ากับขั้นตอนวิธีเชิงพันธุกรรมมาตรฐาน

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ปัญหา Onemax เป็นปัญหาสมมุติ (toy problem) มักถูกใช้เพื่อทดสอบพฤติกรรมของอัลกอริทึมเพื่อเปรียบเทียบความสามารถในการแก้ปัญหาอย่างง่าย สำหรับปัญหานี้โครโมโซมคำตอบถูกแทนด้วยเวกเตอร์ที่มีความยาว n บิต เมื่อ n คือจำนวนตัวแปร และค่าความเหมาะสมของโครโมโซมคำนวณจากจำนวนบิตในเวกเตอร์ที่มีค่าเป็น 1 และ คำตอบที่ดีที่สุดของปัญหานี้คือทุกบิตในเวกเตอร์มีค่าเป็น 1 ทั้งหมด

เนื่องจากการแทนคำตอบของขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับอยู่ในรูปแบบของเวกเตอร์ความน่าจะเป็น (probability vector) ที่มีขนาดเท่ากับจำนวนตัวแปรของปัญหา และแต่ละสมาชิกในเวกเตอร์แทนความน่าจะเป็นที่แต่ละสมาชิกจะมีค่าเป็น 1 ดังนั้น สำหรับปัญหา Onemax ค่าเริ่มต้นในเวกเตอร์ความน่าจะเป็นของแต่ละสมาชิกจะถูกกำหนดค่าเป็น 0.5 หมายความว่าแต่ละบิตมีโอกาสเป็น 0 และ 1 เท่ากัน

ในการทำงานของอัลกอริทึมจะสุ่มสร้างโครโมโซมจำนวน 2 ตัว โอกาสที่แต่ละบิตในโครโมโซมจะมีค่าเป็น 1 หรือ 0 อ้างอิงตามค่าในเวกเตอร์ความน่าจะเป็น จากนั้นทำการประเมินค่าความเหมาะสมของแต่ละโครโมโซมที่สร้างได้อ้างอิงตามฟังก์ชันวัตถุประสงค์ของปัญหา ในที่นี้คือจำนวนบิตที่มีค่าเป็น 1 จากนั้นเปรียบเทียบสองโครโมโซมนั้นว่าโครโมโซมใดมีค่าความเหมาะสมดีกว่าให้เป็นผู้ชนะและอีกโครโมโซมเป็นผู้แพ้ จากนั้นปรับค่าในเวกเตอร์ความน่าจะเป็นไปในทิศทางของผู้ชนะ กล่าวคือ ในตำแหน่งบิตที่โครโมโซมผู้ชนะและผู้แพ้มีค่าไม่ตรงกัน ถ้าในโครโมโซมผู้ชนะบิตนั้นมีค่าเป็น 1 ค่าของเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับบิตในตำแหน่งนั้นจะถูกปรับให้เข้าใกล้ 1 มากขึ้นอ้างอิงตามอัตราการเรียนรู้ที่กำหนดซึ่งโดยปกติมีค่าเท่ากับ $1/np$ เมื่อ np แทนขนาดประชากร ในทางตรงกันข้าม ถ้าในโครโมโซมผู้ชนะ บิตนั้นมีค่าเป็น 0 ค่าของเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับบิตในตำแหน่งนั้นจะถูกปรับค่าให้ลดลง และจะส่งผลให้ในรอบถัดไป บิตนี้จะมีโอกาสสร้างได้บิตที่มีค่าเป็น 0 มากยิ่งขึ้น จากนั้นก็วนทำซ้ำในลักษณะเช่นนี้ไปจนกระทั่งทุกบิตในเวกเตอร์ความน่าจะเป็นลู่เข้าสู่ค่า 0.0 หรือ 1.0 จึงหยุดการทำงาน ดังนั้นสำหรับปัญหา Onemax ความคาดหวังคือเมื่อจบการทำงานของอัลกอริทึมทุกสมาชิกในเวกเตอร์ความน่าจะเป็นมีค่าเป็น 1 ทั้งหมด การทำงานของขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับสรุปได้ดังรูปที่ 2.6

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

1. กำหนดค่าเริ่มต้นให้ทุกสมาชิกในเวกเตอร์ความน่าจะเป็นเท่ากับ 0.5
2. สุ่มสร้างโครโมโซมคำตอบจำนวน 2 ตัวจากเวกเตอร์ความน่าจะเป็น
3. ประเมินค่าความเหมาะสมของทั้งสองโครโมโซมเพื่อตัดสินหาผู้ชนะ และ ผู้แพ้
4. ปรับปรุงเวกเตอร์ความน่าจะเป็นตามโครโมโซมผู้ชนะ โดยพิจารณาแต่ละบิตของโครโมโซม ถ้าบิตที่ *i* ของโครโมโซมผู้ชนะมีค่าไม่ตรงกับบิตที่ *i* ของโครโมโซมผู้แพ้
   - ถ้าบิตนั้นของโครโมโซมผู้ชนะมีค่าเป็น 1 ให้เพิ่มค่าความน่าจะเป็นในตำแหน่งที่ *i* ของเวกเตอร์ความน่าจะเป็นจากเดิมไปอีก $1/np$
   - ถ้าบิตนั้นของโครโมโซมผู้ชนะมีค่าเป็น 0 ให้ลดค่าความน่าจะเป็นในตำแหน่งที่ *i* ของเวกเตอร์ความน่าจะเป็นจากเดิมไปอีก $1/np$
5. ตรวจสอบเวกเตอร์ความน่าจะเป็นหากยังไม่ลู่เข้าสู่คำตอบกลับไปทำซ้ำขั้นตอน 2 – 5

รูปที่ 2.6 ขั้นตอนการทำงานของขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ

2. ตัวอย่างการแก้ปัญหาการเดินทางของพนักงานขายด้วยขั้นตอนวิธีคอยน์

ขั้นตอนวิธีคอยน์เป็นขั้นตอนวิธีการประมาณการแจกแจงในกลุ่มที่มีการขึ้นต่อกันของตัวแปรที่เป็นคู่ โดยขั้นตอนวิธีคอยน์ประสบความสำเร็จในการประยุกต์ใช้กับปัญหาการหาค่าเหมาะสมสุดเชิงการจัด (combinatorial optimization problem) ทั้งแบบวัตถุประสงค์เดี่ยวและหลายวัตถุประสงค์ [50] แนวคิดที่ขยายเพิ่มเติมของขั้นตอนวิธีคอยน์ คือ ยอมให้เกิดการเรียนรู้จากทั้งคำตอบด้อย (poor solution) ร่วมกับคำตอบดี (good solution)

ขั้นตอนวิธีคอยน์ตั้งข้อสังเกตว่าการค้นหาคำตอบที่ดีของขั้นตอนวิธีเชิงพันธุกรรมผ่านตัวดำเนินการไขว้เปลี่ยนและการกลายพันธุ์ไม่มีการแสวงหาประโยชน์ของข้อมูลภายในคำตอบดีเหล่านั้น ลักษณะเช่นนี้ไม่เพียงแต่ทำให้เกิดการสร้างคำตอบที่ไม่มีประสิทธิภาพจำนวนมากมายมหาศาลในปริภูมิค้นหาแต่ยังใช้ต้นทุนการคำนวณที่สูงอีกด้วย ในทางตรงกันข้าม ขั้นตอนวิธีคอยน์มีการพิจารณาข้อมูลภายในกลุ่มคำตอบดีและจดจำเส้นทางที่นำไปสู่คำตอบที่ดีนั้นซึ่งขั้นตอนวิธีคอยน์แทนที่ต้นทุนการคำนวณที่สูงในตัวดำเนินการทางพันธุกรรมของขั้นตอนวิธีเชิงพันธุกรรมด้วยเมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกัน (joint probability matrix) แล้วใช้เมทริกซ์นี้เพื่อสร้างประชากรคำตอบในรุ่นถัด ๆ ไป

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ผลจากการหลีกเลี่ยงรูปแบบการเรียนรู้แบบดั้งเดิมจากโครโมโซมที่มีคุณภาพดีเพียงอย่างเดียว ขั้นตอนวิธีคอยน์ยอมให้มีการเรียนรู้จากโครโมโซมที่มีค่าความเหมาะสมต่ำกว่าค่าความเหมาะสมเฉลี่ยร่วมด้วย ซึ่งหากเป็นขั้นตอนวิธีเชิงวิวัฒนาการแบบดั้งเดิมมักจะละทิ้งโครโมโซมคุณภาพด้อยเหล่านี้โดยปราศจากการใช้ประโยชน์จากข้อมูลใด ๆ แต่ขั้นตอนวิธีคอยน์มีการเรียนรู้จากข้อมูลในโครโมโซมด้อยและใช้ข้อมูลนี้เพื่อหลีกเลี่ยงเหตุการณ์เช่นนั้นที่อาจจะเกิดขึ้นอีกในอนาคต ในขณะเดียวกัน สิ่งที่พบในโครโมโซมที่มีคุณภาพดีก็ยังคงถูกนำไปใช้เพื่อสร้างคำตอบที่มีค่าความเหมาะสมดียิ่งขึ้น ผลที่ตามมาคือโอกาสที่เส้นทางการค้นหาจะถูกชักนำไปสู่โครโมโซมที่มีคุณภาพด้อยที่อาจเกิดในรุ่นถัดไปจะลดลง จำนวนของคำตอบที่เป็นไปได้ที่ถูกพิจารณาลดลง และนำไปสู่การลู่เข้าคำตอบที่เพิ่มขึ้น การทำงานของขั้นตอนวิธีคอยน์สรุปได้ดังรูปที่ 2.7

1. กำหนดค่าเริ่มต้นให้กับเมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกัน
2. สุ่มสร้างประชากรจากเมทริกซ์
3. ประเมินประชากร
4. คัดเลือกโครโมโซมบางส่วนจากประชากร
5. ปรับปรุงค่าในเมทริกซ์โดยใช้โครโมโซมที่ถูกคัดเลือกในขั้นตอนที่ 4
6. ทำซ้ำขั้นตอน 2-5 จนกระทั่งพบเงื่อนไขสิ้นสุดการทำงาน

รูปที่ 2.7 ขั้นตอนการทำงานของขั้นตอนวิธีคอยน์

รายละเอียดโดยสังเขปของแต่ละขั้นตอนย่อยเป็นดังนี้

2.1 กำหนดค่าเริ่มต้นให้เมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกัน

ขั้นตอนวิธีคอยน์ใช้เมทริกซ์ที่เก็บความน่าจะเป็นที่จะเกิดร่วมกันของ 2 ตัวแปรใด ๆ ในกระบวนการสร้างประชากร เมทริกซ์นี้มีขนาด $n \times n$ เมื่อ n คือจำนวนตัวแปรหรือขนาดของปัญหา โดยที่ Mxy แทนสมาชิกในเมทริกซ์แถวที่ x คอลัมน์ที่ y มีค่าอยู่ในช่วง [0, 1] สมาชิกของเมทริกซ์ในแนวทแยงมุม ($x = y$) มีค่าเป็น 0 และสมาชิกตำแหน่งอื่น ๆ มีค่าเป็น $1/(n\text{-}1)$ ตัวอย่างการกำหนดค่าเริ่มต้นให้เมทริกซ์สำหรับ 5 ตัวแปรเป็นดังรูปที่ 2.8

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

	1	2	3	4	5
1	0	0.25	0.25	0.25	0.25
2	0.25	0	0.25	0.25	0.25
3	0.25	0.25	0	0.25	0.25
4	0.25	0.25	0.25	0	0.25
5	0.25	0.25	0.25	0.25	0

รูปที่ 2.8 การกำหนดค่าเริ่มต้นให้เมทริกซ์

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.2 สุ่มสร้างประชากร

หลังจากกำหนดค่าเริ่มต้นให้กับเมทริกซ์แล้ว ขั้นตอนวิธีคอยน์จะสร้างประชากร โดยแต่ละโครโมโซมถูกสุ่มโดยอ้างอิงความน่าจะเป็นจากเมทริกซ์ เนื่องจากตัวอย่างนี้เป็นการแก้ปัญหาการเดินทางของพนักงานขายซึ่งเป็นปัญหาการเรียงสับเปลี่ยนของหมายเลขเมืองที่พนักงานขายจะต้องเดินทางไป ดังนั้น โครโมโซมจะแทนลำดับตัวเลขของเมือง ในตอนเริ่มต้นสตริงที่เก็บโครโมโซมจะว่างเปล่า จากนั้นทำการสุ่ม 1 ตัวแปร เช่น ได้เมืองหมายเลข 2 และเพื่อป้องกันการสุ่มได้เมืองซ้ำ เมื่อเมืองใดถูกสุ่มขึ้นมาหมายเลขคอลัมน์ที่ตรงกับเมืองนั้นจะถูกปิดไป จากนั้นก็ดำเนินการในลักษณะเช่นนี้ไปเรื่อย ๆ กล่าวคือ ปิดคอลัมน์ที่ตรงกับเมืองที่เพิ่งสุ่มได้และสุ่มเมืองลำดับถัดไป เมื่อทุกคอลัมน์ถูกปิดหมดแสดงว่าเดินทางไปครบทุกเมืองแล้วได้ 1 โครโมโซมที่แทน 1 การเรียงสับเปลี่ยนของเส้นทางการเดินทางของพนักงานขาย จากนั้นสร้างโครโมโซมในลักษณะนี้จนได้จำนวนโครโมโซมครบตามขนาดประชากรที่กำหนด

2.3 ประเมินประชากร

เมื่อสร้างโครโมโซมได้ครบตามขนาดประชากรที่กำหนด แต่ละโครโมโซมจะถูกประเมินด้วยฟังก์ชันค่าความเหมาะสมสำหรับปัญหานั้น ๆ ในตัวอย่างนี้คือระยะทางรวมของการเดินทางจากเมืองแรกไปยังเมืองอื่น ๆ จนครบทุกเมืองและวนกลับมาที่เมืองเดิมอ้างอิงเส้นทางตามข้อมูลที่จัดเก็บในโครโมโซม จากนั้นทำการเรียงลำดับโครโมโซมตามค่าความเหมาะสมที่ประเมินได้ ตัวอย่างผลการประเมินค่าความเหมาะสมของ 4 โครโมโซมแสดงดังรูปที่ 2.9 และตัวอย่างนี้เป็นการดำเนินการกับปัญหาการหาค่าต่ำสุด ดังนั้น โครโมโซม *C1* มีค่าความเหมาะสมดีสุด

	โครโมโซม						ค่าความเหมาะสม
C1	1	3	2	4	5	1	12
C2	4	3	1	2	5	4	13
C3	1	2	3	5	4	1	16
C4	2	3	1	4	5	2	17

รูปที่ 2.9 การประเมินค่าความเหมาะสมของแต่ละโครโมโซม

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

2.4 คัดเลือกโครโมโซมบางส่วนจากประชากร

เนื่องจากขั้นตอนวิธีคอยน์มีการเรียนรู้จากทั้งโครโมโซมดี และ โครโมโซมด้อย ดังนั้น จากโครโมโซมที่ถูกเรียงลำดับจากขั้นตอนก่อนหน้า ในขั้นตอนนี้จะจำแนกโครโมโซมทั้งหมดในกลุ่มประชากรออกเป็น 3 กลุ่มย่อย คือ กลุ่มโครโมโซมที่มีคุณภาพดี กลุ่มโครโมโซมที่มีคุณภาพด้อย และกลุ่มโครโมโซมที่ไม่ถูกนำมาพิจารณา โดยโครโมโซม g% จากด้านบนสุดของประชากรจะถูกพิจารณาว่าเป็นโครโมโซมที่ดี และโครโมโซม b% ด้านล่างสุดของประชากรจะถูกพิจารณาว่าเป็นโครโมโซมที่ด้อย โดย g และ b อาจมีค่าเท่ากันหรือไม่ก็ได้ สมมุติตัวอย่างนี้กำหนดไว้เท่ากันคือ 25% จะได้ผลลัพธ์ดังรูปที่ 2.10

	โครโมโซม						ค่าความเหมาะสม	ผลการจำแนก
C1	1	3	2	4	5	1	12	คำตอบดี
C2	4	3	1	2	5	4	13	
C3	1	2	3	5	4	1	16	
C4	2	3	1	4	5	2	17	คำตอบด้อย

รูปที่ 2.10 ตัวอย่างการจำแนกกลุ่มของประชากร

โครโมโซมที่ถูกจำแนกอยู่ในกลุ่มคำตอบดีจะถูกใช้ปรับปรุงเมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกันในทิศทางที่เพิ่มขึ้น ส่วนโครโมโซมที่ถูกจำแนกอยู่ในกลุ่มคำตอบด้อยจะถูกใช้ปรับปรุงเมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกันในทิศทางที่ลดลง รายละเอียดของการปรับปรุงค่าในเมทริกซ์ความน่าจะเป็นจะกล่าวในหัวข้อถัดไป และโครโมโซมอื่น ๆ ที่ไม่จัดอยู่ใน 2 กลุ่มนี้ก็จะถูกทิ้งไปไม่ต้องนำมาคำนวณ

2.5 การปรับปรุงค่าในเมทริกซ์ที่เก็บความน่าจะเป็นที่เกิดร่วมกัน

จากผลการจำแนกโครโมโซมในประชากรจากขั้นตอนที่แล้ว ข้อมูลจากทั้งกลุ่มโครโมโซมที่เป็นคำตอบดี และ กลุ่มโครโมโซมที่เป็นคำตอบด้อยจะถูกใช้ปรับปรุงความน่าจะเป็นในเมทริกซ์

จากโครโมโซม *C1* ซึ่งเก็บข้อมูลการเดินทาง ดังนี้ [1, 3, 2, 4, 5, 1] ถูกพิจารณาว่าเป็นโครโมโซมดี ดังนั้น ข้อมูลที่อยู่ในโครโมโซมนี้จะถูกนำไปใช้ในการเพิ่มค่าความน่าเป็นในเมทริกซ์ แรกสุดข้อมูลภายในโครโมโซมจะถูกแยกออกเป็นคู่ ๆ ดังนั้นในตัวอย่างนี้จะได้ [1,3], [3,2], [2,4], [4,5] และ [5,1] โดยที่ [1,3] แทนเหตุการณ์ที่พนักงานเดินทางจากเมือง 1 ไปเมือง 3 ถูกพบในโครโมโซมที่ดี ดังนั้น สมาชิกของเมทริกซ์ในแถวที่ 1 คอลัมน์ที่ 3 จะถูกปรับความน่าจะเป็นเพิ่มขึ้นอ้างอิงตามอัตราการเรียนรู้ (learning rate) แทนด้วย k ซึ่งเป็นพารามิเตอร์หนึ่งในขั้นตอนวิธีคอยน์ เช่น กำหนดค่า k = 0.2 และคำนวณความน่าจะเป็นที่เพิ่มได้จาก k / (n-1) จะได้ 0.2/4 = 0.05 หมายความว่า [1,3] จะได้ความน่าจะเป็นเพิ่มขึ้นจากเดิมโดยได้มาจากสมาชิกตำแหน่งอื่น ๆ รวมค่าทั้งหมดเป็น 0.15 ซึ่งความน่าจะเป็นนี้ถูกหักมาจาก [1,2], [1,4] และ [1,5] สมาชิกละ 0.05 ผลลัพธ์เป็นดังรูปที่ 2.11 โดยตัดมาเฉพาะข้อมูลของเมทริกซ์แถวที่ 1 และเปรียบเทียบให้เห็นค่าความน่าจะเป็นก่อนปรับปรุง และหลังปรับปรุง

	1	2	3	4	5
ก่อนปรับ	0	0.25	0.25	0.25	0.25
หลังปรับ	0	0.20	0.40	0.20	0.20

รูปที่ 2.11 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับสมาชิก [1,3]

จากนั้นก็ทำในลักษณะเดียวกันนี้สำหรับสมาชิกคู่อื่น ๆ ที่ถูกแยกออกจากโครโมโซมคำตอบดี จากตัวอย่างนี้ ได้แก่ [3,2], [2,4], [4,5] และ [5,1] ในท้ายที่สุดค่าในเมทริกซ์หลังจากปรับปรุงด้วยข้อมูลจากโครโมโซม *C1* จะเป็นดังรูปที่ 2.12

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

	1	2	3	4	5
1	0	0.20	0.40	0.20	0.20
2	0.20	0	0.20	0.40	0.20
3	0.20	0.40	0	0.20	0.20
4	0.20	0.20	0.20	0	0.40
5	0.40	0.20	0.20	0.20	0

รูปที่ 2.12 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับโครโมโซมที่เป็นคำตอบดี

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

การปรับปรุงเมทริกซ์โดยใช้ข้อมูลจากกลุ่มโครโมโซมด้อยทำในทิศทางตรงกันข้าม กล่าวคือ จากตัวอย่างโครโมโซม *C4* ซึ่งเก็บข้อมูลการเดินทาง ดังนี้ [2, 3, 1, 4, 5, 2] ถูกพิจารณาว่าเป็นโครโมโซมด้อย ทำการแยกข้อมูลภายในโครโมโซมออกมาเป็นคู่ ๆ ได้ ดังนี้ [2,3], [3,1], [1,4], [4,5] และ [5,2] โดยสมาชิกในเมทริกซ์ที่ตรงกับข้อมูลที่แยกได้เหล่านี้จะถูกปรับลดความน่าจะเป็นลงเพื่อไปเพิ่มให้กับสมาชิกอื่น ๆ อ้างอิงตามอัตราการเรียนรู้ที่คำนวณโดยใช้สมการเดียวกันคือ $k/(n\text{-}1)$ ดังนั้น อ้างอิงค่าในเมทริกซ์ดังรูป 2.11 สมาชิก [2,3] จะต้องถูกลดค่าลงจากเดิม 0.15 เพื่อนำไปเพิ่มให้กับ [2,1], [2,4] และ [2,5] สมาชิกละ 0.05 ตัวอย่างค่าในเมทริกซ์หลังจากการปรับปรุงด้วยข้อมูลจากโครโมโซม *C4* เป็นดังรูปที่ 2.13

	1	2	3	4	5
1	0	0.25	0.45	0.05	0.25
2	0.25	0	0.05	0.45	0.25
3	0.05	0.45	0	0.25	0.25
4	0.25	0.25	0.25	0	0.25
5	0.45	0.05	0.25	0.25	0

รูปที่ 2.13 ตัวอย่างการปรับปรุงค่าในเมทริกซ์สำหรับโครโมโซมที่เป็นคำตอบด้อย

2.2 งานวิจัยที่เกี่ยวข้อง

หัวข้อนี้นำเสนองานวิจัยที่เกี่ยวข้องกับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอทั้งในกลุ่มการทำนายโครงสร้างด้วย 1 สายลำดับ และ กลุ่มการทำนายโครงสร้างจากหลาย ๆ สายลำดับที่มีการจัดตำแหน่ง รายละเอียดเป็นดังนี้

2.2.1 งานวิจัยเกี่ยวกับการทำนายโครงสร้างด้วย 1 สายลำดับ

งานวิจัยที่เกี่ยวข้องกับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจาก 1 สายลำดับที่นำเสนอในที่นี้แบ่งออกเป็น 2 กลุ่ม โดยกลุ่มแรกเป็นงานวิจัยที่ทำนายโครงสร้างพื้นฐานในส่วนของฮีลิกและลูปชนิดต่าง ๆ แต่ไม่รองรับการทำนายโครงสร้างในส่วนของซูโดนอท และกลุ่มที่สองเป็นงานวิจัยที่สามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่มีซูโดนอทได้

2.2.1.1 งานวิจัยที่ทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่ไม่รวมซูโดนอท

วิธีการแรกสุดและได้รับความนิยมมากสุดในการทำนายโครงสร้างคือกำหนดการพลวัต โดยความพยายามแรกสุดในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอถูกเสนอโดย Nussinov และคณะ [96] ซึ่งใช้กำหนดการพลวัตเพื่อหาจำนวนคู่เบสมากสุด ต่อมา Zuker และคณะ [3] ได้ปรับปรุงกำหนดการพลวัตเพิ่มเติมที่สามารถโมเดลปฏิสัมพันธ์ของค่าพลังงานเพื่อนบ้านใกล้เคียงสุด (nearest neighbor energy interaction) กล่าวโดยสรุป การทำนายโครงสร้างด้วยกำหนดการพลวัต เริ่มต้นค่าพลังงานที่ต่ำสุดถูกกำหนดสำหรับแต่ละชิ้นส่วนย่อย (fragment) ที่เป็นไปได้ของสายลำดับโดยเริ่มจากชิ้นที่สั้นสุดก่อน จากนั้นทำการเรียกตัวเองซ้ำ (recursion) เพื่อสร้างชิ้นส่วนที่ใหญ่ขึ้นเรื่อย ๆ จนท้ายที่สุดเมื่อคำนวณครบทั้งสายลำดับจะได้โครงสร้างของอาร์เอ็นเอที่มีค่าพลังงานต่ำสุด ท่ามกลางขั้นตอนวิธีกำหนดการพลวัตที่ถูกพัฒนาขึ้น โปรแกรมที่อาศัยหลักการของกำหนดการพลวัตเพื่อหาโครงสร้างที่มีค่าพลังงานต่ำสุด เช่น Mfold [4] และ RNAfold [97] เป็นวิธีการที่ได้รับความนิยมในปัจจุบัน กำหนดการพลวัตมีข้อเสียคือวิธีการนี้ให้ผลลัพธ์เฉพาะโครงสร้างที่มีค่าพลังงานต่ำสุด [12]

นอกเหนือจากกำหนดการพลวัต เทคนิคทางด้านการสุ่มเชิงสถิติ (statistical sampling) ถูกนำมาประยุกต์ใช้ในการทำนายโครงสร้าง แรงจูงใจมาจากการที่ถึงแม้ว่าเทคนิคค่าพลังงานต่ำสุดเป็นวิธีการที่นิยมมากในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจาก 1 สายลำดับ แต่แบบจำลองค่าพลังงานก็ยังคงไม่สมบูรณ์ ความคลาดเคลื่อนเพียงเล็กน้อยของพารามิเตอร์ในการคำนวณค่าพลังงานนำไปสู่การกำหนดโครงสร้างที่มีค่าพลังงานต่ำสุดที่แตกต่างกันอย่างมาก [98] นอกจากนี้โครงสร้างที่มีค่าพลังงานต่ำสุดที่ถูกกำหนดโดยขั้นตอนวิธีเหล่านั้นอาจจะไม่ใช่โครงสร้างที่ตรงกับโครงสร้างที่เป็นคำตอบ และโครงสร้างที่เป็นคำตอบอาจเป็นโครงสร้างที่มีค่าพลังงานต่ำรองลงมา งานวิจัย [66] จึงถูกนำเสนอขึ้น สำหรับอาร์เอ็นเอหนึ่ง โครงสร้างทุติยภูมิต่าง ๆ ในโบลทซ์มันน์

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

(Boltzmann) มีความน่าจะเป็นไม่เท่ากัน ทุก ๆ โครงสร้างที่เป็นไปได้จะถูกกำหนดความน่าจะเป็นอ้างอิงตามการแจกแจงความน่าจะเป็นสมดุลของโบลทซ์มันน์ (Boltzmann equilibrium probability distribution) จากนั้นใช้ขั้นตอนวิธีที่เรียกตัวเองซ้ำเพื่อสร้างโครงสร้างที่เป็นตัวแทนจากการแจกแจงนั้น โดยในปี 2005 ทีมวิจัยนี้ได้นำเสนอการใช้โครงสร้างเซนทรอยด์ (centroid structure) เป็นตัวแทนของโครงสร้างใน 1 เซต [99] ซึ่งโครงสร้างเซนทรอยด์คือโครงสร้างที่มีระยะห่างรวมของคู่เบสเมื่อเทียบกับโครงสร้างต่าง ๆ ภายในเซตนั้นต่ำสุด ผลการทดสอบประสิทธิภาพด้วย 81 สายลำดับจากอาร์เอ็นเอ 9 ชนิดเปรียบเทียบกับโครงสร้างที่มีค่าพลังงานต่ำสุด พบว่าโครงสร้างเซนทรอยด์มีความใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากกว่าโครงสร้างที่มีค่าพลังงานต่ำสุด และมีค่าความผิดพลาดในการทำนายที่ต่ำกว่า

ขั้นตอนวิธีอีกกลุ่มหนึ่งที่ได้รับความนิยมในการประยุกต์ใช้สำหรับการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอคือวิธีเมตาฮิวริสติก เช่น RnaPredict [12] โดยพื้นฐานเป็นขั้นตอนวิธีเชิงพันธุกรรมซึ่งทำการเข้ารหัสโครงสร้างทุติยภูมิของอาร์เอ็นเอเป็นการเรียงสับเปลี่ยนเซตของฮีลิกที่เป็นไปได้ทั้งหมด และเพื่อความถูกต้องของโครงสร้างที่ทำนายได้จึงต้องมีการถอดรหัสเพื่อกำจัดฮีลิกที่ขัดแย้งกันทิ้งไป โดยงานวิจัยนี้ใช้ 3 ตัวดำเนินการไขว้เปลี่ยน ได้แก่ การกลายพันธุ์แบบวัฏจักร (Cycle Crossover, CX) การกลายพันธุ์แบบลำดับเวอร์ชัน 2 (Order Crossover #2 ,OX2) และการไขว้เปลี่ยนที่จับคู่บางส่วน (Partially Mapped Crossover, PMX) ประสิทธิภาพของ RnaPredict ถูกทดสอบบน 19 สายลำดับอาร์เอ็นเอเปรียบเทียบกับโครงสร้างคำตอบและโครงสร้างที่ถูกทำนายโดยใช้โปรแกรม Mfold พบว่า RNAPredict ทำนายได้โครงสร้างที่มีค่าพลังงานต่ำกว่าโครงสร้างที่ทำนายได้จาก Mfold และมีประสิทธิภาพที่เทียบเคียงได้กับโครงสร้างที่มีค่าพลังงานต่ำรองลงมาที่คำนวณได้จาก Mfold

SARNA-Predict [13] เป็นขั้นตอนวิธีที่อยู่บนพื้นฐานของการเรียงสับเปลี่ยนที่อาศัยหลักการของแบบจำลองการอบเหนียว โดยทำการเข้ารหัสโครงสร้างทุติยภูมิของอาร์เอ็นเอเป็น 1 การเรียงสับเปลี่ยน จากนั้นใช้ตัวดำเนินการกลายพันธุ์ต่าง ๆ เช่น การกลายพันธุ์แบบสลับ (swap mutation) และ การกลายพันธุ์แบบผกผัน (inversion mutation) และใช้ฟังก์ชันวัตถุประสงค์เป็นการคำนวณค่าพลังงานด้วยแบบจำลอง INNHB และมีการใช้การแจกแจงแบบโบลทซ์มันน์เพื่อกำหนดความน่าจะเป็นที่จะปฏิเสธหรือยอมรับโครงสร้างที่สร้างใหม่เมื่อค่าพลังงานสูงกว่าโครงสร้างก่อนหน้า การจัดการการอบเหนียว (annealing schedule) เปรียบเสมือนเป็นฟังก์ชันสำหรับลดค่าอุณหภูมิจากอุณหภูมิตั้งต้นมีหลายชนิด เช่น monotonic, adaptive, geometric และ quadratic ขั้นตอนวิธีที่นำเสนอถูกทดสอบกับ 13 สายลำดับอาร์เอ็นเอเปรียบเทียบกับโครงสร้างที่เป็นคำตอบ พบว่าการใช้ตัวดำเนินการกลายพันธุ์แบบสลับร่วมกับการจัดการการอบเหนียวแบบปรับเปลี่ยนได้ (adaptive annealing schedule) ให้ค่าความถูกต้องในการทำนายสูงสุด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

TL-PSOfold [14] เป็นขั้นตอนวิธีที่หาค่าเหมาะสมสุดแบบกลุ่มอนุภาคซึ่งแบ่งการทำงานออกเป็น 2 ระดับแต่ละระดับใช้ฟังก์ชันวัตถุประสงค์แตกต่างกัน กล่าวคือ ระดับแรกใช้การหาผลรวมคะแนนของคู่เบสทั้งหมดใน 1 โครงสร้างอ้างอิงตามแบบจำลองพันธะไฮโดรเจน (hydrogen bond model) [100] ซึ่งแต่ละคู่เบสจะถูกกำหนดคะแนนที่แตกต่างกัน (CG = GC = -3, AU = UA = -2 และ GU = UG = -1 ส่วนคู่เบสอื่น ๆ นอกเหนือจากนี้มีค่าเป็น 0) และระดับที่สองใช้ฟังก์ชันวัตถุประสงค์เป็นผลรวมของค่าพลังงานอ้างอิงตามพารามิเตอร์ในฐานข้อมูล NNDB [64] โดยระดับแรกขั้นตอนวิธีที่นำเสนอทำงานกับทั้งปริภูมิค้นหาเพื่อหาผลเฉลยที่ดีสุดของแต่ละกลุ่มอนุภาค (swarm) ในขณะที่ระดับที่สองขั้นตอนวิธีที่นำเสนอจะทำงานกับคำตอบที่ดีที่สุดในแต่ละกลุ่ม (gbest solution) ที่ได้จากระดับที่หนึ่ง ประสิทธิภาพของขั้นตอนวิธีนี้ถูกเปรียบเทียบกับขั้นตอนวิธีที่อาศัยหลักการของการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค ได้แก่ HelixPSO v.1, HelixPSo v.2, PSOfold, SetPSO, IPSO, FPSO และซอฟต์แวร์ที่ได้รับความนิยมในการทำนายโครงสร้างทุติยภูมิของอาร์เอเอ็น ได้แก่ RNAfold และ Mfold นอกจากนี้ยังเปรียบเทียบผลลัพธ์กับวิธีเมตาฮิวริสติกอื่น ๆ ได้แก่ RNAPredict และ SARNA-Predict โดยใช้ตัวชี้วัดเป็นค่าความอ่อนไหว ค่าความจำเพาะ และ F-measure พบว่า ผลลัพธ์ของ TL-PSOfold มีค่าความถูกต้องในการทำนายดีกว่าทุกวิธีที่นำมาเปรียบเทียบ

2.2.1.2 งานวิจัยที่ทำนายโครงสร้างอาร์เอ็นเอที่มีซูโดนอท

pknotsRE [101] เป็นกำหนดการพลวัตสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่เหมาะสมสุดที่มีซูโดนอท ใช้เวลาเป็น $O(N^6)$ และหน่วยความจำ $O(N^4)$ เมื่อ N คือ ความยาวของสายลำดับอาร์เอ็นเอ วิธีการที่นำเสนอใช้พารามิเตอร์ทางอุณหพลศาสตร์ (thermodynamic parameter) มาตรฐานที่มีการเพิ่มเติมพารามิเตอร์เล็กน้อยสำหรับอธิบายค่าพลังงานของซูโดนอท เนื่องจากวิธีการที่นำเสนอมีความต้องการด้านเวลาและหน่วยความจำที่สูงทำให้สามารถประมวลผลได้แค่ในโมเลกุลสายสั้น ๆ แต่งานวิจัยนี้ถือเป็นขั้นตอนวิธีแรกที่สามารถพับโครงสร้างอาร์เอ็นเอที่มีซูโดนอทด้วยแบบจำลองทางอุณหพลศาสตร์มาตรฐาน

HotKnots [102] เป็นขั้นตอนวิธีฮิวริสติกเพื่อทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่มีซูโดนอท วิธีการที่นำเสนอจะค่อย ๆ เลือกโครงสร้างย่อยมาประกอบกันเป็นโครงสร้างที่ใหญ่ขึ้นเรื่อย ๆ และใช้แบบจำลองการคำนวณค่าพลังงานมาตรฐานที่ดำเนินการกับโครงสร้างที่ไม่มีซูโดนอทก่อน จากนั้นจึงขยายให้สามารถดำเนินการกับซูโดนอทได้ ผลการทดสอบขั้นตอนวิธีที่นำเสนอด้วย 43 สายลำดับอาร์เอ็นเอจากฐานข้อมูล Pseudobase และวรรณกรรมต่าง ๆ ที่ดำเนินการกับโครงสร้างที่มีซูโดนอท โดยแบ่งออกเป็น 2 กลุ่ม คือ สายลำดับที่สั้นมีความยาวในช่วง 28 – 108 นิวคลีโอไทด์ และ สายลำดับที่ยาวขึ้นมีความยาวในช่วง 210 – 400 นิวคลีโอไทด์เปรียบเทียบกับ 5 ขั้นตอนวิธี

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ได้แก่ pknotsRE [101] , NUPACK [103], pknotsRG-mfe [104], ILM [70] และ STAR [105] โดยใช้ตัวชี้วัดเป็นค่าความอ่อนไหว และ ค่าความจำเพาะ พบว่า กรณีเฉลี่ยจากทุกสายลำดับอาร์เอ็นเอในกลุ่มสายลำดับที่สั้นขั้นตอนวิธีที่นำเสนอได้ผลลัพธ์ดีสุด และได้ผลลัพธ์เท่ากับ pknotsRG-mfe ด้วยค่าความอ่อนไหวและค่าความจำเพาะเป็น 76% และ 77% ตามลำดับ แต่วิธีการที่นำเสนอใช้เวลารันที่น้อยกว่าในกลุ่มสายลำดับที่ยาวขึ้น และทำผลลัพธ์ได้ดีเป็นอันดับสองรองจาก STAR โดยได้ค่าความอ่อนไหวน้อยกว่าวิธีที่ได้ผลลัพธ์ดีสุดที่นำมาเปรียบเทียบ 5% และ ค่าความจำเพาะต่ำกว่าวิธีที่ได้ผลลัพธ์ดีสุด 3%

งานวิจัย [106] นำเสนอวิธีการใหม่เพื่อทำนายโครงสร้างที่มีซูโดนอท โดยมีแรงจูงใจจากสมมุติฐานที่ว่าโครงสร้างอาร์เอ็นเอพับตัวแบบลำดับชั้น ด้วยการสร้างคู่เบสในลักษณะที่ไม่มีซูโดนอทก่อนแล้วค่อยสร้างซูโดนอทที่มีค่าพลังงานต่ำสุดที่สัมพันธ์กับโครงสร้างที่สร้างไปแล้ว วิธีการที่นำเสนอใช้เวลา $O(N^3)$ ซึ่งมีความซับซ้อนเท่ากับขั้นตอนวิธีที่ดีสุดที่สามารถทำนายโครงสร้างทุติยภูมิแบบไม่มีซูโดนอทด้วยเทคนิคการทำนายโครงสร้างที่มีค่าพลังงานต่ำสุด นอกจากนี้ วิธีการที่นำเสนอสามารถจัดการโครงสร้างทางชีววิทยาต่าง ๆ ได้แก่ kissing hairpins และ nested kissing hairpin ซึ่งก่อนหน้านี้ใช้เวลา $\Theta(N^6)$

งานวิจัย [107] เป็นขั้นตอนวิธีที่ทำการผสานขั้นตอนวิธี P-RnaPredict [108] ที่สามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอแบบไม่มีซูโดนอทเข้ากับแบบจำลองอุณหพลศาสตร์ที่สามารถคำนวณค่าพลังงานของโครงสร้างที่มีซูโดนอทจาก HotKnots [102] ประสิทธิภาพของขั้นตอนวิธีที่นำเสนอเมื่อทดสอบด้วย 8 สายลำดับอาร์เอ็นเอถูกเปรียบเทียบกับวิธี HotKnots แบบดั้งเดิมและโครงสร้างที่เป็นคำตอบ พบว่าวิธีการที่นำเสนอมีค่าความอ่อนไหวและค่าความจำเพาะดีขึ้นอย่างมีนัยสำคัญ

ProbKnot [56] เป็นขั้นตอนวิธีที่สามารถทำนายซูโดนอทที่มีรูปแบบโครงสร้างใด ๆ ในเวลา $O(N^3)$ เมื่อ N เป็นความยาวของสายลำดับอาร์เอ็นเอ เริ่มต้นด้วยการคำนวณความน่าจะเป็นที่เบสจะจับคู่กันด้วยฟังก์ชันพาร์ทิชัน [64] ซึ่งยังไม่รองรับโครงสร้างที่มีซูโดนอท จากนั้นทำการสร้างโครงสร้างที่มีค่าความถูกต้องที่คาดหวังมากสุดในเวลา $O(N^2)$ ประสิทธิภาพของวิธีการนี้ถูกเปรียบเทียบกับโครงสร้างที่เป็นคำตอบในฐานข้อมูลขนาดใหญ่ด้วยสายลำดับที่ยาวน้อยกว่า 700 นิวคลีโอไทด์ได้ค่าความอ่อนไหวและค่าความจำเพาะเป็น 69.3% และ 61.3% ตามลำดับ ซึ่งวิธีการที่นำเสนอให้ผลลัพธ์ดีสุดในบรรดาขั้นตอนวิธีที่นำมาเปรียบเทียบ ได้แก่ pknotsRG [104], ILM [70], Hotknots [102]

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

IPknot [109] เป็นวิธีการที่อยู่บนพื้นฐานของการโปรแกรมเชิงจำนวนเต็มสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่มีซูโดนอท วิธีการนี้แยก 1 โครงสร้างที่มีซูโดนอทออกเป็นเซตของโครงสร้างย่อยที่ยังไม่มีซูโดนอทและคำนวณความน่าจะเป็นที่เบสจะเข้าคู่กันโดยคำนึงถึงซูโดนอทด้วย จากนั้นใช้การโปรแกรมเชิงจำนวนเต็มเพื่อทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ วิธีการที่นำเสนอสามารถทำนายโครงสร้างที่มีซูโดนอทชนิดต่าง ๆ ได้หลากหลาย และใช้เวลาในการประมวลผลที่ค่อนข้างน้อย นอกจากนี้ ผู้วิจัยยังนำเสนอขั้นตอนวิธีฮิวริสติกเพื่อปรับปรุงความน่าจะเป็นของคู่เบสทำให้สามารถปรับปรุงความถูกต้องในการทำนายของ IPknot ให้ดีขึ้น ยิ่งไปกว่านั้นงานวิจัยนี้ยังรองรับการทำนายโครงสร้างที่มีซูโดนอทในกรณีที่มีการระบุสายลำดับที่มีการจัดตำแหน่งมาให้ได้อีกด้วย ขั้นตอนวิธีที่นำเสนอถูกประเมินโดยใช้ข้อมูล 3 ชุด ชุดแรกเรียกว่า RS-pk388 ได้จากฐานข้อมูล RNA STRAND โดยเลือกสายลำดับที่มีอย่างน้อย 1 ซูโดนอท และมีความยาวระหว่าง 150 – 500 นิวคลีโอไทด์รวมทั้งหมด 388 อาร์เอ็นเอ ชุดที่สองเรียกว่า pk168 ที่นำเสนอใน [110] ซึ่งประกอบด้วยซูโดนอททั้งหมด 16 ชนิดความยาวน้อยกว่า 140 นิวคลีโอไทด์รวมทั้งหมด 168 อาร์เอ็นเอ และชุดที่สามเรียกว่า Rfam-PK จำนวน 67 ข้อมูล โดยเลือกจาก Rfam families ที่เป็นไปตาม 3 เงื่อนไข คือ 1) มีอย่างน้อย 1 ซูโดนอท 2) มีความยาวไม่เกิน 500 นิวคลีโอไทด์ 3) มีการจัดตำแหน่งมาจากอย่างน้อย 5 สายลำดับ ในภาพรวมพบว่าขั้นตอนวิธีที่นำเสนอให้ค่าความถูกต้องในการทำนายที่ดีกว่าและเร็วกว่าเมื่อเทียบกับขั้นตอนวิธีที่นำมาเปรียบเทียบ

2.2.2 งานวิจัยเกี่ยวกับการทำนายโครงสร้างโดยใช้หลายสายลำดับ

หลาย ๆ ขั้นตอนวิธีในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอใช้เพียง 1 สายลำดับ การดำเนินการเช่นนี้เพียงพอในบางสถานการณ์ แต่เมื่อไหร่ก็ตามที่สายลำดับอาร์เอ็นเอที่สัมพันธ์กันสามารถหาได้ สารสนเทศเช่นนั้นก็ควรถูกรวมในการวิเคราะห์เชิงโครงสร้างเพื่อให้ผลลัพธ์ที่ดียิ่งขึ้น

2.2.2.1 งานวิจัยที่จัดตำแหน่งสายลำดับก่อนแล้วจึงพับโครงสร้าง

งานวิจัยในกลุ่มนี้ใช้สายลำดับที่มีการจัดตำแหน่งเป็นข้อมูลนำเข้าและทำนายโครงสร้างของแต่ละสายลำดับโดยการหาโครงสร้างที่มีร่วมกัน ข้อจำกัดของงานวิจัยในกลุ่มนี้คือค่าความถูกต้องของการทำนายโครงสร้างขึ้นอยู่กับคุณภาพของการจัดตำแหน่ง [75]

Pfold [8] นำเสนอแนวทางที่ทำงานได้จริงในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่สายลำดับต่าง ๆ ที่เกี่ยวข้องกันถูกกำหนดมาให้ วิธีการนี้ทำการปรับปรุงขั้นตอนวิธีก่อนหน้าคือ KH-99 [111] ที่รวมแบบจำลองการวิวัฒนาการที่ชัดเจน (explicit evolutionary model) ของหลายสายลำดับอาร์เอ็นเอเข้ากับแบบจำลองเชิงสถิติ เมื่อกำหนดสายลำดับที่มีการจัดตำแหน่งมาให้วิธีการนี้จะทำนายโครงสร้างที่มีร่วมกันของทุกสายลำดับ Pfold มีเป้าหมายที่จะปรับปรุงขั้นตอนวิธี KH-99

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ให้ทำงานเร็วขึ้น สามารถวิเคราะห์สายลำดับได้จำนวนมากขึ้น และทนทานต่อความผิดพลาดมากยิ่งขึ้น Pfold ใช้ SCFGs เพื่อทำนายโครงสร้างจาก 1 การจัดตำแหน่งและคำนวณระยะห่างระหว่าง 2 สายลำดับใด ๆ โดยใช้ความเป็นไปได้สูงสุด (maximum likelihood) ผลการทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอพบว่าค่าความถูกต้องในการทำนายเพิ่มขึ้นตามจำนวนสายลำดับที่ถูกจัดตำแหน่งซึ่งเป็นผลมาจากข้อมูลความแปรปรวนร่วม (covariance) ที่มากขึ้นสามารถถูกใช้ แสดงให้เห็นว่าเมื่อสายลำดับที่สัมพันธ์กันสามารถใช้ได้สารสนเทศดังกล่าวควรถูกใช้ในการทำนายโครงสร้างเพื่อเพิ่มประสิทธิภาพในการทำนาย

RNAalifold [3] เวอร์ชันแรกเป็นการประยุกต์ใช้กำหนดการพลวัตมาตรฐานในการทำนายโครงสร้างของอาร์เอ็นเอที่มีการจัดตำแหน่ง โดย RNAalifold เวอร์ชันใหม่ [74] มีการปรับปรุงค่าความถูกต้องในการทำนายให้ดีขึ้นด้วยการนำเสนอการจัดการช่องว่าง (alignment gap) ที่ดีขึ้นเปลี่ยนจากคะแนนความแปรปรวนร่วมที่ง่าย ๆ เป็นคะแนน RIBOSUM-like ที่มีความซับซ้อนมากขึ้น ผลจากการปรับปรุงสิ่งเหล่านี้ทำให้ RNAalifold ไม่เพียงแต่ดีกว่าเวอร์ชันดั้งเดิมแต่ยังสามารถแข่งขันได้กับวิธีการอื่น ๆ ได้แก่ วิธีที่อาศัยหลักการ SCFGs วิธีที่ใช้เทคนิคการทำนายโครงสร้างที่มีค่าความถูกต้องคาดหวังสูงสุด หรือ การจำแนกความใกล้เคียงสุดเชิงลำดับชั้น (hierarchical nearest classifiers)

2.2.2.2 งานวิจัยที่จัดตำแหน่งสายลำดับและพับโครงสร้างไปพร้อมกัน

ในทางปฏิบัติการจัดตำแหน่งสายลำดับเพียงอย่างเดียวไม่เหมาะสมสำหรับการทำนายโครงสร้างที่ 2 สายลำดับมีความคล้ายคลึงกันแค่ประมาณ 50 % แต่อย่างไรก็ตาม วิธีการจัดตำแหน่งในเชิงโครงสร้างก็สามารถช่วยปรับปรุงประสิทธิภาพของการจัดตำแหน่งให้ดีขึ้นได้จึงเกิดขั้นตอนวิธีในกลุ่มนี้ขึ้น

ขั้นตอนวิธี Sankoff [76] เป็นวิธีการแรกที่นำเสนอการจัดตำแหน่งและการพับของสายลำดับอาร์เอ็นเอจำนวนหนึ่งไปพร้อมกันใน 1 การคำนวณ ในอดีตวิธีการนี้ถูกมองว่าเป็นไปได้ยากในทางปฏิบัติเนื่องจากใช้เวลา $O(N^{3s})$ และ ใช้หน่วยความจำ $O(N^{2s})$ เมื่อ s คือจำนวนสายลำดับอาร์เอ็นเอและ N คือความยาวของสายลำดับ ซึ่งค่อนข้างแพงโดยเฉพาะอย่างยิ่งเมื่อทำการวิเคราะห์ข้อมูลมากกว่า 2 สายลำดับ แต่ในปัจจุบันวิธีการนี้ค่อนข้างเป็นที่แพร่หลายและถูกนำไปดัดแปลงเกิดเป็นงานวิจัยต่าง ๆ มากมายดังตัวอย่างที่จะนำเสนอต่อไปนี้

Dynalign [112] ปรับปรุงค่าความถูกต้องในการทำนายโครงสร้างโดยรวมการคำนวณค่าพลังงานต่ำสุดเข้ากับการวิเคราะห์สายลำดับที่นำมาเปรียบเทียบกันเพื่อหา 1 โครงสร้างที่มีเหมือนกันใน 2 สายลำดับที่ให้ค่าพลังงานต่ำสุด วิธีการที่นำเสนอใช้กำหนดการพลวัตที่เสนอโดย Sankoff [76] โดยวิธีการนี้มีการจำกัดความยาวของสายลำดับที่จะทำการวิเคราะห์ว่าห้ามต่างกันเกิน M ผลก็คือได้

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

เวลาเป็น $O(M^3N^3)$ เมื่อ N คือ ความยาวของสายลำดับที่สั้นสุด ค่าความถูกต้องของวิธีการที่นำเสนอถูกทดสอบกับ tRNA ในกรณีเฉลี่ยวิธีการนี้ได้ค่าความถูกต้อง 86.1% เมื่อเทียบกับโครงสร้างคำตอบ แต่หากใช้แค่การคำนวณค่าพลังงานต่ำสุดเพียงอย่างเดียวได้ค่าความถูกต้อง 59.7% และเมื่อทดสอบกับ 5S rRNA ค่าความถูกต้องเฉลี่ยเพิ่มขึ้นจาก 47.8% เป็น 86.4% นอกจากนี้ เมื่อเปรียบเทียบกับเทคนิคการทำนายโครงสร้างที่มีค่าพลังงานต่ำสุดที่ใช้แค่เพียง 1 สายลำดับค่าความถูกต้องสำหรับ rRNA มีค่าความอ่อนไหวเพิ่มขึ้นจาก 47.4% เป็น 73.3% และค่าความจำเพาะเพิ่มขึ้นจาก 47.5% เป็น 73.1% [113] การปรับปรุง Dynalign ที่สามารถลดต้นทุนการคำนวณ ลดเวลาในการคำนวณ และมีค่าความถูกต้องที่มากขึ้นถูกนำเสนอใน [114] ซึ่งมีการคำนวณความน่าเป็นของคู่เบสของแต่ละสายลำดับและยอมให้เฉพาะคู่เบสที่มีความน่าจะเป็นเกินค่าขีดแบ่งค่าหนึ่ง (threshold) ถูกนำไปสร้างเป็นส่วนหนึ่งของโครงสร้าง

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

โครงสร้างที่ถูกทำนายด้วยวิธีการจัดตำแหน่งเชิงโครงสร้างมีความถูกต้องมากกว่าการทำนายโครงสร้างที่ใช้เพียง 1 สายลำดับเนื่องจากสามารถใช้สารสนเทศที่ได้จากการเปรียบเทียบสายลำดับเหล่านั้นมาร่วมในการพิจารณาได้ ปัญหาหลักของวิธีการจัดตำแหน่งเชิงโครงสร้างส่วนใหญ่คือใช้ต้นทุนการคำนวณสูงเกินไป งานวิจัย [78] จึงนำเสนอวิธีฮิวริสติกในการตัดเล็ม (pruning heuristic) ที่ทำให้ FOLDALIGN version 1.0 [77] เร็วขึ้นและใช้หน่วยความจำน้อยลง กล่าวคือ ขั้นตอนวิธีดังกล่าวทำการจัดตำแหน่งเชิงโครงสร้างของ 2 สายลำดับอาร์เอ็นเอ มีการใช้แบบจำลองค่าพลังงานที่มีน้ำหนักเบา (lightweight energy model) และค่าความคล้ายคลึงของสายลำดับเพื่อทำการพับโครงสร้างและจัดตำแหน่งสายลำดับไปพร้อมกัน ขั้นตอนวิธีที่นำเสนอสามารถจัดตำแหน่งของสายลำดับที่มีความแตกต่างกันมาก ๆ ได้เร็วขึ้นอย่างมีนัยสำคัญโดยไม่ทำให้ประสิทธิภาพการทำนายลดลง นอกจากนี้ ความต้องการหน่วยความจำก็ลดลงด้วย ทำให้สามารถทำการวิเคราะห์สายลำดับที่ยาวขึ้นได้

จากตัวอย่างงานวิจัยในข้างต้น พบว่า ขั้นตอนวิธีในกลุ่มนี้ค่อนข้างช้าเมื่อเทียบกับขั้นตอนวิธีในกลุ่มแรก ส่งผลให้ขั้นตอนวิธีเหล่านี้ถูกจำกัดอยู่ที่การดำเนินการแค่กับ 2 สายลำดับ [75]

2.2.2.3 งานวิจัยที่ใช้หลายสายลำดับและรองรับการทำนายซูโดนอท

ขั้นตอนวิธีเกือบทั้งหมดที่นำเสนอในหัวข้อ 2.2.2.1 – 2.2.2.2 พิจารณาเฉพาะโครงสร้างที่ไม่มีซูโดนอท ในหัวข้อนี้ นำเสนองานวิจัยที่เกี่ยวข้องกับการทำนายโครงสร้างจากหลาย ๆ สายลำดับสำหรับทำนายโครงสร้างที่มีซูโดนอท

งานวิจัย [70] นำเสนอขั้นตอนวิธีการจับคู่ลูปแบบวนทำซ้ำ (iterated loop matching) สำหรับทำนายโครงสร้างอาร์เอ็นเอที่มีซูโดนอท วิธีการที่นำเสนอสามารถใช้ประโยชน์จากอุณหพลศาสตร์หรือข้อมูลจากการเปรียบเทียบ (comparative information) หรือทั้งคู่ เพื่อให้สามารถ

ทำนายซูโดนอทได้จากทั้งหลายสายลำดับที่มีการจัดตำแหน่งและสายลำดับเดี่ยว ผลการทดสอบด้วยอาร์เอ็นเอชนิดต่าง ๆ โดยใช้ 8-12 สายลำดับที่มีความคล้ายคลึงกันเปรียบเทียบกับวิธีการจับคู่ที่มีค่าน้ำหนักมากสุด (maximum weighted matching, MWM) [115] พบว่า ในสายลำดับที่มีความยาวน้อยกว่า 300 นิวคลีโอไทด์ ขั้นตอนวิธีที่นำเสนอสามารถระบุคู่เบสได้ถูกต้องเกิน 90% ในขณะที่วิธี MWM ได้ความถูกต้อง 60-85% และ กรณีเฉลี่ยขั้นตอนวิธีที่นำเสนอได้ความถูกต้อง 80% ในขณะที่ MWM ได้ความถูกต้อง 59.2% แสดงให้เห็นว่าวิธีการที่นำเสนอให้ค่าความถูกต้องสูงขึ้น นอกจากนี้ในสายลำดับเดี่ยวเมื่อเปรียบเทียบขั้นตอนวิธีที่นำเสนอกับขั้นตอนวิธี PKNOTS [101] พบว่าวิธีการที่นำเสนอมีค่าความถูกต้องในการทำนายที่สูงกว่าและใช้เวลาในการคำนวณต่ำกว่ามาก

SimulFold [116] ใช้วิธีมอนติคาร์โลลูกโซ่มาร์คอฟแบบเบย์ (Bayesian Markov chain Monte Carlo) เพื่อสุ่มโครงสร้างจากการแจกแจงร่วมภายหลัง (joint posterior distribution) ของโครงสร้างอาร์เอ็นเอต่าง ๆ ใช้การจัดตำแหน่งของสายลำดับ และ ต้นไม้วิวัฒนาการ (evolutionary tree) ที่สัมพันธ์กับสายลำดับเหล่านั้น เมื่อเปรียบเทียบวิธีการที่นำเสนอกับโปรแกรมอื่น ๆ ได้แก่ RNAalifold [117], Hxmatch [118], Pfold [8] และ CARNAC [119] พบว่า ในภาพรวมขั้นตอนวิธีที่นำเสนอให้คุณภาพการทำนายที่สูงกว่าหลาย ๆ วิธีที่นำมาเปรียบเทียบในการตรวจจับโครงสร้างอาร์เอเอ็นที่สอดคล้องกันซึ่งหมายรวมไปถึงโครงสร้างที่มีซูโดนอทด้วย

TurboKnot [120] ทำการปรับปรุงการทำนายโครงสร้างที่มีซูโดนอทด้วยการทำนายโครงสร้างที่สอดคล้องกันโดยใช้ 2 สายลำดับขึ้นไปที่มีความคล้ายคลึงกัน ทำการหาบริเวณที่สอดคล้องกันของสายลำดับเหล่านั้น TurboKnot สร้างโครงสร้างโดยใช้เทคนิคการทำนายโครงสร้างที่มีค่าความถูกต้องที่คาดหวังสูงสุดในลักษณะเดียวกับที่ใช้ใน ProbKNot [56] แต่ TurboKnot คำนวณความน่าจะเป็นของคู่เบสจากหลาย ๆ สายลำดับ ผลการเปรียบเทียบกับ ILM [70], Hxmatch [118] , ProbKnot [56], TurboFold [121] และ MEA [62] บนอาร์เอ็นเอ 7 ชนิด พบว่า กรณีเฉลี่ยขั้นตอนวิธีที่นำเสนอมีค่าความอ่อนไหวและค่าความจำเพาะเป็น 79.8 และ 72.9 ตามลำดับซึ่งค่าความอ่อนไหวกรณีเฉลี่ยของวิธีที่นำเสนอดีกว่าทุกวิธีที่นำมาเปรียบเทียบ แต่ในกรณีของค่าความจำเพาะกรณีเฉลี่ย TurboFold เป็นขั้นตอนวิธีที่ได้ผลลัพธ์ดีสุด ในขณะที่ขั้นตอนวิธีที่นำเสนอทำผลลัพธ์ได้ดีรองลงมา

แทนการคำนวณว่าสายลำดับที่กำหนดให้มีความเป็นไปได้ที่จะพับเป็นโครงสร้างรูปร่างใดบางงานวิจัยใช้การสุ่มโครงสร้าง เช่น PhyloQFold [122] นำเสนอวิธีการที่ใช้ข้อดีของประวัติการวิวัฒนาการของกลุ่มสายลำดับอาร์เอ็นเอที่ถูกจัดตำแหน่งเพื่อสุ่มโครงสร้างทุติยภูมิที่สอดคล้องกันที่มีซูโดนอทอ้างอิงตามค่าความน่าจะเป็นภายหลัง (posterior probability) ที่ประมาณได้จากสายลำดับเหล่านั้น ซึ่งขั้นตอนวิธีที่นำเสนอทำการปรับปรุง McQFold [123] โดยใช้ข้อมูลการจัดตำแหน่งของหลาย ๆ สายลำดับ และ 1 ต้นไม้วิวัฒนาการเป็นข้อมูลนำเข้า เมื่อเปรียบเทียบกับขั้นตอนวิธี

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

อื่น ๆ ได้แก่ RNAalifold [117], Pfold [8], KNetFold [124], SimulFold [116] และ IPKnot [109] บนสายลำดับของอาร์เอ็นเอชนิด RNase P โดยใช้ตัวชี้วัด MMC พบว่า วิธีการที่นำเสนอให้ค่ามัธยฐานสูงสุด (0.739) ซึ่งมากกว่าค่ามัธยฐานของวิธีการอื่น ๆ 11.8-28.7% นอกจากนี้ วิธีการที่นำเสนอให้ค่าส่วนเบี่ยงเบนมาตรฐานต่ำสุด (0.112)

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้นำเสนอขั้นตอนวิธี Hybrid-EDAfold ซึ่งเป็นขั้นตอนวิธีเชิงวิวัฒนาการที่อยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจงสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอชนิดต่าง ๆ ภาพรวมของขั้นตอนวิธีที่นำเสนอเป็นดังรูปที่ 3.1

สายลำดับอาร์เอ็นเอ

GGGGAUGAAUUAGUUUAGAUAUUUAAGUAAUCCAUGAAAAUUCAUCAGAAUUUUGUAAUAUUGCUUAGAUAAAACCAUGG
GCAGUACAUGGCAUCUCCACCA

1.ระบุบริเวณฮีลิกที่เป็นไปได้ทั้งหมด

หมายเลขฮีลิก	ตำแหน่งเบสเริ่มต้น	ตำแหน่งเบสสุดท้าย	ความยาวฮีลิก
1	1	116	12
2	6	42	3
...	7	59	3
130	93	101	2

2. ทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold

ลำดับที่	โครโมโซม
1	{1, 22, 42, 46, 89, 106, 124}
2	{1, 19, 22, 42, 46, 94, 103, 111, 116}
...	{7, 22, 42, 46, 91, 101, 118, 124}
N	{1, 19, 22, 41, 45, 48, 94, 103, 111, 116}

แสดงผลการทำนายโครงสร้างในรูปแบบเครื่องหมายจุด-วงเล็บ

((((((............(((((((((((.....((((((....)))))).....)))))))))))...(((((.......)))))))))))....

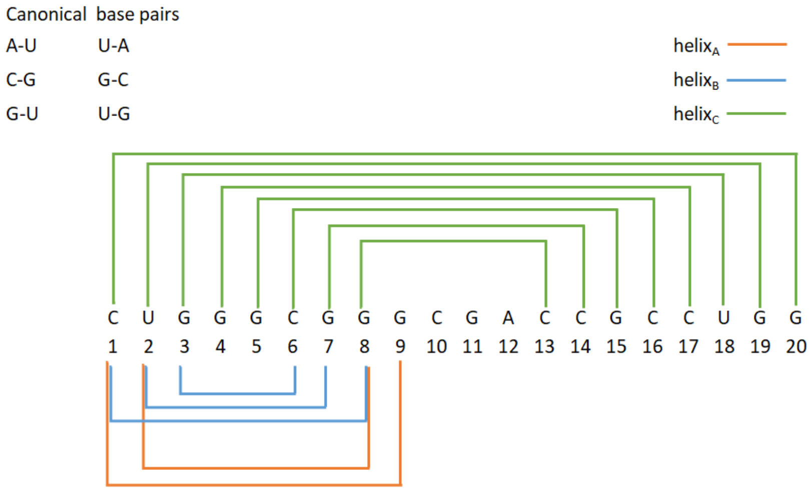
รูปที่ 3.1 ภาพรวมของขั้นตอนวิธีการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 3.1 ขั้นตอนการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอสามารถแบ่งออกเป็น 2 ส่วนหลัก ๆ ส่วนแรกคือการเตรียมเซตของฮีลิกซึ่งเป็นบริเวณที่เบสมีการจับคู่เรียงต่อเนื่องกันที่สามารถพบได้ในสายลำดับอาร์เอ็นเอที่เป็นข้อมูลนำเข้า ขั้นตอนนี้เป็นการเตรียมข้อมูลก่อนเข้าสู่กระบวนการทำนายโครงสร้างซึ่งเนื้อหาในส่วนนี้อธิบายในหัวข้อ 3.1 และส่วนที่สองคือการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold อธิบายในหัวข้อ 3.2 การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่งานวิจัยนี้นำเสนอดำเนินการอยู่ภายใต้แนวคิดของการเลือกเซตย่อยของฮีลิกที่เตรียมไว้ในขั้นตอนแรกมาประกอบกันและทำการวิวัฒนาการเพื่อปรับปรุงโครงสร้างที่ทำนายได้เหล่านี้ให้มีค่าที่ดีขึ้นอ้างอิงตามฟังก์ชันวัตถุประสงค์ที่เลือกใช้ และหัวข้อ 3.3 อธิบายวิธีการประเมินค่าความถูกต้องของโครงสร้างที่ทำนายได้ รายละเอียดเป็นดังนี้

3.1 ระบุฮีลิกที่เป็นไปได้ทั้งหมดใน 1 สายลำดับที่เป็นข้อมูลนำเข้า

ก่อนเข้าสู่กระบวนการทำนายโครงสร้างด้วยขั้นตอนวิธี Hybrid-EDAfold ในขั้นตอนนี้จะทำการระบุบริเวณฮีลิกที่เป็นไปได้ทั้งหมดที่สามารถพบได้ในสายลำดับอาร์เอ็นเอเตรียมไว้ก่อน โดยฮีลิกคือบริเวณที่เบสมีการจับคู่เรียงต่อเนื่องกันไป ในที่นี้สนใจเฉพาะการจับคู่กันของคู่เบสคาร์โนนิคอลได้แก่ AU, CG, GC, GU, UA, UG เช่น สายลำดับที่มีความยาว 20 นิวคลีโอไทด์ สามารถระบุบริเวณที่เป็นฮีลิกได้ดังรูปที่ 3.2



รูปที่ 3.2 ตัวอย่างการระบุบริเวณที่เป็นฮีลิกในสายลำดับอาร์เอ็นเอยาว 20 นิวคลีโอไทด์

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 3.2 เป็นตัวอย่างการระบุฮีลิกโดยใช้ความรู้ของคู่เบสคาร์โนนิคอล จากตัวอย่างสามารถสร้างฮีลิกได้ทั้งหมด 3 ชิ้น ได้แก่ $helix_A$, $helix_B$ และ $helix_C$ โดย $helix_A$ แทนความหมายว่าเบสตำแหน่งที่ 1 จับคู่กับเบสตำแหน่งที่ 9 (CG) และ เบสตำแหน่งที่ 2 จับคู่กับเบสตำแหน่งที่ 8 (UG) ข้อมูลฮีลิกจะถูกเข้ารหัสโดยใช้ 3 พารามิเตอร์ในรูปแบบของ [*start*; *end*; *len*] ในทำนองเดียวกับวิธีการสร้างฮีลิกที่ถูกนำเสนอใน [80] โดยพารามิเตอร์ *start* แทนตำแหน่งเบสเริ่มต้นในฮีลิก พารามิเตอร์ *end* แทนตำแหน่งเบสที่จับคู่กับเบสในตำแหน่ง *start* และพารามิเตอร์ *len* แทนความยาวของฮีลิกหรือจำนวนคู่เบสที่พบในฮีลิกนั้น ดังนั้น $helix_A$ จะถูกเข้ารหัสเป็น [1; 9; 2] $helix_B$ จะถูกเข้ารหัสเป็น [1; 8; 3] และ $helix_C$ จะถูกเข้ารหัสเป็น [1; 20; 8]

วิธีการระบุฮีลิกที่ต่างกันส่งผลต่อจำนวนชิ้นของฮีลิกที่สร้างได้ หากในขั้นตอนการจัดเตรียมฮีลิกสามารถระบุตำแหน่งของฮีลิกได้ตรงกับฮีลิกที่เกิดขึ้นจริงในโครงสร้างที่เป็นคำตอบและมีจำนวนฮีลิกที่ใกล้เคียงกับจำนวนฮีลิกที่พบจริงในโครงสร้างที่เป็นคำตอบก็จะยิ่งส่งผลให้ขั้นตอนวิธีที่นำเสนอสามารถทำนายโครงสร้างได้ถูกต้องแม่นยำมากยิ่งขึ้น

จากการทบทวนวรรณกรรมต่าง ๆ พบว่า มี 2 แนวทางในการระบุบริเวณฮีลิกที่สามารถพบได้ในสายลำดับอาร์เอ็นเอ ได้แก่ การใช้ความรู้เกี่ยวกับคู่เบสคาร์โนนิคอล [125] และการใช้ข้อมูลความน่าจะเป็นของคู่เบส [126] ซึ่งผลจากการเปรียบเทียบประสิทธิภาพในแง่ของจำนวนและความถูกต้องในการระบุตำแหน่งของคู่เบส พบว่าการใช้ข้อมูลความน่าจะเป็นของคู่เบสเป็นเกณฑ์ในการระบุตำแหน่งของฮีลิกให้ผลลัพธ์ที่ดีกว่า กล่าวคือ มีจำนวนชิ้นของฮีลิกน้อยกว่าและฮีลิกเหล่านั้นประกอบด้วยคู่เบสที่มีตำแหน่งตรงกับตำแหน่งคู่เบสที่พบจริงในโครงสร้างคำตอบมากกว่าอีกวิธีการหนึ่ง ดังนั้น ในขั้นตอนการจัดเตรียมฮีลิกงานวิจัยนี้จึงเลือกใช้ข้อมูลความน่าจะเป็นของคู่เบสในการดำเนินการ รายละเอียดเป็นดังนี้

3.1.1 การระบุฮีลิกโดยใช้ข้อมูลความน่าจะเป็นของคู่เบส

การระบุฮีลิกด้วยการใช้ข้อมูลความน่าจะเป็นของคู่เบสถูกนำเสนอใน [126] ดำเนินการโดยสร้างเมทริกซ์ขนาด $n \times n$ เมื่อ n คือ ความยาวของสายลำดับอาร์เอ็นเอ แต่ละสมาชิกในเมทริกซ์จะเก็บความน่าจะเป็นที่เบสตำแหน่งที่ตรงกับแถวที่ i จะจับคู่กับเบสตำแหน่งที่ตรงกับคอลัมน์ที่ j ซึ่งในงานวิจัยนี้ใช้ความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold [6] จากนั้นจัดกลุ่มคู่เบสที่มีค่าความน่าจะเป็นมากกว่า 0 ที่เรียงต่อเนื่องกันเป็น 1 ฮีลิกและทำการเข้ารหัสไว้ รายละเอียดเป็นดังนี้

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

1. นำสายลำดับอาร์เอ็นเอที่เป็นข้อมูลนำเข้าไปคำนวณค่าความน่าจะเป็นของคู่เบสด้วยโปรแกรม RNAfold

2. สร้างเมทริกซ์ขนาด $n \times n$ เมื่อ n คือความยาวของสายลำดับอาร์เอ็นเอที่เป็นข้อมูลนำเข้า แต่ละสมาชิกในเมทริกซ์จะเก็บความความน่าจะเป็นที่เบสตำแหน่งที่ตรงกับหมายเลขแถวจับคู่กับเบสตำแหน่งที่ตรงกับหมายเลขคอลัมน์นั้น ๆ

2.1. ถ้าผลลัพธ์ในข้อ 1 พบข้อมูลความน่าจะเป็นที่เบสแถวที่ *i* จับคู่กับเบสคอลัมน์ที่ *j* กำหนดค่าสมาชิกของ matrix[i][j] เท่ากับความน่าจะเป็นที่ได้

2.2. ไม่เช่นนั้น matrix[i][j] มีค่าเป็น 0

3. พิจารณาบริเวณสามเหลี่ยมครึ่งบนของเมทริกซ์ ถ้าพบว่ามีสมาชิกที่มีค่ามากกว่า 0 เรียงต่อเนื่องกันในแนวทแยงมุมจำนวนตั้งแต่ 2 ตัวขึ้นไปให้ระบุบริเวณนั้นเป็น 1 ฮีลิก จากนั้นทำการเข้ารหัสข้อมูลฮีลิกในรูปแบบ [หมายเลขแถว; หมายเลขคอลัมน์; ความยาว] ดังเช่นที่นำเสนอใน [127] และจัดเก็บข้อมูล (งานวิจัยนี้กำหนดความยาวของฮีลิกสั้นสุดเท่ากับ 2 คู่เบส)

ตารางที่ 3.1 ตัวอย่างความน่าจะเป็นของคู่เบสที่ได้จากโปรแกรม RNAfold

เบสลำดับที่ *i*	**เบสลำดับที่ *j***	**ความน่าจะเป็น**
1	20	0.92
2	19	0.95
3	17	0.01
3	18	0.99
4	16	0.01
4	17	1.00
5	16	1.00
6	15	1.00
6	19	0.01
7	14	1.00
7	18	0.01
8	13	1.00
8	14	0.01
8	17	0.01
9	13	0.01
9	16	0.01
10	15	0.01

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตัวอย่างการระบุเซตของฮีลิกที่เป็นไปได้และการเข้ารหัสแสดงดังรูปที่ 3.3 โดยกำหนดให้ความน่าจะเป็นของคู่เบสที่ได้จากโปรแกรม RNAfold สำหรับสายลำดับอาร์เอ็นเอยาว 20 นิวคลีโอไทด์เป็นดังตารางที่ 3.1

จากตารางที่ 3.1 คอลัมน์ที่ 1 แสดงตำแหน่งของเบส *i* คอลัมน์ที่ 2 แสดงตำแหน่งของเบส *j* และ คอลัมน์ที่ 3 แสดงความน่าจะเป็นที่เบสตำแหน่งที่ *i* จับคู่กับเบสตำแหน่งที่ *j* เช่น ข้อมูลในแถวแรกของตารางที่ 3.1 คือ ความน่าจะเป็นที่เบสตำแหน่งที่ 1 จะจับคู่กับเบสตำแหน่งที่ 20 มีค่าเท่ากับ 0.92 และเมื่อนำความน่าจะเป็นในตารางที่ 3.1 ไประบุค่าในเมทริกซ์โดยหมายเลขแถวสัมพันธ์กับตำแหน่งเบสที่ *i* และ หมายเลขคอลัมน์สัมพันธ์กับตำแหน่งเบสที่ *j* จะได้ผลลัพธ์ดังรูปที่ 3.3

	1	2	...	12	13	14	15	16	17	18	19	20
1												0.92
2											0.95	
3									0.01	0.99		
4								0.01	1.00			
5								1.00				
6							1.00				0.01	
7						1.00				0.01		
8					1.00	0.01			0.01			
9					0.01			0.01				
10							0.01					
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												

หมายเลขฮีลิก	หมายเลขแถว	หมายเลขคอลัมน์	ความยาว
1	1	20	8
2	3	17	2
3	6	19	5
4	8	14	2

รูปที่ 3.3 ตัวอย่างการสร้าง และการเข้ารหัสฮีลิก

จากรูปที่ 3.3 เมื่อดำเนินการตามขั้นตอนวิธีที่อธิบายไปข้างต้น สำหรับสายลำดับอาร์เอ็นเอนี้จะระบุฮีลิกได้ทั้งหมด 4 ชิ้น ข้อมูลการเข้ารหัสฮีลิกแสดงทางด้านขวามือของเมทริกซ์ เช่น ฮีลิกหมายเลข 2 คือ ฮีลิกบริเวณที่มีการแรเงาในรูปถูกเข้ารหัสเป็น [3; 17; 2] แทนความหมายว่าเบสตำแหน่งที่ 3 จับคู่กับเบสตำแหน่งที่ 17 และเบสตำแหน่งที่ 4 จับคู่กับเบสตำแหน่งที่ 16 และฮีลิกนี้มีความยาว 2 คู่เบส เป็นต้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.1.2 ประเมินประสิทธิภาพของขั้นตอนการจัดเตรียมฮีลิก

ในหัวข้อนี้นำเสนอการทดสอบประสิทธิภาพของขั้นตอนวิธีจัดเตรียมของฮีลิกในแง่ของจำนวนฮีลิกที่สร้างได้และความถูกต้องเมื่อนำไปตรวจสอบกับโครงสร้างที่เป็นคำตอบ โดยทดสอบกับข้อมูลสายลำดับอาร์เอ็นเอจำนวน 20 สาย ได้ผลลัพธ์แสดงดังตารางที่ 3.2 (รายละเอียดของข้อมูลที่นำมาทดสอบนำเสนอในตารางที่ 4.2 ของบทที่ 4)

ตารางที่ 3.2 การประเมินประสิทธิภาพของขั้นตอนวิธีการระบุฮีลิก

ลำดับ	รหัส โมเลกุล	ความยาว	จำนวนคู่เบสเฉลย	จำนวนฮีลิกเฉลย	จำนวนฮีลิกที่สร้างได้	จำนวนฮีลิกที่สร้างได้ตรงกับเฉลย	จำนวนตำแหน่งคู่เบสที่ระบุได้ถูกต้อง
1	CRW_00557	117	38	7	102	9	33
2	CRW_00570	118	37	6	37	6	35
3	CRW_01516	120	40	7	117	8	37
4	CRW_00548	122	38	8	36	8	33
5	CRW_00567	123	40	7	83	8	37
6	CRW_00555	124	40	7	120	9	35
7	CRW_00016	394	120	20	436	23	107
8	CRW_00010	454	126	20	420	22	119
9	CRW_00013	456	115	22	925	25	102
10	CRW_00006	468	113	23	634	21	96
11	CRW_00012	543	141	26	787	27	135
12	CRW_00004	556	131	24	553	21	116
13	CRW_00018	605	121	24	701	25	115
14	CRW_00423	697	189	41	1928	38	125
15	CRW_00429	784	233	49	2283	40	177
16	CRW_00418	940	260	61	1321	46	165
17	CRW_00463	945	254	63	1668	56	211
18	CRW_00438	954	268	61	2007	56	206
19	CRW_00419	964	265	62	1738	59	215
20	CRW_00039	1495	468	91	1348	86	389

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากตารางที่ 3.2 คอลัมน์ที่ 2 แสดงรหัสโมเลกุลอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3 แสดงความยาวของสายลำดับอาร์เอ็นเอ คอลัมน์ที่ 4 แสดงจำนวนคู่เบสที่พบในโครงสร้างอาร์เอเอ็นที่เป็นคำตอบ คอลัมน์ที่ 5 แสดงจำนวนฮีลิกที่พบในโครงสร้างอาร์เอ็นเอที่เป็นคำตอบ คอลัมน์ที่ 6 แสดงจำนวนฮีลิกที่ระบุได้จากขั้นตอนวิธีจัดเตรียมฮีลิก คอลัมน์ที่ 7 แสดงจำนวนฮีลิกที่สร้างได้ถูกต้องตรงกับฮีลิกที่พบในโครงสร้างคำตอบ (พิจารณาจากการที่ฮีลิกนั้นมีตำแหน่งของคู่เบสบางส่วนตรงกับฮีลิกที่พบในโครงสร้างคำตอบ) คอลัมน์ที่ 8 แสดงจำนวนตำแหน่งคู่เบสที่ระบุได้ตรงกับตำแหน่งคู่เบสที่พบจริงในโครงสร้างคำตอบ

จากการประเมินประสิทธิภาพของขั้นตอนวิธีสำหรับระบุฮีลิกที่งานวิจัยนี้เลือกใช้ เมื่อทดสอบกับข้อมูลสายลำดับอาร์เอ็นเอ 20 รายการดังตารางที่ 3.2 พบว่า ข้อมูล 13 รายการแรกสามารถระบุจำนวนฮีลิกได้ใกล้เคียงกับจำนวนฮีลิกที่พบจริงในโครงสร้างคำตอบ นอกจากนี้ ตำแหน่งคู่เบสที่ระบุได้ก็มีความใกล้เคียงกับตำแหน่งคู่เบสที่พบจริงในโครงสร้างคำตอบ บริเวณคู่เบสที่ทำนายผิดพลาดไปบางส่วนเป็นบริเวณของคู่เบสที่ไม่ใช่คาร์โนนิคอล (non-canonical base pair) และบางส่วนเป็นบริเวณคู่เบสเดี่ยว เนื่องจากงานวิจัยนี้กำหนดขนาดความยาวฮีลิกสั้นสุดไว้ที่ 2 คู่เบส ทำให้ไม่สามารถระบุคู่เบสในบริเวณเหล่านี้ได้ โดยข้อมูล 13 รายการนี้เป็นข้อมูลอาร์เอ็นเอในกลุ่มของ 5S Ribosomal RNA และ Group I Intron แต่ข้อมูลอีก 7 รายการที่เหลือซึ่งเป็นข้อมูลอาร์เอ็นเอจากกลุ่มของ 16S Ribosomal RNA ผลการระบุฮีลิกให้ค่าความถูกต้องลดลงทั้งในแง่ของจำนวนฮีลิกและความถูกต้องของตำแหน่งคู่เบสที่ระบุได้ สาเหตุอาจเนื่องมาจากข้อมูลในกลุ่มนี้มีความยาวของสายลำดับที่ค่อนข้างมากจึงทำให้ความแม่นยำในส่วนความน่าจะเป็นของคู่เบสลดลง แต่ในภาพรวมถือว่าความน่าจะเป็นของคู่เบสสามารถนำมาใช้เป็นเกณฑ์ในการระบุฮีลิกได้ดีเพียงพอ

อย่างไรก็ตาม แม้ว่าขั้นตอนวิธีการเตรียมฮีลิกที่งานวิจัยนี้เลือกใช้จะให้ผลลัพธ์ที่ดี แต่จำนวนฮีลิกที่สร้างได้ยังมีจำนวนค่อนข้างมากเมื่อเทียบกับจำนวนฮีลิกที่พบจริงในโครงสร้างคำตอบ นอกจากนี้ ฮีลิกที่ระบุได้บางชิ้นมีความยาวไม่พอดีกับฮีลิกชิ้นที่เป็นคำตอบ กล่าวคือ ตำแหน่งคู่เบสบางส่วนถูกต้องแต่บางส่วนอาจระบุมากเกิน แท้จริงแล้วบริเวณนั้นเป็นเบสอิสระไม่มีการจับคู่กับเบสอื่น งานวิจัยนี้จึงนำเสนอวิธีการปรับปรุงฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกที่ได้นำเสนอไปเพื่อสามารถตัดทอนฮีลิกบางชิ้นให้มีขนาดสั้นลง ทำให้ข้อมูลฮีลิกที่สร้างได้มีความใกล้เคียงกับฮีลิกชิ้นที่เป็นคำตอบมากขึ้น ส่งผลให้กระบวนการทำนายโครงสร้างอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold มีความแม่นยำมากยิ่งขึ้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.1.3 การปรับปรุงเซตของฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิก

ปัญหาที่พบจากขั้นตอนการจัดเตรียมฮีลิกที่ได้นำเสนอไป คือ ฮีลิกที่สร้างได้มีจำนวนคู่เบสมากเกินกว่าจำนวนคู่เบสที่พบจริงในโครงสร้างที่เป็นคำตอบ เช่น โมเลกุลอาร์เอ็นเอรหัส CRW_00557 มีฮีลิกที่พบในโครงสร้างที่เป็นคำตอบและฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกเป็นดังตารางที่ 3.3 โดยข้อมูลที่แสดงในตารางคัดเลือกมาเฉพาะฮีลิกที่สามารถระบุตำแหน่งคู่เบสตรงกับฮีลิกที่พบในโครงสร้างคำตอบอย่างน้อย 1 คู่

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 3.3 การเปรียบเทียบเซตของฮีลิกที่พบในโครงสร้างคำตอบกับเซตของฮีลิกที่สร้างได้

เซตของฮีลิกที่เป็นคำตอบ	**เซตของฮีลิกที่สร้างได้**	**คำอธิบาย**
1) 1 ; 116 ; 8	1) 1 ; 116 ; 12	ฮีลิกชิ้นนี้มีคู่เบสเกินจากคำตอบ 4 คู่
2) 14 ; 66 ; 2	2) 14 ; 66 ; 2	ฮีลิกชิ้นนี้สร้างได้ถูกต้องตรงคำตอบทุกคู่เบส
3) 16 ; 63 ; 6	3) 16 ; 63 ; 6	ฮีลิกชิ้นนี้สร้างได้ถูกต้องตรงคำตอบทุกคู่เบส
4) 26 ; 54 ; 3	4) 27 ; 53 ; 2	ฮีลิกชิ้นนี้ระบุคู่เบสขาดไปจากคำตอบ 1 คู่
5) 29 ; 49 ; 4	5) 29 ; 49 ; 4	ฮีลิกชิ้นนี้สร้างได้ถูกต้องตรงคำตอบทุกคู่เบส
6) 68 ; 105 ; 7	6) 67 ; 106 ; 3	คู่เบสในฮีลิกชิ้นที่ 6 และ 7 เป็นส่วนหนึ่งของฮีลิกคำตอบชิ้นที่ 6 แต่ระบุคู่เบสเกินจากคำตอบ 1 คู่ และขาดไป 3 คู่
	7) 71 ; 102 ; 2	
7) 77 ; 96 ; 8	8) 77 ; 96 ; 2	คู่เบสในฮีลิกชิ้นที่ 8 และ 9 เป็นส่วนหนึ่งของฮีลิกคำตอบชิ้นที่ 7 แต่ฮีลิกที่สร้างได้ระบุคู่เบสขาดไป 1 คู่
	9) 79 ; 92 5	

ตารางที่ 3.3 คอลัมน์ที่ 1 แสดงเซตของฮีลิกที่พบในโครงสร้างที่เป็นคำตอบมีทั้งหมด 7 ชิ้น แต่ละชิ้นถูกเข้ารหัสด้วย 3 พารามิเตอร์ดังที่ได้นำเสนอไปในหัวข้อก่อนหน้า คอลัมน์ที่ 2 แสดงเซตของฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกโดยคัดเลือกเฉพาะฮีลิกชิ้นที่ระบุตำแหน่งของคู่เบสบางส่วนได้ตรงกับตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบอย่างน้อย 1 คู่ พบว่า ฮีลิกชิ้นที่สร้างได้ถูกต้อง 100% ได้แก่ ฮีลิกชิ้นที่ 2, 3 และ 5 ในขณะที่ฮีลิกชิ้นอื่น ๆ ระบุตำแหน่งของคู่เบสได้ถูกต้องบางส่วน โดยรายละเอียดอธิบายในคอลัมน์ที่ 3

จากตัวอย่างที่นำเสนอในตารางที่ 3.3 พบว่า วิธีการเตรียมฮีลิกที่เลือกใช้สามารถสร้างฮีลิกที่ระบุตำแหน่งของคู่เบสได้ถูกต้องตรงกับตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบจำนวน 33 คู่ และตำแหน่งคู่เบสอีก 5 คู่ถูกพบในโครงสร้างที่เป็นคำตอบแต่ไม่ถูกระบุโดยวิธีการจัดเตรียมฮีลิก ซึ่งเมื่อพิจารณาในรายละเอียด พบว่า คู่เบสทั้ง 5 คู่นั้นเป็นบริเวณของคู่เบสที่ไม่ใช่คาร์โนนิคอลจึงทำให้วิธีการจัดเตรียมฮีลิกไม่สามารถระบุตำแหน่งคู่เบสเหล่านั้นได้

ภายใต้หลักการที่ว่าเบสใด ๆ สามารถจับคู่กับเบสอื่น ๆ ในสายลำดับเดียวกันได้แค่ 1 ตำแหน่งเท่านั้น [125] นั่นคือ ถ้าเบสตำแหน่งที่ *i* จับคู่กับเบสตำแหน่งที่ *j* แล้วต้องไม่พบว่าเบสตำแหน่งที่ *i* จับคู่กับเบสตำแหน่งที่ *k* อีก แต่จากตัวอย่างที่นำเสนอไปในตารางที่ 3.3 พบว่า ฮีลิกที่สร้างได้บางชิ้นมีการแชร์ตำแหน่งคู่เบสร่วมกัน คือ ฮีลิกชิ้นที่ 1 กับชิ้นที่ 6 กล่าวคือ เมื่อถอดรหัสฮีลิกชิ้นที่ 1 จะประกอบด้วยตำแหน่งเบสที่เข้าคู่กัน ดังนี้ {(1-116), (2-115), (3-114), (4-113), (5-112), (6-111), (7-110), (8-109), (9-108), (10-107), (11-106), (12-105)} ในขณะที่ฮีลิกชิ้นที่ 6 เมื่อถอดรหัสจะได้ตำแหน่งเบสที่เข้าคู่กัน ดังนี้ {(67-106), (68-105), (69-104)} หากอ้างอิงตามหลักการที่ได้กล่าวไปฮีลิกทั้ง 2 ชิ้นนี้จะไม่สามารถปรากฏร่วมกันในโครงสร้างเดียวกันได้ เนื่องจากทำให้โครงสร้างที่ทำนายได้ไม่ถูกต้องเพราะเบสตำแหน่งที่ 106 จะจับคู่กับเบสทั้งในตำแหน่งที่ 11 และ 67 ดังนั้น ถ้าไม่ดำเนินการใด ๆ กับฮีลิกเหล่านี้และต้องการให้โครงสร้างที่ทำนายได้มีความถูกต้อง ในขั้นตอนการทำนายโครงสร้างจะสามารถเลือกได้แค่เพียงฮีลิกชิ้นใดชิ้นหนึ่งเท่านั้น แต่จากการตรวจสอบกับโครงสร้างคำตอบ พบว่า ฮีลิกทั้งคู่ต่างก็มีคู่เบสที่พบในโครงสร้างที่เป็นคำตอบดังนั้นทั้งคู่ควรถูกเลือกมาสร้างโครงสร้าง

เนื่องจากความไม่แม่นยำของขั้นตอนการเตรียมฮีลิกทำให้เกิดปัญหาดังที่ได้นำเสนอไป งานวิจัยนี้จึงนำเสนอวิธีการปรับปรุงข้อมูลการเข้ารหัสของฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกเพื่อเพิ่มความยืดหยุ่นในขั้นตอนการทำนายโครงสร้างอาร์เอ็นเอให้สามารถปรับลดจำนวนคู่เบสของฮีลิกได้ในระหว่างกระบวนการปรับปรุงคำตอบ จุดมุ่งหมายก็เพื่อแก้ไขความผิดพลาดที่เกิดขึ้นในขั้นตอนการเตรียมฮีลิกที่มีการระบุข้อมูลฮีลิกบางส่วนคลาดเคลื่อนไปจากฮีลิกที่พบในโครงสร้างคำตอบ และมุ่งหวังว่าผลจากการแก้ไขข้อมูลฮีลิกด้วยวิธีการที่นำเสนอจะช่วยให้ผลลัพธ์การทำนายโครงสร้างมีความถูกต้องมากยิ่งขึ้น เช่น จากตัวอย่างถ้าสามารถแก้ไขการเข้ารหัสของฮีลิกชิ้นที่ 1 เป็น [1 ; 116; 10] และ การเข้ารหัสของฮีลิกชิ้นที่ 6 คงไว้เหมือนเดิมก็จะไม่เกิดปัญหาที่คู่เบสจากทั้งสองฮีลิกมีการแชร์ตำแหน่งเบสร่วมกัน สามารถเลือกฮีลิกทั้งคู่มาประกอบร่วมกันในโครงสร้าง นอกจากนั้นผลจากการแก้ไขนี้ยังช่วยลดจำนวนคู่เบสที่ทำนายผิดได้ 2 คู่เบส (FP ลดลง)

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

วิธีการปรับปรุงข้อมูลการเข้ารหัสฮีลิกที่งานวิจัยนี้นำเสนอ ประกอบด้วย 2 ขั้นตอนย่อย ได้แก่ 1) พิจารณากรณีที่ 2 ฮีลิกใด ๆ มีการแชร์ตำแหน่งเบสร่วมกันว่าอยู่ในเงื่อนไขที่จะทำการแก้ไขข้อมูลการเข้ารหัสหรือไม่ 2) ดำเนินการแก้ไขข้อมูลการเข้ารหัสของฮีลิกคู่นั้นหากพบว่าเป็นไปตามเงื่อนไขที่กำหนด รายละเอียดเป็นดังนี้

3.1.3.1 การพิจารณาว่าคู่ของฮีลิกที่มีตำแหน่งเบสบางส่วนตรงกันผ่านเงื่อนไขที่จะนำไปสู่การแก้ไขข้อมูลการเข้ารหัสหรือไม่

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

เฉพาะในบางกรณีเท่านั้นเมื่อพบว่าคู่ของฮีลิกใด ๆ มีการแชร์ตำแหน่งของคู่เบสบางส่วนร่วมกันและเข้าสู่กระบวนการแก้ไขข้อมูลที่ทำการเข้ารหัสไว้ในขั้นตอนต่อไป หลักการคือ ตำแหน่งของคู่เบสในฮีลิกหนึ่งจะต้องไม่เป็นซับเซตแท้ของตำแหน่งคู่เบสของอีกฮีลิกหนึ่ง ไม่เช่นนั้นผลการแก้ไขข้อมูลการเข้ารหัสก็จะเทียบเท่ากับการเลือกแค่ฮีลิกใดฮีลิกหนึ่งเท่านั้น

กำหนดให้ฮีลิกหมายเลข A ($helix_A$) ถูกเข้ารหัสด้วย [i ; j ; l] และ ฮีลิกหมายเลข B ($helix_B$) ถูกเข้ารหัสด้วย [m ; n ; k] เมื่อ i และ m แทนตำแหน่งเบสเริ่มต้นของ $helix_A$ และ $helix_B$ ตามลำดับ j และ n แทนตำแหน่งเบสที่เข้าคู่กับตำแหน่งเบสเริ่มต้นของ $helix_A$ และ $helix_B$ ตามลำดับ และ l และ k แทนความยาวของ $helix_A$ และ $helix_B$ ตามลำดับ กำหนดให้ตำแหน่งเบสที่ i น้อยกว่าหรือเท่ากับตำแหน่งเบสที่ m เสมอ และกำหนดให้คู่เบสของ $helix_A$ แทนด้วย () และคู่เบสของ $helix_B$ แทนด้วย []

<u>**กรณีที่ 1**</u>

ตำแหน่งคู่เบสของ $helix_B$ อยู่ภายในขอบเขตตำแหน่งเบสของ $helix_A$ ($m>= i$ และ $n<= j$) ตัวอย่างกรณีนี้แสดงดังตารางที่ 3.4 ซึ่งเงื่อนไขในการพิจารณาว่าจะมีการแก้ไขข้อมูลการเข้ารหัสของฮีลิกคู่นี้หรือไม่เป็นดังนี้ : $m>i$ และ $(m+k-1) > (i+l-1)$ และ $(n-k+1) < (j-l+1)$ และ $n < j$

หากไม่เป็นไปตามเงื่อนไขนี้จะไม่ทำการแก้ไขข้อมูลใด ถือว่าฮีลิกทั้งคู่ขัดแย้งกันไม่สามารถเกิดร่วมกันในโครงสร้างเดียวกันได้

ตารางที่ 3.4 ตัวอย่างการแชร์ตำแหน่งเบสร่วมกันของสองฮีลิกที่ตรงกับกรณีที่ 1

i *j*

$helix_A$	.	(	(	(	.	.	.	.	.	)	)	)	.	.	.
$helix_B$	.	.	[	[	[	.	.	]	]	]	.	.	.	.	.

m *n*

กรณีที่ 2

ตำแหน่งคู่เบสของ $helix_B$ ในส่วนของวงเล็บเปิดมีการแชร์ตำแหน่งคู่เบสกับ $helix_A$ ในส่วนของวงเล็บปิด ตัวอย่างกรณีนี้แสดงดังตารางที่ 3.5 ซึ่งเงื่อนไขในการพิจารณาว่าจะมีการแก้ไขข้อมูลการเข้ารหัสของฮีลิกคู่นี้หรือไม่เป็นดังนี้ : $m > (j-l+1)$ และ $(m+k-1) > j$

หากไม่เป็นไปตามเงื่อนไขนี้ไม่ต้องทำการแก้ไขข้อมูลใด ถือว่าฮีลิกทั้งคู่ขัดแย้งกันไม่สามารถเกิดร่วมกันในโครงสร้างเดียวกันได้

ตารางที่ 3.5 ตัวอย่างการแชร์ตำแหน่งเบสร่วมกันของสองฮีลิกที่ตรงกับกรณีที่ 2

	i								*j*						
$helix_A$	(	(	(	.	.	.	)	)	)	.	.	.	.	.	.
$helix_B$	.	.	.	.	.	.	.	[	[	[	.	.	]	]	]
								m							*n*

3.1.3.2 การปรับปรุงข้อมูลการเข้ารหัสของคู่ฮีลิกที่มีตำแหน่งเบสบางส่วนตรงกัน

เมื่อ 2 ฮีลิกใด ๆ เป็นไปตามเงื่อนไขในหัวข้อ 3.1.3.1 ขั้นตอนต่อไปคือการแก้ไขข้อมูลของฮีลิกที่ถูกเข้ารหัสไว้ชั่วคราว สาเหตุที่ใช้คำว่าชั่วคราวเนื่องจากการปรับแก้จะเกิดขึ้นเฉพาะเมื่อมีการเลือกฮีลิกคู่นี้มาประกอบร่วมกันในโครงสร้างที่ทำนายได้ในรุ่นของการวิวัฒนาการนั้น ๆ แต่ข้อมูลต้นฉบับที่ถูกเข้ารหัสไว้ตั้งแต่ขั้นตอนการจัดเตรียมฮีลิกจะคงไว้เหมือนเดิม

วิธีการคือพิจารณาบริเวณที่ทั้งสองฮีลิกมีการแชร์ตำแหน่งเบสร่วมกันและแข่งขันกันโดยใช้ความน่าจะเป็นของคู่เบส กล่าวคือ คู่เบสที่มีความน่าจะเป็นสูงกว่าจะถูกคงข้อมูลไว้เหมือนเดิม ส่วนคู่เบสที่มีความน่าจะเป็นต่ำกว่าจะถูกแก้ไขข้อมูลการเข้ารหัสในบริเวณนั้นให้เป็นเบสอิสระ เช่น จากตัวอย่าง ฮีลิกชิ้นที่ 1 และ ฮีลิกชิ้นที่ 6 มีการเข้ารหัสเป็น [1 ; 116; 12] และ [67 ; 106 ; 3] ตามลำดับ จะตรงกับกรณีที่ 1 ที่ตำแหน่งคู่เบสของ $helix_6$ อยู่ภายในขอบเขตของ $helix_1$ และมีการแชร์ตำแหน่งเบสบางส่วนร่วมกันแสดงดังตารางที่ 3.6 โดย 3 คอลัมน์แรกจะเป็นข้อมูลตำแหน่งเบสและความน่าจะเป็นของคู่เบสจาก $helix_1$ และ 3 คอลัมน์ท้ายจะเป็นข้อมูลตำแหน่งเบสและความน่าจะเป็นของคู่เบสจาก $helix_6$ และบริเวณที่มีการแรเงาในตารางคือบริเวณที่เบสจากทั้ง 2 ฮีลิกมีตำแหน่งตรงกัน และตัวเลขตำแหน่งเบสที่มีเครื่องหมายดอกจันเป็นคู่เบสที่ถูกแก้ไขให้เป็นเบสอิสระเนื่องจากมีความน่าจะเป็นต่ำกว่าอีกคู่เบสที่นำมาเปรียบเทียบ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากตารางที่ 3.6 ทั้ง 2 ฮีลิกมีคู่เบสที่แชร์ตำแหน่งร่วมกัน 2 คู่ ดังนี้

- จากการเข้ารหัสใน $helix_1$ เบสตำแหน่งที่ 11 จับคู่กับเบสตำแหน่งที่ 106 ในขณะที่การเข้ารหัสใน $helix_6$ เบสตำแหน่งที่ 67 จับคู่กับเบสตำแหน่งที่ 106 และผลการเปรียบเทียบพบว่าคู่เบสจาก $helix_1$ มีความน่าจะเป็นสูงกว่า ดังนั้น คู่เบสตำแหน่ง 67 และ 106 ใน $helix_6$ จะถูกแก้ไขเป็นเบสอิสระ

- จากการเข้ารหัสใน $helix_1$ เบสตำแหน่งที่ 12 จับคู่กับเบสตำแหน่งที่ 105 ในขณะที่การเข้ารหัสใน $helix_6$ เบสตำแหน่งที่ 68 จับคู่กับเบสตำแหน่งที่ 105 และผลการเปรียบเทียบพบว่าคู่เบสจาก $helix_6$ มีความน่าจะเป็นสูงกว่า ดังนั้น คู่เบสตำแหน่ง 12 และ 105 ใน $helix_1$ จะถูกแก้ไขเป็นเบสอิสระ

ดังนั้น ข้อมูลการเข้ารหัสของ $helix_1$ จะถูกแก้ไขจากเดิม [1 ; 116 ; 12] เป็น [1 ; 116 ; 11] และข้อมูลการเข้ารหัสของ $helix_6$ จะถูกแก้ไขจากเดิม [67 ; 106 ; 3] เป็น [68 ; 105 ; 2]

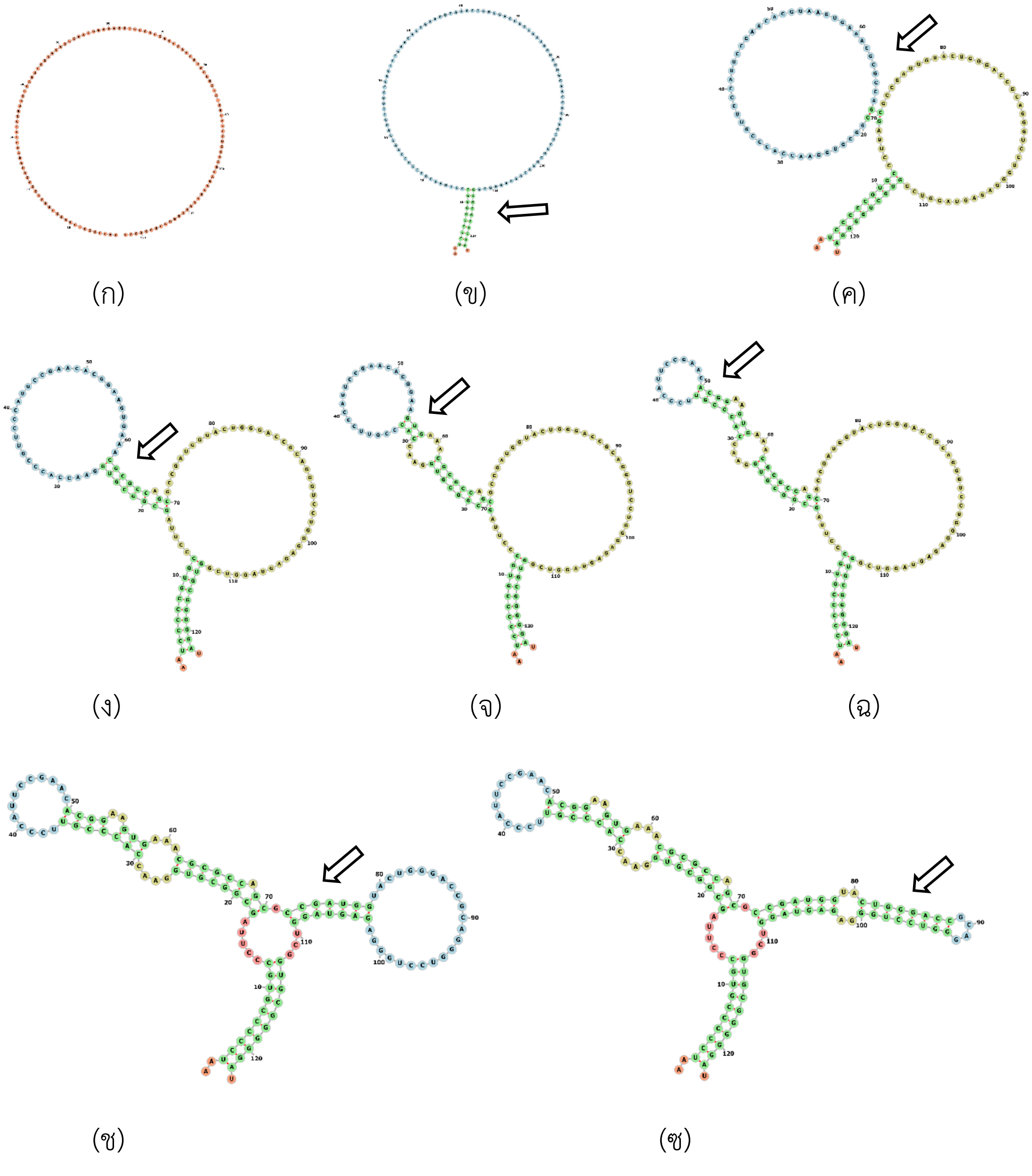
ตารางที่ 3.6 การพิจารณาตำแหน่งของคู่เบสที่มีการทับซ้อนกันของ $helix_1$ และ $helix_6$

$helix_1$			**$helix_6$**		
เบส ตำแหน่งที่ *i*	**เบส ตำแหน่งที่ *j***	**ความน่าจะเป็นของคู่เบส**	**เบส ตำแหน่งที่ *i***	**เบส ตำแหน่งที่ *j***	**ความน่าจะเป็นของคู่เบส**
1	116	0.99			
2	115	1.00			
3	114	1.00			
4	113	1.00			
5	112	1.00			
6	111	1.00			
7	110	0.99			
8	109	1.00			
9	108	1.00			
10	107	1.00			
11	106	0.95	67*	106*	0.83
12*	105*	0.79	68	105	0.96
			69	104	0.96

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2 การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold

การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่งานวิจัยนี้นำเสนออยู่ภายใต้แนวความคิดที่ว่า “1 โครงสร้างประกอบไปด้วยบริเวณที่เป็นฮีลิกกับบริเวณที่เป็นลูป” ดังนั้น เมื่อมีการระบุบริเวณของฮีลิกที่เป็นไปได้ทั้งหมดเตรียมไว้แล้วจากขั้นตอนก่อนหน้า (หัวข้อ 3.1) งานส่วนที่เหลือคือการเลือกว่าในโครงสร้างควรประกอบไปด้วยฮีลิกหมายเลขใดบ้าง แนวคิดเช่นนี้คล้ายกับการดำเนินการในงานวิจัย [128] ตัวอย่างโครงสร้างอาร์เอ็นเอที่สร้างโดยอาศัยแนวความคิดนี้เป็นดังรูปที่ 3.4



รูปที่ 3.4 ตัวอย่างการสร้างโครงสร้างอาร์เอ็นเอด้วยการเลือกฮีลิกครั้งละชิ้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

โดยรูปที่ 3.4 (ก) แสดงสายลำดับอาร์เอ็นเอตั้งต้นทุกเบสยังไม่มีการจับคู่กับเบสอื่น ภาพ 3.4 (ข – ช) แสดงโครงสร้างที่เกิดจากการเลือกฮีลิกเพิ่มขึ้นครั้งละชิ้นโดยบริเวณของฮีลิกที่ถูกเพิ่มในโครงสร้างของแต่ละภาพย่อยแสดงด้วยบริเวณที่ลูกศรชี้ และภาพ 3.4 (ซ) แสดงโครงสร้างที่เสร็จสมบูรณ์แล้ว เมื่อสร้างโครงสร้างเสร็จขั้นตอนต่อไปคือการประเมินค่าความเหมาะสมของโครงสร้างที่สร้างได้ด้วยวิธีการทางอุณหพลศาสตร์ (thermodynamics) ภายใต้สมมุติฐานที่ว่าค่าพลังงานของโครงสร้างใดยิ่งต่ำยิ่งมีโอกาสสูงที่เป็นโครงสร้างที่ตรงกับโครงสร้างที่เป็นคำตอบ

ภายใต้กรอบการทำงานของขั้นตอนวิธีประมาณการแจกแจงมาตรฐาน คำตอบของปัญหาจะถูกเข้ารหัสเป็นโครโมโซมซึ่งในที่นี้คือโครงสร้างทุติยภูมิของอาร์เอ็นเอที่อยู่ในรูปแบบเซตของหมายเลขฮีลิกที่ถูกเลือกมาประกอบกันเป็น 1 โครงสร้าง และใช้เวกเตอร์ความน่าจะเป็นสำหรับเก็บข้อมูลเชิงสถิติว่าฮีลิกแต่ละชิ้นที่สกัดได้จากขั้นตอนการจัดเตรียมฮีลิกมีโอกาสเป็นฮีลิกชิ้นที่ปรากฎอยู่ในโครงสร้างที่เป็นคำตอบมากน้อยแค่ไหน ระหว่างกระบวนการวิวัฒนาการของขั้นตอนวิธีประมาณการแจกแจงค่าในเวกเตอร์ความน่าจะเป็นนี้จะถูกปรับปรุงอ้างอิงจากการที่ฮีลิกนั้น ๆ ประสบความสำเร็จหรือล้มเหลวเมื่อถูกนำไปประกอบร่วมกันเป็นโครงสร้างทุติยภูมิของอาร์เอ็นเอ

ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอชื่อว่า Hybrid-EDAfold ได้รับแรงบันดาลใจจากขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับ และ ขั้นตอนวิธีคอยน์มาตรฐาน ซึ่งเป็นขั้นตอนวิธีเชิงวิวัฒนาการที่ประสบความสำเร็จในการแก้ปัญหาการหาค่าเหมาะที่สุดเชิงการจัด ความน่าสนใจของวิธีการที่นำเสนอคือมีการเรียนรู้จากทั้งคำตอบดีและคำตอบด้อย ซึ่งแตกต่างจากขั้นตอนวิธีประมาณการแจกแจงมาตรฐานทั่วไปที่จะใช้เฉพาะคำตอบที่ดีเท่านั้น นอกจากนี้ วิธีการที่นำเสนอประกอบด้วย 2 ขั้นตอนวิธีประมาณการแจกแจงที่มีพฤติกรรมการค้นหาที่แตกต่างกัน ขั้นตอนวิธีประมาณการแจกแจงตัวที่หนึ่งเป็นตัวแทนของการค้นหาแบบโกลบอล (global search) ในขณะที่ขั้นตอนวิธีประมาณการแจกแจงตัวที่สองเป็นตัวแทนการค้นหาแบบโลคอล (local search) และตราบใดที่ยังไม่ครบตามจำนวนรุ่นที่กำหนดทั้งสองขั้นตอนวิธีประมาณการแจกแจงนี้จะสลับการทำงานกันเมื่อพบว่าขั้นตอนวิธีประมาณการแจกแจงที่กำลังทำงานอยู่ไม่มีความก้าวหน้า (ประเมินจากไม่สามารถหาคำตอบที่มีค่าความเหมาะสมดีขึ้นได้ติดต่อกัน m รุ่น)

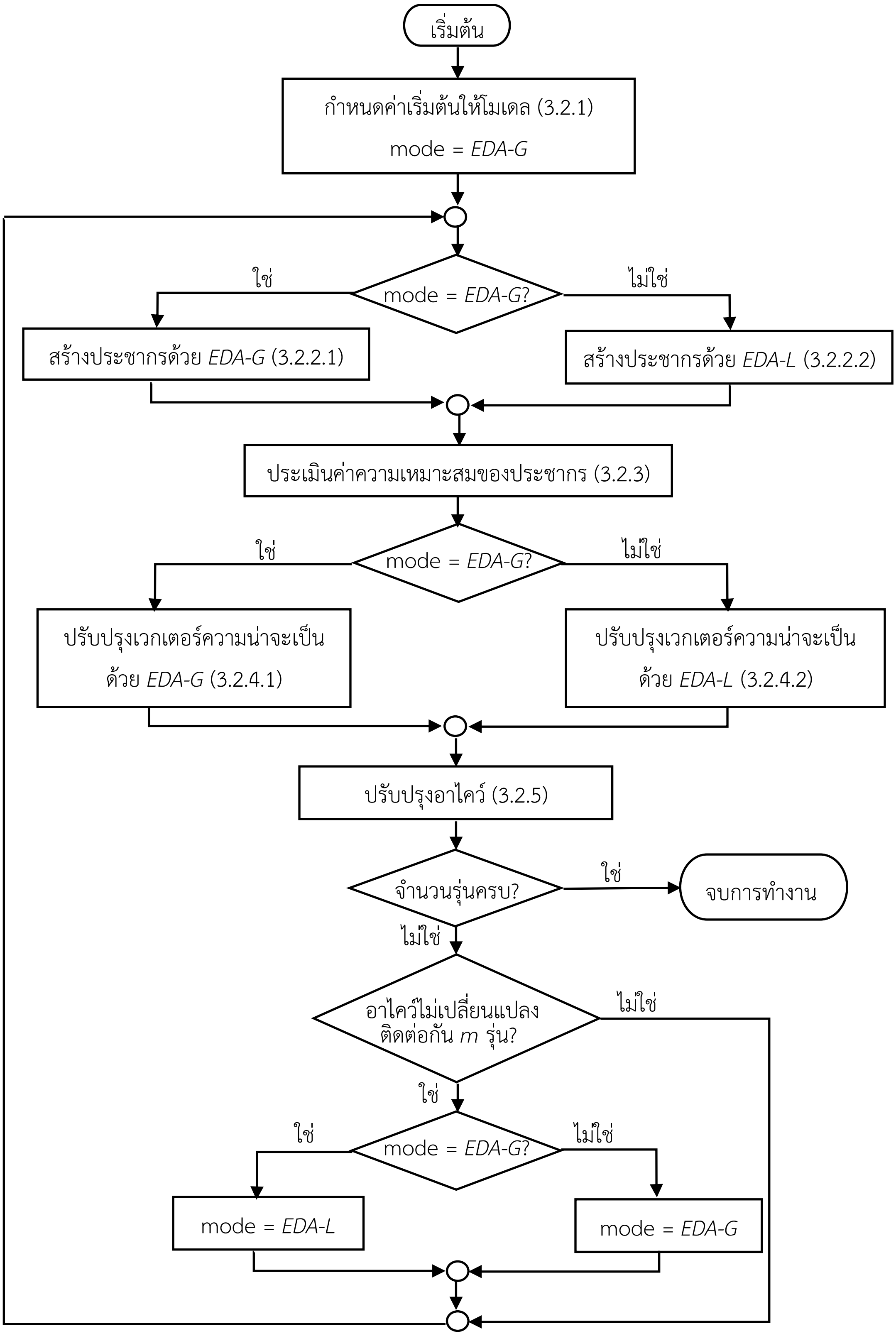
ขั้นตอนวิธี Hybrid-EDAfold ประกอบด้วย 2 ขั้นตอนวิธีประมาณการแจกแจงย่อยและมีโครงสร้างข้อมูลที่สำคัญ 2 ส่วน คือ เวกเตอร์ความน่าจะเป็นใช้สำหรับควบคุมโอกาสที่ฮีลิกแต่ละชิ้นจะถูกเลือกมาสร้างโครงสร้าง และเมทริกซ์ความเข้ากันได้ของฮีลิกใช้สำหรับพิจารณาว่าฮีลิกใดสามารถเกิดร่วมกันในโครงสร้างได้ ถ้าฮีลิกหนึ่งถูกเลือกฮีลิกชิ้นอื่น ๆ ที่มีตำแหน่งเบสขัดแย้งกับฮีลิกนั้นจะถูกตัดทิ้งไม่สามารถเลือกมาประกอบร่วมกันในโครงสร้างได้เพื่อทำให้โครงสร้างที่ทำนายได้มีความถูกต้องอยู่เสมอ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

การทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอด้วยขั้นตอนวิธีที่งานวิจัยนี้นำเสนอเริ่มต้นด้วยการกำหนดค่าเริ่มต้นให้กับโครงสร้างข้อมูลทั้ง 2 ส่วน (อธิบายในหัวข้อ 3.2.1) จากนั้นเข้าสู่กระบวนการมาตรฐานของขั้นตอนวิธีประมาณการแจกแจง ได้แก่ การสร้างประชากร (อธิบายในหัวข้อ 3.2.2) การประเมินค่าความเหมาะสมของประชากร (อธิบายในหัวข้อ 3.2.3) การปรับปรุงเวกเตอร์ความน่าจะเป็น (อธิบายในหัวข้อ 3.2.4) และการปรับปรุงอาไคว์ (archive) ซึ่งถูกใช้สำหรับเก็บตัวแทนคำตอบของขั้นตอนวิธีที่นำเสนอ (อธิบายในหัวข้อ 3.2.5)

แต่ละขั้นตอนวิธีประมาณการแจกแจงทำงานภายใต้กรอบการทำงานมาตรฐาน แต่มีขั้นตอนการทำงานที่แตกต่างกันในส่วนของการสร้างประชากรและการปรับปรุงเวกเตอร์ความน่าจะเป็น ในตอนต้นขั้นตอนวิธีประมาณการแจกแจงตัวที่หนึ่งแทนด้วย *EDA-G* ทำงานก่อน เนื่องจากความรู้ที่ถูกเก็บอยู่ในเวกเตอร์ความน่าจะเป็นยังอยู่ในระหว่างกระบวนการเรียนรู้ ดังนั้น ในขั้นตอนของการสร้างประชากรสำหรับ *EDA-G* ใช้การสุ่มเลือกหมายเลขฮีลิกจากเซตของฮีลิกที่เตรียมไว้มาประกอบร่วมกันเป็น 1 โครงสร้าง จากนั้นประเมินค่าความเหมาะสมของประชากรด้วยการคำนวณค่าพลังงานของโครงสร้างที่ทำนายได้ ประชากรในแต่ละรุ่นจะถูกจำแนกเป็นคำตอบคุณภาพดีและคำตอบคุณภาพด้อยแล้วนำข้อมูลส่วนนี้กลับไปปรับปรุงค่าในเวกเตอร์ความน่าจะเป็น จากนั้นประเมินคำตอบใน แต่ละรุ่นเทียบกับคำตอบที่สร้างได้ในอดีตเพื่อทำการปรับปรุงข้อมูลในอาไคว์ (อาไคว์เก็บคำตอบดีสุด n ตัวแรกที่ถูกพบระหว่างกระบวนการวิวัฒนาการ เมื่อ n เป็นพารามิเตอร์ที่กำหนดโดยผู้ใช้) ตราบใดที่ยังไม่ครบตามจำนวนรุ่นของการวิวัฒนาการ (generation) ที่กำหนดก็จะทำงานซ้ำอยู่ที่ *EDA-G* ไปจนกระทั่งไม่พบความก้าวหน้าของคำตอบที่สร้างได้ซึ่งประเมินจากข้อมูลในอาไคว์ไม่เปลี่ยนแปลงติดต่อกัน m รุ่น (m เป็นพารามิเตอร์ที่กำหนดโดยผู้ใช้) ก็จะสลับไปทำงานด้วยขั้นตอนวิธีประมาณการแจกแจงตัวที่สองแทนด้วย *EDA-L* ซึ่งจะเปลี่ยนวิธีการสร้างประชากรคำตอบไปเป็นการกลายพันธุ์โครโมโซมบรรพบุรุษเพื่อสร้างโครโมโซมลูก กล่าวคือ สุ่มเลือกโครงสร้างที่ถูกจัดเก็บในอาไคว์มาเป็นบรรพบุรุษจากนั้นทำการกลายพันธุ์เพื่อผลิตลูก (บรรพบุรุษ 1 ตัว กลายพันธุ์ได้ลูก 1 ตัว) วิธีการกลายพันธุ์ทำโดยการสุ่มลบฮีลิกบางชิ้นในโครงสร้างพ่อแม่ทิ้ง และสุ่มเลือกฮีลิกอื่น ๆ ที่เข้ากันได้กับฮีลิกที่เหลืออยู่มาประกอบในโครงสร้างเพิ่มเติม (สัดส่วนจำนวนฮีลิกที่ถูกสุ่มทิ้งเป็นพารามิเตอร์ที่กำหนดโดยผู้ใช้) จากนั้นประเมินค่าความเหมาะสมด้วยฟังก์ชันวัตถุประสงค์เดียวกันกับ *EDA-G* จากนั้นแข่งขันกันระหว่างบรรพบุรุษกับลูกที่ผลิตได้เพื่อนำข้อมูลส่วนนี้กลับไปปรับปรุงค่าในเวกเตอร์ความน่าจะเป็น และตรวจสอบประชากรที่สร้างได้ หากพบว่ามีคำตอบที่ดีกว่าที่เคยเก็บในอาไคว์จะทำการปรับปรุงข้อมูลในอาไคว์ และวนทำซ้ำอยู่ที่ *EDA-L* ไปจนกว่าข้อมูลในอาไคว์ไม่มีการเปลี่ยนแปลงติดต่อกัน m รุ่นก็จะสลับกลับไปทำงานด้วย *EDA-G* และทำเช่นนี้ไปจนกระทั่งครบตามจำนวนรุ่นที่กำหนด ภาพรวมของขั้นตอนวิธี Hybrid-EDAfold แสดงดังรูปที่ 3.5

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



รูปที่ 3.5 ภาพรวมของขั้นตอนวิธี Hybrid-EDAfold

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 3.5 ขั้นตอนวิธี Hybrid-EDAfold เริ่มต้นการทำงานด้วยการกำหนดค่าเริ่มต้นให้กับโครงสร้างข้อมูลทั้งสองส่วน คือ เวกเตอร์ความน่าจะเป็นและเมทริกซ์ความเข้ากันได้ของฮีลิก จากนั้นมีตัวแปรพิเศษชื่อว่า *mode* เก็บข้อมูลว่าขณะนี้ขั้นตอนวิธีประมาณการแจกแจงใดกำลังทำงาน เริ่มต้น *mode* มีค่าเป็น *EDA-G* จากรูปเป็นการทำงานทางด้านซ้ายของผังงาน โดยกระบวนการทำงานของ *EDA-G* ประกอบด้วยการสร้างประชากรจากเซตของฮีลิกที่จัดเตรียมไว้ จากนั้นเข้าสู่ขั้นตอนการประเมินค่าความเหมาะสมและใช้ค่าความเหมาะสมนี้ในการจำแนกกลุ่มประชากรออกเป็นกลุ่มโครโมโซมดีกับกลุ่มโครโมโซมด้อย เพื่อสกัดฮีลิกชิ้นที่คาดว่าเป็นฮีลิกดีและฮีลิกชิ้นที่คาดว่าเป็นฮีลิกด้อยและนำข้อมูลนี้กลับไปปรับปรุงเวกเตอร์ความน่าจะเป็น โดยสมาชิกของเวกเตอร์ที่สอดคล้องกับฮีลิกดีจะได้ความน่าจะเป็นเพิ่มขึ้น และในทางกลับกันสมาชิกของเวกเตอร์ที่สอดคล้องกับฮีลิกด้อยจะถูกลดความน่าจะเป็นลง จากนั้นปรับปรุงข้อมูลในอาไคว์ถ้าพบว่าโครโมโซมที่สร้างได้ในรุ่นนี้ดีกว่าโครโมโซมที่เคยถูกจัดเก็บในอาไคว์ และทำการพิจารณาว่าตรงกับเงื่อนไขใดต่อไปนี้ 1) ถ้าครบตามจำนวนรุ่นที่กำหนดก็จบการทำงานของขั้นตอนวิธี Hybrid-EDAfold 2) ถ้ายังไม่ครบตามจำนวนรุ่นที่กำหนดและข้อมูลในอาไคว์ยังมีการเปลี่ยนแปลงก็วนกลับไปทำงานตามกระบวนการทั้งหมดใหม่โดยยังอยู่ที่ *EDA-G* ตัวเดิม 3) ถ้ายังไม่ครบตามจำนวนรุ่นที่กำหนดและข้อมูลในอาไคว์ไม่มีการเปลี่ยนแปลงติดต่อกัน *m* รอบแล้วก็เปลี่ยนแปลงค่าในตัวแปร *mode* เพื่อสลับให้ *EDA-L* ทำงาน

กระบวนการทำงานของ *EDA-L* ซึ่งจากรูปที่ 3.5 จะอยู่ทางด้านขวาของผังงานก็มีลำดับขั้นตอนการทำงานเหมือน *EDA-G* คือ สร้างประชากรโดยสุ่มเลือกโครโมโซมที่ถูกจัดเก็บในอาไคว์มาเป็นต้นแบบในการกลายพันธุ์เพื่อสร้างลูก เมื่อได้โครโมโซมครบตามขนาดประชากรที่กำหนดก็ทำการประเมินค่าความเหมาะสมของประชากรที่สร้างได้ เปรียบเทียบค่าความเหมาะสมระหว่างโครโมโซมบรรพบุรุษและโครโมโซมลูก และคัดเลือกเฉพาะคู่ที่ดำเนินการกลายพันธุ์แล้วได้ลูกที่มีค่าความเหมาะสมดีขึ้น จากนั้นฮีลิกจากกลุ่มที่ถูกสุ่มเพิ่มเติมในการสร้างโครโมโซมลูกจะถูกพิจารณาเป็นฮีลิกดีและฮีลิกจากกลุ่มที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษจะถูกพิจารณาว่าเป็นฮีลิกด้อยแล้วนำข้อมูลสองส่วนนี้ไปปรับปรุงเวกเตอร์ความน่าจะเป็นในทำนองเดียวกับ *EDA-G* จากนั้นตรวจสอบประชากรที่สร้างได้ในรุ่นนี้เปรียบเทียบกับข้อมูลในอาไคว์ถ้าพบโครโมโซมที่มีค่าความเหมาะสมดีกว่าให้ทำการแทนที่ข้อมูลในอาไคว์และพิจารณาเงื่อนไขการจบการทำงานหรือสลับการทำงานดังเช่นที่ได้นำเสนอไป รายละเอียดของแต่ละขั้นตอนย่อยเป็นดังนี้

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2.1 กำหนดค่าเริ่มต้นให้กับโมเดล

เพื่อให้โครงสร้างอาร์เอ็นเอที่ทำนายได้มีความถูกต้อง กล่าวคือ ฮีลิกใด ๆ ที่ประกอบร่วมกันในโครงสร้างไม่มีตำแหน่งเบสขัดแย้งกันและไม่มีการแชร์ตำแหน่งเบสร่วมกัน ดังนั้น การทำนายโครงสร้างที่งานวิจัยนี้นำเสนอเกี่ยวข้องกับ 2 โครงสร้างข้อมูล ได้แก่ เวกเตอร์ความน่าจะเป็น และเมทริกซ์ความเข้ากันได้ของฮีลิก

3.2.1.1 เวกเตอร์ความน่าจะเป็น

เวกเตอร์ความน่าจะเป็นมีขนาดเท่ากับจำนวนฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิก โดยแต่ละสมาชิกของเวกเตอร์แทนความน่าจะเป็นที่ฮีลิกหมายเลขนั้นจะถูกเลือกมาสร้างโครงสร้าง ดังนั้น แต่ละสมาชิกของเวกเตอร์นี้มีค่าอยู่ในช่วง [0, 1] ค่ายิ่งมากยิ่งมีโอกาสสูงที่จะถูกเลือกไปใช้เป็นส่วนหนึ่งของโครงสร้างที่ทำนายได้

โดยทั่วไปหากเป็นขั้นตอนวิธีประมาณการแจกแจงมาตรฐาน ค่าเริ่มต้นของแต่ละสมาชิกในเวกเตอร์จะถูกกำหนดเท่ากับ 0.5 แต่ในงานวิจัยนี้ค่าเริ่มต้นของแต่ละสมาชิกจะไม่เท่ากัน โดยจะใช้ข้อมูลความน่าจะเป็นของคู่เบสซึ่งถูกใช้ในขั้นตอนของการจัดเตรียมฮีลิกสำหรับกำหนดค่าเริ่มต้นให้กับสมาชิกในเวกเตอร์ ดังนั้น ค่าเริ่มต้นของแต่ละสมาชิกมีค่าเท่ากับความน่าจะเป็นเฉลี่ยของคู่เบสที่ปรากฏในฮีลิกนั้น

ตัวอย่างการกำหนดค่าเริ่มต้นของเวกเตอร์ความน่าจะเป็นแสดงดังตารางที่ 3.7 กำหนดให้สายลำดับอาร์เอ็นเอนี้มีจำนวนฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกจำนวน N ชิ้น และฮีลิกแต่ละชิ้นมีหมายเลขกำกับตั้งแต่ 1 ถึง N

ตารางที่ 3.7 ตัวอย่างการกำหนดค่าเริ่มต้นให้เวกเตอร์ความน่าจะเป็น

หมายเลขฮีลิก	1	2	3	4	5	6	7	8	...	N
ความน่าจะเป็น	0.97	0.09	0.09	0.35	0.01	0.19	0.03	0.02	...	0.04

3.2.1.2 เมทริกซ์ความเข้ากันได้ของฮีลิก

เมทริกซ์นี้ถูกใช้สำหรับพิจารณาว่าฮีลิกต่าง ๆ สามารถเกิดร่วมกันในโครงสร้างอาร์เอ็นเอได้หรือไม่ โดยเมทริกซ์นี้มีขนาด $N \times N$ เมื่อ N คือจำนวนฮีลิกที่สร้างได้จากขั้นตอนการเตรียมฮีลิกซึ่ง matrix[i][j] จะมีค่าเป็น 1 ถ้าฮีลิกหมายเลข i สามารถปรากฏร่วมกับฮีลิกหมายเลข j โดยไม่มีคู่เบสใดที่มีตำแหน่งเบสตรงกันหรือขัดแย้งกัน และ matrix[i][j] จะมีค่าเป็น 2 ถ้าฮีลิกหมายเลข i ปรากฏร่วมกับฮีลิกหมายเลข j ได้แต่มีการแชร์ตำแหน่งเบสบางส่วนร่วมกันอ้างอิงตามการพิจารณาที่นำเสนอไปในหัวข้อ 3.1.3.1 ไม่เช่นนั้น matrix[i][j] จะมีค่าเป็น 0 ตัวอย่างการกำหนดค่าในเมทริกซ์ความเข้ากันได้ของฮีลิกแสดงดังตารางที่ 3.8

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 3.8 ตัวอย่างการกำหนดค่าในเมทริกซ์ตรวจสอบความเข้ากันได้ของฮีลิก

i \ *j*	1	2	3	4	5	6	7	8	9	10	…	*N*
1	0	0	0	1	1	1	1	1	1	1	1	1
2		0	0	1	1	1	1	1	1	1	1	1
3			0	0	0	0	0	1	1	1	1	1
4				0	0	0	0	0	0	0	0	1
5					0	1	1	1	1	1	1	1
6						0	0	2	1	1	1	1
7							0	1	1	1	1	1
8								0	1	1	1	1
9									0	1	1	1
10										0	0	1
…											0	1
N												0

การตรวจสอบการขัดแย้งกันของตำแหน่งเบสในฮีลิกตรวจสอบเหมือนการเช็คคู่ของวงเล็บ เช่น กำหนดให้ตำแหน่งคู่เบสของ $helix_A$ แทนด้วย () และตำแหน่งคู่เบสของ $helix_B$ แทนด้วย [] คู่ของฮีลิกใด ๆ จะถูกพิจารณาว่าขัดแย้งกันและมีค่าในเมทริกซ์ความเข้ากันได้ของฮีลิกเป็น 0 ก็ต่อเมื่อถอดรหัสตำแหน่งคู่เบสที่พบในฮีลิกคู่นั้นแล้วพบว่ามี (จับคู่กับ]

จากข้อมูลตัวอย่างในตารางที่ 3.8 มีกลุ่มของฮีลิกที่ตำแหน่งเบสบางส่วนขัดแย้งกันไม่สามารถเลือกฮีลิกคู่นั้นมาประกอบร่วมกันในโครงสร้างได้และค่าในเมทริกซ์เป็น 0 เช่น ฮีลิกหมายเลข 1 ($helix_1$) ซึ่งถูกเข้ารหัสคือ [3 ; 118 ; 6] กับ ฮีลิกหมายเลข 2 ($helix_2$) ถูกเข้ารหัสคือ [6 ; 118 ; 3] ดังนั้น matrix[1][2] = 0 นอกจากนี้ กลุ่มของฮีลิกที่ปรากฏร่วมกันในโครงสร้างโดยไม่มีตำแหน่งเบสขัดแย้งหรือตรงกันจะมีค่าในเมทริกซ์เป็น 1 เช่น $helix_1$ กับ $helix_5$ (ซึ่งถูกเข้ารหัสเป็น [10 ; 68 ; 2]) ดังนั้น matrix[1][5] = 1 และกลุ่มฮีลิกที่สามารถเกิดร่วมกันในโครงสร้างได้แต่มีการแชร์ตำแหน่งเบสบางส่วนร่วมกันค่าในเมทริกซ์จะเป็น 2 เช่น $helix_6$ ซึ่งถูกเข้ารหัสเป็น [12 ; 69 ; 4] กับ $helix_8$ ซึ่งถูกเข้ารหัสเป็น [15 ; 63 ; 3] ดังนั้น matrix[6][8] = 2

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2.2 การสร้างประชากร

หลังจากกำหนดค่าเริ่มต้นให้กับโมเดลทั้งในส่วนของเวกเตอร์ความน่าจะเป็นและเมทริกซ์ความเข้ากันได้ของฮีลิกเรียบร้อยแล้ว ขั้นตอนต่อไปก็จะเริ่มเข้าสู่กระบวนการของขั้นตอนวิธีประมาณการแจกแจง โดยขั้นตอนแรกคือการสุ่มสร้างประชากรซึ่งเป็นตัวแทนโครงสร้างทุติยภูมิอาร์เอ็นเอที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ โดยมีจำนวนโครงสร้างอ้างอิงตามค่าที่กำหนดไว้ในพารามิเตอร์ขนาดประชากร (population size)

แต่ละโครงสร้างถูกสร้างโดยสุ่มเลือกหมายเลขฮีลิกจากเซตของฮีลิกที่เป็นไปได้ทั้งหมด โดยฮีลิกหนึ่งมีโอกาสถูกสุ่มมาสร้างโครงสร้างก็ต่อเมื่อมันสามารถเข้ากันได้กับฮีลิกอื่น ๆ ที่ถูกเลือกไปสร้างโครงสร้างแล้วเพื่อทำให้โครงสร้างที่ทำนายได้มีความถูกต้องอยู่เสมอ กล่าวอีกนัยหนึ่งคือฮีลิกชิ้นแรกสุดจะเป็นฮีลิกหมายเลขใดก็ได้แต่ฮีลิกชิ้นถัดไปจะต้องมีค่าในเมทริกซ์ความเข้ากันได้เทียบกับฮีลิกที่ผ่านการคัดเลือกไปแล้วไม่ใช่ 0

โดยขั้นตอนวิธีที่นำเสนอจะทำการสุ่มเลือกทีละฮีลิกมาประกอบรวมกันในโครงสร้างไปจนกระทั่งไม่มีฮีลิกที่สามารถเข้ากันได้แล้วหรือความยาวของโครโมโซมเป็นไปตามที่กำหนดก็จะเสร็จสิ้นการทำนาย 1 โครงสร้าง ดังนั้น ความยาวของแต่ละโครโมโซมไม่จำเป็นต้องเท่ากัน

3.2.2.1 การสร้างประชากรสำหรับ *EDA-G*

จุดมุ่งหมายของ *EDA-G* คือความพยายามในการสำรวจให้ทั่วทั้งปริภูมิค้นหา ดังนั้นนอกเหนือจากการสุ่มฮีลิกหมายเลขต่าง ๆ มาประกอบกันอ้างอิงตามเวกเตอร์ความน่าจะเป็นแล้วกลไกอีกอย่างที่ใส่เพิ่มเติมเข้าไปคือการพยายามเลือกชิ้นของฮีลิกที่มาประกอบกันในโครงสร้างให้มีความหลากหลายมากที่สุด กล่าวคือ ในแต่ละรุ่นของการสร้างประชากรการสุ่มเลือกจะเป็นในลักษณะของการสุ่มออกมาแบบไม่ใส่คืนจนกระทั่งเหลือจำนวนฮีลิกที่สามารถสุ่มเลือกได้ไม่ถึง 40% จากทั้งหมดก็ทำการคืนค่าฮีลิกที่มีโอกาสถูกสุ่มเลือกได้กลับไปเต็มจำนวนเช่นเดิม การทำเช่นนี้เพื่อเพิ่มโอกาสให้ฮีลิกที่มีความน่าจะเป็นต่ำมีโอกาสถูกเลือกมาประกอบโครงสร้างมากยิ่งขึ้น ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-G* แสดงดังรูปที่ 3.6

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60
2430366845

<u>ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-G*</u>

กำหนดให้ เซต *H* แทนหมายเลขฮีลิกที่สามารถเลือกมาสร้างโครงสร้างได้

เซต *R* แทนหมายเลขฮีลิกที่ขัดกับฮีลิกชิ้นที่ถูกเลือกไปแล้ว

1. ถ้าเป็นโครโมโซมตัวแรก $|H|_{\text{เริ่มต้น}} = N$ เมื่อ *N* แทนจำนวนฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิก ไม่เช่นนั้น $|H|_{\text{เริ่มต้น}} = N$ – จำนวนฮีลิกที่ปรากฏอยู่ในโครโมโซมอื่น และ ถ้า $|H|_{\text{เริ่มต้น}} < 0.4*N$ กำหนดให้ $|H|_{\text{เริ่มต้น}} = N$
2. สุ่มเลือกฮีลิกจำนวน 1 ชิ้นจาก *H* โดยโอกาสที่ฮีลิกแต่ละชิ้นจะถูกเลือกอ้างอิงตามความน่าจะเป็นในเวกเตอร์ความน่าจะเป็น
3. ปรับปรุงสมาชิกใน *H* ให้เหลือเฉพาะฮีลิกที่สามารถเข้ากันได้กับฮีลิกชิ้นที่ถูกเลือกไปแล้วจะได้ $H_{\text{ใหม่}} = H_{\text{เดิม}} - R$
4. วนทำซ้ำข้อ 2–3 จน *H* เป็นเซตว่างหรือความยาวของโครโมโซมเป็นไปตามที่กำหนด
5. วนทำซ้ำข้อ 1 – 4 จนสร้างโครโมโซมได้ครบตามขนาดประชากรที่กำหนด

รูปที่ 3.6 ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-G*

ตัวอย่างการสุ่มสร้างประชากรขนาด 5 โครโมโซมจากเซตของฮีลิกที่มีจำนวน 20 ชิ้น และกำหนดความยาวโครโมโซมสูงสุดเท่ากับ 5 เป็นดังนี้

กำหนดให้ *C1* แทนโครโมโซมแรกของประชากร

1. เนื่องจาก *C1* เป็นโครโมโซมแรก ดังนั้น $|H|_{\text{เริ่มต้น}} = 20$ และสมาชิกในเซต *H* เป็นดังนี้

   $H = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$

2. สุ่มเลือกฮีลิกชิ้นแรกจาก *H* ได้หมายเลข 19 จัดเก็บใน *C1*

   $C1 = \{19\}$

3. ปรับปรุงสมาชิกใน *H*

   กำหนดให้ *R* แทนเซตของฮีลิกที่ขัดกับ 19 ซึ่ง $R = \{16, 17, 18, 19\}$

   $H_{\text{ใหม่}} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20\}$

4. สุ่มเลือกฮีลิกชิ้นถัดไปได้หมายเลข 8 จัดเก็บใน *C1*

   $C1 = \{19, 8\}$

5. ปรับปรุงสมาชิกใน *H*

กำหนดให้ *R* แทนเซตของฮีลิกที่ขัดกับฮีลิกใน *C1* ซึ่ง *R* = {8, 16, 17, 18, 19}

$H_{ใหม่}$ = {1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 20}

6. สุ่มเลือกฮีลิกชิ้นถัดไปได้หมายเลข 9 จัดเก็บใน C1

C1 = {19, 8, 9}

7. ปรับปรุงสมาชิกใน *H*

กำหนดให้ *R* แทนเซตฮีลิกที่ขัดกับฮีลิกใน *C1* ซึ่ง *R* = {8, 9, 16, 17, 18, 19}

$H_{ใหม่}$ = {1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 15, 20}

8. ทำเช่นนี้ไปเรื่อย ๆ จนท้ายที่สุด *H* เป็นเซตว่างจบการทำงานได้โครโมโซม *C1* คือ {19, 8, 9, 12, 10, 15, 1, 20, 7}

กำหนดให้ *C2* แทนโครโมโซมตัวที่สองของประชากร

1. เนื่องจาก *C2* ไม่ใช่โครโมโซมแรก ดังนั้น $|H|_{เริ่มต้น}$ = 11 และสมาชิกในเซต *H* เป็นดังนี้

H = {2, 3, 4, 5, 6, 11, 13, 14, 16, 17, 18}

2. สุ่มเลือกฮีลิกชิ้นแรกจาก *H* ได้หมายเลข 18 จัดเก็บใน *C2*

C2 = {18}

3. ปรับปรุงสมาชิกใน *H*

กำหนดให้ *R* แทนเซตของฮีลิกที่ขัดกับ 18 ซึ่ง *R* = {14, 18}

$H_{ใหม่}$ = {2, 3, 4, 5, 6, 11, 13, 16, 17}

4. สุ่มเลือกฮีลิกชิ้นถัดไปได้หมายเลข 17 จัดเก็บใน *C2*

C2 = {18, 17}

5. ปรับปรุงสมาชิกใน *H*

กำหนดให้ *R* แทนเซตของฮีลิกที่ขัดกับฮีลิกใน *C2* ซึ่ง *R* = {4, 14, 17, 18}

$H_{ใหม่}$ = {2, 3, 5, 6, 11, 13, 16}

6. สุ่มเลือกฮีลิกชิ้นถัดไปได้หมายเลข 16 จัดเก็บใน *C2*

C2 = {18, 17, 16}

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

7. ปรับปรุงสมาชิกใน *H*

   กำหนดให้ *R* แทนเซตของฮีลิกที่ขัดกับฮีลิกใน *C2* ซึ่ง *R* = {4, 14, 16, 17, 18}

   $H_{\text{ใหม่}}$ = {2, 3, 5, 6, 11, 13}

8. ทำเช่นนี้ไปจนกระทั่ง *H* เป็นเซตว่าง เมื่อจบการทำงานได้โครโมโซม *C2* คือ {18, 17, 16, 11, 3, 13}

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากนั้นวนทำซ้ำเช่นนี้จนได้โครโมโซมครบตามขนาดประชากรที่กำหนด เนื่องจาก $|H|_{\text{เริ่มต้น}}$ สำหรับการสร้างโครโมโซมตัวที่สามเมื่อหักฮีลิกชิ้นที่ถูกเลือกไปสร้างโครโมโซม *C1*, *C2* แล้วเหลือเพียง 5 ชิ้น ซึ่งน้อยกว่า 40% ของจำนวนฮีลิกทั้งหมด ดังนั้นจำนวนสมาชิกใน *H* จะถูกคืนค่ากลับไปเต็มจำนวนที่ 20 ชิ้นเหมือนเดิม ดังนั้น เมื่อเสร็จสิ้นขั้นตอนการสร้างประชากรของ *EDA-G* ได้ผลลัพธ์เป็นดังนี้

C1 = {19, 8, 9, 12, 10, 15, 1, 20, 7}

C2 = {18, 17, 16, 11, 3, 13}

C3 = {17, 10, 16, 7, 12, 9, 8, 18, 5, 1}

C4 = {19, 15, 3, 20, 11, 13}

C5 = {19, 9, 5, 8, 12, 7, 1, 10, 20}

3.2.2.2 การสร้างประชากรสำหรับ *EDA-L*

EDA-L ถูกสลับมาทำงานก็ต่อเมื่อ *EDA-G* ไม่สามารถหาคำตอบที่มีค่าความเหมาะสมดีขึ้นได้ ซึ่งประเมินจากการที่ข้อมูลในอาไคว์ไม่มีการเปลี่ยนแปลงติดต่อกัน *m* รุ่น (*m* เป็นพารามิเตอร์) ขั้นตอนวิธี Hybrid-EDAfold จะสลับการทำงานไปยัง *EDA-L* และเปลี่ยนวิธีการสร้างประชากรที่แตกต่างไปจาก *EDA-G* นั่นคือ *EDA-L* สร้างประชากรโดยการสุ่มเลือกโครโมโซมที่ถูกจัดเก็บในอาไคว์มาเป็นบรรพบุรุษแล้วทำการกลายพันธุ์เพื่อสร้างโครโมโซมลูก 1 ตัว ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-L* แสดงดังรูปที่ 3.7

<u>ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-L*</u>

กำหนดให้ *C* แทนเซตของฮีลิกที่ปรากฏในโครโมโซมลูก

R แทนเซตของฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษ

H แทนเซตของฮีลิกที่เข้ากันได้กับฮีลิกที่เหลืออยู่หลังจากลบฮีลิกทิ้งตาม *R*

1. สุ่มเลือก 1 โครโมโซมจากอาไคว์ กำหนดให้เป็นโครโมโซมบรรพบุรุษ (ทุกโครโมโซมในอาไคว์มีโอกาสถูกเลือกเท่ากัน) จัดเก็บหมายเลขฮีลิกที่พบในโครโมโซมบรรพบุรุษในเซต *C*
2. คำนวณจำนวนฮีลิกที่จะต้องสุ่มทิ้งอ้างอิงตามพารามิเตอร์ *per_Remove* โดยมีจำนวนเท่ากับ |*C*| x *per_Remove*
3. คำนวณสมาชิกของ *R* จากการสุ่มฮีลิกจากโครโมโซมบรรพบุรุษทิ้งตามจำนวนที่คำนวณได้ในข้อ 2 โอกาสที่ฮีลิกใด ๆ จะถูกสุ่มทิ้งเป็นส่วนกลับของความน่าจะเป็นของฮีลิกนั้นในเวกเตอร์ความน่าจะเป็น (ความน่าจะเป็นสูงโอกาสถูกสุ่มทิ้งต่ำ)
4. $C_{\text{ใหม่}} = C_{\text{เดิม}} - R$
5. คำนวณสมาชิกของ *H* ซึ่งเป็นฮีลิกที่เข้ากันได้กับฮีลิกที่ปรากฏในเซต *C*
6. สุ่มเลือก 1 ฮีลิก จากเซต *H* และจัดเก็บในเซต *C*
7. ทำซ้ำข้อ 5 – 6 จนกระทั่งได้โครโมโซมยาวตามที่กำหนดหรือ *H* เป็นเซตว่างได้ 1 โครโมโซม
8. ทำซ้ำข้อ 1 – 7 จนได้โครโมโซมครบตามขนาดประชากรที่กำหนด

รูปที่ 3.7 ขั้นตอนวิธีการสร้างประชากรสำหรับ *EDA-L*

ตัวอย่างการสร้างประชากรด้วย *EDA-L* เมื่อกำหนดค่าพารามิเตอร์ *per_Remove* เป็น 50% และข้อมูลโครโมโซมในอาไคว์เป็นดังนี้

A1 = {1, 7, 8, 9, 10, 12, 14, 19, 20}

A2 = {3, 8, 9, 10, 12, 15, 16, 17, 18}

A3 = {1, 5, 7, 8, 9, 10, 12, 17, 18, 20}

A4 = {3, 11, 13, 14, 19, 20}

A5 = {3, 11, 13, 15, 16, 17, 18}

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

กำหนดให้ *R* แทนเซตของฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษ

H แทนเซตของฮีลิกที่เข้ากันได้กับฮีลิกที่เหลืออยู่หลังจากลบฮีลิกทิ้งตาม *R*

A แทนเซตของฮีลิกที่ถูกสุ่มจาก *H* ในการสร้างโครโมโซมลูกหลาน

ตัวอย่างการสร้างโครโมโซมลูก *C1* โดยใช้โครโมโซมบรรพบุรุษ *A5* เป็นต้นแบบ

1. สุ่มเลือกโครโมโซมในอาไคว์ได้โครโมโซม *A5* เป็นโครโมโซมบรรพบุรุษสำหรับสร้างโครโมโซมลูก *C1* ดังนั้น *C1* = {3, 11, 13, 15, 16, 17, 18}
2. คำนวณจำนวนฮีลิกที่จะสุ่มทิ้งจากโครโมโซม *A5* ได้ 9 x 0.5 = 4 ชิ้น
3. สุ่มฮีลิกจำนวน 4 ชิ้นทิ้งจาก *A5* ได้ *R* = {11, 13, 15, 16}
4. ปรับปรุงสมาชิกในเซต *C1* โดยลบสมาชิกที่ปรากฏใน *R* ทิ้ง ดังนั้น *C1* = {3, 17, 18}
5. คำนวณสมาชิกในเซต *H* ซึ่งเป็นฮีลิกชิ้นที่เข้ากันได้กับฮีลิกใน *C1* ดังนั้น *H* = {8, 9, 10, 11, 12, 13, 15, 16, 20}
6. สุ่มเลือกฮีลิก 1 ชิ้นจาก *H* ได้หมายเลข 8 จัดเก็บใน *C1* ดังนั้น *C1* = {3, 8, 17, 18}
7. ปรับปรุงสมาชิกในเซต *H* ได้เป็น {9, 10, 11, 12, 13, 15, 16, 20}
8. ทำเช่นนี้ไปจนกระทั่ง *H* เป็นเซตว่างหรือโครโมโซมลูกมีความยาวตามที่กำหนด เมื่อจบการทำงานได้โครโมโซม *C1* คือ {3, 8, 9, 10, 12, 15, 16, 17, 18}

ทำในลักษณะเช่นนี้จนได้โครโมโซมครบตามขนาดประชากรที่กำหนด เมื่อเสร็จสิ้นขั้นตอนการสร้างประชากรของ *EDA-L* ได้ผลลัพธ์ดังนี้

C1 = {3, 8, 9, 10, 12, 15, 16, 17, 18}

C2 = {1, 7, 8, 9, 10, 13, 14, 19, 20}

C3 = {3, 8, 9, 10, 12, 14, 19, 20}

C4 = {1, 7, 8, 9, 10, 12, 15, 16, 17, 18}

C5 = {1, 5, 7, 8, 9, 10, 12, 19, 20}

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2.3 การประเมินคุณภาพของประชากร

เมื่อสร้างโครงสร้างได้ครบตามขนาดประชากรที่กำหนดแล้ว ขั้นตอนนี้จะเป็นการประเมินว่าแต่ละโครงสร้างที่สร้างได้มีแนวโน้มเป็นโครงสร้างที่ตรงกับโครงสร้างที่เป็นคำตอบมากน้อยเพียงใด โดยการทำนายโครงสร้างของงานวิจัยนี้อยู่บนพื้นฐานวิธีการหาค่าพลังงานต่ำสุด โดยทำการคำนวณค่าพลังงานของแต่ละโครงสร้างอ้างอิงกฎและค่าพารามิเตอร์จาก nearest-neighbor parameter (NNDB) [50] ภายใต้สมมุติฐานที่ว่าโครงสร้างใดยิ่งมีค่าพลังงานต่ำจะยิ่งมีโอกาสสูงที่จะเป็นโครงสร้างที่เหมือนหรือใกล้เคียงกับโครงสร้างที่เป็นคำตอบ

NNDB เป็นฐานข้อมูลที่รวบรวมกฎหรือสมการและพารามิเตอร์สำหรับทำนายความเสถียรของโครงสร้างทุติยภูมิของอาร์เอ็นเอ โดยพารามิเตอร์เหล่านี้ถูกใช้แพร่หลายในโปรแกรมคอมพิวเตอร์ที่อาศัยหลักการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอที่มีค่าพลังงานต่ำ เช่น Mfold, RNAfold และ RNAstructure เป็นต้น สำหรับอาร์เอ็นเอสมการที่ใช้ทำนายความเสถียรจะแยกตามลักษณะรูปร่างพื้นฐานที่ปรากฏในโครงสร้าง ได้แก่ ฮีลิก ลูปชนิดแฮร์พิน ลูปชนิดอินเทอร์นอล ลูปชนิดบัลจ์ ลูปชนิดมัลติบรานซ์ และ ลูปชนิดเอ็กเทียเรียร์ ดังนั้น ฟังก์ชันวัตถุประสงค์ที่งานวิจัยนี้เลือกใช้สำหรับประเมินค่าความเหมาะสมของโครงสร้างที่ทำนายได้คือผลรวมค่าพลังงานของโครงสร้างพิจารณาแยกตามรูปร่างที่ปรากฏ โดยตัวอย่างสมการสำหรับการคำนวณความเสถียรของรูปร่างฮีลิกและลูปชนิดแฮร์พินเป็นดังสมการที่ 3.1 – 3.3

การคำนวณค่าพลังงานสำหรับบริเวณที่เป็นฮีลิก เป็นดังสมการที่ 3.1

$$\begin{aligned}\Delta G^{\circ}_{37\,Watson-Crick} = {} & \Delta G^{\circ}_{37\,init} + \Delta G^{\circ}_{37\,AU\,penalty}(per\ AU) \\ & + \Delta G^{\circ}_{37\,symmetry}(self - complementary) \\ & + \sum\left[\Delta G^{\circ}_{37\,stacking}\right]\end{aligned} \tag{3.1}$$

เมื่อ $\Delta G^{\circ}_{37\,init}$ แทนค่าพลังงานตั้งต้นในการสร้างโครงสร้างโมเลกุลมีค่าเท่ากับ 4.09

$\Delta G^{\circ}_{37\,AU\,penalty}(per\ AU)$ แทนค่าพลังงานเมื่อปลายฮีลิกมีเบส A จับคู่กับเบส U พิจารณาที่ปลายทั้งสองด้านของฮีลิก คิดค่าพลังงานเพิ่มด้านละ 0.45

$\Delta G^{\circ}_{37\,symmetry}(self - complementary)$ แทนกรณีที่ฮีลิกมีลำดับแบบพาลินโดรมเมื่อพิจารณาจากทิศทางจากปลาย 5’ ไปปลาย 3’ หรือ ทิศทางจากปลาย 3’ ไปปลาย 5’ หากพบว่าเป็นไปตามเงื่อนไขนี้ค่าพลังงานจะถูกบวกเพิ่มไปอีก 0.43 เช่น $\begin{matrix}5'AGCGCU3' \\ 3'UCGCGA5'\end{matrix}$

$\sum\left[\Delta G^{\circ}_{37\,stacking}\right]$ แทนค่าพลังงานของคู่เบสที่พบในฮีลิก คำนวณไล่จากซ้ายไปขวาทีละ 2 คู่เบสขยับครั้งละ 1 ตำแหน่ง ถ้าฮีลิกยาว *N* คู่เบสจะได้ทั้งหมด *N*-1 พจน์ เช่น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

$$\Delta G^\circ{}_{37}\begin{bmatrix}5'AGCGCU3'\\3'UCGCGA5'\end{bmatrix} = \Delta G^\circ{}_{37}\begin{pmatrix}AG\\UC\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}GC\\CG\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}CG\\GC\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}GC\\CG\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}CU\\GA\end{pmatrix}$$

และค่าพลังงานของแต่ละพจน์อ้างอิงตามค่าพารามิเตอร์ใน NNDB

การคำนวณค่าพลังงานสำหรับบริเวณที่เป็นลูปชนิดแฮร์พิน เนื่องจากแฮร์พินคือบริเวณของลูปที่ติดกับหนึ่งฮีลิก ดังนั้น การคำนวณขึ้นกับจำนวนเบสอิสระที่พบในลูป ดังนี้

- ถ้าจำนวนเบสที่พบในลูปมีจำนวนต่ำกว่า 3 ค่าพลังงานเป็น 0
- ถ้าจำนวนเบสที่พบในลูปมีจำนวน 3 เบส คำนวณค่าพลังงานดังสมการที่ 3.2

$$\Delta G^\circ{}_{37\,hairpin} = \Delta G^\circ{}_{37\,init}(3) + \Delta G^\circ{}_{37\,penalty}(all\ C\ loops) \quad (3.2)$$

เมื่อ $\Delta G^\circ{}_{37\,init}(3)$ แทนค่าพลังงานตั้งต้นมีค่าเป็น 5.4

$\Delta G^\circ{}_{37\,penalty}(all\ C\ loops)$ แทนค่าพลังงานที่บวกเพิ่มถ้าทุกเบสในลูปเป็น C

- ถ้าจำนวนเบสที่พบในลูปมีจำนวนมากกว่า 3 คำนวณค่าพลังงานดังสมการที่ 3.3

$$\begin{aligned}\Delta G^\circ{}_{37\,hairpin} = {} & \Delta G^\circ{}_{37\,init}(n) + \Delta G^\circ{}_{37}(terminal\ mismatch)\\ & + \Delta G^\circ{}_{37}\,(UU\ or\ GA\ first\ mismatch)\\ & + \Delta G^\circ{}_{37}\,(GG\ first\ mismatch)\\ & + \Delta G^\circ{}_{37}\,(special\ GU\ closure)\\ & + \Delta G^\circ{}_{37\,penalty}\,(all\ C\ loops)\end{aligned} \quad (3.3)$$

เมื่อ n คือ จำนวนเบสที่พบในลูป

$\Delta G^\circ{}_{37\,init}(n)$ คือ ค่าพลังงานตั้งต้นอ้างอิงตามจำนวนเบสที่พบในลูป

terminal mismatches คือ คู่เบสที่ไม่ใช่คาร์โนนิคอลที่ติดกับปลายแต่ละด้านของฮีลิก

first mismatch คือ เบส 2 ตัวบริเวณเริ่มต้นลูปถัดจากส่วนที่เป็นฮีลิก

- นอกจากนี้ยังมีลูปชนิดแฮร์พินแบบพิเศษ หากพบโครงสร้างที่มีข้อมูลตรงตามนี้ให้ใช้ค่าพลังงานตามที่ระบุใน NNDB ได้เลย ไม่ต้องคำนวณตามสมการที่ 3.2 หรือ 3.3

โดยงานวิจัยนี้ไม่ได้รองรับการคำนวณค่าพลังงานของโครงสร้างในส่วนของลูปชนิดเอ็กเทียเรียร์และทั้งสองขั้นตอนวิธีประมาณการแจกแจงใช้วิธีการประเมินค่าความเหมาะสมเหมือนกัน ตัวอย่างการประเมินค่าความเหมาะสมของประชากรแสดงดังตารางที่ 3.9

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 3.9 การประเมินค่าความเหมาะสมสำหรับประชากร

โครโมโซม	**free energy**
C1 = {1, 7, 8, 9, 10, 12, 15, 16, 17, 18}	-55.10
C2 = {1, 6, 8, 9, 10, 12, 16, 17, 18}	-50.40
C3 = {3, 8, 9, 10, 12, 15, 16, 17, 18}	-47.10
C4 = {2, 7, 8, 9, 10, 12, 15, 16, 17, 18}	-46.00
C5 = {1, 5, 7, 8, 9, 10, 12, 19, 20}	-41.60

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากตารางที่ 3.9 โครโมโซม *C1* มีค่าความเหมาะสมดีสุดตามด้วยโครโมโซม *C2*, *C3*, *C4* และ *C5* ตามลำดับ เนื่องจากปัญหานี้เป็นปัญหาการหาค่าความเหมาะสมต่ำสุด ในเบื้องต้นได้มีการประเมินเพื่อศึกษาว่าฟังก์ชันวัตถุประสงค์ที่งานวิจัยนี้เลือกใช้สอดคล้องกับคุณภาพคำตอบที่ได้จริงหรือไม่ โดยการนำแต่ละโครโมโซมที่เป็นตัวแทนของโครงสร้างที่ทำนายได้ไปเปรียบเทียบกับโครงสร้างที่เป็นคำตอบและนับจำนวนตำแหน่งคู่เบสที่ทำนายได้ถูกต้อง ผลลัพธ์แสดงดังตารางที่ 3.10

ตารางที่ 3.10 ความสอดคล้องกันระหว่างค่าความเหมาะสมกับคุณภาพคำตอบ

โครโมโซม	**จำนวนคู่เบสที่ทำนายได้ถูกต้อง**
C1 = {1, 7, 8, 9, 10, 12, 15, 16, 17, 18}	24
C2 = {1, 6, 8, 9, 10, 12, 16, 17, 18}	20
C3 = {3, 8, 9, 10, 12, 15, 16, 17, 18}	14
C4 = {2, 7, 8, 9, 10, 12, 15, 16, 17, 18}	14
C5 = {1, 5, 7, 8, 9, 10, 12, 19, 20}	14

จากตารางที่ 3.10 คอลัมน์ที่ 2 แสดงจำนวนตำแหน่งคู่เบสที่ทำนายได้อย่างถูกต้องเมื่อนำข้อมูลที่เข้ารหัสไว้ตามหมายเลขฮีลิกที่ปรากฏในแต่ละโครโมโซมไปถอดรหัสได้เป็นโครงสร้างที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ตรงกับตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบ แสดงให้เห็นว่าฟังก์ชันวัตถุประสงค์ที่งานวิจัยนี้เลือกใช้มีความสอดคล้องกับคุณภาพคำตอบในระดับที่ดีเพียงพอ นั่นคือ ค่าพลังงานที่ได้ยิ่งต่ำจำนวนตำแหน่งคู่เบสที่ทำนายได้ถูกต้องยิ่งมากขึ้น

3.2.4 การปรับปรุงค่าในเวกเตอร์ความน่าจะเป็น

เมื่อประเมินค่าความเหมาะสมของประชากรเรียบร้อยแล้ว ขั้นตอนต่อไปคือการปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นโดยใช้ข้อมูลจากคุณภาพคำตอบที่สร้างได้ในแต่ละรุ่น วิธีการที่งานวิจัยนี้นำเสนอแตกต่างจากขั้นตอนวิธีประมาณการแจกแจงมาตรฐานที่ใช้แต่โครโมโซมที่มีค่าความเหมาะสมดีเพียงอย่างเดียวเท่านั้น โดยงานวิจัยนี้จะใช้ทั้งโครโมโซมที่มีค่าความเหมาะสมดีและค่าความเหมาะสมด้อยร่วมกันในการปรับปรุงเวกเตอร์ความน่าจะเป็น และขั้นตอนวิธีประมาณการแจกแจงทั้งคู่มีวิธีการปรับปรุงเวกเตอร์ความน่าจะเป็นแตกต่างกัน รายละเอียดเป็นดังนี้

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2.4.1 การปรับปรุงเวกเตอร์ความน่าจะเป็นด้วย *EDA-G*

เนื่องจากขั้นตอนวิธี Hybrid-EDAfold มีการเรียนรู้จากทั้งโครโมโซมดีและโครโมโซมด้อย ดังนั้น โครโมโซมทั้งหมดในประชากรจะถูกจำแนกออกเป็น 3 กลุ่ม คือ โครโมโซมกลุ่มดี โครโมโซมกลุ่มด้อย และโครโมโซมกลุ่มที่ไม่นำมาพิจารณา ซึ่งในกระบวนการจำแนกกลุ่มโครโมโซมเกี่ยวข้องกับ 2 พารามิเตอร์ได้แก่ *perG* และ *perP* โดยที่ *perG* คือ สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่ามีคุณภาพดีในประชากรและ *perP* คือ สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่ามีคุณภาพด้อยในประชากร

โครโมโซมที่ถูกพิจารณาว่ามีคุณภาพดีประเมินจากค่าพลังงานที่คำนวณได้จากขั้นตอนการประเมินค่าความเหมาะสมของประชากรต่ำสุด n ตัวแรกในกลุ่มประชากรรุ่นที่กำลังพิจารณา (n คำนวณจาก *perG* * ขนาดประชากร) และในทางตรงกันข้ามโครโมโซมที่ถูกพิจารณาว่ามีคุณภาพด้อยประเมินจากการมีค่าพลังงานสูงสุด m ตัวแรกในกลุ่มประชากรรุ่นที่กำลังพิจารณา (m คำนวณจาก *perP* * ขนาดประชากร) เช่น ถ้ากำหนดขนาดประชากรเป็น 10 โครโมโซม และกำหนดค่า *perG* และ *perP* เท่ากับ 0.3 เมื่อทำการเรียงลำดับโครโมโซมในประชากรด้วยค่าความเหมาะสมจากน้อยไปมากได้ผลลัพธ์ดังตารางที่ 3.11 จะได้ว่าโครโมโซม 3 ลำดับแรกถูกจำแนกว่าเป็นโครโมโซมดีและโครโมโซม 3 ลำดับท้ายจะถูกจำแนกว่าเป็นโครโมโซมด้อย และโครโมโซมอื่น ๆ นอกเหนือจากนี้จะอยู่ในกลุ่มโครโมโซมที่ไม่ถูกนำมาพิจารณาในการปรับปรุงเวกเตอร์ความน่าจะเป็น

ตารางที่ 3.11 การจำแนกโครโมโซมดีและโครโมโซมด้อยในกลุ่มประชากร

โครโมโซม	free energy	ผลการจำแนก
C1 = {1,5,7,8,9, 11, 12, 16, 17, 18}	-49.40	ดี
C2 = {1,5,7,8,9,10, 13, 16, 17, 18}	-48.00	ดี
C3 = {3, 8, 9, 10, 12, 15, 16, 17, 18}	-47.10	ดี
C4 = {2, 8, 9, 10, 12, 14, 19}	-45.90	
C5 = {3, 11, 13, 15, 16, 17, 18}	-45.90	
C6 = {2, 5, 7, 11, 13, 16, 17, 18}	-41.10	
C7 = {1, 5, 7, 8, 9, 10, 12, 18, 20}	-40.50	
C8 = {1, 5, 7, 11, 13, 16, 17, 18}	-35.80	ด้อย
C9 = {1, 6, 11, 12, 16, 17, 18}	-34.70	ด้อย
C10 = {2, 5, 7, 8, 9, 10, 13, 19}	-34.20	ด้อย

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

เมื่อจำแนกประชากรออกเป็นกลุ่มโครโมโซมดีและกลุ่มโครโมโซมด้อยเรียบร้อยแล้ว ข้อมูลจากโครโมโซมทั้ง 2 กลุ่มนี้จะถูกใช้ในการปรับปรุงเวกเตอร์ความน่าจะเป็น กล่าวคือ กลุ่มฮีลิกที่พบในกลุ่มโครโมโซมดีถือว่าประสบความสำเร็จในการนำมาสร้างโครงสร้างเนื่องจากให้ค่าพลังงานที่ต่ำเมื่อเทียบกับโครโมโซมอื่น ๆ ในประชากร ดังนั้นสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับฮีลิกเหล่านี้ควรได้ความน่าจะเป็นเพิ่มขึ้นจากเดิม ในทางกลับกัน กลุ่มฮีลิกที่ปรากฏในกลุ่มโครโมโซมด้อยถือว่าไม่ส่งเสริมต่อการนำมาสร้างโครงสร้างเนื่องจากเมื่อนำมาประกอบรวมกันในโครงสร้างแล้วได้ค่าพลังงานที่ค่อนข้างสูงเมื่อเทียบกับโครโมโซมอื่น ๆ ในประชากร ดังนั้นสมาชิกในเวกเตอร์ที่สอดคล้องกับฮีลิกเหล่านี้ควรถูกปรับลดความน่าจะเป็นลง และปริมาณการเพิ่มขึ้นหรือลดลงของความน่าจะเป็นในงานวิจัยนี้อ้างอิงตามค่าอัตราการเรียนรู้ (learning rate) ซึ่งเป็นพารามิเตอร์ที่กำหนดโดยผู้ใช้ นอกจากนี้ หากหมายเลขฮีลิกใดที่นับความถี่ได้แค่ 1 (พบหมายเลขฮีลิกนี้แค่ในโครโมโซมเดียว) จะไม่ถูกนำมาพิจารณาในขั้นตอนการปรับปรุงความน่าจะเป็น เพื่อลดความเสี่ยงจากการปรับปรุงความน่าจะเป็นของฮีลิกผิดชิ้น (ถ้าฮีลิกนั้นเป็นฮีลิกที่ดีหรือฮีลิกด้อยควรถูกพบในหลาย ๆ โครโมโซมที่อยู่ในกลุ่มคุณภาพคำตอบเดียวกัน) และควบคุมให้ทุกสมาชิกในเวกเตอร์ความน่าจะเป็นมีค่าอยู่ในช่วง 0.0 – 1.0 เท่านั้น โดยขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-G* แสดงดังรูป 3.8

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-G*

กำหนดให้ *Good* แทนเซตของหมายเลขฮีลิกในกลุ่มดี

Poor แทนเซตของหมายเลขฮีลิกในกลุ่มด้อย

1. นับความถี่ของหมายเลขฮีลิกที่พบอยู่ในกลุ่มโครโมโซมดี
2. นับความถี่ของหมายเลขฮีลิกที่พบอยู่ในกลุ่มโครโมโซมด้อย
3. พิจารณาความถี่ของแต่ละฮีลิกในข้อ 1 ถ้าหมายเลขฮีลิกใดมีความถี่มากกว่า 1 เก็บหมายเลขฮีลิกนั้นใน *Good*
4. พิจารณาความถี่ของแต่ละฮีลิกในข้อ 2 ถ้าหมายเลขฮีลิกใดมีความถี่มากกว่า 1 เก็บหมายเลขฮีลิกนั้นใน *Poor*
5. เพิ่มความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับฮีลิกที่พบในเซต *Good* แต่ไม่พบในเซต *Poor* จากเดิมด้วยค่าอัตราการเรียนรู้ที่กำหนด
6. ลดความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับฮีลิกที่พบในเซต *Poor* แต่ไม่พบในเซต *Good* จากเดิมด้วยค่าอัตราการเรียนรู้ที่กำหนด

รูปที่ 3.8 ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-G*

ตัวอย่างการปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-G* โดยอ้างอิงผลการจำแนกโครโมโซมดีและโครโมโซมด้อยจากตารางที่ 3.11 และกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.01 เป็นดังนี้

1. นับความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมดี (จากตัวอย่างนี้ ได้แก่ โครโมโซม *C1, C2* และ *C3*) ได้ผลลัพธ์ดังตารางที่ 3.12 ดังนั้น *Good* = {1, 5, 7, 8, 9, 10, 12, 16, 17, 18}

ตารางที่ 3.12 ความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมดี

หมายเลขฮีลิก	1	3	5	7	8	9	10	11	12	13	15	16	17	18
ความถี่	2	1	2	2	3	3	2	1	2	1	1	3	3	3

2. นับความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมด้อย (จากตัวอย่างนี้ ได้แก่ โครโมโซม *C8, C9* และ *C10*) ได้ผลลัพธ์ดังตารางที่ 3.13 ดังนั้น *Poor* = {1, 5, 7,11, 13, 16, 17, 18}

ตารางที่ 3.13 ความถี่ของฮีลิกที่พบในกลุ่มโครโมโซมด้อย

หมายเลขฮีลิก	1	2	5	6	7	8	9	10	11	12	13	16	17	18	19
ความถี่	2	1	2	1	2	1	1	1	2	1	2	2	2	2	1

3. *Good* – *Poor* = {8, 9, 10, 12}

4. *Poor* – *Good* = {11, 13}

5. เพิ่มความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับหมายเลขฮีลิกในข้อ 3 ด้วยค่าอัตราการเรียนรู้ที่กำหนด ได้ผลลัพธ์ดังตารางที่ 3.14 ใน 4 คอลัมน์แรก

6. ลดความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับหมายเลขฮีลิกในข้อ 4 ด้วยค่าอัตราการเรียนรู้ที่กำหนด ได้ผลลัพธ์ดังตารางที่ 3.14 ใน 2 คอลัมน์สุดท้าย

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 3.14 การปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-G*

สมาชิกตำแหน่งที่	8	9	10	12	11	13
ความน่าจะเป็นก่อนปรับ	0.98	0.04	0.99	0.42	0.01	0.43
ความน่าจะเป็นหลังปรับ	0.99	0.05	1.00	0.43	0.00	0.42

3.2.4.2 การปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-L*

สำหรับ *EDA-L* คำตอบคุณภาพดีและคำตอบคุณภาพด้อยเกิดจากการแข่งกันระหว่างบรรพบุรุษกับลูกที่เกิดจากการกลายพันธุ์ กล่าวคือ ถ้าผลจากการกลายพันธุ์บรรพบุรุษผลิตลูกที่มีค่าความเหมาะสมดีขึ้น (ค่าพลังงานต่ำลง) ข้อมูลของโครโมโซมคู่นั้นจะถูกใช้ในการปรับปรุงเวกเตอร์ความน่าจะเป็นโดยมีสมมุติฐานว่ากลุ่มหมายเลขฮีลิกที่ถูกสุ่มลบทิ้งจากโครโมโซมบรรพบุรุษเป็นฮีลิกที่ไม่ควรพบในโครงสร้าง และกลุ่มหมายเลขฮีลิกที่ถูกสุ่มเพิ่มเติมในการสร้างโครโมโซมลูกทำให้ได้โครงสร้างที่มีค่าพลังงานต่ำลง ดังนั้น สมาชิกของเวกเตอร์ที่สอดคล้องกับหมายเลขฮีลิกกลุ่มที่ถูกลบทิ้งควรถูกลดความน่าจะเป็นลง และสมาชิกของเวกเตอร์ที่สอดคล้องกับหมายเลขฮีลิกกลุ่มที่ถูกเพิ่มเข้ามาควรได้ความน่าจะเป็นเพิ่ม ปริมาณความน่าจะเป็นที่เพิ่มหรือลดอ้างอิงตามค่าอัตราการเรียนรู้เช่นเดียวกัน สำหรับในกรณีที่ผลการกลายพันธุ์ไม่ได้ผลิตลูกที่มีค่าความเหมาะสมดีขึ้นก็ไม่ต้องดำเนินการใด

เพื่อความสอดคล้องกับการปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นของ *EDA-G* ดังที่ได้นำเสนอไปในหัวข้อก่อนหน้า เมื่อพบว่าบรรพบุรุษตัวใดกลายพันธุ์ได้ลูกที่มีค่าความเหมาะสมดีขึ้นให้สกัดกลุ่มของฮีลิกที่ถูกลบทิ้งและกลุ่มฮีลิกที่ถูกสุ่มเพิ่มเติมเข้ามาเก็บสะสมไว้ก่อน เมื่อครบทุกคู่ของการพิจารณาแล้วค่อยปรับปรุงเวกเตอร์ความน่าจะเป็นครั้งเดียว นั่นคือ นับความถี่ของกลุ่มหมายเลขฮีลิกที่ถูกลบทิ้งและนับความถี่ของกลุ่มหมายเลขฮีลิกที่ถูกเพิ่ม จากนั้นดำเนินการปรับปรุงความน่าจะเป็นของหมายเลขฮีลิกที่มีความถี่มากกว่า 1 ดังเช่นที่ได้นำเสนอไป ขั้นตอนวิธีการปรับปรุงเวกเตอร์ความน่าจะเป็นของ *EDA-L* แสดงดังรูปที่ 3.9

ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-L*

กำหนดให้ *Good* แทนเซตของหมายเลขฮีลิกในกลุ่มดี

Poor แทนเซตของหมายเลขฮีลิกในกลุ่มด้อย

1. คัดเลือกเฉพาะคู่ของโครโมโซมบรรพบุรุษที่ทำการกลายพันธุ์แล้วได้ลูกที่ดีขึ้น
2. รวบรวมหมายเลขฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษจากโครโมโซมคู่ที่ผ่านการคัดเลือกในข้อ 1 เก็บหมายเลขฮีลิกเหล่านั้นในเซต *Poor*
3. นับความถี่ของหมายเลขฮีลิกในเซต *Poor*
4. รวบรวมหมายเลขฮีลิกที่ถูกเพิ่มเติมในการสร้างโครโมโซมลูกหลานจากโครโมโซมคู่ที่ผ่านการคัดเลือกในข้อ 1 เก็บหมายเลขฮีลิกเหล่านั้นในเซต *Good*
5. นับความถี่ของหมายเลขฮีลิกในเซต *Good*
6. เพิ่มความน่าจะเป็นของสมาชิกในเวกเตอร์ที่สอดคล้องกับหมายเลขฮีลิกที่พบในเซต *Good* แต่ไม่พบในเซต *Poor* จากเดิมด้วยค่าอัตราการเรียนรู้ที่กำหนด
7. ลดความน่าจะเป็นของสมาชิกในเวกเตอร์ที่สอดคล้องกับหมายเลขฮีลิกที่พบในเซต *Poor* แต่ไม่พบในเซต *Good* จากเดิมด้วยค่าอัตราการเรียนรู้ที่กำหนด

รูปที่ 3.9 ขั้นตอนวิธีปรับปรุงเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-L*

ตัวอย่างการปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-L* อ้างอิงจากตัวอย่างการสร้างประชากรของ *EDA-L* ที่นำเสนอไปในหัวข้อ 3.2.2.2 และผลการประเมินค่าความเหมาะสมของโครโมโซมบรรพบุรุษและลูกที่กลายพันธุ์ได้แสดงดังตารางที่ 3.15

ตารางที่ 3.15 การเปรียบเทียบค่าความเหมาะสมของบรรพบุรุษกับลูกหลานที่สร้างได้

ลำดับ	free energy ของบรรพบุรุษ	free energy ของลูก
1	-24.30	-47.10
2	-23.20	-43.60
3	-23.20	-46.00
4	-47.10	-55.10
5	-48.90	-41.60

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

1. จากตารางที่ 3.15 พบว่ามีโครโมโซมจำนวน 4 คู่ที่ทำการกลายพันธุ์แล้วได้ลูกหลานที่มีค่าความเหมาะสมดีขึ้น (ค่าพลังงานต่ำลง) ดังนั้น ใช้ข้อมูลจากโครโมโซมคู่ที่ 1 – 4 สำหรับปรับปรุงค่าในเวกเตอร์ความน่าจะเป็น

2. รวบรวมกลุ่มของฮีลิกที่ถูกลบทิ้งออกจากโครโมโซมบรรพบุรุษ โดยกำหนดให้
   หมายเลขฮีลิกที่ถูกลบจากโครโมโซมบรรพบุรุษตัวที่ 1 คือ {11, 13, 15, 16}
   หมายเลขฮีลิกที่ถูกลบจากโครโมโซมบรรพบุรุษตัวที่ 2 คือ {3, 11, 19}
   หมายเลขฮีลิกที่ถูกลบจากโครโมโซมบรรพบุรุษตัวที่ 3 คือ {11, 13, 19}
   หมายเลขฮีลิกที่ถูกลบจากโครโมโซมบรรพบุรุษตัวที่ 4 คือ {3, 9, 15, 16, 18}
   ดังนั้น *Poor* = {3, 9, 11, 13, 15, 16, 18, 19}

3. นับความถี่ของหมายเลขฮีลิกที่พบในเซต *Poor* ได้ผลลัพธ์ดังตารางที่ 3.16

ตารางที่ 3.16 ความถี่ของหมายเลขฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษ

หมายเลขฮีลิก	3	9	11	13	15	16	18	19
ความถี่	2	1	3	2	2	2	1	2

4. รวบรวมกลุ่มของฮีลิกที่ถูกสุ่มเพิ่มเติมในการสร้างโครโมโซมลูก โดยกำหนดให้
   หมายเลขฮีลิกที่ถูกสุ่มเพิ่มเพื่อสร้างลูกตัวที่ 1 คือ {8, 9, 10, 12, 15, 16}
   หมายเลขฮีลิกที่ถูกสุ่มเพิ่มเพื่อสร้างลูกตัวที่ 2 คือ {1, 7, 8, 9, 10,19}
   หมายเลขฮีลิกที่ถูกสุ่มเพิ่มเพื่อสร้างลูกตัวที่ 3 คือ {8, 9, 10, 12,19}
   หมายเลขฮีลิกที่ถูกสุ่มเพิ่มเพื่อสร้างลูกตัวที่ 4 คือ {1, 7, 9, 15, 16,18}
   ดังนั้น *Good* = {1, 7, 8, 9, 10, 12, 15, 16, 18, 19}

5. นับความถี่ของหมายเลขฮีลิกที่พบในเซต *Good* ได้ผลลัพธ์ดังตารางที่ 3.17

ตารางที่ 3.17 ความถี่ของหมายเลขฮีลิกที่ถูกสุ่มเพิ่มเติมในการสร้างโครโมโซมลูก

หมายเลขฮีลิก	1	7	8	9	10	12	15	16	18	19
ความถี่	2	2	3	4	3	2	2	2	1	2

6. Good – Poor = {1, 7, 8, 9, 10, 12} ดังนั้น เพิ่มความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับฮีลิกหมายเลขเหล่านี้ด้วยอัตราการเรียนรู้ที่กำหนด ได้ผลลัพธ์ดังตารางที่ 3.18 เนื่องจากสมาชิกในตำแหน่งที่ 7, 8 และ 10 มีความน่าจะเป็นสูงสุดแล้วคือ 1 ก็คงไว้ตามเดิม

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

7. Poor – Good = {3, 11, 13} ดังนั้น ลดความน่าจะเป็นของสมาชิกในเวกเตอร์ความน่าจะเป็นที่สอดคล้องกับฮีลิกหมายเลขเหล่านี้ด้วยอัตราการเรียนรู้ที่กำหนดได้ผลลัพธ์ดังตารางที่ 3.18 เนื่องจากฮีลิกหมายเลข 11 มีความน่าจะเป็นต่ำสุดแล้วคือ 0 ก็คงไว้ตามเดิม

ตารางที่ 3.18 การปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นสำหรับ *EDA-L*

สมาชิกตำแหน่งที่	1	7	8	9	10	12	3	11	13
ความน่าจะเป็นก่อนปรับ	0.98	1.00	1.00	0.03	1.00	0.42	0.19	0.00	0.42
ความน่าจะเป็นหลังปรับ	0.99	1.00	1.00	0.04	1.00	0.43	0.18	0.00	0.41

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

3.2.5 การปรับปรุงข้อมูลในอาไคว์

อ้างอิงจากหลาย ๆ งานวิจัย [16, 63-66] พบว่าโครงสร้างอาร์เอ็นเอที่เป็นคำตอบมักมีค่าพลังงานที่ต่ำแต่อาจไม่ต่ำที่สุด งานวิจัยนี้จึงใช้ประโยชน์จากกลุ่มประชากรของโครงสร้างที่ทำนายได้ในระหว่างกระบวนการวิวัฒนาการเพื่อรองรับการทำนายหลายโครงสร้าง

ขั้นตอนนี้เกี่ยวข้องกับพารามิเตอร์ *N_archive* ซึ่งแทนจำนวนโครงสร้างอาร์เอ็นเอที่เป็นตัวแทนคำตอบเมื่อจบการทำงานของขั้นตอนวิธี Hybrid-EDAfold หลักการคือ ในรุ่นแรกหลังจากผ่านขั้นตอนประเมินค่าความเหมาะสมของประชากรเรียบร้อยแล้วข้อมูลในอาไคว์จะถูกกำหนดค่าเริ่มต้นโดยคัดเลือกเฉพาะโครโมโซมที่มีค่าความเหมาะสมดีสุด *N_archive* ตัวแรก จากนั้นในรุ่นถัด ๆ ไปหลังจากผ่านขั้นตอนการประเมินค่าความเหมาะสมของประชากรแล้วให้ทำการตรวจสอบว่ามีโครโมโซมใดในรุ่นนั้นที่มีค่าความเหมาะสมดีกว่าโครโมโซมที่ถูกจัดเก็บในอาไคว์หรือไม่ ถ้ามีก็ทำการเพิ่มโครโมโซมที่ดีกว่านั้นแทนที่โครโมโซมตัวที่แย่สุดในอาไคว์ กล่าวโดยสรุป ข้อมูลที่เก็บในอาไคว์คือโครโมโซมที่มีค่าความเหมาะสมดีสุดที่ถูกพบในระหว่างกระบวนการวิวัฒนาการคำตอบของขั้นตอนวิธี Hybrid-EDAfold นั่นเอง

3.3 การประเมินค่าความถูกต้องของโครงสร้างที่ทำนายได้

เมื่อสิ้นสุดกระบวนการทำนายโครงสร้างของขั้นตอนวิธีที่งานวิจัยนี้นำเสนอจะได้โครงสร้างทุติยภูมิที่มีค่าพลังงานต่ำสุดและโครงสร้างทุติยภูมิที่มีค่าพลังงานต่ำรองลงมาซึ่งถูกเก็บอยู่ในอาไคว์จำนวน *N_archive* โครงสร้างและถือเป็นตัวแทนคำตอบที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้

ในการประเมินประสิทธิภาพของขั้นตอนวิธีการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอดำเนินการผ่าน 6 ตัวชี้วัด ได้แก่ true positive (TP) , false positive (FP) และ false negative (FN), ค่าความอ่อนไหว (sensitivity) , ค่าความจำเพาะ (specificity) และ F-measure

โดยที่ true positive (TP) แทนจำนวนตำแหน่งคู่เบสที่ทำนายได้ถูกต้องตรงกับตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบ

false positive (FP) แทนจำนวนตำแหน่งคู่เบสที่ทำนายผิด ไม่พบคู่เบสเหล่านั้นในโครงสร้างคำตอบ

false negative (FN) แทนจำนวนตำแหน่งคู่เบสที่ไม่ได้ทำนาย แต่พบคู่เบสเหล่านั้นในโครงสร้างคำตอบ

ค่าความอ่อนไหว คือ สัดส่วนจำนวนตำแหน่งคู่เบสที่ทำนายได้ถูกต้องเทียบกับจำนวนตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบ

ค่าความจำเพาะ คือ สัดส่วนจำนวนตำแหน่งคู่เบสที่ทำนายได้ถูกต้องเทียบกับจำนวนตำแหน่งคู่เบสทั้งหมดที่พบในโครงสร้างที่ทำนาย

F-measure คือ ค่าเฉลี่ยฮาร์โมนิก (harmonic mean) ของค่าความอ่อนไหว และ ค่าความจำเพาะ คำนวณได้ดังสมการที่ 3.4

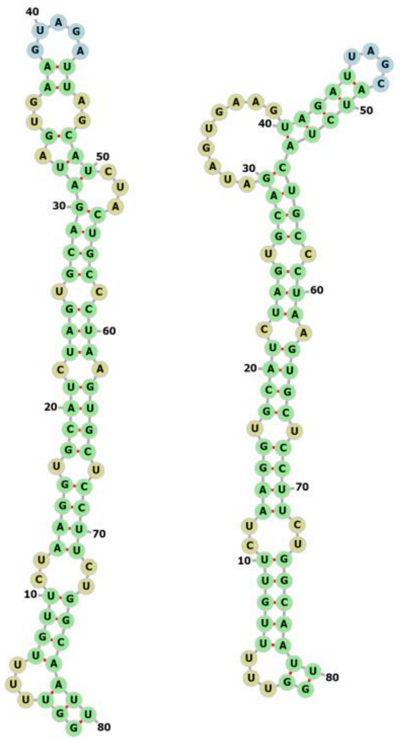
$$F - measure = 2\,x\,\frac{sensitivity\ x\ specificity}{sensitivity + specificity} \quad (3.4)$$

ตัวอย่างการประเมินค่าความถูกต้องของขั้นตอนวิธีการทำนายโครงสร้างแสดงดังรูปที่ 3.10 ซึ่งเป็นข้อมูลอาร์เอ็นเอ pre-miRNA-18 ความยาว 80 นิวคลีโอไทด์ โดยรูป (ก) แสดงโครงสร้างที่เป็นคำตอบ และ รูป (ข) แสดงโครงสร้างที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ ตัวเลขที่ปรากฏในภาพคือตำแหน่งของเบส

จากรูปที่ 3.10 โครงสร้างที่ทำนายได้จากขั้นตอนวิธีที่นำเสนอระบุได้ตำแหน่งคู่เบสจำนวน 27 คู่ และคู่เบสเหล่านั้นมีตำแหน่งตรงกับตำแหน่งคู่เบสที่ปรากฏในโครงสร้างคำตอบทั้งหมด 21 คู่ ดังนั้น TP = 21 เช่น ในโครงสร้างคำตอบตำแหน่งที่ 7 จับคู่กับเบสตำแหน่งที่ 77 และโครงสร้างที่ทำนายได้ก็พบว่าเบสในตำแหน่งดังกล่าวมีการจับคู่กัน ค่า FP = 6 ประเมินจากขั้นตอนวิธีที่นำเสนอทำนายว่าเบสตำแหน่งนี้จับคู่กันแต่ในโครงสร้างคำตอบเบสบริเวณนี้ไม่ได้จับคู่กัน เช่น ทำนายว่าเบส

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตำแหน่งที่ 6 จับคู่กับเบสตำแหน่งที่ 78 แต่ในโครงสร้างคำตอบเบสตำแหน่งที่ 6 เป็นเบสอิสระ และค่า FN = 6 ประเมินจากขั้นตอนวิธีที่นำเสนอไม่ได้ทำนายตำแหน่งคู่เบสบริเวณนั้นแต่ในโครงสร้างคำตอบพบว่าเบสบริเวณนี้มีการจับคู่กัน เช่น ในโครงสร้างคำตอบเบสตำแหน่งที่ 3 จับคู่กับเบสตำแหน่งที่ 78 แต่ขั้นตอนวิธีที่นำเสนอระบุว่าเบสตำแหน่งที่ 3 เป็นเบสอิสระ ความอ่อนไหวมีเท่ากับ 0.78 ประเมินจากโครงสร้างคำตอบมีจำนวนคู่เบสทั้งหมด 27 คู่ และขั้นตอนวิธีที่นำเสนอทำนายตำแหน่งคู่เบสเหล่านั้นได้ถูกต้อง 21 คู่ ความจำเพาะมีเท่ากับ 0.78 ประเมินจากโครงสร้างที่ทำนายได้ระบุจำนวนคู่เบสทั้งหมด 27 คู่และทำนายได้ถูกต้องจำนวน 21 คู่ และ F-measure มีค่าเท่ากับ $2 \times (0.78 \times 0.78) / (0.78 + 0.78) = 0.78$



(A) โครงสร้างที่เป็นคำตอบ (B) โครงสร้างที่ได้จากการทำนาย

รูปที่ 3.10 การเปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบ

2430366845

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บทที่ 4

ผลการวิจัย

ในบทนี้นำเสนอการทดสอบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เริ่มต้นด้วยการศึกษาพารามิเตอร์ของขั้นตอนวิธีที่นำเสนอ เป้าหมายเพื่อทำการศึกษาว่าค่าพารามิเตอร์ที่แตกต่างกันส่งผลต่อค่าความถูกต้องในการทำนายโครงสร้างของวิธีการที่นำเสนออย่างไร และชุดของพารามิเตอร์ใดที่ให้ค่าความถูกต้องในการทำนายโครงสร้างสูงสุด โดยทดสอบกับข้อมูลสายลำดับอาร์เอ็นเอจำนวน 20 รายการที่มีความแตกต่างกันทั้งในแง่ของความยาวและชนิดของอาร์เอ็นเอ รายละเอียดในส่วนนี้นำเสนอในหัวข้อ 4.1 จากนั้นศึกษาประสิทธิภาพการทำนายหลายโครงสร้างของขั้นตอนวิธีที่นำเสนอโดยใช้ชุดของพารามิเตอร์ที่ให้ค่าความถูกต้องสูงสุดจากหัวข้อที่แล้วโดยทดสอบกับข้อมูลสายลำดับอาร์เอ็นเอกลุ่มเดิมเปรียบเทียบกับโปรแกรม Mfold และ RNAstructure ที่รองรับการทำนายหลายโครงสร้างเช่นกัน เป้าหมายเพื่อศึกษาว่าวิธีการทำนายหลายโครงสร้างโดยการเก็บคำตอบไว้ในอาไคว์ที่งานวิจัยนี้นำเสนอมีประสิทธิภาพเทียบเคียงได้กับวิธีการทำนายหลายโครงสร้างที่ถูกใช้ในโปรแกรมอื่น ๆ หรือไม่รายละเอียดในส่วนนี้นำเสนอในหัวข้อ 4.2 จากนั้นทำการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold บนข้อมูลสายลำดับอาร์เอ็นเอจำนวน 3 กลุ่ม ดังนี้ กลุ่มแรกเป็นข้อมูลสายลำดับอาร์เอ็นเอ pre-miRNA ของมนุษย์จำนวน 10 รายการเปรียบเทียบกับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตจำนวน 3 โปรแกรม ได้แก่ Mfold, RNAfold, และ RNAstructure ซึ่งข้อมูลในกลุ่มนี้มีความยาวไม่มากนักและมีรูปร่างโครงสร้างใกล้เคียงกันถือเป็นกลุ่มปัญหาง่ายรายละเอียดในส่วนนี้นำเสนอในหัวข้อ 4.3 กลุ่มที่สองเป็นข้อมูลสายลำดับอาร์เอ็นเอ 20 รายการดังที่เคยนำเสนอไปซึ่งถูกรวบรวมจากวรรณกรรมต่าง ๆ ของขั้นตอนวิธีในกลุ่มเมตาฮิวริสติก โดยขั้นตอนวิธี Hybrid-EDAfold จะถูกเปรียบเทียบกับขั้นตอนวิธีการเมตาฮิวริสติกอื่น ๆ จำนวน 3 วิธี ได้แก่ RnaPredict, SARNA-Predict และ TL-PSO โดยข้อมูลในกลุ่มนี้จะมีความซับซ้อนของโครงสร้างมากกว่าข้อมูลกลุ่มที่หนึ่งเนื่องจากมีความยาวมากกว่าและมาจากอาร์เอ็นเอที่แตกต่างกัน 3 ชนิด เป้าหมายเพื่อศึกษาว่าขั้นตอนวิธีที่นำเสนอมีประสิทธิภาพอย่างไรเมื่อดำเนินการกับข้อมูลสายลำดับอาร์เอ็นเอที่มีความซับซ้อนของโครงสร้างมากยิ่งขึ้นรายละเอียดในส่วนนี้นำเสนอในหัวข้อ 4.4 และข้อมูลกลุ่มที่สามเป็นข้อมูลที่รวบรวมจากฐานข้อมูล RNA STARND 2.0 [33] ซึ่งประกอบด้วยอาร์เอ็นเอทั้งหมด 14 ชนิด โดยคัดเลือกข้อมูลอาร์เอ็นเอที่มีความยาวแตกต่างกันจำนวน 750 รายการ เป้าหมายเพื่อศึกษาประสิทธิภาพของขั้นตอนวิธีที่นำเสนอเมื่อต้องดำเนินการกับข้อมูลที่มีความหลากหลายค่อนข้างมากทั้งในแง่ของความยาวและชนิดของอาร์เอ็นเอรายละเอียดในส่วนนี้นำเสนอในหัวข้อ 4.5 และสรุปสิ่งที่ได้จากการศึกษาประสิทธิภาพของขั้นตอนวิธีที่นำเสนออธิบายในหัวข้อ 4.6

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

4.1 ประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เมื่อกำหนดค่าพารามิเตอร์แตกต่างกัน

ขั้นตอนวิธี Hybrid-EDAfold มีพารามิเตอร์ที่เกี่ยวข้องจำนวน 8 รายการ ได้แก่ ขนาดประชากร จำนวนรอบของการวิวัฒนาการ ความยาวโครโมโซม จำนวนคำตอบที่เก็บในอาไคว์ สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่าเป็นโครโมโซมดีในกลุ่มประชากร สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่าเป็นโครโมโซมด้อยในกลุ่มประชากร สัดส่วนจำนวนฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษในขั้นตอนการสร้างประชากรสำหรับ *EDA-L* และ อัตราการเรียนรู้ โดยในที่นี้เลือกทำการศึกษา 2 พารามิเตอร์ ได้แก่ ขนาดประชากรและจำนวนรอบการวิวัฒนาการซึ่งจะทดสอบขั้นตอนวิธีที่นำเสนอโดยใช้ค่าพารามิเตอร์ที่แตกต่างกัน ในขณะที่อีก 6 พารามิเตอร์ที่เหลือทำการกำหนดค่าเป็นดังนี้

- ความยาวโครโมโซม :	ความยาวสายลำดับอาร์เอ็นเอที่เป็นข้อมูลนำเข้า/15
- จำนวนคำตอบที่เก็บในอาไคว์	20
- สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่าเป็นโครโมโซมดีในประชากร	20%
- สัดส่วนจำนวนโครโมโซมที่ถูกพิจารณาว่าเป็นโครโมโซมด้อยในประชากร	20%
- สัดส่วนจำนวนฮีลิกที่ถูกลบทิ้งจากโครโมโซมบรรพบุรุษสำหรับ *EDA-L*	50%
- อัตราการเรียนรู้	0.001

การทดลองในส่วนนี้ทำการศึกษาการกำหนดค่าพารามิเตอร์ของขั้นตอนวิธี Hybrid-EDAfold โดยเลือกพารามิเตอร์ 2 รายการ คือ ขนาดประชากรซึ่งกำหนดค่าแตกต่างกัน ดังนี้ 50, 100, 200 และจำนวนรอบของการวิวัฒนาการซึ่งกำหนดค่าแตกต่างกัน ดังนี้ 100, 200, 500 โดยทำการประเมินในทุกรูปแบบที่เป็นไปได้จะได้พารามิเตอร์ทั้งหมด 9 ชุด เป้าหมายเพื่อศึกษาว่าพารามิเตอร์ดังกล่าวส่งผลต่อการทำนายโครงสร้างของขั้นตอนวิธีที่งานวิจัยนี้นำเสนออย่างไร โดยเลือกทำการทดสอบบนข้อมูลสายลำดับอาร์เอ็นเอจำนวน 20 สายที่รวบรวมจากวรรณกรรมของวิธีการทางเมตาฮิวริสติกต่าง ๆ [12 - 14] และเป็นข้อมูลจากฐานข้อมูล RNA STRAND v2.0 เนื่องจากมีความหลากหลายในแง่ของความยาวและชนิดของอาร์เอ็นเอ รายละเอียดของข้อมูลที่นำมาทดสอบแสดงดังตารางที่ 4.1 และ ผลลัพธ์การประเมินประสิทธิภาพแสดงดังตารางที่ 4.2

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.1 คุณลักษณะของ 20 สายลำดับอาร์เอ็นเอ

ลำดับ	ชื่อโมเลกุล	รหัสโมเลกุล	ชนิดอาร์เอ็นเอ	ความยาว	จำนวนคู่เบสที่พบในโครงสร้างคำตอบ
1	d.5.b.G.stearothermophilus.2	CRW_00557	5S rRNA	117	38
2	d.5.e.S.cerevisiae	CRW_00570	5S rRNA	118	37
3	d.5.b.E.coli	CRW_01516	5S rRNA	120	40
4	d.5.a.H.marismortui	CRW_00548	5S rRNA	122	38
5	d.5.b.T.aquaticus	CRW_00567	5S rRNA	123	40
6	d.5.b.D.radiodurans.rrnB	CRW_00555	5S rRNA	124	40
7	b.I1.e.M.anisopliae.3.C1.LSU.1921	CRW_00016	Group I Intron	394	120
8	b.I1.e.C.saccharophila.C1.SSU.156	CRW_00010	Group I Intron	454	126
9	b.I1.e.M.anisopliae.2.C1.LSU.1921	CRW_00013	Group I Intron	456	115
10	b.I1.e.A.lagunensis.C1.SSU.516	CRW_00006	Group I Intron	468	113
11	b.I1.e.H.rubra.1.C1.SSU.1506	CRW_00012	Group I Intron	543	141
12	b.I1.e.A.griffini.1.C1.SSU.516	CRW_00004	Group I Intron	556	131
13	b.I1.e.P.leucosticta.C1.SSU.516	CRW_00018	Group I Intron	605	121
14	d.16.m.C.elegans	CRW_00423	16S rRNA	697	189
15	d.16.m.D.virilis	CRW_00429	16S rRNA	784	233
16	d.16.m.A.cahirinus	CRW_00418	16S rRNA	940	260
17	d.16.m.X.laevis	CRW_00463	16S rRNA	945	254
18	d.16.m.H.sapiens.5	CRW_00438	16S rRNA	954	268
19	d.16.m.A.fulgens	CRW_00419	16S rRNA	964	265
20	d.16.a.S.acidocaldarius	CRW_00039	16S rRNA	1495	468

จากตารางที่ 4.1 คอลัมน์ที่ 2 แสดงชื่อโมเลกุลอาร์เอ็นเอ คอลัมน์ที่ 3 แสดงรหัสโมเลกุลที่ใช้อ้างอิงในฐานข้อมูล RNA STARND v2.0 คอลัมน์ที่ 4 แสดงชนิดของอาร์เอ็นเอ คอลัมน์ที่ 5 แสดงความยาวของสายลำดับอาร์เอ็นเอ และ คอลัมน์ที่ 6 แสดงจำนวนคู่เบสที่พบในโครงสร้างที่เป็นคำตอบของสายลำดับนั้นซึ่งเป็นข้อมูลที่ต้องการทำนายให้ถูกต้องมากที่สุด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.2 การทดสอบพารามิเตอร์ของขั้นตอนวิธี Hybrid-EDAfold บน 20 อาร์เอ็นเอ

ลำดับ	ความยาว	จำนวนคู่เบสเฉลย	ตัวชี้วัด	ค่าพารามิเตอร์ที่ทำการทดสอบ (ขนาดประชากร x จำนวนรอบ)								
				50 x 100	50 x 200	50 x 500	100 x 100	100 x 200	100 x 500	200 x 100	200 x 200	200 x 500
1	117	38	*Predict*	**37**	39	39	39	39	39	39	39	36
			TP	**32**	30	30	30	30	30	30	30	29
			F-measure	**85.3**	77.9	77.9	77.9	77.9	77.9	77.9	77.9	78.4
2	118	37	*Predict*	**34**	34	34	34	34	34	34	34	34
			TP	**33**	33	33	33	33	33	33	33	33
			F-measure	**93.0**	93.0	93.0	93.0	93.0	93.0	93.0	93.0	93.0
3	120	40	*Predict*	**38**	38	38	38	38	38	38	38	38
			TP	**35**	35	35	35	35	33	35	33	35
			F-measure	**89.7**	89.7	89.7	89.7	89.7	84.6	89.7	84.6	89.7
4	122	38	*Predict*	**38**	38	38	38	38	34	38	38	34
			TP	**33**	33	33	33	33	31	33	33	31
			F-measure	**86.8**	86.8	86.8	86.8	86.8	86.1	86.8	86.8	86.1
5	123	40	*Predict*	**40**	40	40	40	40	40	40	40	40
			TP	**37**	37	37	37	37	37	37	37	37
			F-measure	**92.5**	92.5	92.5	92.5	92.5	92.5	92.5	92.5	92.5
6	124	40	*Predict*	**36**	36	36	36	36	36	36	36	36
			TP	**35**	35	35	35	35	35	35	35	35
			F-measure	**92.1**	92.1	92.1	92.1	92.1	92.1	92.1	92.1	92.1
7	394	120	*Predict*	119	121	118	114	**112**	121	117	122	122
			TP	98	95	98	97	**98**	96	96	97	98
			F-measure	82.0	78.8	82.4	82.9	**84.5**	79.7	81.0	80.2	81.0
8	454	126	*Predict*	**130**	**125**	129	128	129	128	127	133	132
			TP	**108**	**106**	106	107	107	107	107	109	107
			F-measure	**84.4**	**84.5**	83.1	84.3	83.9	84.3	84.6	84.2	82.6
9	456	115	*Predict*	**136**	132	133	132	134	130	134	131	134
			TP	**57**	53	53	54	52	55	55	55	52
			F-measure	**45.4**	42.9	42.7	43.7	41.8	44.9	44.2	44.7	41.8
10	468	113	*Predict*	**128**	133	130	132	134	134	133	135	128
			TP	**75**	75	75	75	75	75	75	75	72
			F-measure	**62.2**	61.0	61.7	61.2	60.7	60.7	61.0	60.5	59.8
11	543	141	*Predict*	**158**	176	168	165	167	164	173	158	157
			TP	**107**	103	105	105	108	99	103	101	99
			F-measure	**71.6**	65.0	68.0	68.6	70.1	64.9	65.6	67.6	66.4

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ลำดับ	ความยาว	จำนวนคู่เบสเฉลี่ย	ตัวชี้วัด	ค่าพารามิเตอร์ที่ทำการทดสอบ (ขนาดประชากร x จำนวนรอบ)								
				50 x 100	50 x 200	50 x 500	100 x 100	100 x 200	100 x 500	200 x 100	200 x 200	200 x 500
12	556	131	*Predict*	**176**	169	170	173	170	168	170	172	170
			TP	**95**	89	91	91	89	89	91	92	89
			F-measure	**61.9**	59.3	60.5	59.9	59.1	59.5	60.5	60.7	59.1
13	605	121	*Predict*	**172**	171	176	174	170	169	169	174	175
			TP	**80**	79	80	79	79	79	79	79	79
			F-measure	**54.6**	54.1	53.9	53.6	54.3	54.5	54.5	53.6	53.4
14	697	189	*Predict*	218	200	211	**192**	202	201	205	217	220
			TP	57	55	51	**55**	52	53	53	54	53
			F-measure	28.0	28.3	25.5	**28.9**	26.6	27.2	26.9	26.6	25.9
15	784	233	*Predict*	246	229	240	236	223	**230**	243	241	241
			TP	68	57	62	60	60	**66**	59	59	59
			F-measure	28.4	24.7	26.2	25.6	26.3	**28.5**	24.8	24.9	24.9
16	940	260	*Predict*	**262**	268	264	266	260	263	263	258	265
			TP	**63**	61	62	59	60	61	59	59	58
			F-measure	**24.1**	23.1	23.7	22.4	23.1	23.3	22.6	22.8	22.1
17	945	254	*Predict*	**278**	256	275	272	276	269	274	263	275
			TP	**124**	111	113	118	117	116	117	110	111
			F-measure	**46.6**	43.5	42.7	44.9	44.2	44.4	44.3	42.6	42.0
18	954	268	*Predict*	258	253	246	255	**248**	257	254	252	251
			TP	96	96	100	98	**107**	96	97	99	95
			F-measure	36.5	36.9	38.9	37.5	**41.5**	36.6	37.2	38.1	36.6
19	964	265	*Predict*	277	286	268	267	278	261	276	277	**266**
			TP	81	86	78	84	82	80	79	77	**89**
			F-measure	29.9	31.2	29.3	31.6	30.2	30.4	29.2	28.4	**33.5**
20	1495	468	*Predict*	486	487	478	487	485	478	484	487	**481**
			TP	271	277	273	271	273	271	272	278	**277**
			F-measure	56.8	58.0	57.7	56.8	57.3	57.3	57.1	58.2	**58.4**
เฉลี่ย	549	152	*Predict*	**163**	162	162	161	161	160	162	162	162
			TP	**79**	77	78	78	78	77	77	77	77
			F-measure	**62.6**	61.2	61.4	61.7	61.8	61.1	61.3	61.0	61.0

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากตารางที่ 4.2 คอลัมน์ที่ 2 แสดงความยาวของสายลำดับอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3 แสดงจำนวนคู่เบสที่พบในโครงสร้างที่เป็นคำตอบ คอลัมน์ที่ 4 แสดงตัวชี้วัดที่ทำการประเมิน โดยที่ *Predict* แทนจำนวนคู่เบสที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้เมื่อทดสอบด้วยพารามิเตอร์แต่ละชุด *TP* แทนจำนวนคู่เบสที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ถูกต้องตรงกับโครงสร้างคำตอบ และ *F-measure* แทนค่าความถูกต้องของผลการทำนายโครงสร้าง คอลัมน์ที่ 5 – 13 แสดงผลลัพธ์ที่ได้เมื่อทดสอบด้วยพารามิเตอร์แต่ละชุดในแต่ละตัวชี้วัด และบริเวณที่แรเงาในตารางแทนชุดของพารามิเตอร์ที่ให้ผลการทำนายดีสุดในแต่ละข้อมูลที่นำมาทดสอบ

ผลลัพธ์จากตารางที่ 4.2 พบว่าสำหรับอาร์เอ็นเอที่มีความยาวไม่มากนักในที่นี้คือข้อมูลลำดับที่ 1 – 6 ซึ่งเป็นข้อมูลจากกลุ่มของ 5S Ribosomal RNA ชุดของพารามิเตอร์ที่แตกต่างกันไม่ส่งผลต่อค่าความถูกต้องของการทำนายทั้งในส่วนของ *TP* และ *F-measure* ยกเว้นข้อมูลในลำดับที่ 1 ที่เมื่อขนาดของประชากรและจำนวนรอบในการวิวัฒนาการมากขึ้นแล้วทำให้ค่าความถูกต้องในการทำนายโครงสร้างลดลง และชุดของพารามิเตอร์ที่ให้ค่าความถูกต้องมากสุดสำหรับข้อมูลในกลุ่มนี้ คือ ขนาดประชากรที่ 50 และ จำนวนรอบการวิวัฒนาการที่ 100

สำหรับข้อมูลลำดับที่ 7 – 13 ซึ่งเป็นข้อมูลจากกลุ่มของ Group I Intron ผลลัพธ์ที่ได้ก็เป็นไปในทิศทางเดียวกัน ชุดของพารามิเตอร์ที่ให้ค่าความถูกต้องสูงสุดส่วนใหญ่คือ ขนาดประชากรที่ 50 และ จำนวนรอบการวิวัฒนาการที่ 100 ในภาพรวมค่าพารามิเตอร์ที่เปลี่ยนแปลงไปส่งผลให้ค่าความถูกต้องในการทำนายแตกต่างกันเล็กน้อยและมีแนวโน้มว่าขนาดประชากรหรือจำนวนรอบการวิวัฒนาการที่มากเกินไปอาจทำให้ค่าความถูกต้องลดลงสังเกตจากชุดพารามิเตอร์ที่มีขนาดประชากรเป็น 200 และ จำนวนการวิวัฒนาการเป็น 500

สำหรับข้อมูลลำดับที่ 14 – 20 ซึ่งเป็นข้อมูลจากกลุ่มของ 16S Ribosomal RNA พบว่าชุดพารามิเตอร์ที่ให้ค่าความถูกต้องสูงสุดแตกต่างกันไปในแต่ละข้อมูลที่นำมาทดสอบ และข้อมูลในกลุ่มนี้ค่อนข้างมีความอ่อนไหวต่อค่าพารามิเตอร์ที่เปลี่ยนแปลงไปมากกว่าข้อมูลอีก 2 ชุดที่ได้ทำการวิเคราะห์ไป โดยสรุป ค่าเฉลี่ยจาก 20 สายลำดับทั้งในส่วนของ *TP* และ *F-measure* พบว่าชุดพารามิเตอร์ที่ให้ค่าความถูกต้องสูงสุด คือ ขนาดประชากร 50 และ จำนวนรอบการวิวัฒนาการ 100 ผู้วิจัยจึงเลือกใช้พารามิเตอร์ชุดนี้ในการทดสอบขั้นตอนวิธี Hybrid-EDAfold กับข้อมูลอาร์เอ็นเอต่าง ๆ ทั้งจากฐานข้อมูล RNA STRAND v2.0 และ pre-miRNA ของมนุษย์เปรียบเทียบกับขั้นตอนวิธีอื่น ๆ ทั้งในกลุ่มกำหนดการพลวัตและกลุ่มวิธีทางเมตาฮิวริสติกดังที่จะนำเสนอในหัวข้อถัดไป

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

4.2 เปรียบเทียบประสิทธิภาพของวิธีการทำนายหลายโครงสร้างที่งานวิจัยนี้นำเสนอกับวิธีการที่ใช้ในโปรแกรมอื่น ๆ

ในหัวข้อนี้ทำประเมินประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold ในส่วนของการรองรับการทำนายหลายโครงสร้าง โดยวิธีที่งานวิจัยนี้นำเสนอคือเก็บคำตอบที่มีค่า free energy ต่ำสุดที่พบในระหว่างกระบวนการวิวัฒนาการไว้ในอาไคว์จำนวน *n* คำตอบ โดยทำการเปรียบเทียบกับวิธีการอื่น ๆ ที่มีการรองรับการทำนายหลายโครงสร้างเช่นกัน ได้แก่ Mfold [4] และ RNAstructure [60] โดยเลือกทดสอบกับข้อมูล 20 สายลำดับอาร์เอ็นเอดังที่ได้นำเสนอไปในหัวข้อก่อนหน้า รายละเอียดของข้อมูลที่นำมาทดสอบได้นำเสนอไปแล้วในตารางที่ 4.1

ผลลัพธ์จากโปรแกรม Mfold คำนวณจาก http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form โดยใช้ค่าพารามิเตอร์เริ่มต้นและกำหนดพารามิเตอร์ที่ควบคุมจำนวนโครงสร้างสูงสุดที่โปรแกรมทำนายได้ไว้ที่ 20 โครงสร้าง

ผลลัพธ์จากโปรแกรม RNAstructure คำนวณจาก https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html โดยใช้ค่าพารามิเตอร์เริ่มต้น

ผลลัพธ์จากขั้นตอนวิธี Hybrid-EDAfold ในการทดสอบกับทุกสายลำดับอาร์เอ็นเอใช้พารามิเตอร์ชุดเดียวกันดังรายละเอียดที่ได้นำเสนอไปแล้วในหัวข้อ 4.1 แต่ละสายลำดับถูกรันจำนวน 30 ครั้ง และรายงานผลการรันครั้งที่ให้ค่า F-measure สูงสุด

การเปรียบเทียบผลการทำนายโครงสร้างเมื่อแต่ละขั้นตอนวิธีรายงานผลการทำนายเฉพาะโครงสร้างที่มีค่า free energy ต่ำสุด แสดงดังตารางที่ 4.3 เพื่อความยุติธรรมในการเปรียบเทียบโครงสร้างที่แต่ละขั้นตอนวิธีทำนายได้จะถูกนำไปคำนวณค่าพลังงานด้วยโปรแกรม RNAeval [6] โดยคอลัมน์ที่ 2 แสดงความยาวของแต่ละสายลำดับอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 4 – 6 แสดงค่า free energy ของโครงสร้างคำตอบ และ โครงสร้างที่ทำนายได้จากโปรแกรม Mfold, RNAstructure และ Hybrid-EDAfold ตามลำดับ และคอลัมน์ที่ 7-9 แสดงผลต่างค่า free energy ของโครงสร้างที่ทำนายได้จากแต่ละขั้นตอนวิธีเปรียบเทียบกับค่า free energy ของโครงสร้างคำตอบของแต่ละอาร์เอ็นเอ (คำนวณจาก free energy$_{\text{โครงสร้างคำตอบ}}$ – free energy$_{\text{ทำนายได้จากโปรแกรม}}$) ถ้าค่าผลต่างเป็น 0 หมายความว่าโปรแกรมทำนายโครงสร้างได้ค่า free energy เท่ากับโครงสร้างคำตอบ ถ้าค่าผลต่างเป็นบวกหมายความว่าโปรแกรมทำนายโครงสร้างได้ค่า free energy ต่ำกว่าโครงสร้างคำตอบ และ ถ้าค่าผลต่างเป็นลบหมายความว่าโปรแกรมทำนายโครงสร้างได้ค่า free energy สูงกว่าโครงสร้างคำตอบ บริเวณที่มีการแรเงาแสดงขั้นตอนวิธีที่ทำนายโครงสร้างได้ค่า free energy ใกล้เคียงกับโครงสร้างคำตอบมากที่สุดสำหรับแต่ละอาร์เอ็นเอ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.3 เปรียบเทียบค่า free energy ของโครงสร้างที่ทำนายได้จากโปรแกรม Mfold, RNAstructure และ Hybrid-EDAfold กับโครงสร้างคำตอบบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ

ลำดับ	ความยาว	free energy				ผลต่าง free energy		
		known structure	Mfold	RNA structure	Hybrid-EDA	Mfold	RNA structure	Hybrid-EDA
1	117	-41.5	-40.7	-47.4	-46.5	**-0.8**	5.9	5.0
2	118	-41.5	-48.0	-48.2	-44.7	6.5	6.7	**3.2**
3	120	-47.8	-50.7	-50.5	-47.6	2.9	2.7	**-0.2**
4	122	-48.7	-48.8	-53.5	-53.3	**0.1**	4.8	4.6
5	123	-52.6	-46.6	-57.5	-56.6	-6.0	4.9	**4.0**
6	124	-49.2	-43.8	-43.7	-45.7	-5.4	-5.5	**-3.5**
7	394	-100.8	-118.4	-125.9	-122.9	**17.6**	25.1	22.1
8	454	-157.1	-181.8	-185.7	-175.9	24.7	28.6	**18.8**
9	456	-92.8	-148.3	-149.3	-133.8	55.5	56.5	**41.0**
10	468	-86.1	-125.4	-132.7	-125.3	39.3	46.6	**39.2**
11	543	-142.0	-187.1	-195.0	-178.3	45.1	53.0	**36.3**
12	556	-110.1	-171.6	-177.3	-167.9	61.5	67.2	**57.8**
13	605	-116.6	-212.4	-220.9	-206.1	95.8	104.3	**89.5**
14	697	3.2	-118.6	-121.8	-64.5	121.8	125.0	**67.7**
15	784	-8.6	-125.9	-132.0	-42.5	117.3	123.4	**33.9**
16	940	-89.5	-174.9	-186.5	-137.0	85.4	97.0	**47.5**
17	945	-131.3	-216.9	-227.5	-147.5	85.6	96.2	**16.2**
18	954	-113.8	-210.4	-219.2	-145.8	96.6	105.4	**32.0**
19	964	-95.2	-183.8	-192.5	-101.3	88.6	97.3	**6.1**
20	1495	-606.5	-757.6	-774.8	-719.1	151.1	168.3	**112.6**

จากตารางที่ 4.3 แสดงให้เห็นว่า Hybrid-EDAfold เป็นขั้นตอนวิธีที่ทำนายโครงสร้างส่วนใหญ่ได้ค่า free energy ใกล้เคียงกับโครงสร้างคำตอบ มีเพียง 3 อาร์เอ็นเอ คือ อาร์เอ็นเอลำดับที่ 1, 4 และ 7 ที่โปรแกรม Mfold ทำนายได้โครงสร้างที่มีค่า free energy ใกล้เคียงกับโครงสร้างคำตอบมากกว่า

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากนั้นเมื่อนำโครงสร้างที่มีค่า free energy ต่ำสุดเหล่านี้ที่แต่ละโปรแกรมทำนายได้ไปเปรียบเทียบกับโครงสร้างคำตอบได้ผลลัพธ์แสดงดังตารางที่ 4.4 โดยคอลัมน์ที่ 2 แสดงความยาวของสายลำดับอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3 แสดงจำนวนคู่เบสที่พบในโครงสร้างคำตอบ คอลัมน์ที่ 4 - 6 แสดงค่า F-measure ของโปรแกรม Mfold, RNAstructure และ Hybrid-EDAfold ตามลำดับ พื้นที่ที่แรเงาแสดงขั้นตอนวิธีที่ได้ค่า F-measure สูงสุดสำหรับแต่ละสายลำดับอาร์เอ็นเอที่นำมาทดสอบ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.4 การเปรียบเทียบโครงสร้างที่มีค่า free energy ต่ำสุดที่ทำนายด้วยโปรแกรม Mfold, RNAstructure และ Hybrid-EDA กับโครงสร้างคำตอบบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ

ลำดับ	ความยาว	จำนวนคู่เบสในโครงสร้าง	F-measure		
			Mfold	RNAstructure	Hybrid-EDAfold
1	117	38	69.5	70.1	64.1
2	118	37	84.6	70	82.2
3	120	40	25.3	25	24.7
4	122	38	80.6	72	81.6
5	123	40	46	69.1	81.5
6	124	40	82.7	24	81.6
7	394	120	62.5	70	82
8	454	126	73.9	59.6	77.9
9	456	115	15.4	27.9	45.4
10	468	113	50.9	53.8	49.18
11	543	141	47.3	44.7	57.5
12	556	131	40.8	49.7	46.6
13	605	121	45.8	46.8	42
14	697	189	9.9	14	16.4
15	784	233	15.9	**16.8**	15.29
16	940	260	15.5	15.1	**19.0**
17	945	254	30.5	36.4	**39.3**
18	954	268	**35.6**	19.9	28.9
19	964	265	**21.7**	16.5	20.6
20	1495	468	49.4	48.7	**49.8**
เฉลี่ย	549	152	45.2	42.5	**50.3**

จากตารางที่ 4.4 ผลลัพธ์ที่ได้สอดคล้องกับหัวข้อก่อนหน้าคือ ขั้นตอนวิธี Hybrid-EDA ทำนายได้โครงสร้างส่วนใหญ่มีค่า free energy ใกล้เคียงกับโครงสร้างคำตอบมากที่สุด ดังนั้นเมื่อนำโครงสร้างที่ทำนายได้เหล่านี้ไปเปรียบเทียบกับโครงสร้างคำตอบจึงได้ค่า F-measure เฉลี่ยสูงกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบ โดยมีค่า F-measure เฉลี่ยสูงกว่าโปรแกรม Mfold ประมาณ 5% และ มีค่า F-measure เฉลี่ยสูงกว่าโปรแกรม RNAstructure ประมาณ 8%

ผลการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างเมื่อทั้ง 3 ขั้นตอนวิธีที่นำมาเปรียบเทียบให้ผลการทำนายในแต่ละสายลำดับอาร์เอ็นเอเป็นชุดของโครงสร้าง เรียกว่า suboptimal structures การเปรียบเทียบดำเนินการดังนี้ ในแต่ละสายลำดับแต่ละวิธีที่นำมาเปรียบเทียบจะให้ผลการทำนายโครงสร้างเป็นจำนวนที่แตกต่างกันออกไป ในที่นี้จำกัดจำนวนโครงสร้างที่ทำนายได้สูงสุดของทุกขั้นตอนวิธีไว้ที่ 20 โครงสร้าง จากนั้นนำทุกโครงสร้างที่แต่ละขั้นตอนวิธีทำนายได้ในแต่ละข้อมูลอาร์เอ็นเอไปเปรียบเทียบกับโครงสร้างที่เป็นคำตอบของอาร์เอ็นเอนั้น และรายงานผลการเปรียบเทียบเฉพาะโครงสร้างที่มีค่า F-measure สูงสุดที่แต่ละขั้นตอนวิธีทำนายได้ ผลลัพธ์แสดงดังตารางที่ 4.5

จากตารางที่ 4.5 คอลัมน์ที่ 2 แสดงความยาวของแต่ละสายลำดับอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3-5 แสดงค่า F-measure สูงสุดจากกลุ่มของโครงสร้างที่แต่ละขั้นตอนวิธีทำนายได้ และคอลัมน์ที่ 6 – 8 แสดงค่า F-measure ที่เพิ่มขึ้น เปรียบเทียบระหว่างการทำนายแค่ 1 โครงสร้างที่มีค่า free energy ต่ำสุดกับการทำนายหลายโครงสร้าง พื้นที่แรเงาแสดงขั้นตอนวิธีที่ได้ค่า F-measure สูงสุดสำหรับแต่ละสายลำดับอาร์เอ็นเอที่นำมาทดสอบ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.5 ประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Mfold, RNAstructure และ Hybrid-EDAfold เมื่อแต่ละขั้นตอนวิธีรองรับการทำนายหลายโครงสร้างบนชุดข้อมูลอาร์เอ็นเอ 20 รายการ

ลำดับ	ความยาว	F-measure			ค่า F-measure ที่เพิ่มขึ้น		
		Mfold	RNA structure	Hybrid-EDA	Mfold	RNA structure	Hybrid-EDA
1	117	69.5	80.5	**85.3**	0.0	10.4	21.2
2	118	84.6	84.2	**93.0**	0.0	14.2	10.8
3	120	25.6	75.3	**89.7**	0.3	50.3	65.0
4	122	80.6	81.6	**86.8**	0.0	9.6	5.2
5	123	66.7	90.2	**92.5**	20.7	21.2	11.0
6	124	82.7	81.6	**92.1**	0.0	57.6	10.5
7	394	78.0	74.7	**82.0**	15.5	4.7	0.0
8	454	73.9	74.1	**84.4**	0.0	14.5	6.5
9	456	41.9	32.8	**45.4**	26.6	4.9	0.0
10	468	**62.5**	53.8	62.2	11.6	0.0	13.0
11	543	53.9	56.9	**71.5**	6.6	12.2	14.0
12	556	62.3	**71.4**	61.9	21.5	21.8	15.3
13	605	47.6	50.0	**54.6**	1.7	3.2	12.6
14	697	19.9	20.1	**28.0**	10.0	6.1	11.6
15	784	**33.8**	20.4	26.1	17.9	3.5	10.8
16	940	**38.2**	16.7	24.1	22.7	1.6	5.1
17	945	40.9	38.2	**45.3**	10.4	1.8	6.0
18	954	36.9	**49.3**	37.1	1.3	29.5	8.2
19	964	29.3	**36.5**	29.9	7.6	20.0	9.3
20	1495	**57.0**	55.4	56.8	7.6	6.6	7.0
เฉลี่ย	549	54.3	57.2	**62.4**	9.1	14.7	12.2

ผลลัพธ์จากตารางที่ 4.5 ยังคงสอดคล้องกับผลการประเมินประสิทธิภาพที่ได้นำเสนอไปในข้างต้น กล่าวคือ ขั้นตอนวิธี Hybrid-EDAfold ยังเป็นขั้นตอนวิธีที่ให้ผลการทำนายส่วนใหญ่ดีกว่าวิธีการอื่นๆ ที่นำมาเปรียบเทียบโดยมีค่า F-measure เฉลี่ยคือ 62.4 ซึ่งสูงกว่า F-measure เฉลี่ยจากการทำนายแค่ 1 โครงสร้างที่มีค่า free energy ต่ำสุดประมาณ 12% รองลงมาเป็นโปรแกรม RNAstructure เมื่อรองรับการทำนายหลายโครงสร้างได้ค่า F-measure เฉลี่ย 57.2 ซึ่งสูงกว่าค่า F-measure เฉลี่ยกรณีที่ทำนายแค่ 1 โครงสร้างประมาณ 15% และ โปรแกม Mfold เมื่อรองรับ

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

การทำนายหลายโครงสร้างได้ค่า F-measure เฉลี่ยคือ 54.3 ซึ่งสูงกว่า F-measure เฉลี่ยกรณีทำนายแค่ 1 โครงสร้างประมาณ 9%

การประเมินประสิทธิภาพในหัวข้อนี้แสดงให้เห็นว่า เมื่อแต่ละขั้นตอนวิธีที่นำมาเปรียบเทียบให้ผลการทำนายเป็นชุดของโครงสร้างในลักษณะของ suboptimal structures ค่า F-measure ที่ได้มีแนวโน้มเพิ่มสูงขึ้นกว่าการทำนายแค่ 1 โครงสร้างที่มีค่า free energy ต่ำสุด ดังนั้น ข้อสรุปในเบื้องต้นคือการทำนายหลายโครงสร้างสามารถช่วยเพิ่มโอกาสให้ขั้นตอนวิธีต่าง ๆ พบโครงสร้างที่ใกล้เคียงกับโครงสร้างคำตอบมากยิ่งขึ้น และสามารถบรรเทาข้อผิดพลาดอันเกิดจากความไม่สมบูรณ์ของพารามิเตอร์ที่ใช้ในแบบจำลองการคำนวณค่าพลังงานได้

นอกจากนี้ วิธีการทำนายหลายโครงสร้างที่งานวิจัยนี้นำเสนอในลักษณะของการเก็บคำตอบที่มีค่าความเหมาะสมดีสุดที่พบในระหว่างกระบวนการค้นหาคำตอบจำนวน *n* โครโมโซมไว้ในอาไคว์ เมื่อ *n* คือพารามิเตอร์ที่กำหนดโดยผู้ใช้ ให้ผลการทำนายโครงสร้างที่ดีเมื่อเทียบกับวิธีการทำนายหลายโครงสร้างแบบที่ใช้ในโปรแกรม Mfold และ RNAstructure ซึ่งดำเนินการในลักษณะของการกำหนดพารามิเตอร์โดยผู้ใช้เช่นเดียวกัน แต่เป็นพารามิเตอร์ของเปอร์เซ็นค่าพลังงานที่เพิ่มขึ้นเมื่อเทียบกับโครงสร้างที่มีค่าพลังงานต่ำสุด นอกจากนี้ ในบางข้อมูลที่ขั้นตอนวิธี Hybrid-EDAfold ให้ผลการทำนายโครงสร้างยังไม่ดีนักสามารถปรับปรุงให้ดีขึ้นได้โดยเพิ่มจำนวนโครงสร้างที่ถูกเก็บในอาไคว์ให้สูงขึ้นซึ่งจะทำให้พบโครงสร้างที่มีความใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากยิ่งขึ้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

4.3 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตบนข้อมูล pre-miRNA ของมนุษย์จำนวน 10 รายการ

หัวข้อนี้นำเสนอประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold เปรียบเทียบกับโปรแกรมสำหรับทำนายโครงสร้างทุติยภูมิที่ได้รับความนิยมใช้งานกันอยู่ในปัจจุบันซึ่งอยู่บนพื้นฐานของกำหนดการพลวัต ได้แก่ Mfold, RNAfold และ RNAstructure ด้วยข้อมูล pre-miRNA ของมนุษย์จำนวน 10 สายลำดับอาร์เอ็นเอซึ่งถูกรวบรวมจากการทดลอง (experimental method) และนำเสนอใน [34] รายละเอียดของข้อมูลในส่วนนี้แสดงดังตารางที่ 4.6

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.6 คุณลักษณะของสายลำดับ pre-miRNA ของมนุษย์

ลำดับ	ชื่อ	ความยาว	จำนวนคู่เบสที่พบในโครงสร้างคำตอบ
1	pre-let-7c	85	30
2	pre-let-7f-2	87	37
3	pre-miR-15a	87	30
4	pre-miR-16-1	91	30
5	pre-miR-17	86	32
6	pre-miR-18	80	27
7	pre-miR-19a	84	34
8	pre-miR-25	84	29
9	pre-miR-29a	68	26
10	pre-miR-30a	73	30

จากตารางที่ 4.6 คอลัมน์ที่ 2 แสดงชื่อของสายลำดับอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3 แสดงความยาวของสายลำดับอาร์เอ็นเอและคอลัมน์สุดท้ายแสดงจำนวนคู่เบสที่พบในโครงสร้างคำตอบ

ผลลัพธ์การเปรียบเทียบประสิทธิภาพที่ได้แสดงดังตารางที่ 4.7 และรายละเอียดการกำหนดค่าต่าง ๆ ของแต่ละโปรแกรมเป็นดังนี้

ผลลัพธ์จากโปรแกรม RNAfold คำนวณจาก http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi โดยใช้ค่าพารามิเตอร์เริ่มต้น

ผลลัพธ์จากโปรแกรม Mfold คำนวณจาก http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form โดยใช้ค่าพารามิเตอร์เริ่มต้นและกำหนดพารามิเตอร์ในส่วนของจำนวนโครงสร้างสูงสุดที่โปรแกรมทำนายได้คือ 20 โครงสร้าง

ผลลัพธ์จากโปรแกรม RNAstructure คำนวณจาก https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html โดยใช้ค่าพารามิเตอร์เริ่มต้น

ผลลัพธ์จากขั้นตอนวิธี Hybrid-EDAfold ในการทดสอบกับทุกสายลำดับอาร์เอ็นเอใช้พารามิเตอร์ชุดเดียวกัน รายละเอียดของการกำหนดค่าพารามิเตอร์นำเสนอไปแล้วในหัวข้อ 4.1 แต่ละสายลำดับถูกรันจำนวน 30 ครั้ง และรายงานผลการรันครั้งที่ให้ค่า F-measure สูงสุด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.7 การเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold กับวิธีในกลุ่มกำหนดการพลวัตเมื่อทดสอบกับข้อมูล pre-miRNA ของมนุษย์

ลำดับ	ชื่อ	ความยาว	จำนวนคู่เบส	ขั้นตอนวิธี	Predict	TP	Sent.	Spec.	F-measure
1	pre-let-7c	85	30	Mfold	31	**25**	**83.33**	**80.65**	**81.97**
				RNAfold	33	**25**	**83.33**	75.76	79.37
				RNAstructure	31	**25**	**83.33**	**80.65**	**81.97**
				hEDAfold	31	**25**	**83.33**	**80.65**	**81.97**
2	pre-let-7f-2	87	37	Mfold	33	33	89.19	**100.00**	94.29
				RNAfold	36	28	75.68	77.78	76.71
				RNAstructure	36	36	97.30	**100.00**	98.63
				hEDAfold	37	**37**	**100.00**	**100.00**	**100.00**
3	pre-miR-15a	87	30	Mfold	28	27	90.00	96.43	93.10
				RNAfold	25	25	83.33	**100.00**	90.91
				RNAstructure	25	25	83.33	**100.00**	90.91
				hEDAfold	32	**30**	**100.00**	93.75	**96.77**
4	pre-miR-16-1	91	30	Mfold	33	**24**	**80.00**	70.59	**76.19**
				RNAfold	34	**24**	**80.00**	70.59	75.00
				RNAstructure	34	**24**	**80.00**	70.59	75.00
				hEDAfold	32	23	76.67	**71.88**	74.19
5	pre-miR-17	86	32	Mfold	32	**32**	**100.00**	**100.00**	**100.00**
				RNAfold	32	**32**	**100.00**	**100.00**	**100.00**
				RNAstructure	32	**32**	**100.00**	**100.00**	**100.00**
				hEDAfold	33	**32**	**100.00**	**100.00**	**100.00**
6	pre-miR-18	80	27	Mfold	24	18	66.67	75.00	70.59
				RNAfold	25	19	70.37	76.00	73.08
				RNAstructure	25	19	70.37	76.00	73.08
				hEDAfold	24	**22**	**81.48**	**91.67**	**86.27**

ลำดับ	ชื่อ	ความยาว	จำนวนคู่เบส	ขั้นตอนวิธี	Predict	TP	Sent.	Spec.	F-measure
7	pre-miR-19a	84	34	Mfold	34	**33**	**97.06**	**97.06**	**97.06**
				RNAfold	34	**33**	**97.06**	**97.06**	**97.06**
				RNAstructure	34	**33**	**97.06**	**97.06**	**97.06**
				hEDAfold	34	**33**	**97.06**	**97.06**	**97.06**
8	pre-miR-25	84	29	Mfold	29	21	72.41	72.41	72.41
				RNAfold	30	**28**	**96.55**	**93.33**	**94.92**
				RNAstructure	31	24	82.76	77.42	80.00
				hEDAfold	30	**28**	**96.55**	**93.33**	**94.92**
9	pre-miR-29a	68	26	Mfold	25	21	80.77	84.00	82.35
				RNAfold	25	21	80.77	84.00	82.35
				RNAstructure	25	21	80.77	84.00	82.35
				hEDAfold	26	**26**	**100.00**	**100.00**	**100.00**
10	pre-miR-30a	73	30	Mfold	30	29	96.67	96.67	96.67
				RNAfold	30	**30**	**100.00**	**100.00**	**100.00**
				RNAstructure	30	**30**	**100.00**	**100.00**	**100.00**
				hEDAfold	30	**30**	**100.00**	**100.00**	**100.00**
ค่าเฉลี่ย		83	31	Mfold	30	26	85.61	87.49	86.46
				RNAfold	30	27	86.71	87.45	86.94
				RNAstructure	30	27	87.49	88.57	87.90
				hEDAfold	**31**	**29**	**93.51**	**92.83**	**93.12**

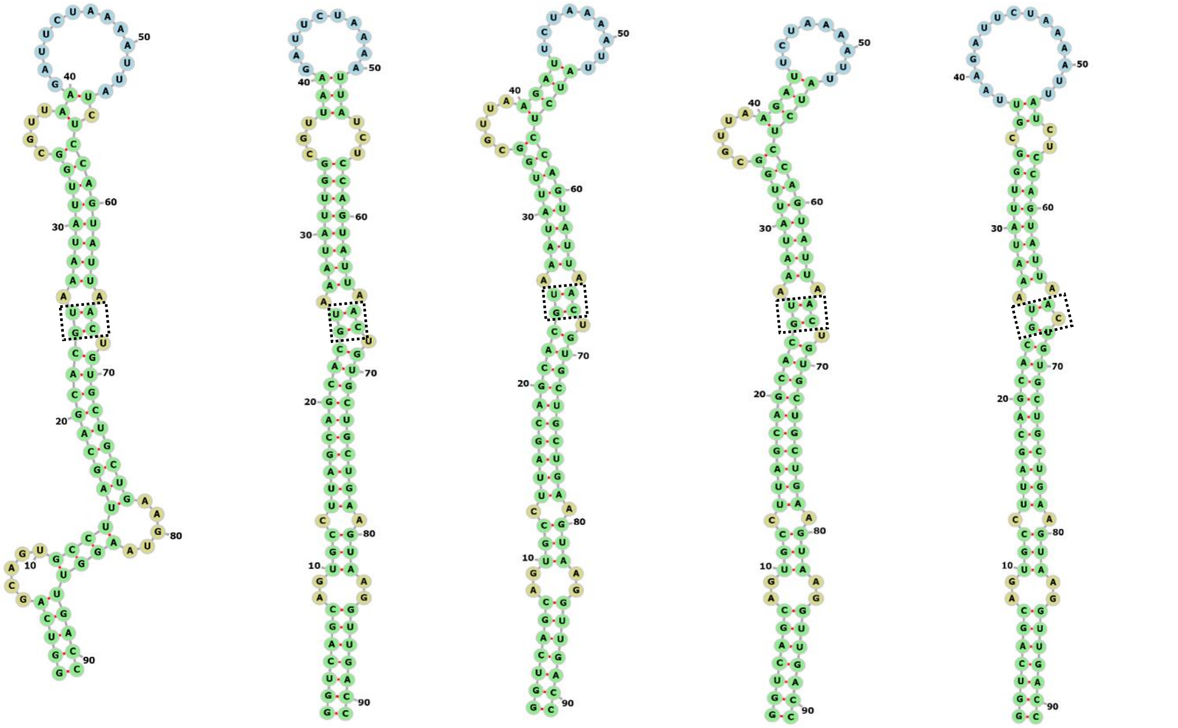
ตารางที่ 4.7 คอลัมน์ที่ 2 แสดงชื่อของอาร์เอ็นเอ คอลัมน์ที่ 3 แสดงความยาวของสายลำดับอาร์เอ็นเอ คอลัมน์ที่ 4 แสดงจำนวนคู่เบสที่พบในโครงสร้างคำตอบของข้อมูลที่นำมาทดสอบ คอลัมน์ที่ 5 แสดงวิธีการที่นำมาเปรียบเทียบ โดยขั้นตอนวิธี Hybrid-EDAfold ในตารางจะแทนด้วย hEDAfold คอลัมน์ที่ 6 แสดงจำนวนคู่เบสทั้งหมดที่แต่ละวิธีทำนายได้ คอลัมน์ที่ 7 แสดงจำนวนคู่เบสที่แต่ละวิธีทำนายได้ถูกต้อง คอลัมน์ที่ 8 - 10 แสดงผลการทำนายของแต่ละวิธีเมื่อประเมินด้วยค่าความอ่อนไหว (Sent.) ค่าความจำเพาะ (Spec.) และ F-measure ตามลำดับ และบริเวณที่มีการแรเงาในตารางแสดงขั้นตอนวิธีที่ทำผลลัพธ์ดีสุดสำหรับแต่ละอาร์เอ็นเอในแต่ละตัวชี้วัด

จากตารางที่ 4.7 ขั้นตอนวิธี Hybrid-EDAfold ทำผลลัพธ์ได้ดีกว่าหรือเท่ากับวิธีการอื่น ๆ ที่นำมาเปรียบเทียบโดยประเมินจาก F-measure จำนวน 9 รายการ มีเพียง pre-miR-16-1 เท่านั้นที่ได้ค่า F-measure ต่ำกว่าวิธีการอื่น ๆ ที่นำมาเปรียบเทียบ แต่ในภาพรวมเฉลี่ยจากทั้ง 10 ข้อมูลขั้นตอนวิธี Hybrid-EDAfold ได้ผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ในทุกตัวชี้วัด โดยมีค่าเฉลี่ยของ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ค่าความอ่อนไหว ค่าความจำเพาะ และ F-measure คือ 93.51, 92.83 และ 93.12 ตามลำดับ โดยมีค่า F-measure เฉลี่ยสูงกว่าโปรแกรม Mfold, RNAfold และ RNAstructure คือ 6.66, 6.18 และ 5.22 ตามลำดับ นอกจากนี้ ขั้นตอนวิธี Hybrid-EDAfold ยังสามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอสำหรับข้อมูลชุดนี้ได้ถูกต้อง 100% ใน 4 อาร์เอ็นเอ คือ pre-let-7f-2, pre-miR-17, pre-miR-29a และ pre-miR-30a ในขณะที่ Mfold ทำนายโครงสร้างได้ถูกต้อง 100% ใน 1 อาร์เอ็นเอ คือ pre-miR-17 และ RNAfold กับ RNAstructure ทำนายโครงสร้างได้ถูกต้อง 100% ใน 2 อาร์เอ็นเอ คือ pre-miR-17 และ pre-miR-30a

ในกรณีของ pre-miR-16-1 ที่ขั้นตอนวิธี Hybrid-EDAfold ให้ผลการทำนายโครงสร้างที่ต่ำกว่าวิธีการอื่นเล็กน้อย เมื่อประเมินในรายละเอียดแสดงดังรูปที่ 4.1 ซึ่ง (ก) แสดงโครงสร้างที่เป็นคำตอบ และ (ข-ง) แสดงโครงสร้างที่ทำนายได้จากวิธีที่นำมาเปรียบเทียบ และ (จ) แสดงโครงสร้างที่ทำนายโดยขั้นตอนวิธี Hybrid-EDAfold



(ก) คำตอบ (ข) Mfold (ค) RNAfold (ง) RNAstructure (จ) Hybrid-EDAfold

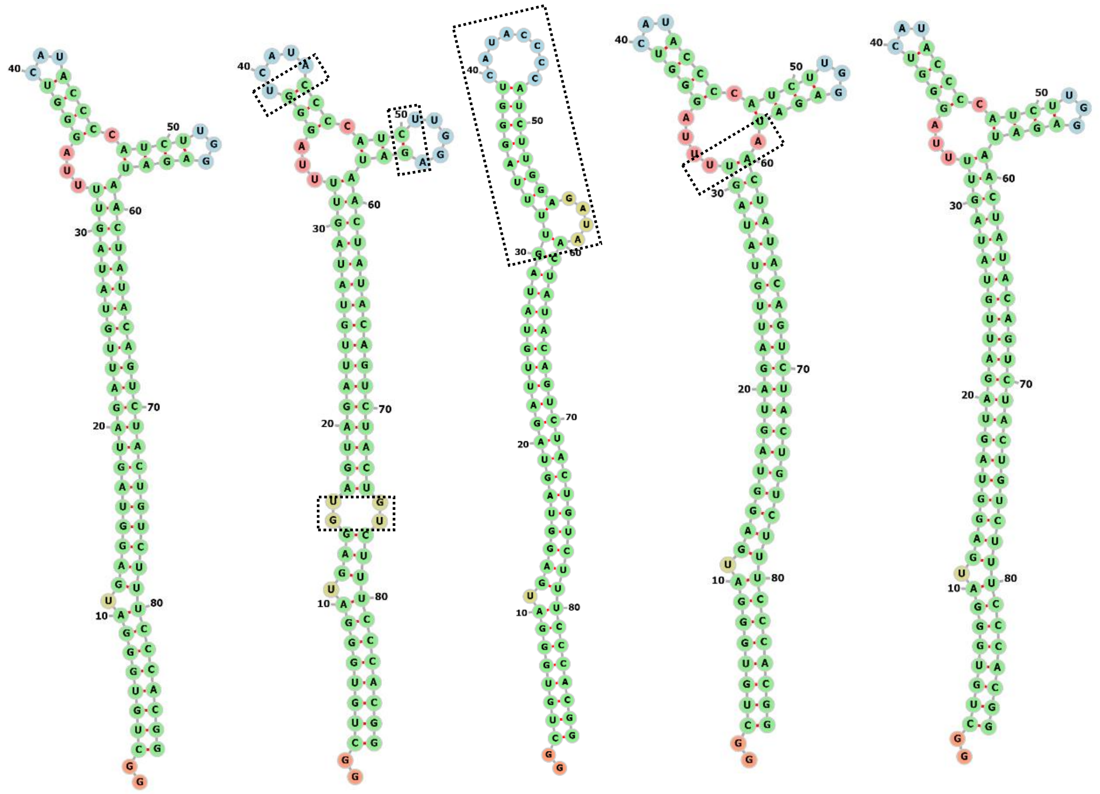
รูปที่ 4.1 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-16-1

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 4.1 พบว่าทุกวิธีที่นำมาเปรียบเทียบให้ผลการทำนายใกล้เคียงกัน บริเวณที่ตีกรอบในรูปเป็นบริเวณของฮีลิกที่มีความยาว 2 คู่เบส วิธีอื่น ๆ ทำนายได้ถูกต้องตรงกับโครงสร้างคำตอบแต่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ถูกต้องเพียง 1 คู่เบส จึงทำให้ผลการทำนายอาร์เอ็นเอนี้ของวิธีการที่นำเสนอมีค่า TP ต่ำกว่าวิธีการอื่นๆ 1 คู่เบส สาเหตุของความผิดพลาดนี้เกิดจากขั้นตอนของการจัดเตรียมฮีลิกและขั้นตอนการปรับปรุงฮีลิกที่งานวิจัยนี้นำเสนอ กล่าวคือ ฮีลิกคำตอบที่อยู่บริเวณที่ตีกรอบมีการเข้ารหัสเป็นดังนี้ [15 ; 77 ; 9] (กำหนดให้เป็น $helix_A$) และ [24 ; 67 ; 2] (กำหนดให้เป็น $helix_B$) ในขณะที่ฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกมีการเข้ารหัสเป็น [14 ; 78 ; 11] (กำหนดให้เป็น $helix_C$) และ [24 ; 67 ; 2] ($helix_D$) หมายความว่า $helix_D$ สร้างได้ตรงกับ $helix_B$ ที่เป็นคำตอบ และ $helix_C$ สร้างได้ใกล้เคียงกับ $helix_A$ ที่เป็นคำตอบแต่สร้างได้คู่เบสมากเกินกว่าคำตอบ 2 คู่ คือ เบสในลำดับที่ 14 จับคู่กับ 78 และ เบสในลำดับที่ 24 จับคู่กับ 68 สังเกตว่า $helix_C$ และ $helix_D$ มีตำแหน่งเบสบางส่วนตรงกันคือ $helix_C$ มีเบสในลำดับที่ 24 จับกับ 68 ส่วน $helix_D$ มีเบสในลำดับที่ 24 จับกับ 67 ดังนั้น ในกรณีนี้ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอจะยอมให้ฮีลิกทั้ง 2 ชิ้นนี้เกิดร่วมกันในโครงสร้างได้แต่ต้องมีการแก้ไขข้อมูลบริเวณที่มีการแชร์ตำแหน่งเบสร่วมกัน ผลปรากฏว่าคู่เบส (24 – 67) ซึ่งเป็นคำตอบมีความน่าจะเป็นต่ำกว่าคู่เบส (24 - 68) ซึ่งไม่ใช่คำตอบ ผลก็คือ คู่เบส (24 – 68) จะถูกเก็บไว้ในโครงสร้างแทน ดังนั้น สำหรับ Hybrid-EDAfold ขั้นตอนของการจัดเตรียมฮีลิกและวิธีการปรับปรุงฮีลิกเมื่อคู่เบสบางส่วนมีการแชร์ตำแหน่งเบสร่วมกันยังคงต้องมีการปรับปรุงต่อไปเพื่อทำให้ฮีลิกที่ถูกสร้างมีตำแหน่งและความยาวใกล้เคียงกับฮีลิกที่พบในโครงสร้างคำตอบมากที่สุด และกรณีที่ฮีลิกมีการแชร์ตำแหน่งเบสบางส่วนร่วมกันเกณฑ์ในการพิจารณาเพื่อเลือกเก็บคู่เบสไว้ในโครงสร้างที่ทำนายได้อาจต้องใช้เกณฑ์อื่น ๆ มาร่วมพิจารณาเพิ่มเติมนอกเหนือจากการพิจารณาแค่ค่าความน่าจะเป็นของคู่เบสเพียงอย่างเดียว

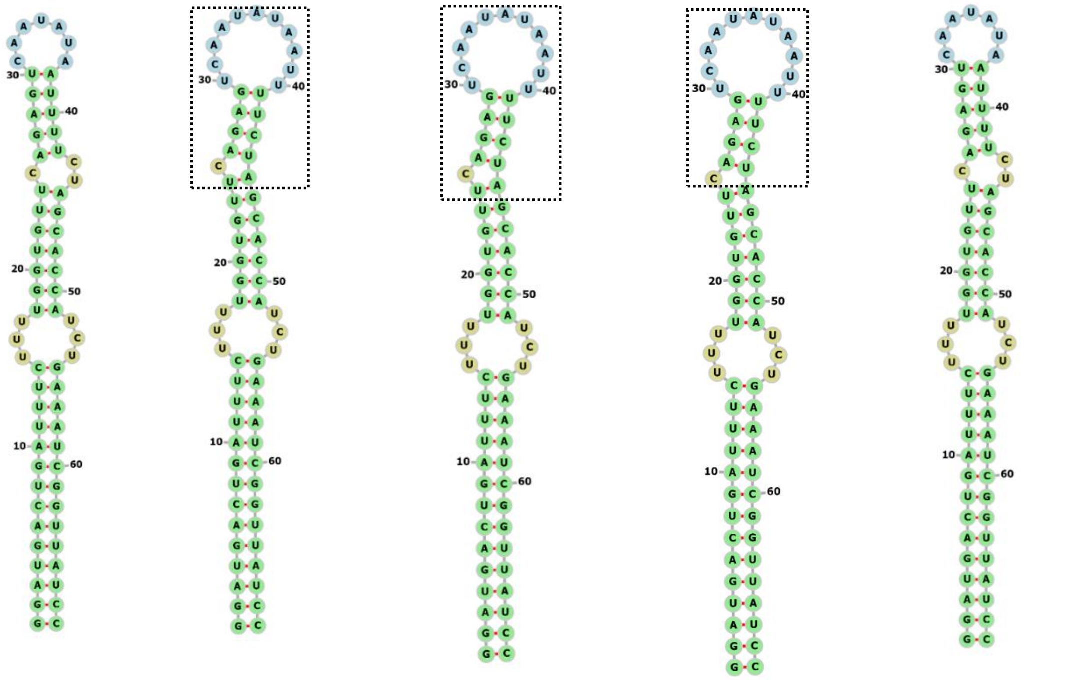
สำหรับ 2 อาร์เอ็นเอที่ขั้นตอนวิธี Hybrid-EDAfold สามารถทำนายโครงสร้างได้ถูกต้อง 100% ในขณะที่วิธีการอื่น ๆ ที่นำมาเปรียบเทียบมีการทำนายตำแหน่งคู่เบสผิดพลาดเล็กน้อย ได้แก่ pre-let-7f-2 และ pre-miR-29a โดยโครงสร้างคำตอบและโครงสร้างที่แต่ละขั้นตอนวิธีทำนายได้แสดงดังรูปที่ 4.2 และ รูปที่ 4.3 ตามลำดับ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



(ก) คำตอบ (ข) Mfold (ค) RNAfold (ง) RNAstructure (จ) Hybrid-EDAfold

รูปที่ 4.2 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-let-7f-2



(ก) คำตอบ (ข) Mfold (ค) RNAfold (ง) RNAstructure (จ) Hybrid-EDAfold

รูปที่ 4.3 เปรียบเทียบโครงสร้างที่ทำนายได้กับโครงสร้างคำตอบของ pre-miR-29a

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 4.2 แสดงผลการทำนายโครงสร้างของอาร์เอ็นเอ pre-let-7f-2 พบว่า Mfold ทำนายตำแหน่งคู่เบสผิดไป 4 คู่ แสดงด้วยบริเวณที่ตีกรอบ RNAfold ทำนายผิดไป 9 คู่ บริเวณด้านบนของโครงสร้างตั้งแต่เบสตำแหน่งที่ 32 – 59 และ RNAstructure ทำนายผิดไป 1 คู่ คือขาดการทำนายคู่เบสตำแหน่งที่ 32 ซึ่งต้องจับคู่กับตำแหน่งที่ 59

จากรูปที่ 4.3 แสดงผลการทำนายโครงสร้างของ pre-miR-29a พบว่าทั้ง 3 ขั้นตอนวิธีที่นำมาเปรียบเทียบทำนายได้ผลลัพธ์เหมือนกัน โดยทำนายคู่เบสผิดไป 5 คู่ แทนด้วยบริเวณที่ตีกรอบในรูป เมื่อพิจารณาในรายละเอียดพบว่าปัจจัยที่ทำให้ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอสามารถทำนายโครงสร้างของสองอาร์เอ็นเอนี้ได้ถูกต้อง 100% มาจาก 2 ส่วน คือ 1) ความแม่นยำในขั้นตอนการจัดเตรียมฮีลิก เมื่อพิจารณาในเซตของฮีลิกที่สร้างได้จาก 2 อาร์เอ็นเอนี้พบว่ามีฮีลิกที่พบอยู่ในโครงสร้างคำตอบครบทุกชิ้นและทุกชิ้นระบุตำแหน่งคู่เบสได้ถูกต้อง 100% 2) เมื่อเข้าสู่กระบวนการทำนายโครงสร้างด้วยขั้นตอนวิธี Hybrid-EDAfold ฮีลิกทุกชิ้นที่เป็นคำตอบก็ถูกเลือกมาประกอบร่วมกันเป็นโครงสร้างที่ขั้นตอนวิธีทำนายได้ อีกประเด็นหนึ่งที่น่าสนใจคือ ในอาร์เอ็นเอ pre-let-7f-2 แม้ว่าโปรแกรม RNAfold จะเป็นขั้นตอนวิธีที่ทำนายได้ผลลัพธ์แย่สุดในบรรดาขั้นตอนวิธีที่นำมาเปรียบเทียบ แต่การใช้ความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรมดังกล่าวในขั้นตอนการสร้างฮีลิกก็สามารถระบุตำแหน่งของฮีลิกได้ถูกต้อง 100% ดังนั้นการประเมินประสิทธิภาพในหัวข้อนี้สอดคล้องกับสมมุติฐานที่ได้นำเสนอไปว่าหากขั้นตอนการจัดเตรียมฮีลิกสามารถระบุตำแหน่งของฮีลิกได้อย่างถูกต้องแล้วจะส่งผลทำให้การทำนายโครงสร้างด้วยขั้นตอนวิธี Hybrid-EDAfold มีความถูกต้องมากยิ่งขึ้น

4.4 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มเมตาฮิวริสติกด้วยข้อมูลอาร์เอ็นเอจำนวน 20 รายการ

ในหัวข้อนี้นำเสนอการประเมินประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เปรียบเทียบกับขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ ได้แก่ RnaPredict [12], SARNA-Predict [13] และ TL-PSOfold [14] ซึ่งมีพื้นฐานมาจากขั้นตอนวิธีเชิงพันธุกรรม แบบจำลองการอบเหนียว และขั้นตอนวิธีหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค ตามลำดับ โดยทำการทดสอบกับสายลำดับอาร์เอ็นเอ 20 รายการดังที่ได้นำไปแล้วในตารางที่ 4.1 และผลการเปรียบเทียบประสิทธิภาพนำเสนอในตารางที่ 4.8 โดยผลลัพธ์ของขั้นตอนวิธีที่นำมาเปรียบเทียบรวบรวมจากข้อมูลที่แต่ละขั้นตอนวิธีรายงานในบทความและผลลัพธ์จากขั้นตอนวิธี Hybrid-EDAfold ในการทดสอบกับทุกสายลำดับอาร์เอ็นเอใช้พารามิเตอร์ชุดเดียวกันและกำหนดค่าพารามิเตอร์ต่าง ๆ ดังที่ได้นำเสนอไปในหัวข้อ 4.1 แต่ละสายลำดับถูกรันจำนวน 30 ครั้ง และรายงานผลการรันครั้งที่ให้ค่า F-measure สูงสุด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.8 การเปรียบเทียบผลการทำนายโครงสร้างของ Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มเมตาฮิวริสติก

รหัสโมเลกุล	ความยาว	จำนวนคู่เบสเฉลย	จำนวนคู่เบสที่ทำนายถูก				F-measure			
			GA	SA	PSO	hEDA	GA	SA	PSO	hEDA
CRW_00557	117	38	25	-	27	**32**	68.5	-	75.0	**85.3**
CRW_00570	118	37	**33**	**33**	**33**	**33**	86.8	89.2	88.0	**93.0**
CRW_01516	120	40	10	-	-	**35**	25.3	-	-	**89.7**
CRW_00548	122	38	27	27	31	**33**	79.4	79.4	83.8	**86.8**
CRW_00567	123	40	33	-	36	**37**	86.8	-	91.1	**92.5**
CRW_00555	124	40	25	-	-	**35**	68.5	-	-	**92.1**
CRW_00016	394	120	75	67	-	**98**	62.2	56.1	-	**82.0**
CRW_00010	454	126	86	-	-	**108**	65.4	-	-	**84.4**
CRW_00013	456	115	55	48	-	**57**	44.0	37.9	-	**45.4**
CRW_00006	468	113	68	67	-	**75**	55.7	54.9	-	**62.2**
CRW_00012	543	141	79	74	-	**107**	52.3	48.8	-	**71.6**
CRW_00004	556	131	81	79	-	**95**	55.5	51.0	-	**61.9**
CRW_00018	605	121	63	-	-	**80**	46.0	-	-	**54.6**
CRW_00423	697	189	55	43	**88**	57	28.1	21.9	**46.4**	28.0
CRW_00429	784	233	65	55	**104**	68	27.4	23.0	**44.8**	28.4
CRW_00418	940	260	**74**	-	-	63	**30.3**	-	-	24.1
CRW_00463	945	254	93	103	122	**124**	37.7	42.0	**48.9**	46.6
CRW_00438	954	268	89	111	**132**	96	34.4	42.3	**49.0**	36.5
CRW_00419	964	265	82	92	**106**	81	32.4	35.5	**42.6**	29.9
CRW_00039	1495	468	-	219	**276**	271	-	46.6	**60.5**	56.8
ค่าเฉลี่ย	549	152	59	78	**96**	79	51.9	48.4	**63.0**	62.6

ตารางที่ 4.8 คอลัมน์ที่ 1 แสดงรหัสโมเลกุลที่ใช้อ้างอิงในฐานข้อมูล RNA STRAND v2.0 คอลัมน์ที่ 2 แสดงความยาวของแต่ละสายลำดับอาร์เอ็นเอ คอลัมน์ที่ 3 แสดงจำนวนคู่เบสที่พบในโครงสร้างที่เป็นคำตอบ คอลัมน์ที่ 4-7 แสดงจำนวนคู่เบสที่แต่ละขั้นตอนวิธีทำนายได้ถูกต้อง โดยในตารางแทนขั้นตอนวิธี RnaPredict, SARNA-Predict, TL-PSOfold และ Hybrid-EDAfold ด้วย GA, SA, PSO และ hEDA ตามลำดับ และคอลัมน์ที่ 8-11 แสดง F-measure ที่แต่ละขั้นตอนวิธีทำนายได้ บริเวณที่มีการแรเงาในตารางแสดงขั้นตอนวิธีที่ทำผลลัพธ์ได้ดีที่สุดสำหรับแต่ละอาร์เอ็นเอในแต่ละตัวชี้วัด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ผลการประเมินประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เปรียบเทียบกับขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ ดังแสดงในตารางที่ 4.8 พบว่าขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายทั้งในส่วนของ TP และ F-measure ดีกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบใน 13 ข้อมูล คือ ข้อมูลลำดับที่ 1 – 13 ในขณะที่ขั้นตอนวิธี RnaPredict มีผลการทำนายดีกว่าขั้นตอนวิธีที่งานวิจัยนี้นำเสนอในข้อมูลลำดับที่ 16 และ ขั้นตอนวิธี TL-PSOfold มีผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ใน 6 ข้อมูล คือ ข้อมูลลำดับที่ 14 - 15 และ 18 – 20 และ F-measure โดยเฉลี่ยจากทั้ง 20 สายลำดับ TL-PSOfold ได้ผลลัพธ์ดีสุด คือ 63.0 และขั้นตอนวิธี Hybrid-EDAfold ได้ผลลัพธ์ดีรองลงมา คือ 62.6 ตามด้วยขั้นตอนวิธี RnaPredict ซึ่งมีค่า F-measure เป็น 51.9 และ ขั้นตอนวิธี SARNA-Predict มีค่า F-measure เป็น 48.4

แม้ว่าในภาพรวมค่าเฉลี่ย F-measure ของขั้นตอนวิธีที่งานวิจัยนี้นำเสนอจะไม่ใช่ขั้นตอนวิธีที่ให้ผลลัพธ์ดีสุดในบรรดาขั้นตอนวิธีที่นำมาเปรียบเทียบแต่ก็ต่ำกว่าขั้นตอนวิธี TL-PSOfold ซึ่งเป็นขั้นตอนวิธีที่ให้ F-measure เฉลี่ยดีสุดเพียงเล็กน้อยไม่ถึง 1% และหากประเมินผลการทำนายแยกตามชนิดอาร์เอ็นเอ ข้อมูลลำดับที่ 1- 6 เป็นอาร์เอ็นเอจากกลุ่มของ 5S Ribosomal RNA วิธีการที่งานวิจัยนี้นำเสนอมีผลการทำนายดีกว่า TL-PSOfold ในทุกสายลำดับที่มีผลการเปรียบเทียบ ข้อมูลลำดับที่ 7 – 13 เป็นอาร์เอ็นเอจากกลุ่มของ Group I intron ขั้นตอนวิธี TL-PSOfold ไม่มีรายงานผลการทดลองในกลุ่มนี้จึงทำได้เพียงเปรียบเทียบขั้นตอนวิธีที่งานวิจัยนี้นำเสนอกับอีก 2 ขั้นตอนวิธีที่เหลือซึ่งขั้นตอนวิธี Hybrid-EDAfold ได้ผลลัพธ์ดีกว่าขั้นตอนวิธีอื่น ๆ ในทุกสายลำดับที่มีผลการเปรียบเทียบ และข้อมูลลำดับที่ 14 – 20 เป็นอาร์เอ็นเอจากกลุ่มของ 16S Ribosomal RNA พบว่าขั้นตอนวิธี TL-PSOfold เป็นขั้นตอนวิธีที่ทำผลลัพธ์ได้ดีที่สุดในบรรดาขั้นตอนวิธีที่นำมาเปรียบเทียบ จึงอาจกล่าวได้ว่าการประเมินผลการทำนายแค่เพียง 20 สายลำดับนี้อาจยังไม่สามารถตัดสินได้อย่างชัดเจนว่าขั้นตอนวิธีใดมีประสิทธิภาพดีที่สุด แต่อย่างไรก็ตาม ผลการศึกษาในส่วนนี้สามารถใช้เป็นข้อมูลเบื้องต้นในการประเมินประสิทธิภาพการทำนายโครงสร้างของวิธีที่งานวิจัยนี้นำเสนอเปรียบเทียบกับขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ

ในบรรดาขั้นตอนวิธีทางเมตาฮิวริสติกที่นำมาเปรียบเทียบ ขั้นตอนวิธี RnaPredict เป็นขั้นตอนวิธีที่ใกล้เคียงกับขั้นตอนวิธี Hybrid-EDAfold มากที่สุด เนื่องจากโดยพื้นฐาน RnaPredict เป็นขั้นตอนวิธีเชิงพันธุกรรม ส่วนวิธีการที่งานวิจัยนี้นำเสนอมีความคล้ายกับขั้นตอนเชิงพันธุกรรมแบบกระชับแต่มี 2 ขั้นตอนวิธีย่อยทำงานสลับกันและมีการใช้กลุ่มโครโมโซมด้อยร่วมในการปรับปรุงเวกเตอร์ความน่าจะเป็นด้วยซึ่งแตกต่างจากขั้นตอนเชิงพันธุกรรมแบบกระชับทั่วไปที่จะใช้แต่โครโมโซมที่ดีเท่านั้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

โดยทั่วไปประสิทธิภาพของขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธีเชิงพันธุกรรมแบบกระชับมีความใกล้เคียงกัน แต่จากผลการทดลองในตารางที่ 4.8 ประเมินจาก F-measure เฉลี่ยพบว่าขั้นตอนวิธี Hybrid-EDAfold ได้ค่าเฉลี่ย F-measure สูงกว่า RnaPredict 10.7 แสดงให้เห็นว่าการใช้ 2 ขั้นตอนวิธีประมาณการแจกแจงที่มีพฤติกรรมในการค้นหาที่แตกต่างกันมาทำงานร่วมกัน (*EDA-G* พยายามค้นหาให้ทั่วทั้งปริภูมิ ส่วน *EDA-L* ทำการค้นหาบริเวณใกล้เคียงจากตำแหน่งปัจจุบัน) เสริมด้วยการใช้ความรู้จากทั้งกลุ่มโครโมโซมดีและด้อยร่วมกันในการปรับปรุงเวกเตอร์ความน่าจะเป็นมีส่วนช่วยส่งเสริมให้ขั้นตอนวิธีที่นำเสนอมีผลลัพธ์การทำนายโครงสร้างที่มีความถูกต้องมากยิ่งขึ้น

4.5 เปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตด้วยข้อมูลอาร์เอ็นเอจำนวน 750 รายการจากฐานข้อมูล RNA STARND v2.0

การเปรียบเทียบประสิทธิภาพในหัวข้อนี้ดำเนินการกับข้อมูลสายลำดับอาร์เอ็นเอจำนวน 750 สาย รวบรวมจาก 14 ชนิดอาร์เอ็นเอดังที่ปรากฎในฐานข้อมูล RNA STRAND v2.0 [33] เข้าถึงได้จาก http://www.rnasoft.ca/strand/ สำหรับทดสอบผลการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold เปรียบเทียบกับขั้นตอนวิธีในกลุ่มของกำหนดการพลวัตได้แก่ Mfold, RNAfold และ RNAstructure คุณลักษณะของข้อมูลที่นำมาทดสอบโดยสรุปแสดงดังตารางที่ 4.9

ตารางที่ 4.9 ข้อมูลสรุปของอาร์เอ็นเอ 14 ชนิดจากฐานข้อมูล RNA STRAND v2.0

ลำดับ	ชนิดอาร์เอ็นเอ	จำนวนสายลำดับทั้งหมด	จำนวนสายลำดับที่ถูกเลือก	ความยาว (nt.) สั้นสุด – ยาวสุด
1	Transfer Messenger RNA	726	86	102 - 437
2	16S Ribosomal RNA	723	200	612 - 1995
3	Transfer RNA	707	6	144 - 152
4	Ribonuclease P RNA	470	163	189 - 486
5	Synthetic RNA	450	14	101 - 302
6	Signal Recognition Particle RNA	394	93	101 - 533
7	23S Ribosomal RNA	205	11	953 - 1915
8	5S Ribosomal RNA	161	22	117 - 135
9	Group I Intron	152	106	210 - 1860
10	Hammerhead Ribozyme	146	6	114 - 119
11	Other Ribosomal RNA	64	8	116 - 500
12	Other Ribozyme	53	10	139 - 968
13	Group II Intron	42	22	619 - 1979
14	Cis-regulatory element	41	3	100 - 102

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากตารางที่ 4.9 คอลัมน์ที่ 2 แสดงชนิดของอาร์เอ็นเอ คอลัมน์ที่ 3 แสดงจำนวนสายลำดับทั้งหมดที่พบในฐานข้อมูล คอลัมน์ที่ 4 แสดงจำนวนสายลำดับที่ถูกเลือกมาทดสอบ โดยเกณฑ์ในการคัดเลือกคือทำการเรียงลำดับข้อมูลสายลำดับอาร์เอ็นเอแต่ละชนิดตามความยาวจากสั้นสุดไปยาวสุด หากความยาวเท่ากันให้เรียงลำดับตามรหัสโมเลกุลจากน้อยไปมากและเลือกเฉพาะสายลำดับที่ผ่านเงื่อนไขดังต่อไปนี้เป็นตัวแทนในการทดสอบ 1) มีความยาวแตกต่างกัน 2) ข้อมูลนิวคลีโอไทน์ของสายลำดับนั้นมีเฉพาะ 'A', 'C', 'G' และ 'U' และ 3) ความยาวของสายลำดับอยู่ในช่วง 100 – 2000 นิวคลีโอไทด์ และคอลัมน์ที่ 5 แสดงช่วงความยาวของสายลำดับที่ถูกเลือกมาเป็นตัวแทนของอาร์เอ็นเอแต่ละชนิด

เนื่องจากสายลำดับอาร์เอ็นเอที่ถูกเลือกมาเป็นตัวแทนในการทดสอบประสิทธิภาพบางชนิดอาร์เอ็นเอมีจำนวนค่อนข้างมาก การรายงานผลในหัวข้อนี้จึงเลือกนำเสนอผลการทำนายโครงสร้างในภาพรวม โดยแบ่งออกเป็น 2 ส่วนย่อย คือ ผลการเปรียบเทียบค่าความถูกต้องโดยเฉลี่ยสำหรับแต่ละชนิดอาร์เอ็นเอซึ่งนำเสนอในหัวข้อ 4.5.1 และผลการจัดอันดับค่า F-measure ที่ได้จากขั้นตอนวิธีที่นำมาเปรียบเทียบแยกการพิจารณาในแต่ละชนิดอาร์เอ็นเอนำเสนอในหัวข้อ 4.5.2 รายละเอียดเป็นดังนี้

4.5.1 การเปรียบเทียบค่าความถูกต้องโดยเฉลี่ย

การประเมินประสิทธิภาพในส่วนนี้ทำการคำนวณค่าเฉลี่ยของผลการทำนายโครงสร้างในแต่ละชนิดอาร์เอ็นเอสำหรับทุกตัวชี้วัด ผลลัพธ์เป็นดังตารางที่ 4.10

ตารางที่ 4.10 คอลัมน์ที่ 2 แสดงชื่อชนิดอาร์เอ็นเอที่นำมาทดสอบ คอลัมน์ที่ 3 แสดงความยาวเฉลี่ยของสายลำดับอาร์เอ็นเอในแต่ละชนิด คอลัมน์ที่ 4 แสดงจำนวนคู่เบสเฉลี่ยที่พบในโครงสร้างอาร์เอ็นเอที่เป็นคำตอบในอาร์เอ็นเอแต่ละชนิด คอลัมน์ที่ 5 แสดงขั้นตอนวิธีที่นำมาเปรียบเทียบ โดยขั้นตอนวิธี Hybrid-EDAfold ในตารางจะแทนด้วย hEDAfold คอลัมน์ที่ 6 แสดงจำนวนคู่เบสเฉลี่ยที่แต่ละขั้นตอนวิธีทำนายได้ทั้งหมดบนข้อมูลที่ถูกเลือกมาทดสอบ คอลัมน์ที่ 7 แสดงจำนวนคู่เบสเฉลี่ยที่แต่ละขั้นตอนวิธีทำนายได้ถูกต้องบนข้อมูลที่ถูกเลือกมาทดสอบ คอลัมน์ที่ 8 – 10 แสดงค่าเฉลี่ยของค่าความอ่อนไหว ค่าความจำเพาะ และ F-measure ตามลำดับ ที่แต่ละขั้นตอนวิธีทำนายได้บนข้อมูลที่ถูกเลือกมาทดสอบ และบริเวณที่มีการแรเงาในตารางแสดงขั้นตอนวิธีที่ทำผลลัพธ์ได้ดีที่สุดสำหรับแต่ละชนิดอาร์เอ็นเอในแต่ละตัวชี้วัด

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ตารางที่ 4.10 การเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตบนข้อมูลสายลำดับอาร์เอ็นเอ 14 ชนิด

ลำดับ	ชนิดอาร์เอ็นเอ	ความยาว (เฉลี่ย)	จำนวนคู่เบสเฉลย (เฉลี่ย)	ขั้นตอนวิธี	จำนวนคู่เบสที่ทำนาย (เฉลี่ย)	TP (เฉลี่ย)	ความอ่อนไหว (เฉลี่ย)	ความจำเพาะ (เฉลี่ย)	F-measure (เฉลี่ย)
1	Transfer Messenger RNA	366	93	Mfold	106	50	54.08	47.09	49.72
				RNAfold	111	41	44.89	37.43	40.27
				RNAstructure	108	52	56.46	48.53	51.59
				hEDAfold	105	**53**	**58.19**	**51.07**	**53.72**
2	16S Ribosomal RNA	1443	411	Mfold	442	**181**	**43.03**	**40.79**	**41.75**
				RNAfold	455	145	34.20	31.52	32.72
				RNAstructure	454	163	38.98	35.87	37.24
				hEDAfold	421	164	39.01	38.56	38.67
3	Transfer RNA	148	40	Mfold	45	18	47.31	40.21	43.39
				RNAfold	47	12	32.39	26.46	29.07
				RNAstructure	43	**28**	**71.29**	**65.00**	**67.69**
				hEDAfold	44	**28**	71.10	63.25	66.65
4	Ribonuclease P RNA	338	103	Mfold	102	65	63.60	63.59	63.30
				RNAfold	106	56	54.98	52.92	53.68
				RNAstructure	100	64	61.97	63.32	62.30
				hEDAfold	99	**69**	**66.93**	**68.62**	**67.45**
5	Synthetic RNA	170	55	Mfold	54	27	47.75	51.03	48.91
				RNAfold	54	23	42.01	44.62	42.91
				RNAstructure	54	28	48.99	53.18	50.50
				hEDAfold	51	**29**	**51.20**	**57.01**	**53.57**
6	Signal Recognition Particle RNA	276	86	Mfold	90	**62**	72.71	70.31	71.33
				RNAfold	93	51	60.52	56.70	58.41
				RNAstructure	92	61	72.35	68.79	70.37
				hEDAfold	88	**62**	**73.30**	**71.96**	**72.47**
7	23S Ribosomal RNA	1298	298	Mfold	387	91	30.24	23.30	26.11
				RNAfold	379	71	24.24	18.77	20.95
				RNAstructure	393	99	33.01	24.90	28.16
				hEDAfold	368	**104**	**35.08**	**28.34**	**31.11**
8	5S Ribosomal RNA	124	40	Mfold	38	26	65.01	69.28	66.93
				RNAfold	40	25	61.45	61.19	61.22
				RNAstructure	40	29	73.22	74.00	73.54
				hEDAfold	37	**32**	**79.53**	**86.77**	**82.85**
9	Group I Intron	572	104	Mfold	172	67	63.04	47.82	52.24
				RNAfold	178	58	54.74	40.43	44.74
				RNAstructure	170	65	61.33	48.28	51.94
				hEDAfold	166	**71**	**67.91**	**52.55**	**57.07**

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ลำดับ	ชนิด อาร์เอ็นเอ	ความยาว (เฉลี่ย)	จำนวนคู่เบสเฉลย (เฉลี่ย)	ขั้นตอนวิธี	จำนวนคู่เบสที่ทำนาย (เฉลี่ย)	TP (เฉลี่ย)	ความอ่อนไหว (เฉลี่ย)	ความจำเพาะ (เฉลี่ย)	F-measure (เฉลี่ย)
10	Hammerhead Ribozyme	117	14	Mfold	35	8	57.14	22.86	32.63
				RNAfold	35	7	51.19	20.05	28.77
				RNAstructure	36	8	57.14	22.69	32.45
				hEDAfold	27	**9**	**60.71**	**31.56**	**41.49**
11	Other Ribosomal RNA	278	91	Mfold	87	48	47.18	51.98	49.26
				RNAfold	88	43	43.47	47.54	45.19
				RNAstructure	87	47	49.51	54.30	51.55
				hEDAfold	88	**53**	**57.46**	**63.11**	**59.81**
12	Other Ribozyme	334	112	Mfold	104	70	64.40	73.24	68.36
				RNAfold	108	63	58.08	66.11	61.50
				RNAstructure	106	71	**64.86**	**73.90**	**68.80**
				hEDAfold	107	**72**	64.65	72.67	68.17
13	Group II Intron	974	148	Mfold	292	81	53.34	29.20	36.79
				RNAfold	303	72	45.66	24.84	31.36
				RNAstructure	288	82	53.47	30.52	37.89
				hEDAfold	285	**86**	**55.67**	**31.44**	**39.26**
14	Cis-regulatory element	101	32	Mfold	31	**27**	**85.42**	87.97	86.61
				RNAfold	31	**27**	**85.42**	87.97	86.61
				RNAstructure	32	26	80.21	81.68	80.88
				hEDAfold	31	**27**	**85.42**	**88.00**	**86.65**
ค่าเฉลี่ย		467	116	Mfold	142	59	56.73	51.34	52.67
				RNAfold	145	50	49.52	44.04	45.53
				RNAstructure	143	59	58.77	53.21	54.64
				hEDAfold	137	**61**	**61.91**	**57.75**	**58.63**

ผลการประเมินประสิทธิภาพในภาพรวมจากทั้ง 14 ชนิดอาร์เอ็นเอดังแสดงในตารางที่ 4.10 พบว่า ขั้นตอนวิธี Hybrid-EDAfold ทำผลลัพธ์ได้ดีกว่าขั้นตอนวิธีในกลุ่มกำหนดการพลวัตในทุกตัวชี้วัด กล่าวคือ ทำนายจำนวนคู่เบสได้ใกล้เคียงกับจำนวนคู่เบสที่พบจริงในโครงสร้างคำตอบมากที่สุด (จำนวนคู่เบสโดยเฉลี่ยที่ทำนายได้ทั้งหมด คือ 137 คู่และจำนวนคู่เบสโดยเฉลี่ยที่พบจริงในโครงสร้างคำตอบมีทั้งหมด 116 คู่) มีค่าเฉลี่ยของความอ่อนไหว ความจำเพาะ และ F-measure เป็น 61.91, 57.75 และ 58.63 ตามลำดับ ซึ่งดีกว่าผลลัพธ์จากโปรแกรม Mfold เมื่อประเมินโดยใช้ตัวชี้วัดเดียวกัน คือ 5.18, 6.41 และ 5.96 ตามลำดับ ดีกว่าผลลัพธ์จากโปรแกรม RNAfold คือ 12.39, 13.71 และ 13.1 ตามลำดับ และ ดีกว่าผลลัพธ์จากโปรแกรม RNAstructure คือ 3.14, 4.54 และ 3.99 ตามลำดับ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

หากเปรียบเทียบแค่เฉพาะขั้นตอนวิธีในกลุ่มกำหนดการพลวัต (ยังไม่รวมวิธีการที่งานวิจัยนี้นำเสนอ) พบว่าโปรแกรม Mfold มีผลการทำนายโครงสร้างดีสุดประเมินจากค่าเฉลี่ย F-measure ในอาร์เอ็นเอ 6 ชนิด ได้แก่ 16S Ribosomal RNA, Ribonuclease P RNA, Signal Recognition Particle RNA, Group I Intron, Hammerhead Ribozyme, และ Cis-regulatory element (อาร์เอ็นเอชนิดนี้ Mfold ทำนายผลลัพธ์ได้ดีเทียบเท่ากับ RNAfold) ในขณะที่โปรแกรม RNAstructure มีผลลัพธ์การทำนายโครงสร้างดีสุดในอีก 8 ชนิดอาร์เอ็นเอที่เหลือ และผลการประเมินจากทั้ง 14 ชนิดพบว่า RNAstructure มีค่าเฉลี่ย F-measure สูงกว่า Mfold ประมาณ 2%

เมื่อนำขั้นตอนวิธี Hybrid-EDAfold มาเปรียบเทียบกับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตที่ทำผลลัพธ์ได้ดีสุดในแต่ละชนิดของอาร์เอ็นเอได้ผลลัพธ์ ดังนี้

- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 2.13 เมื่อทดสอบกับ Transfer Messenger
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ต่ำกว่า Mfold เท่ากับ 3.08 เมื่อทดสอบกับ 16S Ribosomal RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ต่ำกว่า RNAstructure เท่ากับ 1.04 เมื่อทดสอบกับ Transfer RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 4.15 เมื่อทดสอบกับ Ribonuclease P RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 3.07 เมื่อทดสอบกับ Synthetic RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า Mfold เท่ากับ 1.14 เมื่อทดสอบกับ Signal Recognition Particle RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 2.95 เมื่อทดสอบกับ 23S Ribosomal RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 9.31 เมื่อทดสอบกับ 5S Ribosomal RNA
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า Mfold เท่ากับ 4.83 เมื่อทดสอบกับ Group I Intron
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า Mfold เท่ากับ 8.86 เมื่อทดสอบกับ Hammerhead Ribozyme
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 8.26 เมื่อทดสอบกับ Other Ribosomal RNA

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 1.2 เมื่อทดสอบกับ Other Ribozyme
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า RNAstructure เท่ากับ 1.37 เมื่อทดสอบกับ Group II Intron
- Hybrid-EDAfold มีค่าเฉลี่ย F-measure ดีกว่า Mfold และ RNAfold เล็กน้อย โดยขั้นตอนวิธีที่นำเสนอมีค่า F-measure เฉลี่ยเป็น 86.65 ในขณะที่ Mfold และ RNAfold ทำผลลัพธ์ได้เท่ากันโดยมีค่า F-measure เฉลี่ยเป็น 86.61

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

โดยสรุป ขั้นตอนวิธี Hybrid-EDAfold ทำนายโครงสร้างอาร์เอ็นเอได้ผลลัพธ์ดีกว่าขั้นตอนวิธีในกลุ่มกำหนดการพลวัตจำนวน 12 ชนิดอาร์เอ็นเอ สำหรับอีก 2 ชนิด ขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายเมื่อประเมินจากค่าเฉลี่ยของ F-measure ที่ต่ำกว่า Mfold ในกลุ่มของ 16S Ribosomal RNA และ ต่ำกว่า RNAstructure ในกลุ่มของ Transfer RNA เมื่อทำการวิเคราะห์ลงในรายละเอียดพบว่าในกรณีของ 16S Ribosomal RNA ผลการทำนายที่ยังไม่ดีนักอาจมีสาเหตุมาจาก 2 ปัจจัย 1) ค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold สำหรับ 16S Ribosomal RNA ค่อนข้างไม่แม่นยำ (ในบรรดา 14 ชนิดอาร์เอ็นเอค่าความน่าจะเป็นของคู่เบสของอาร์เอ็นเอกลุ่มนี้มีความแม่นยำน้อยสุด) ดังนั้นเมื่อนำข้อมูลในส่วนนี้ไปใช้ในขั้นตอนของการจัดเตรียมฮีลิกจึงทำให้ฮีลิกที่สร้างได้ระบุตำแหน่งของคู่เบสได้ถูกต้องตรงกับตำแหน่งคู่เบสที่พบในโครงสร้างคำตอบประมาณ 67% ในขณะที่อาร์เอ็นเอชนิดอื่น ๆ ฮีลิกที่สร้างได้ระบุตำแหน่งของคู่เบสได้ถูกต้องอยู่ในช่วงประมาณ 73% – 92% และ 2) สัดส่วนจำนวนชิ้นฮีลิกที่สร้างได้เมื่อเทียบกับฮีลิกชิ้นที่เป็นคำตอบค่อนข้างต่ำเนื่องจากอาร์เอ็นเอชนิดนี้ประกอบด้วยสายลำดับที่ค่อนข้างยาว หากพิจารณาที่ความยาวเฉลี่ยจะสังเกตเห็นว่าอาร์เอ็นเอชนิดนี้มีความยาวสูงสุดในบรรดาอาร์เอ็นเอทั้ง 14 ชนิด สายลำดับยิ่งยาวจำนวนชิ้นของฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิกก็ยิ่งมาก ซึ่งเมื่อนำฮีลิกที่สร้างได้ในกลุ่มนี้ไปตรวจสอบกับฮีลิกที่พบในโครงสร้างคำตอบพบว่ามีจำนวนฮีลิกที่พบในโครงสร้างคำตอบไม่ถึง 10% หมายความว่า ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอจะต้องพยายามเลือกเซตย่อยของฮีลิกที่คาดว่าน่าจะเป็นชิ้นที่เป็นคำตอบจากเซตของฮีลิกที่เป็นไปได้ทั้งหมดซึ่งมีจำนวนเยอะมาก (คำตอบเพียงเล็กน้อยปะปนอยู่ในกลุ่มของสิ่งที่สามารถเลือกได้จำนวนมาก)

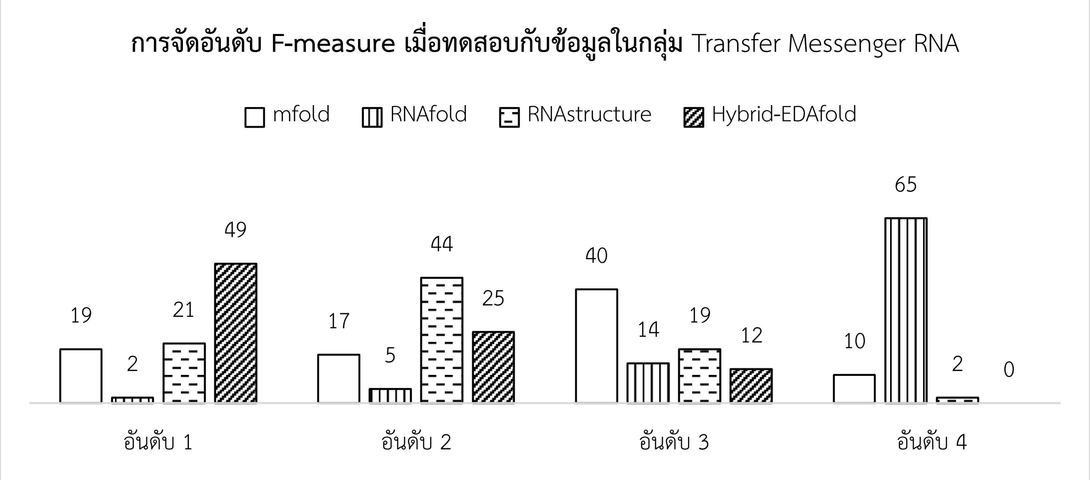
ในกรณีของ Transfer RNA ที่ขั้นตอนวิธี Hybrid-EDAfold ทำนายได้ผลลัพธ์ต่ำกว่าโปรแกรม RNAstructure เล็กน้อย เมื่อพิจารณาในรายละเอียดพบว่าวิธีการที่นำเสนอทำนายจำนวนของคู่เบสได้ถูกต้องใกล้เคียงกับ RNAstructure (ประเมินจากค่าเฉลี่ยของ TP) แต่เนื่องจากวิธีการที่นำเสนอทำนายจำนวนคู่เบสสูงกว่าจึงทำให้ค่าเฉลี่ยของความจำเพาะ และ F-measure ที่ได้ต่ำกว่า RNAstructure เล็กน้อย

แม้ว่าโปรแกรม RNAfold จะให้ผลการทำนายโครงสร้างที่ไม่ดีนักเมื่อเทียบกับขั้นตอนวิธีอื่น ๆ สาเหตุอาจเนื่องมาจากโปรแกรมดังกล่าวรายงานผลลัพธ์แค่โครงสร้างที่มีค่าพลังงานต่ำสุด ไม่ได้รองรับการทำนายหลายโครงสร้างเหมือนวิธีการอื่น ๆ ที่นำมาเปรียบเทียบ แต่เนื่องจากขั้นตอนวิธี Hybrid-EDAfold ใช้ค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold การทดลองในหัวข้อนี้จึงนำขั้นตอนวิธีนี้มาเปรียบเทียบร่วมด้วย ในภาพรวมพบว่าการใช้ข้อมูลค่าความน่าจะเป็นของคู่เบสที่ได้จากโปรแกรม RNAfold ร่วมกับวิธีการค้นหาคำตอบที่งานวิจัยนี้นำเสนอช่วยส่งเสริมให้ขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายโครงสร้างที่ดีขึ้นจากโปรแกรม RNAfold เมื่อประเมินด้วยค่าเฉลี่ยของ F-measure ในแต่ละชนิดอาร์เอ็นเอ คือ 13.5, 5.9, 37.6, 13.8, 10.7, 14.1, 10.2, 21.6, 12.4, 12.7, 14.6, 8.5, 7.9 และ 0.04 ตามลำดับ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

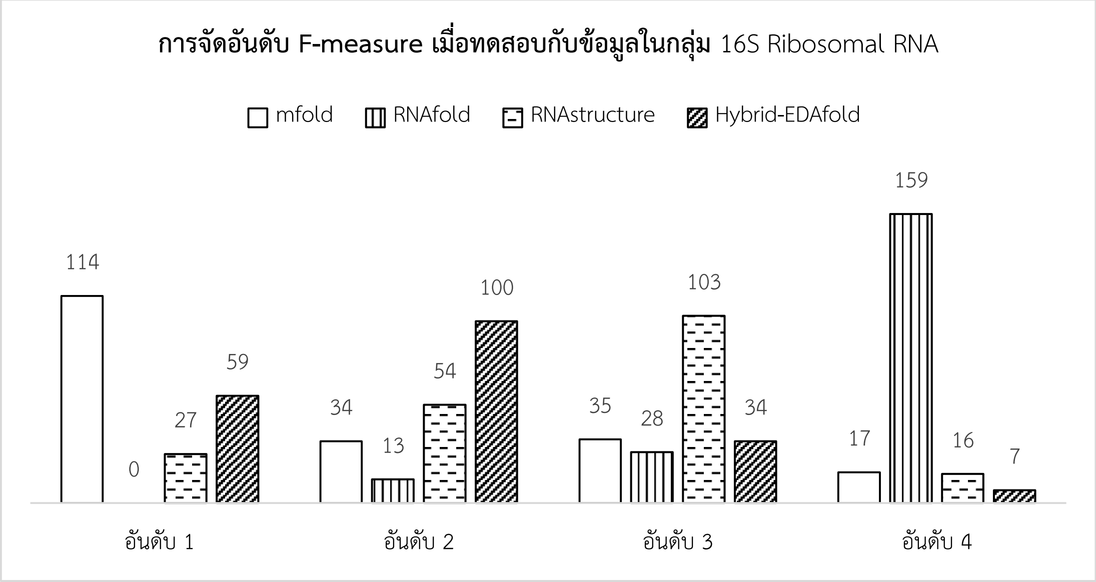
4.5.2 ผลการจัดอันดับ F-measure สำหรับอาร์เอ็นเอแต่ละชนิด

การประเมินประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold เปรียบเทียบกับขั้นตอนวิธีในกลุ่มกำหนดการพลวัตในหัวข้อนี้เป็นการนำผลลัพธ์ในส่วนของ F-measure จากทุกขั้นตอนวิธีมาแข่งขันกันในแต่ละชนิดของอาร์เอ็นเอ ขั้นตอนวิธีใดมีค่า F-measure สูงสุดจะถูกพิจารณาว่าอยู่อันดับ 1 และขั้นตอนวิธีใดที่ได้ค่า F-measure ต่ำรองลงมาก็จะถูกจัดอยู่ในอันดับที่สูงขึ้นไล่ไปเรื่อย ๆ และหากข้อมูลสายลำดับอาร์เอ็นเอใดที่มีหลายขั้นตอนวิธีทำนายได้ค่า F-measure เท่ากันก็จะถูกจัดให้อยู่ในอันดับเดียวกัน ผลลัพธ์แสดงดังรูปที่ 4.4 – 4.17 โดยตัวเลขที่ปรากฏอยู่บนกราฟแสดงจำนวนสายลำดับอาร์เอ็นเอที่ถูกจัดให้อยู่อันดับนั้น ๆ



รูปที่ 4.4 การจัดอันดับ F-measure เมื่อทดสอบกับ Transfer Messenger RNA

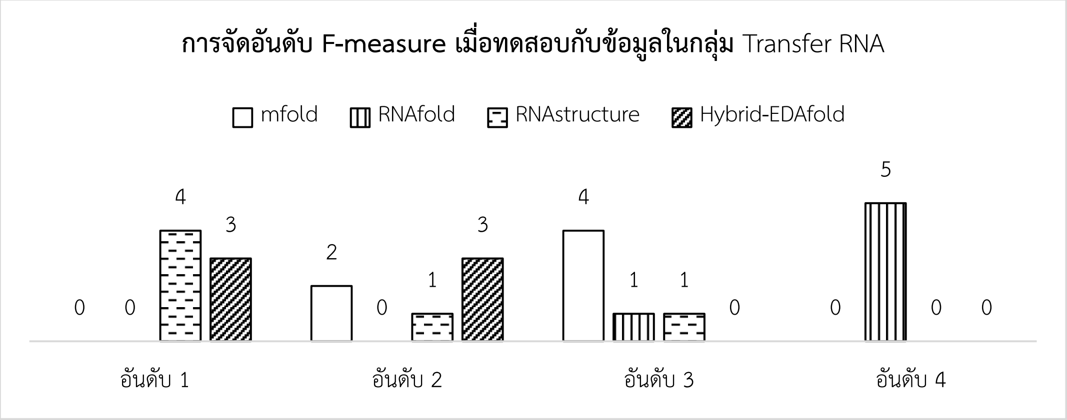
จากรูปที่ 4.4 ข้อมูลในกลุ่มนี้ถูกเลือกมาจำนวน 86 รายการ พบว่าผลการทำนายส่วนใหญ่ขั้นตอนวิธี Hybrid-EDAfold มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 49 รายการ และ ไม่มีข้อมูลรายการใดที่ให้ค่า F-measure อยู่ในอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สามจำนวน 40 รายการ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สี่จำนวน 65 รายการ และโปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สองจำนวน 44 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, RNAstructure, Mfold และ RNAfold ตามลำดับ



รูปที่ 4.5 การจัดอันดับ F-measure เมื่อทดสอบกับ 16S Ribosomal RNA

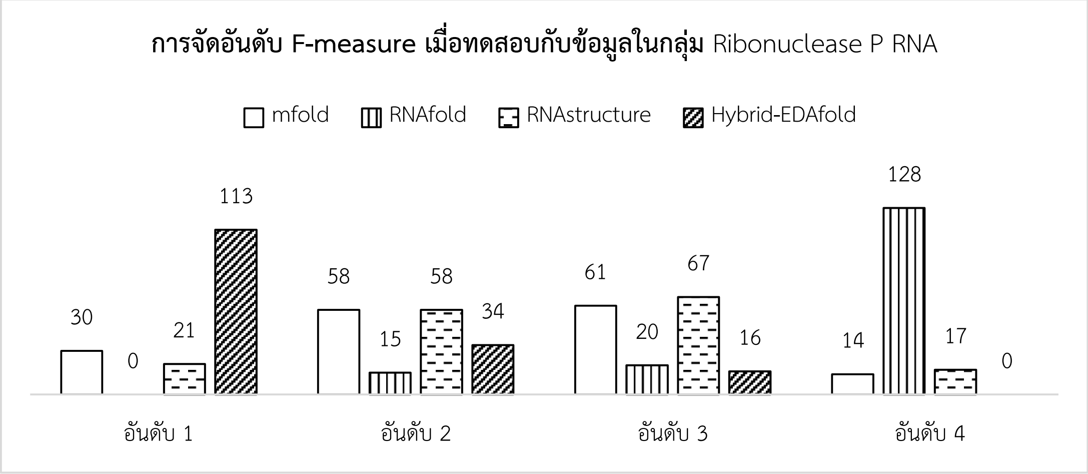
จากรูปที่ 4.5 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 200 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 100 รายการ และมีค่า F-measure เป็นอันดับที่สี่จำนวนน้อยสุดเมื่อเทียบกับวิธีการอื่น ๆ ในขณะที่โปรแกรม Mfold ทำผลลัพธ์ในข้อมูลกลุ่มนี้ได้ดีสุดโดยมีค่า F-measure ส่วนใหญ่สูงเป็นอันดับที่หนึ่งจำนวน 114 รายการ โปรแกรม RNAfold มีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่สี่จำนวน 159 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure อยู่ในอันดับที่หนึ่ง โปรแกรม RNAstructure มีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่สามจำนวน 103 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Mfold, Hybrid-EDAfold, RNAstructure และ RNAfold ตามลำดับ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



รูปที่ 4.6 การจัดอันดับ F-measure เมื่อทดสอบกับ Transfer RNA

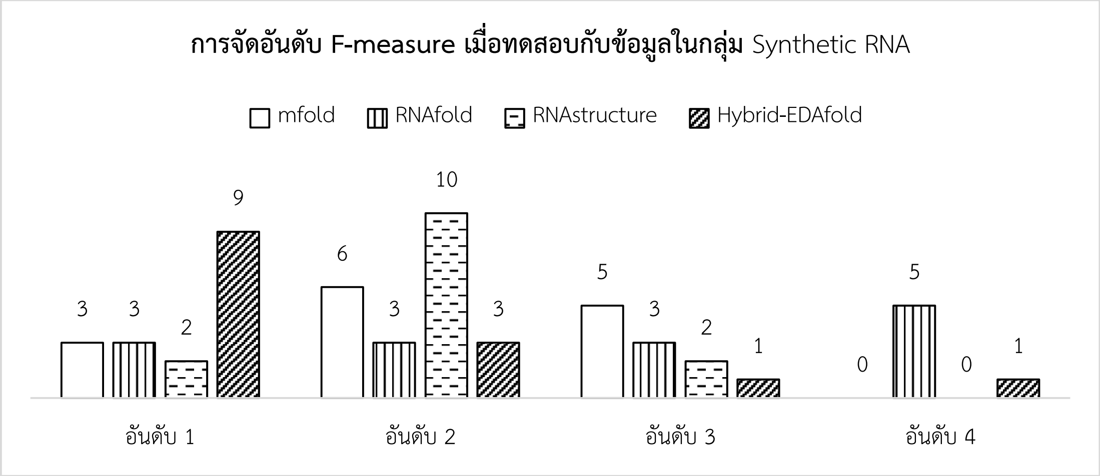
จากรูปที่ 4.6 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 6 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งและอันดับที่สองอย่างละครึ่ง และไม่มีข้อมูลรายการใดที่มีค่า F-measure อยู่ในอันดับที่สามและสี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สามจำนวน 4 รายการ อีก 2 รายการที่เหลือมีค่า F-measure อยู่ในอันดับที่สอง โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สี่จำนวน 5 รายการ โปรแกรม RNAstructure เป็นวิธีที่มีผลลัพธ์ดีที่สุดในข้อมูลกลุ่มนี้ โดยมีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่หนึ่งจำนวน 4 รายการ อีก 2 รายการที่เหลือได้ผลการทำนายอยู่ในอันดับที่สองและสามอย่างละ 1 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น RNAstructure, Hybrid-EDAfold, Mfold และ RNAfold ตามลำดับ



รูปที่ 4.7 การจัดอันดับ F-measure เมื่อทดสอบกับ Ribonuclease P RNA

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 4.7 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 163 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 113 รายการ นอกจากนี้ มีค่า F-measure อยู่ในอันดับที่สามเป็นจำนวนน้อยสุดในบรรดาขั้นตอนวิธีที่นำมาเปรียบเทียบและไม่มีข้อมูลรายการใดที่มีค่า F-measure อยู่ในอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สามจำนวน 61 รายการ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure อยู่ในอันดับที่สี่จำนวน 128 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงสุดเป็นอันดับที่หนึ่ง โปรแกรม RNAstructure ให้ผลลัพธ์เป็นไปในทิศทางเดียวกับ Mfold คือ มีค่า F-measure ส่วนใหญ่สูงเป็นอันดับที่สามจำนวน 67 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ แม้ว่า Mfold กับ RNAstructure จะมีค่า F-measure ส่วนใหญ่อยู่ในอันดับสามทั้งคู่ แต่ Mfold มีจำนวนข้อมูลที่อยู่ในอันดับหนึ่งมากกว่าจึงจัดให้ Mfold อยู่ในอันดับที่ดีกว่า RNAstructure

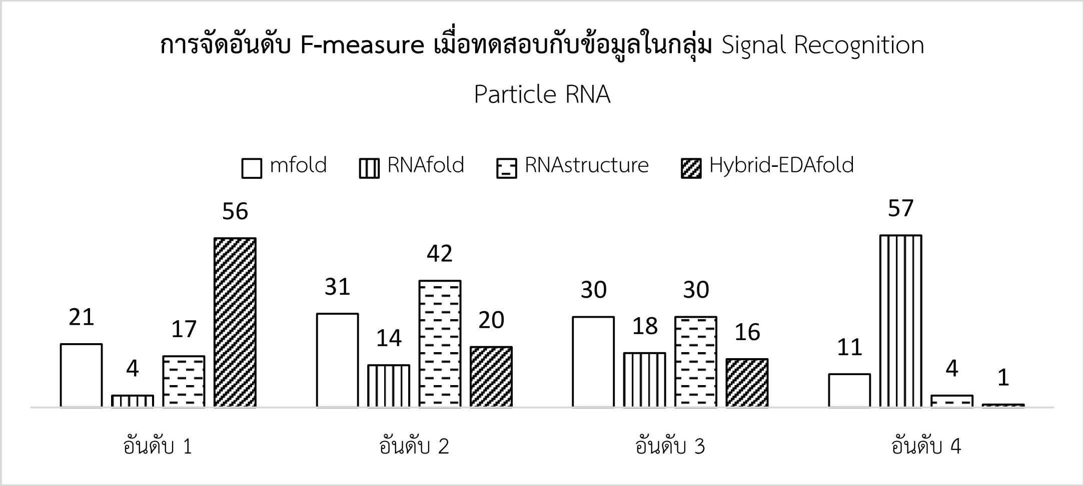


รูปที่ 4.8 การจัดอันดับ F-measure เมื่อทดสอบกับ Synthetic RNA

จากรูปที่ 4.8 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 14 รายการ พบว่าขั้นตอนวิธี Hybrid-EDAfold มีค่า F-measure ส่วนใหญ่สูงเป็นอันดับที่หนึ่งจำนวน 9 รายการ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 6 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure อยู่ในอันดับที่สี่แต่มีค่า F-measure อยู่ในลำดับที่สามเป็นจำนวนมากสุดเมื่อเปรียบเทียบกับวิธีการอื่น ๆ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 5 รายการ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 10 รายการ และ และไม่มีข้อมูลรายการใดที่มีค่า F-measure อยู่ในอันดับที่สี่ ดังนั้นสำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

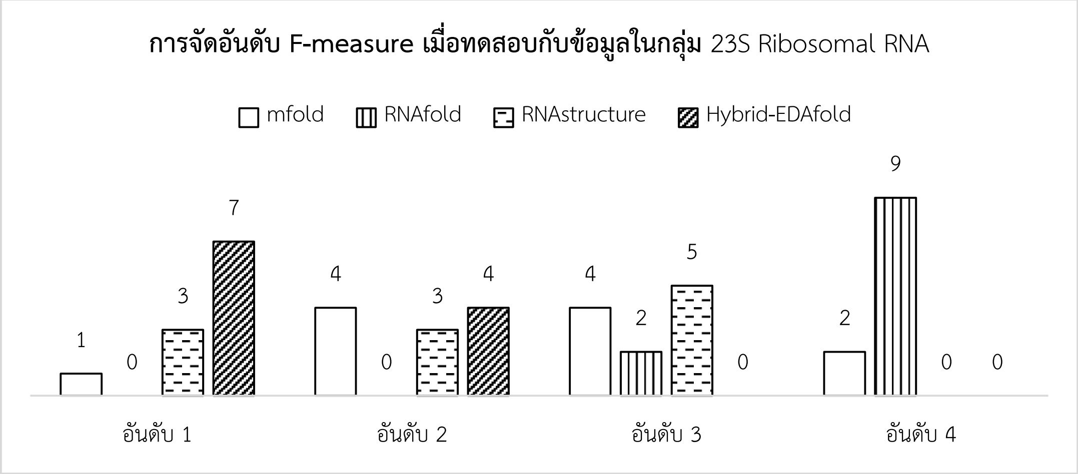
ได้เป็น Hybrid-EDAfold, RNAstructure, Mfold และ RNAfold ตามลำดับ แม้ว่า Mfold กับ RNAstructure จะมีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่สองทั้งคู่ แต่ RNAstructure มีจำนวนข้อมูลที่อยู่ในอันดับที่สองมากกว่าจึงจัดให้ RNAstructure อยู่ในอันดับที่ดีกว่า



รูปที่ 4.9 การจัดอันดับ F-measure เมื่อทดสอบกับ Signal Recognition Particle RNA

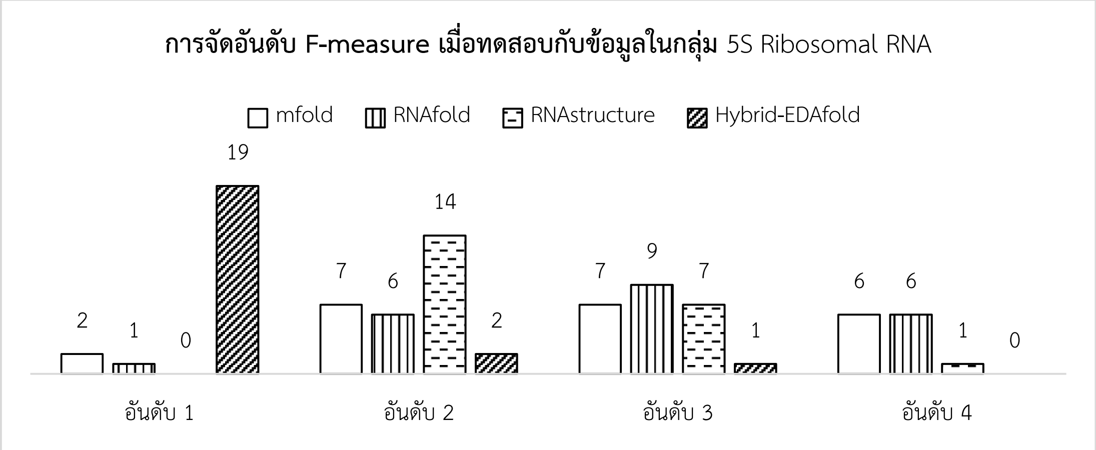
จากรูปที่ 4.9 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 93 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 56 รายการ และมีข้อมูลที่มีค่า F-measure สูงเป็นอันดับที่สามและสี่เป็นจำนวนน้อยสุดเมื่อเทียบกับวิธีการอื่น ๆ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 31 รายการ และมีค่า F-measure สูงเป็นอันดับที่สามมากสุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 57 รายการ และมีค่า F-measure สูงเป็นอันดับที่หนึ่งเป็นจำนวนน้อยสุดเมื่อเทียบกับวิธีการอื่น ๆ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 42 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, RNAstructure, Mfold และ RNAfold ตามลำดับ แม้ว่า Mfold กับ RNAstructure จะมีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่สองทั้งคู่ แต่ RNAstructure มีจำนวนข้อมูลที่อยู่ในอันดับที่สองมากกว่าจึงจัดให้ RNAstructure อยู่ในอันดับที่ดีกว่า

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



รูปที่ 4.10 การจัดอันดับ F-measure เมื่อทดสอบกับ 23S Ribosomal RNA

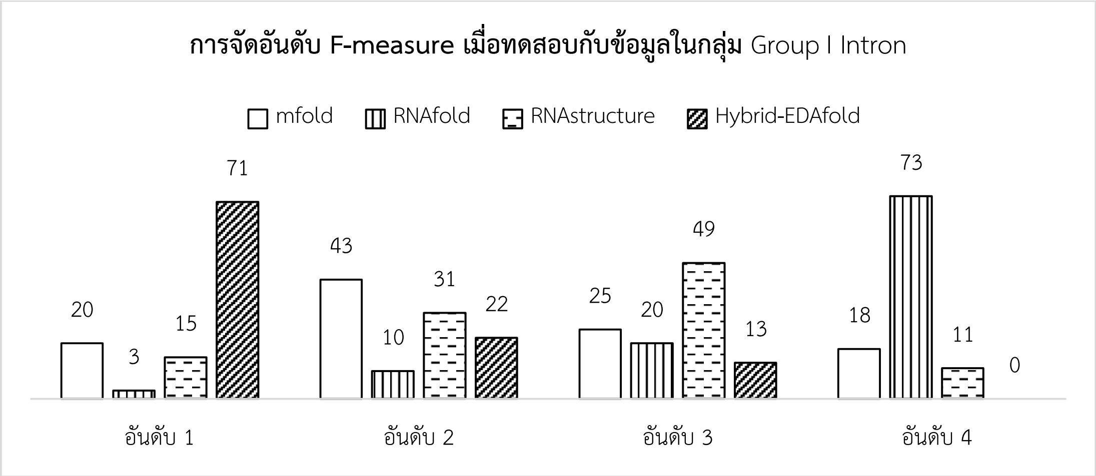
จากรูปที่ 4.10 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 11 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 7 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สามและสี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองและสามเท่ากันจำนวนอันดับละ 4 รายการ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 9 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่หนึ่งหรือสอง โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สามจำนวน 5 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สี่ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ



รูปที่ 4.11 การจัดอันดับ F-measure เมื่อทดสอบกับ 5S Ribosomal RNA

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

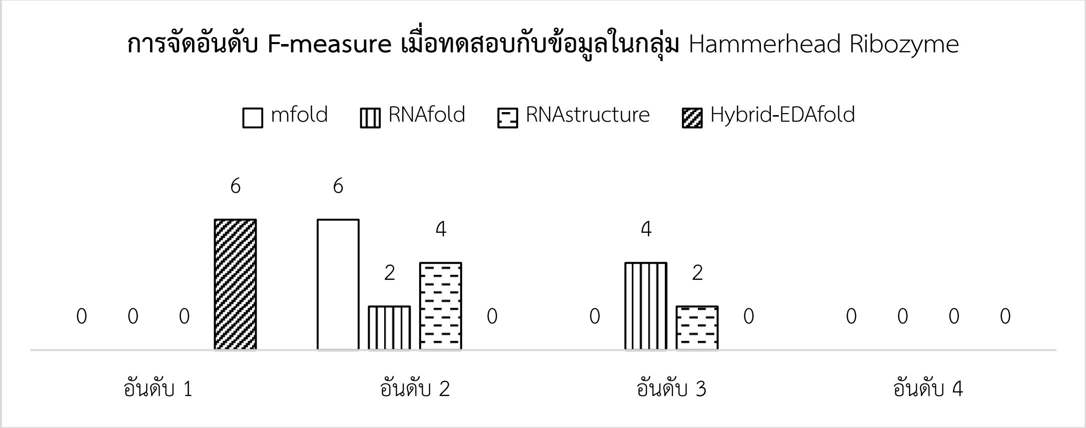
จากรูปที่ 4.11 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 22 รายการ พบว่าขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายประเมินส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 19 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองและสามเท่ากันจำนวนอันดับละ 7 รายการ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สามจำนวน 9 รายการ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 14 รายการและไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่หนึ่ง ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, RNAstructure, RNAfold และ Mfold ตามลำดับ แม้ว่า Mfold กับ RNAstructure จะมีค่า F-measure ส่วนใหญ่อยู่ในอันดับที่สองเหมือนกันแต่ RNAstucture มีจำนวนข้อมูลที่อยู่ในอันดับสองมากกว่า จากนั้นเปรียบเทียบอันดับของ Mfold กับ RNAfold พบว่า RNAfold มีจำนวนข้อมูลในอันดับสามมากกว่าจึงจัดให้ RNAfold อยู่ในอันดับที่ดีกว่า Mfold



รูปที่ 4.12 การจัดอันดับ F-measure เมื่อทดสอบกับ Group I Intron

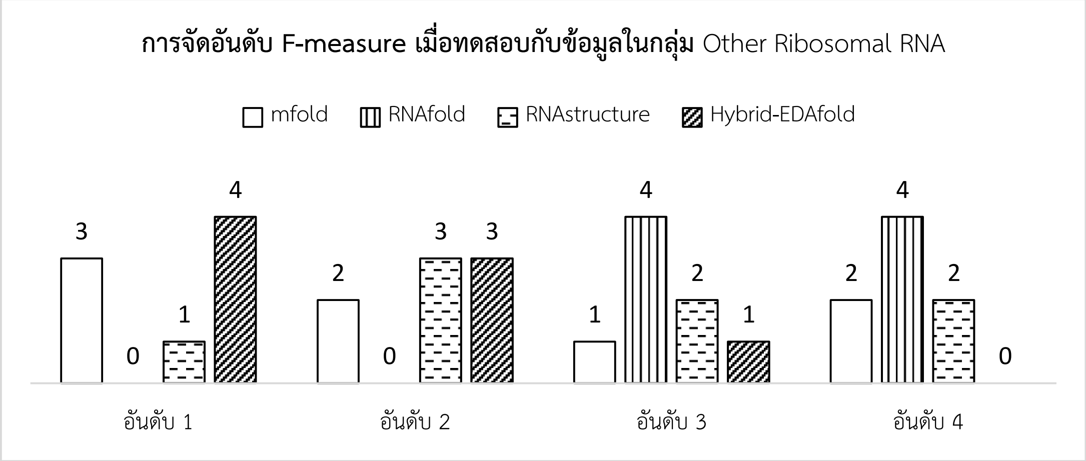
จากรูปที่ 4.12 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 106 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายประเมินส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 71 รายการและไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 43 รายการ โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 73 รายการ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สามจำนวน 49 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



รูปที่ 4.13 การจัดอันดับ F-measure เมื่อทดสอบกับ Hammerhead Ribozyme

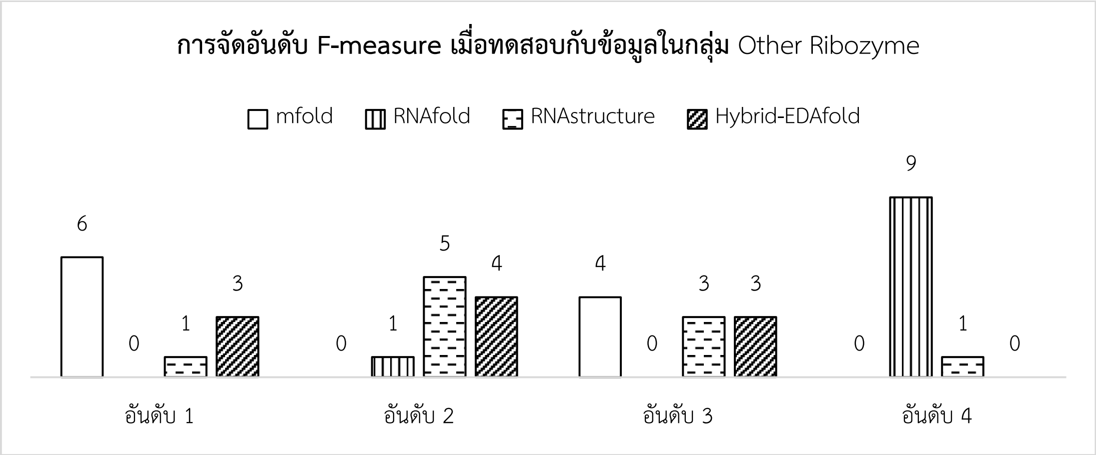
จากรูปที่ 4.13 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 6 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายทั้งหมดมีค่า F-measure สูงเป็นอันดับที่หนึ่ง โปรแกรม Mfold ผลการทำนายทั้งหมดมีค่า F-measure สูงเป็นอันดับที่สอง โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สามจำนวน 4 รายการ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 4 รายการ เนื่องจากผลการทำนายสำหรับข้อมูลกลุ่มนี้แต่ละวิธีการที่นำมาเปรียบเทียบค่อนข้างทำผลลัพธ์ได้เท่ากันหรือใกล้เคียงกันจึงทำให้ไม่ปรากฏข้อมูลที่ให้ค่า F-measure อยู่ในอันดับที่สี่ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ แม้ว่า Mfold กับ RNAstructure จะมี F-measure ส่วนใหญ่อยู่ในอันดับที่สองเหมือนกัน แต่ Mfold มีจำนวนข้อมูลที่อยู่ในอันดับสองมากกว่าจึงจัดให้ Mfold อยู่ในอันดับที่ดีกว่า RNAstructure



รูปที่ 4.14 การจัดอันดับ F-measure เมื่อทดสอบกับ Other Ribosomal RNA

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

จากรูปที่ 4.14 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 8 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 4 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งเช่นกันแต่มีเพียงแค่ 3 รายการ ซึ่งน้อยกว่าวิธีการที่งานวิจัยนี้นำเสนอเล็กน้อย โปรแกรม RNAfold ผลการทำนายทั้งหมดมีค่า F-measure สูงเป็นอันดับที่สามและสี่เท่ากันจำนวนอันดับละ 4 รายการ โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 3 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ แม้ว่า Mfold กับ Hybrid-EDAfold จะมี F-measure ส่วนใหญ่อยู่ในอันดับที่หนึ่งเหมือนกัน แต่ Hybrid-EDAfold มีจำนวนข้อมูลที่อยู่ในอันดับหนึ่งมากกว่า จึงจัดให้ Hybrid-EDAfold อยู่ในอันดับที่ดีกว่า Mfold

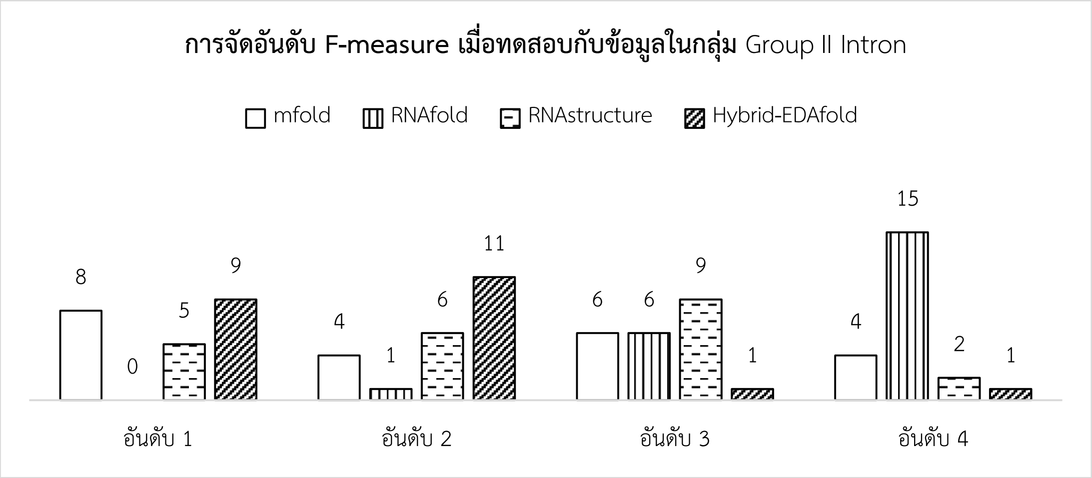


รูปที่ 4.15 การจัดอันดับ F-measure เมื่อทดสอบกับ Other Ribozyme

จากรูปที่ 4.15 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 10 รายการ พบว่าขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 4 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่สี่ โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 6 รายการ อีกสี่รายการที่เหลือมีค่า F-measure สูงเป็นอันดับที่สาม โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 9 รายการ อีก 1 รายการที่เหลืออยู่ในอันดับที่สอง โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 5 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Mfold, Hybrid-EDAfold, RNAstructure และ RNAfold ตามลำดับ แม้ว่า RNAstructure กับ Hybrid-EDAfold จะมี

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

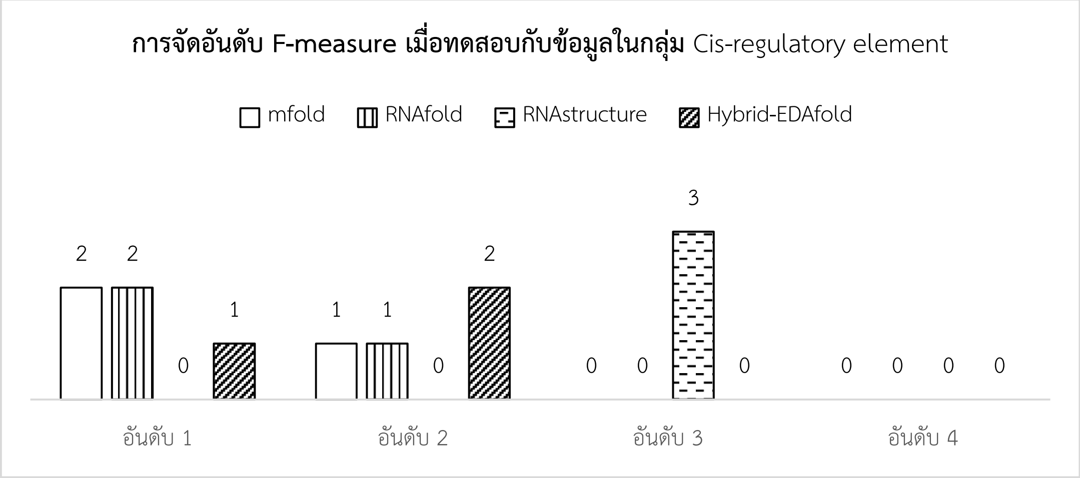
F-measure ส่วนใหญ่อยู่ในอันดับที่สองเหมือนกัน แต่ Hybrid-EDAfold มีจำนวนข้อมูลที่อยู่ในอันดับหนึ่งมากกว่า จึงจัดให้ Hybrid-EDAfold อยู่ในอันดับที่ดีกว่า RNAstructure



รูปที่ 4.16 การจัดอันดับ F-measure เมื่อทดสอบกับ Group II Intron

จากรูปที่ 4.16 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 22 รายการ พบว่า ขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สองจำนวน 11 รายการ มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 9 รายการซึ่งเป็นจำนวนที่มากสุดเมื่อเทียบกับวิธีการอื่น ๆ และ มีค่า F-measure สูงเป็นอันดับที่สามและสี่จำนวนอันดับละ 1 รายการซึ่งเป็นจำนวนที่น้อยสุดเมื่อเทียบกับวิธีการอื่น ๆ จึงทำให้ผลการทำนายบนข้อมูลกลุ่มนี้ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอมีค่า F-measure เฉลี่ยสูงที่สุด โปรแกรม Mfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่หนึ่งจำนวน 8 รายการ ซึ่งน้อยกว่าขั้นตอนวิธี Hybrid-EDAfold โปรแกรม RNAfold ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สี่จำนวน 15 รายการ และไม่มีข้อมูลรายการใดที่มีค่า F-measure สูงเป็นอันดับที่หนึ่ง โปรแกรม RNAstructure ผลการทำนายส่วนใหญ่มีค่า F-measure สูงเป็นอันดับที่สามจำนวน 9 รายการ ดังนั้น สำหรับอาร์เอ็นเอชนิดนี้เรียงลำดับขั้นตอนวิธีที่นำมาเปรียบเทียบจากอันดับหนึ่งไปอันดับสี่ได้เป็น Hybrid-EDAfold, Mfold, RNAstructure และ RNAfold ตามลำดับ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60



รูปที่ 4.17 การจัดอันดับ F-measure เมื่อทดสอบกับ Cis-regulatory element

จากรูปที่ 4.17 ข้อมูลที่ถูกเลือกมาทดสอบในกลุ่มนี้มีจำนวน 3 รายการ สาเหตุที่ข้อมูลในกลุ่มนี้ผ่านการคัดเลือกค่อนข้างน้อยเนื่องจากสายลำดับอาร์เอ็นเอส่วนใหญ่ค่อนข้างสั้นและมีความยาวใกล้เคียงกัน เมื่อใช้เกณฑ์การคัดเลือกที่ได้นำเสนอไปจึงมีข้อมูลที่ผ่านการคัดเลือกน้อยกว่าข้อมูลในกลุ่มอื่น ๆ โดยจากผลการประเมินพบว่าขั้นตอนวิธี Hybrid-EDAfold ผลการทำนายทั้งหมดมีค่า F-measure สูงเป็นอันดับที่หนึ่งและสองจำนวน 1 และ 2 รายการ ตามลำดับ โปรแกรม Mfold และ RNAfold ทั้งสองโปรแกรมให้ผลการทำนายเท่ากันโดยมีค่า F-measure สูงเป็นอันดับที่หนึ่งและสองจำนวน 2 และ 1 รายการตามลำดับ โปรแกรม RNAstructure ผลการทำนายทั้งหมดมีค่า F-measure สูงเป็นอันดับที่สาม ดังนั้น โดยสรุป สำหรับข้อมูลในกลุ่มนี้ขั้นตอนวิธี Hybrid-EDAfold, Mfold และ RNAfold ให้ผลการทำนายโครงสร้างที่ดีใกล้เคียงกัน ส่วน RNAstructure ได้ผลลัพธ์ต่ำกว่าวิธีการอื่น ๆ ประมาณ 6%

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

4.6 สรุปสิ่งที่ได้จากการทดสอบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold

ผลการทดสอบพารามิเตอร์ของขั้นตอนวิธี Hybrid-EDAfold บนอาร์เอ็นเอ 3 ชนิด จำนวน 20 สายลำดับ พบว่า พารามิเตอร์ในส่วนของขนาดประชากรและจำนวนรอบของการวิวัฒนาการมีความอ่อนไหวต่ออาร์เอ็นเอในบางกลุ่ม และขั้นตอนวิธีที่นำเสนอมีแนวโน้มว่าจะลู่เข้าสู่คำตอบเร็ว ประเมินจากชุดพารามิเตอร์ที่ให้ค่าความถูกต้องมากสุดคือ ขนาดประชากรที่ 50 และจำนวนรอบการวิวัฒนาการที่ 100 สาเหตุอาจเนื่องมาจากขั้นตอนวิธีถูกเอนเอียง (bias) ด้วยค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold ส่งผลให้ประสิทธิภาพของขั้นตอนวิธีที่นำเสนอแปรผันตามความแม่นยำของค่าความน่าจะเป็นนี้ นอกจากนี้ ขนาดประชากรและจำนวนรอบที่มากเกินไปอาจทำให้ค่าความถูกต้องในการทำนายต่ำลงอาจเป็นผลมาจากขั้นตอนวิธีที่นำเสนอมีการวิวัฒนาการคำตอบไปสู่โครงสร้างที่มีค่าพลังงานต่ำไปกว่าโครงสร้างที่เป็นคำตอบ

งานวิจัยนี้มีการใช้ประโยชน์จากกลุ่มประชากรของขั้นตอนวิธีเชิงวิวัฒนาการโดยนำเสนอการทำนายหลายโครงสร้างโดยการจัดเก็บโครโมโซมที่มีค่าความเหมาะสมดีสุดที่พบในระหว่างกระบวนการวิวัฒนาการ ซึ่งเมื่อเปรียบเทียบวิธีการที่งานวิจัยนี้นำเสนอกับวิธีการทำนายหลายโครงสร้างแบบที่ใช้ในโปรแกรม Mfold และ RNAstructure พบว่า หากขั้นตอนวิธีที่นำมาเปรียบเทียบเลือกรายงานผลเฉพาะโครงสร้างที่มีค่าพลังงานต่ำสุด ขั้นตอนวิธี Hybrid-EDAfold ให้ผลการทำนายในส่วนของ F-measure เฉลี่ยดีสุด ซึ่งมีผลลัพธ์เพิ่มขึ้นจาก Mfold ประมาณ 11.28% (คำนวณจาก (50.3 - 45.2) / 45.2 * 100) และ เพิ่มขึ้นจาก RNAstructure ประมาณ 18.35% (คำนวณจาก (50.3 – 42.5) / 42.5 * 100) และเมื่อทุกขั้นตอนวิธีที่นำมาเปรียบเทียบมีการรองรับการทำนายหลายโครงสร้าง ทุกขั้นตอนวิธีมีค่า F-measure เฉลี่ยดีขึ้นกว่าการทำนายแค่เพียง 1 โครงสร้างที่มีค่าพลังงานต่ำสุด นั่นคือ Mfold มีค่า F-measure เฉลี่ยดีขึ้นคิดเป็น 20.13% RNAstructute ดีขึ้น 34.59% และ ขั้นตอนวิธี Hybrid-EDAfold ดีขึ้น 24.06% นอกจากนี้ การทดลองในส่วนนี้ยังแสดงให้เห็นว่า แม้ว่า RNAstructure จะเป็นขั้นตอนวิธีที่ทำนายได้โครงสร้างส่วนใหญ่ที่มีค่าพลังงานต่ำสุด แต่กลับมีค่า F-measure เฉลี่ยของโครงสร้างที่ทำนายได้เหล่านั้นแย่กว่าวิธีการอื่น ๆ แสดงให้เห็นว่าในบางอาร์เอ็นเอที่โครงสร้างคำตอบไม่ใช่โครงสร้างที่มีค่าพลังงานต่ำสุด โปรแกรมทำนายโครงสร้างสามารถบรรเทาปัญหานี้ได้โดยการทำนายหลาย ๆ โครงสร้างที่มีค่าพลังงานสูงขึ้นจะทำให้ได้โครงสร้างที่มีความใกล้เคียงกับคำตอบมากยิ่งขึ้น

ผลการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold และขั้นตอนวิธีในกลุ่มของกำหนดการพลวัตกับข้อมูล pre-miRNA ของมนุษย์จำนวน 10 สายลำดับ พบว่า ขั้นตอนวิธีที่นำเสนอมีค่า F-measure ดีกว่าขั้นตอนวิธีที่นำมาเปรียบเทียบจำนวน 9 สายลำดับ มีเพียง pre-miR-16-1 ที่มีค่า F-measure ต่ำกว่าขั้นตอนวิธีที่นำมาเปรียบเทียบเล็กน้อย

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

นอกจากนี้ ขั้นตอนวิธี Hybrid-EDAfold สามารถทำนายโครงสร้างของข้อมูลชุดนี้ได้ถูกต้อง 100% ใน 4 อาร์เอ็นเอ ได้แก่ pre-let-7f-2, pre-miR-17, pre-miR-29a และ pre-miR-30a และ ในภาพรวมประเมินจากค่าเฉลี่ยจากทั้ง 10 ข้อมูล ขั้นตอนวิธี Hybrid-EDAfold ได้ผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ในทุกตัวชี้วัด โดยมีค่าความอ่อนไหวเฉลี่ยเพิ่มขึ้นจากโปรแกรม Mfold, RNAfold และ RNAstructure คิดเป็น 9.23%, 7.84% และ 6.88% ตามลำดับ มีค่าความจำเพาะเฉลี่ยเพิ่มขึ้นจากโปรแกรม Mfold, RNAfold และ RNAstructure คิดเป็น 6.1%, 6.15% และ 4.81% ตามลำดับ และมีค่า F-measure เฉลี่ยเพิ่มขึ้นจากโปรแกรม Mfold, RNAfold และ RNAstructure คิดเป็น 7.7%, 7.11% และ 5.94% ตามลำดับ อย่างไรก็ตาม ผลการทำนายที่ยังไม่ค่อยดีนักในบางอาร์เอ็นเอเป็นผลมาจากข้อผิดพลาดในขั้นตอนของการจัดเตรียมฮีลิกและวิธีการแก้ไขบริเวณของคู่เบสที่มีการแชร์ตำแหน่งเบสบางส่วนร่วมกัน ซึ่งการใช้ค่าความน่าจะเป็นของคู่เบสเพียงอย่างเดียวเป็นเกณฑ์ในการตัดสินว่าจะเลือกเก็บคู่เบสไหนไว้ในโครงสร้างอาจไม่ถูกต้องเสมอไป ประเด็นนี้จะต้องมีการพัฒนาปรับปรุงต่อไปในอนาคต

ผลการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold และขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ ได้แก่ RnaPredict, SARNA-Predict และ TL-PSOfold บนสายลำดับอาร์เอ็นเอ 3 ชนิดจำนวน 20 สายลำดับที่รวบรวมมาจากวรรณกรรมของขั้นตอนวิธีที่นำมาเปรียบเทียบพบว่าวิธีการที่นำเสนอทำผลลัพธ์ได้ดีกว่าวิธีการอื่น ๆ ใน 2 ชนิดอาร์เอ็นเอ คือ 5S Ribosomal RNA และ Group I Intron ในขณะที่ 16S Ribosomal RNA ขั้นตอนวิธี TL-PSOfold เป็นขั้นตอนวิธีที่ทำผลลัพธ์ได้ดีที่สุด โดยได้ผลลัพธ์ดีกว่าวิธีการที่งานวิจัยนี้นำเสนอเล็กน้อย อย่างไรก็ตาม เมื่อเปรียบเทียบขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธี RnaPredict ซึ่งเป็นขั้นตอนวิธีที่มีรากฐานมาจากขั้นตอนวิธีเชิงพันธุกรรมเช่นเดียวกัน พบว่า ขั้นตอนวิธีที่นำเสนอมีค่า F-measure เฉลี่ยเพิ่มขึ้นจาก RnaPredict คิดเป็น 20.62% และขั้นตอนวิธีที่นำเสนอมีค่า F-measure เฉลี่ยเพิ่มขึ้นจาก SARNAR-Predict คิดเป็น 29.34% แสดงให้เห็นว่าการมี 2 ขั้นตอนวิธีประมาณการแจกแจงช่วยกันทำงานและการใช้ทั้งกลุ่มโครโมโซมดีและโครโมโซมด้อยในการปรับปรุงเวกเตอร์ความน่าจะเป็นสามารถช่วยปรับปรุงผลการทำนายโครงสร้างให้ดียิ่งขึ้น

ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีที่อยู่บนพื้นฐานของกำหนดการพลวัตซึ่งเป็นที่นิยมใช้งานในปัจจุบันจำนวน 3 โปรแกรม ได้แก่ Mfold, RNAfold และ RNAstructure โดยทดสอบกับ 14 ชนิดอาร์เอ็นเอซึ่งรวบรวมจากฐานข้อมูล RNA STRAND v2.0 จำนวน 750 อาร์เอ็นเอ การประเมินประสิทธิภาพโดยเฉลี่ยพบว่าขั้นตอนวิธี Hybrid-EDAfold ทำผลลัพธ์ได้ดีกว่าวิธีการอื่น ๆ ที่นำมาเปรียบเทียบใน 12 ชนิดอาร์เอ็นเอ และมีเพียง 2 ชนิด คือ 16S Ribosomal RNA ที่มีค่า F-measure เฉลี่ยต่ำกว่าโปรแกรม Mfold ประมาณ 3 และ Transfer RNA มีค่า F-measure เฉลี่ยต่ำกว่าโปรแกรม RNAstructure ประมาณ 1 นอกจากนี้ ใน

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ภาพรวมซึ่งประเมินจากค่าเฉลี่ยจากทั้ง 14 ชนิด ขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบในทุกตัวชี้วัด กล่าวคือ ขั้นตอนวิธี Hybrid-EDAfold มีค่าเฉลี่ยของความอ่อนไหว ความจำเพาะ และ F-measure เพิ่มขึ้นจาก Mfold คิดเป็น 9.13%, 12.49% และ 11.32% ตามลำดับ เพิ่มขึ้นจาก RNAfold คิดเป็น 25.02%, 31.13% และ 28.77% ตามลำดับ และ เพิ่มขึ้นจาก RNAstructure คิดเป็น 5.34%, 8.53% และ 7.3% ตามลำดับ

นอกจากนี้ การประเมินประสิทธิภาพจากผลการจัดอันดับ F-measure ของขั้นตอนวิธีที่นำมาเปรียบเทียบกันสำหรับอาร์เอ็นเอแต่ละชนิด พบว่า ไม่มีขั้นตอนวิธีใดทำนายโครงสร้างได้ผลลัพธ์ดีสุดในทุกชนิดอาร์เอ็นเอ ในที่นี้ประเมินจากมีค่า F-measure สูงเป็นอันดับที่หนึ่ง กล่าวคือ ขั้นตอนวิธี Hybrid-EDAfold มีผลการจัดอันดับของ F-measure สูงเป็นอันดับที่หนึ่งบนอาร์เอ็นเอ 10 ชนิด และอีก 4 ชนิดอาร์เอ็นเอที่เหลือมีผลการจัดอันดับ F-measure สูงเป็นอันดับที่สอง ได้แก่ 16S Ribosomal RNA, Transfer RNA, Other Ribozyme และ Cis-regulatory element สำหรับโปรแกรม Mfold มีผลการจัดอันดับของ F-measure สูงเป็นอันดับที่หนึ่งบนอาร์เอ็นเอ 3 ชนิด ได้แก่ 16S Ribosomal RNA, Other Ribozyme และ Cis-regulatory element และมีผลการจัดอันดับ F-measure เป็นอันดับที่สี่บนอาร์เอ็นเอในกลุ่ม 5S Ribosomal RNA สำหรับ RNAfold ผลการจัดอันดับของ F-measure ส่วนใหญ่อยู่ในอันดับที่ 4 อาจเนื่องมาจากขั้นตอนวิธีนี้รายงานผลการทำนายโครงสร้างเฉพาะโครงสร้างที่มีค่าพลังงานต่ำสุด แต่อย่างไรก็ตาม สำหรับอาร์เอ็นเอในกลุ่มของ Cis-regulatory element โปรแกรม RNAfold มีผลการจัดอันดับของ F-measure อยู่ในอันดับที่หนึ่ง สำหรับโปรแกรม RNAstructure มีผลการจัดอันดับของ F-measure สูงเป็นอันดับหนึ่งในชนิดอาร์เอ็นเอ Transfer RNA ดังนั้น การเลือกโปรแกรมสำหรับทำนายโครงสร้างให้เหมาะสมกับชนิดของอาร์เอ็นเอที่ต้องการทำนายก็เป็นปัจจัยหนึ่งที่ส่งผลต่อค่าความถูกต้องที่ได้รับ

อย่างไรก็ตาม การทำนายโครงสร้างของอาร์เอ็นเอในกลุ่มของ 16S Ribosomal RNA ยังต้องมีการพัฒนาต่อไป ปัญหาที่พบตอนนี้คือค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold สำหรับข้อมูลในกลุ่มนี้ยังไม่แม่นยำเท่าที่ควร อาจเนื่องมาจากข้อมูลในสายลำดับอาร์เอ็นเอในกลุ่มนี้ค่อนข้างยาวส่งผลให้จำนวนฮีลิกที่สร้างได้มีจำนวนค่อนข้างมาก งานวิจัยสำหรับทำนายโครงสร้างอาร์เอ็นเอในลักษณะที่ทำการตัดสายลำดับอาร์เอ็นเอออกเป็นท่อนสั้น ๆ แล้วทำนายโครงสร้างของแต่ละท่อน จากนั้นนำผลลัพธ์ทั้งหมดมารวมกันเป็น 1 โครงสร้างอาจเป็นอีกแนวทางหนึ่งที่น่าสนใจในการปรับปรุงประสิทธิภาพการทำนายโครงสร้างสำหรับข้อมูลในกลุ่มนี้

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บทที่ 5
สรุปผล

5.1 สรุปผลการวิจัย

เนื่องจากความก้าวหน้าในอุปกรณ์และเทคนิคทางด้านชีวเทคโนโลยีที่มีประสิทธิภาพสูง ชุดข้อมูลมากมายมหาศาลและเป็นข้อมูลที่มีมิติสูงถูกสกัดและรวบรวมจากการวิเคราะห์ยีนและโมเลกุลต่าง ๆ เทคนิคการหาค่าเหมาะสุดแบบคลาสสิกทำการสำรวจได้แค่เฉพาะส่วนที่จำกัดของพื้นที่คำตอบที่เป็นไปได้ทั้งหมด และอาจไม่เพียงพอในการดำเนินการบนพื้นที่การค้นหาขนาดใหญ่เหล่านี้ได้ การใช้เครื่องมือการค้นหาในลักษณะที่อาศัยกลุ่มประชากร (population-based) หรือ การค้นหาเชิงสุ่ม (randomized search) เป็นอีกทางเลือกหนึ่งที่เอาชนะข้อจำกัดเหล่านี้และสามารถสำรวจพื้นที่คำตอบที่มากมายได้ดีขึ้น ขั้นตอนวิธีค้นหาเชิงวิวัฒนาการเป็นอัลกอริทึมที่สำคัญสำหรับแก้ปัญหาการหาค่าเหมาะสุดและถูกประยุกต์ในงานด้านต่าง ๆ เช่น การขนส่ง อุตสาหกรรม รวมทั้งปัญหาทางชีวสารสนเทศ ขั้นตอนวิธีประมาณการแจกแจงเป็นขั้นตอนวิธีเชิงวิวัฒนาการแบบใหม่ที่ใช้แบบจำลองความน่าจะเป็นในการสร้างประชากรคำตอบแทนการใช้ตัวดำเนินการพันธุกรรม เช่น การไขว้เปลี่ยนหรือการกลายพันธุ์ มีการเรียนรู้จากกลุ่มคำตอบคุณภาพดีที่พบในรุ่นก่อนหน้าและใช้ข้อมูลนี้เพื่อปรับปรุงแบบจำลองความน่าจะเป็นดังกล่าวเพื่อนำทางกระบวนการค้นหาไปสู่คำตอบที่คาดว่าน่าจะดีขึ้นในรุ่นถัด ๆ ไป จากความรู้ของผู้วิจัย ขั้นตอนวิธีประมาณการแจกแจงถูกประยุกต์ใช้ในงานทางด้านชีวสารสนเทศทั้งในส่วนของการวิเคราะห์โครงสร้างของยีน รวมทั้งการออกแบบโปรตีนและการทำนายโครงสร้างของโปรตีนแต่ยังไม่พบการนำขั้นตอนวิธีประมาณการแจกแจงไปประยุกต์ใช้ในการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอซึ่งเป็นปัญหาที่สำคัญและได้รับความสนใจอย่างมากในงานวิจัยทางด้านชีวเทคโนโลยีในปัจจุบัน

งานวิจัยนี้นำเสนอขั้นตอนวิธี Hybrid-EDAfold ซึ่งเป็นขั้นตอนวิธีเชิงวิวัฒนาการที่อยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจงสำหรับทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอจาก 1 สายลำดับอาร์เอ็นเอ วิธีการที่นำเสนอประกอบด้วย 2 ขั้นตอนย่อย คือ การระบุฮีลิกที่เป็นไปได้ทั้งหมดจากสายลำดับอาร์เอ็นเอที่เป็นข้อมูลนำเข้า และการทำนายโครงสร้างของอาร์เอ็นเอด้วยขั้นตอนวิธี Hybrid-EDAfold โดยผลลัพธ์ของวิธีการที่นำเสนอจะรายงานเป็นชุดของโครงสร้างที่ทำนายได้ ซึ่งประกอบด้วยโครงสร้างที่มีค่าพลังงานต่ำสุดและโครงสร้างที่มีค่าพลังงานต่ำรองลงมาเพื่อจัดการกับปัญหาความไม่สมบูรณ์ของค่าพารามิเตอร์ของแบบจำลองที่ใช้ในการคำนวณค่าพลังงานของโครงสร้างอาร์เอ็นเอที่ทำให้บางอาร์เอ็นเอมีโครงสร้างคำตอบที่ไม่ใช่โครงสร้างที่มีค่า

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

พลังงานต่ำสุด ผลจากการรองรับการทำนายหลายโครงสร้างทำให้เพิ่มโอกาสที่พบโครงสร้างที่ทำนายได้ใกล้เคียงกับโครงสร้างที่เป็นคำตอบมากยิ่งขึ้น

ในขั้นตอนการจัดเตรียมฮีลิกงานวิจัยนี้ใช้ข้อมูลค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold โดยจัดกลุ่มคู่เบสที่มีค่าความน่าจะเป็นมากกว่า 0 และมีตำแหน่งติดกันเป็น 1 ฮีลิก จากนั้นใช้เฉพาะหมายเลขของฮีลิกในขั้นตอนของการทำนายโครงสร้าง จากการประเมินพบว่าการจัดเตรียมฮีลิกด้วยวิธีการเช่นนี้ให้ผลลัพธ์ที่ดี ภายในเซตของฮีลิกที่สร้างได้มีฮีลิกชิ้นที่พบในโครงสร้างคำตอบอยู่ไม่ต่ำกว่า 80% แต่อย่างไรก็ตาม วิธีการดังกล่าวยังมีข้อจำกัดบางประการคือฮีลิกที่สร้างได้บางชิ้นมีจำนวนคู่เบสมากเกินกว่าคู่เบสของฮีลิกที่พบในโครงสร้างคำตอบส่งผลให้ฮีลิกบางชิ้นมีการแชร์ตำแหน่งของคู่เบสบางส่วนร่วมกัน ซึ่งงานวิจัยนี้ได้นำเสนอวิธีการแก้ปัญหานี้โดยยอมให้ฮีลิกที่มีคู่เบสบางส่วนแชร์ตำแหน่งเบสร่วมกันสามารถถูกเลือกมาสร้างโครงสร้างได้และใช้ความน่าจะเป็นของคู่เบสเป็นเกณฑ์ในการพิจารณาว่าบริเวณที่มีการแชร์ตำแหน่งร่วมกันนั้นคู่เบสไหนจะถูกคงไว้และแก้ไขอีกคู่เบสให้กลายเป็นเบสอิสระเพื่อปรับปรุงข้อผิดพลาดของขั้นตอนการสร้างฮีลิกและทำให้โครงสร้างที่ทำนายได้มีความถูกต้องอยู่เสมอ จากการทดสอบกับข้อมูลอาร์เอ็นเอจำนวนหนึ่งแสดงให้เห็นว่าการดำเนินการแก้ไขในลักษณะนี้มีส่วนช่วยปรับปรุงประสิทธิภาพการทำนายโครงสร้างให้มีความถูกต้องมากยิ่งขึ้น

งานวิจัยนี้ได้แปลงปัญหาการทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอไปเป็นปัญหาการหาค่าเหมาะสมสุดเชิงการจัด โดยทำการเข้ารหัสฮีลิกแต่ละชิ้นที่สร้างได้และใช้เฉพาะหมายเลขฮีลิกในขั้นตอนการทำนายโครงสร้างด้วยขั้นตอนวิธี Hybrid-EDAfold โดยพยายามเลือกชุดย่อยของฮีลิกที่จะนำมาประกอบกันเป็น 1 โครงสร้าง จากนั้นทำการประเมินค่าความเหมาะสมของโครงสร้างที่ทำนายได้ด้วยวิธีอุณหพลศาสตร์ที่นิยมใช้งานกันอย่างแพร่หลายอ้างอิงตาม Turner 2004 ภายใต้สมมุติฐานที่ว่าในสภาวะสมดุลโครงสร้างที่พบในธรรมชาติมักเป็นโครงสร้างที่มีค่าพลังงานต่ำ ขั้นตอนวิธีประมาณการแจกแจงมาตรฐานมีขั้นตอนการทำงาน ดังนี้ 1) กำหนดค่าเริ่มต้นให้แบบจำลองความน่าจะเป็น 2) สุ่มประชากรอ้างอิงตามแบบจำลองความน่าจะเป็น 3) ประเมินค่าความเหมาะสมของประชากร 4) คัดเลือกกลุ่มประชากรย่อย 5) ปรับปรุงแบบจำลองความน่าจะเป็นโดยใช้กลุ่มประชากรย่อยที่ถูกคัดเลือก 6) ทำซ้ำขั้นตอนที่ 2 – 5 ไปจนกระทั่งพบเงื่อนไขจบการทำงาน

โดยทั่วไป การกำหนดค่าเริ่มต้นให้แบบจำลองความน่าจะเป็นของขั้นตอนวิธีประมาณการแจกแจงมาตรฐานค่าในแต่ละสมาชิกจะถูกกำหนดให้เท่ากันเพื่อแทนการแจกแจงแบบสม่ำเสมอ (uniform distribution) แต่ในมุมมองของผู้วิจัย ข้อดีของขั้นตอนวิธีประมาณการแจกแจงที่ต่างจากขั้นตอนวิธีเชิงวิวัฒนาการทั่วไป คือ แบบจำลองความน่าจะเป็นซึ่งเก็บข้อมูลเชิงสถิติและถูกใช้ในการสร้างประชากรคำตอบสามารถถูกกำหนดค่าไว้ล่วงหน้า หากผู้ใช้มีความรู้หรือมีข้อมูลที่เกี่ยวข้องกับปัญหาโดยข้อมูลเหล่านั้นอาจเป็นข้อมูลที่ยังไม่สมบูรณ์ก็ได้ แล้วนำข้อมูลดังกล่าวกำหนดค่าเริ่มต้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ให้กับแบบจำลองและในระหว่างกระบวนการวิวัฒนาการของขั้นตอนวิธีประมาณการแจกแจงแบบจำลองนี้ก็จะถูกปรับปรุงอ้างอิงตามกลุ่มคำตอบที่ถูกคัดเลือก การดำเนินการในลักษณะนี้สามารถช่วยลดต้นทุนในการค้นหาได้ทำให้สามารถจัดการแบบจำลองเพื่อให้ได้คำตอบที่น่าพึงพอใจมากขึ้น แบบจำลองความน่าจะเป็นที่งานวิจัยนี้เลือกใช้คือเวกเตอร์ความน่าจะเป็นที่มีความยาวเท่ากับจำนวนฮีลิกที่สร้างได้จากขั้นตอนการจัดเตรียมฮีลิก แต่ละสมาชิกในเวกเตอร์แทนความน่าจะเป็นที่ฮีลิกชิ้นนี้จะเป็นฮีลิกชิ้นที่พบในโครงสร้างคำตอบ โดยงานวิจัยนี้เลือกใช้ความน่าจะเป็นเฉลี่ยของคู่เบสที่พบในแต่ละฮีลิกเป็นค่าเริ่มต้นในการกำหนดค่าให้กับสมาชิกในเวกเตอร์ความน่าจะเป็น ซึ่งผลจากการทดสอบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold พบว่าการกำหนดค่าเริ่มต้นให้เวกเตอร์ความน่าจะเป็นในลักษณะนี้ให้ผลลัพธ์ที่ดีกว่าการกำหนดค่าเริ่มต้นของแต่ละสมาชิกในเวกเตอร์เป็น 0.5 แบบที่นิยมทำในขั้นตอนวิธีประมาณการแจกแจงมาตรฐานทั่วไป หากความน่าจะเป็นของคู่เบสที่ได้จากโปรแกรม RNAfold มีความแม่นยำขั้นตอนวิธี Hybrid-EDAfold จะลู่เข้าสู่คำตอบโดยเร็วและให้ค่าความถูกต้องในการทำนายโครงสร้างที่สูงกว่าการกำหนดค่าเริ่มต้นเป็น 0.5 ฮีลิกที่มีความน่าจะเป็นสูง ๆ มีโอกาสมากกว่าที่จะถูกเลือกมาประกอบโครงสร้างเป็นลำดับต้น ๆ ของการทำนาย เมื่อฮีลิกเหล่านี้ถูกเลือกฮีลิกที่เข้ากันได้กับฮีลิกที่ถูกเลือกไปแล้วจะมีจำนวนลดน้อยลงเรื่อย ๆ เนื่องจากฮีลิกที่ขัดแย้งกับฮีลิกชิ้นที่ถูกเลือกไปแล้วจะถูกตัดออกจากเซตของฮีลิกที่สามารถเลือกได้และในท้ายที่สุดขั้นตอนวิธี Hybrid-EDAfold ก็มีแนวโน้มเลือกฮีลิกชิ้นที่ถูกต้องมาประกอบร่วมกันจนครบทั้งโครงสร้าง อย่างไรก็ตาม ในบางอาร์เอ็นเอข้อมูลค่าความน่าจะเป็นที่ได้จาก RNAfold ยังมีความคลาดเคลื่อน กล่าวคือ ฮีลิกชิ้นที่เป็นคำตอบมีค่าความน่าจะเป็นต่ำกว่าชิ้นที่ไม่ใช่คำตอบ อนาคตหากมีความรู้ด้านอื่นนอกเหนือจากค่าความน่าจะเป็นของคู่เบสนี้ก็สามารถนำมากำหนดค่าเริ่มต้นให้กับเวกเตอร์ความน่าจะเป็นก็จะช่วยส่งเสริมให้ขั้นตอนวิธีที่นำเสนอมีค่าความถูกต้องในการทำนายที่ดียิ่งขึ้นและลดระยะเวลาในการค้นหาของอัลกอริทึมได้

ขั้นตอนวิธี Hybrid-EDAfold ประกอบด้วย 2 ขั้นตอนวิธีย่อย ทั้งคู่อยู่บนพื้นฐานของขั้นตอนวิธีประมาณการแจกแจง เริ่มต้นการทำงานด้วยขั้นตอนวิธีประมาณการแจกแจงตัวแรกแทนด้วย *EDA-G* เมื่อใดก็ตามที่ *EDA-G* ไม่สามารถทำนายโครงสร้างที่ให้ค่าความเหมาะสมดีขึ้นติดต่อกันเป็นจำนวน m รุ่น การทำงานจะสลับไปที่ขั้นตอนวิธีประมาณการแจกแจงตัวที่สองแทนด้วย *EDA-L* และสลับการทำงานกันไปเช่นนี้จนกว่าจะพบเงื่อนไขจบการทำงาน จึงกล่าวได้ว่ารูปแบบการสลับการทำงานของขั้นตอนวิธีประมาณการแจกแจงทั้งคู่เป็นแบบปรับตัวได้ ข้อดีคือ เมื่อต้องดำเนินการกับข้อมูลสายลำดับอาร์เอ็นเอที่มีความหลากหลายทั้งในแง่ของชนิดอาร์เอ็นเอและความยาวของสายลำดับ ขั้นตอนวิธีที่นำเสนอสามารถปรับตัวให้เข้ากับข้อมูลนำเข้า จากการสังเกตสำหรับข้อมูลสายลำดับอาร์เอ็นเอที่มีความยาวไม่มาก *EDA-G* ถูกเรียกทำงานเพียงไม่กี่รุ่นก็จะสลับไปทำงานด้วย *EDA-L* เนื่องจากไม่พบคำตอบที่มีคุณภาพดีขึ้น แต่ในกรณีสายลำดับอาร์เอ็นเอที่ค่อนข้างยาวการทำงานจะ

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

ดำเนินการอยู่บน *EDA-G* เป็นจำนวนรุ่นที่มากกว่า เนื่องจากเมื่อสายลำดับยาวขึ้นส่งผลให้จำนวนฮีลิกที่สร้างได้มีจำนวนมากขึ้นอัลกอริทึมจึงใช้ระยะเวลาที่นานกว่าที่จะสำรวจทั่วทั้งปริภูมิค้นหาและเมื่อไม่สามารถหาคำตอบที่ดีขึ้นได้แล้วจึงย้ายไปทำการค้นหาในระดับโลคอลด้วย *EDA-L* และเป็นเช่นนี้ไปจนพบเงื่อนไขจบการทำงาน

โดยแต่ละขั้นตอนวิธีประมาณการแจกแจงมีพฤติกรรมการค้นหาคำตอบที่แตกต่างกัน *EDA-G* ถูกออกแบบให้ทำการค้นหาทั่วทั้งปริภูมิการค้นหาให้มากที่สุด ดังนั้นในขั้นตอนการสร้างประชากรจึงใช้การสุ่มเลือกหมายเลขฮีลิกจากเซตของฮีลิกที่สามารถเลือกได้แบบไม่ใส่คืนหมายความว่าในแต่ละรุ่นฮีลิกหมายเลขใดถูกเลือกมาสร้างในโครโมโซมหนึ่งแล้วจะถูกกำกับไว้ไม่ให้มีโอกาสถูกสุ่มเลือกมาสร้างในโครโมโซมตัวต่อไปอีก ในขณะที่ *EDA-L* ถูกออกแบบให้ทำการค้นหาในระดับโลคอลกล่าวคือสร้างประชากรโดยการกลายพันธุ์โครโมโซมบรรพบุรุษด้วยการสุ่มฮีลิกบางชิ้นออกจากโครงสร้างของบรรพบุรุษและสุ่มเลือกฮีลิกชิ้นอื่นที่เข้ากันได้มาใส่เพิ่มเติมเพื่อผลิตโครโมโซมลูก

นอกเหนือจากความแตกต่างในขั้นตอนการสร้างประชากรแล้วทั้ง 2 ขั้นตอนวิธีประมาณการแจกแจงนี้ยังมีความแตกต่างกันในขั้นตอนของการปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นอีกด้วย ซึ่งเวกเตอร์ความน่าจะเป็นถูกใช้เป็นแบบจำลองทางสถิติเพื่อควบคุมโอกาสที่ฮีลิกแต่ละชิ้นจะถูกเลือกมาสร้างโครงสร้าง โดยเวกเตอร์ความน่าจะเป็นจะถูกปรับปรุงทุกครั้งหลังจากขั้นตอนการประเมินค่าความเหมาะสมในแต่ละรุ่น หาก *EDA-G* กำลังทำงานอยู่เวกเตอร์ความน่าจะเป็นจะถูกปรับปรุงโดยการจำแนกโครโมโซมในประชากรออกเป็น 3 กลุ่ม ได้แก่ โครโมโซมกลุ่มดี โครโมโซมกลุ่มด้อย และโครโมโซมกลุ่มที่ไม่นำมาพิจารณา จากนั้นสกัดหมายเลขฮีลิกที่พบในกลุ่มโครโมโซมดีและกลุ่มโครโมโซมด้อย โดยทุกสมาชิกในเวกเตอร์ที่สัมพันธ์กับหมายเลขฮีลิกที่พบในกลุ่มโครโมโซมดีจะได้ความน่าจะเป็นเพิ่มจากเดิมเท่ากับค่าอัตราการเรียนรู้ที่กำหนด ในทางตรงกันข้ามสมาชิกในเวกเตอร์ที่สัมพันธ์กับหมายเลขฮีลิกที่พบในกลุ่มโครโมโซมด้อยจะถูกลดความน่าจะเป็นลงเท่ากับค่าอัตราการเรียนรู้เช่นเดียวกัน สังเกตว่าขั้นตอนวิธีที่งานวิจัยนี้นำเสนอมีความแตกต่างจากขั้นตอนวิธีประมาณการแจกแจงมาตรฐานทั่วไปที่มีการใช้ทั้งโครโมโซมดีและด้อยในการปรับปรุงเวกเตอร์ความน่าจะเป็นและกรณีที่ *EDA-L* ทำงานเวกเตอร์ความน่าจะเป็นจะถูกปรับปรุงโดยแข่งขันกันระหว่างบรรพบุรุษและลูกโดยเลือกเฉพาะคู่ที่บรรพบุรุษทำการกลายพันธุ์ได้ลูกที่มีค่าความเหมาะสมดีขึ้น เนื่องจากลูกเกิดจากการสุ่มฮีลิกบางชิ้นในบรรพบุรุษทิ้งไปและสุ่มฮีลิกชิ้นอื่นมาใส่เพิ่มเติมแปลว่าฮีลิกชิ้นที่ถูกสุ่มทิ้งเป็นกลุ่มฮีลิกที่ไม่ดีและฮีลิกที่ถูกสุ่มเติมเป็นกลุ่มฮีลิกที่ดี จากนั้นทำการรวบรวมและปรับปรุงค่าในเวกเตอร์ความน่าจะเป็นในลักษณะเดียวกับ *EDA-G*

ผลการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold และขั้นตอนวิธีในกลุ่มของกำหนดการพลวัตด้วยข้อมูล pre-miRNA ของมนุษย์จำนวน 10 สายลำดับพบว่าขั้นตอนวิธีที่นำเสนอมีค่า F-measure ดีกว่าขั้นตอนวิธีที่นำมาเปรียบเทียบจำนวน 9 สายลำดับ

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

มีเพียง pre-miR-16-1 ที่มีค่า F-measure ต่ำกว่าขั้นตอนวิธีที่นำมาเปรียบเทียบประมาณ 1% นอกจากนี้ขั้นตอนวิธี Hybrid-EDAfold สามารถทำนายโครงสร้างของข้อมูลชุดนี้ได้ถูกต้อง 100% ใน 4 อาร์เอ็นเอ ได้แก่ pre-let-7f-2, pre-miR-17, pre-miR-29a และ pre-miR-30a ในขณะที่ Mfold ทำนายโครงสร้างได้ถูกต้อง 100% ใน 1 อาร์เอ็นเอ คือ pre-miR-17 และ RNAfold กับ RNAstructure ทำนายโครงสร้างได้ถูกต้อง 100% ใน 2 อาร์เอ็นเอ คือ pre-miR-17 และ pre-miR-30a ในภาพรวมประเมินจากค่าเฉลี่ยจากทั้ง 10 ข้อมูล ขั้นตอนวิธี Hybrid-EDAfold ได้ผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ในทุกตัวชี้วัด โดยมีค่า F-measure เฉลี่ยสูงกว่าโปรแกรม Mfold, RNAfold และ RNAstructure คือ 6.66, 6.18 และ 5.22 ตามลำดับ อย่างไรก็ตาม ผลการทำนายที่ยังไม่ค่อยดีนักในบางอาร์เอ็นเอเป็นผลมาจากข้อผิดพลาดในขั้นตอนของการจัดเตรียมฮีลิกและการแก้ไขบริเวณของคู่เบสที่มีการแชร์ตำแหน่งเบสร่วมกันซึ่งการใช้แค่เพียงค่าความน่าจะเป็นของคู่เบสเป็นเกณฑ์ในการตัดสินว่าจะเลือกเก็บคู่เบสไหนไว้ในโครงสร้างอาจไม่ถูกต้องเสมอไปประเด็นนี้จะต้องมีการพัฒนาปรับปรุงต่อไปในอนาคต

ผลการเปรียบเทียบประสิทธิภาพการทำนายโครงสร้างของขั้นตอนวิธี Hybrid-EDAfold และขั้นตอนวิธีทางเมตาฮิวริสติกอื่น ๆ ได้แก่ RnaPredict, SARNA-Predict และ TL-PSOfold บนสายลำดับอาร์เอ็นเอ 3 ชนิด จำนวน 20 สายลำดับที่รวบรวมมาจากวรรณกรรมของขั้นตอนวิธีที่นำมาเปรียบเทียบพบว่าวิธีการที่นำเสนอทำผลลัพธ์ได้ดีกว่าวิธีการอื่น ๆ ใน 2 ชนิดอาร์เอ็นเอ คือ 15S Ribosomal RNA และ Group I Intron ในขณะที่ 16S Ribosomal RNA ขั้นตอนวิธี TL-PSOfold เป็นขั้นตอนวิธีที่ทำผลลัพธ์ได้ดีที่สุด โดยได้ผลลัพธ์ดีกว่าวิธีการที่งานวิจัยนี้นำเสนอคิดเป็น 0.63% การเปรียบเทียบในหัวข้อนี้เป็นเพียงผลการประเมินในเบื้องต้น เนื่องจากแต่ละขั้นตอนวิธีที่นำมาเปรียบเทียบรายงานผลไม่ครบทั้ง 20 สายลำดับ อย่างไรก็ตาม เมื่อเปรียบเทียบขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธี RnaPredict พบว่าขั้นตอนวิธีที่นำเสนอมีค่า F-measure เฉลี่ยดีกว่า RnaPredict เท่ากับ 10.7 และขั้นตอนวิธีที่นำเสนอมีค่า F-measure เฉลี่ยดีกว่า SARNA-Predict เท่ากับ 14.2 แสดงให้เห็นว่าการมี 2 ขั้นตอนวิธีประมาณการแจกแจงช่วยกันทำงานและการใช้ทั้งกลุ่มโครโมโซมดีและกลุ่มโครโมโซมด้อยในการปรับปรุงเวกเตอร์ความน่าจะเป็นสามารถช่วยปรับปรุงผลการทำนายโครงสร้างให้ดียิ่งขึ้น

ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี Hybrid-EDAfold กับขั้นตอนวิธีที่อยู่บนพื้นฐานของกำหนดการพลวัตซึ่งเป็นที่นิยมใช้งานในปัจจุบันจำนวน 3 โปรแกรม ได้แก่ Mfold, RNAfold และ RNAstructure โดยทดสอบกับ 14 ชนิดอาร์เอ็นเอซึ่งรวบรวมจากฐานข้อมูล RNA STRAND v2.0 จำนวน 750 อาร์เอ็นเอ พบว่าขั้นตอนวิธี Hybrid-EDAfold ทำผลลัพธ์ได้ดีกว่าวิธีการอื่น ๆ ที่นำมาเปรียบเทียบใน 12 ชนิดอาร์เอ็นเอและมีเพียง 2 ชนิด คือ 16S Ribosomal RNA ที่มีค่า F-measure เฉลี่ยต่ำกว่าโปรแกรม Mfold เท่ากับ 3.08 และ Transfer RNA มีค่า F-measure ต่ำ

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

กว่าโปรแกรม RNAstructure เท่ากับ 1.04 นอกจากนี้ในภาพรวมซึ่งประเมินจากค่าเฉลี่ยจากทั้ง 14 ชนิดขั้นตอนวิธี Hybrid-EDAfold มีผลการทำนายดีกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบในทุกตัวชี้วัด กล่าวคือ ขั้นตอนวิธี Hybrid-EDAfold มีค่าเฉลี่ยของค่าความอ่อนไหว ค่าความจำเพาะ และ F-measure ดีกว่า Mfold เท่ากับ 5.18, 6.41 และ 5.96 ตามลำดับ ดีกว่า RNAfold เท่ากับ 12.39, 13.71 และ 13.1 ตามลำดับ และ ดีกว่า RNAstructure เท่ากับ 3.14, 4.54 และ 3.99 ตามลำดับ อย่างไรก็ตาม การทำนายโครงสร้างของอาร์เอ็นเอในกลุ่มของ 16S Ribosomal RNA ยังต้องมีการพัฒนาต่อไป ปัญหาที่พบตอนนี้คือค่าความน่าจะเป็นของคู่เบสที่คำนวณได้จากโปรแกรม RNAfold สำหรับข้อมูลในกลุ่มนี้ยังไม่แม่นยำเท่าที่ควร นอกจากนี้ข้อมูลในสายลำดับอาร์เอ็นเอในกลุ่มนี้ค่อนข้างยาวส่งผลให้จำนวนฮีลิกที่สร้างได้มีจำนวนค่อนข้างมาก หากมีความรู้ที่สามารถใช้เพื่อคัดกรองจำนวนฮีลิกที่สร้างได้ให้มีจำนวนลดน้อยลงหรือนำเสนอวิธีการระบุบริเวณที่เป็นฮีลิกในแนวทางอื่นที่มีประสิทธิภาพมากยิ่งขึ้นน่าจะช่วยปรับปรุงประสิทธิภาพการทำนายโครงสร้างสำหรับข้อมูลในกลุ่มนี้ให้ดียิ่งขึ้น

อ้างอิงจากหลาย ๆ งานวิจัยที่รายงานผลการทำนายเป็นชุดของโครงสร้างแทนการรายงานผลแค่เฉพาะ 1 โครงสร้างที่มีค่าพลังงานต่ำสุดเพื่อลดข้อจำกัดที่เกิดจากความไม่แม่นยำของพารามิเตอร์ที่ใช้ในการคำนวณค่าพลังงาน งานวิจัยนี้จึงใช้ประโยชน์จากกลุ่มประชากรของขั้นตอนวิธีเชิงวิวัฒนาการโดยนำเสนอการทำนายหลายโครงสร้างด้วยการจัดเก็บโครโมโซมที่มีค่าความเหมาะสมดีสุดที่พบในระหว่างกระบวนการวิวัฒนาการจำนวน n โครโมโซมไว้ในอาไคว์ เมื่อเปรียบเทียบวิธีการทำนายหลายโครงสร้างที่งานวิจัยนี้นำเสนอกับวิธีการทำนายหลายโครงสร้างที่ใช้ในโปรแกรม Mfold และ RNAstructure พบว่าหากขั้นตอนวิธีที่นำมาเปรียบเทียบเลือกรายงานผลเฉพาะโครงสร้างที่มีค่าพลังงานต่ำสุด ขั้นตอนวิธี Hybrid-EDAfold ให้ผลการทำนายในส่วนของ F-measure เฉลี่ยดีสุดซึ่งดีกว่า Mfold คิดเป็น 5.1 และดีกว่า RNAstructure คิดเป็น 7.8 และเมื่อทุกขั้นตอนวิธีที่นำมาเปรียบเทียบมีการรองรับการทำนายหลายโครงสร้าง ทุกขั้นตอนวิธีมีค่า F-measure เฉลี่ยดีขึ้นกว่าการทำนายแค่เพียง 1 โครงสร้างที่มีค่าพลังงานต่ำสุดคือ Mfold มีค่า F-measure เฉลี่ยดีขึ้นคิดเป็น 20.13% RNAstructute ดีขึ้นคิดเป็น 34.59% และ ขั้นตอนวิธี Hybrid-EDAfold ดีขึ้นคิดเป็น 24.06% นอกจากนี้การทดลองในส่วนนี้ยังแสดงให้เห็นว่าแม้ว่า RNAstructure จะเป็นขั้นตอนวิธีที่ทำนายได้โครงสร้างส่วนใหญ่ที่มีค่าพลังงานต่ำสุดแต่ก็ไม่ใช่วิธีการที่ทำนายโครงสร้างได้ค่า F-measure เฉลี่ยสูงสุด สอดคล้องกับหลาย ๆ งานวิจัยที่นำเสนอว่าในบางอาร์เอ็นเอโครงสร้างที่มีค่าพลังงานต่ำสุดอาจไม่ใช่โครงสร้างที่ตรงกับคำตอบแต่โปรแกรมทำนายโครงสร้างสามารถบรรเทาปัญหานี้ได้โดยการทำนายหลาย ๆ โครงสร้างที่มีค่าพลังงานสูงขึ้นจะทำให้ได้โครงสร้างที่มีความใกล้เคียงกับคำตอบมากยิ่งขึ้น

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

กล่าวโดยสรุป ขั้นตอนวิธี Hybrid-EDAfold มีข้อดีคือขั้นตอนวิธีที่นำเสนอมีการใช้แบบจำลองความน่าจะเป็นซึ่งสามารถใส่ความรู้ที่เกี่ยวข้องกับปัญหาเพื่อจัดการแบบจำลองในทิศทางที่ได้คำตอบที่น่าพึงพอใจมากขึ้นและเมื่อจบการทำงานของขั้นตอนวิธีที่นำเสนอแบบจำลองความน่าจะเป็นที่ได้สามารถถูกนำมาใช้สำหรับการตีความ หรือ วิเคราะห์ เพื่อเปิดเผยข้อมูลที่เป็นประโยชน์กับการแก้ปัญหานั้น นอกจากนี้ ขั้นตอนวิธีที่นำเสนอมีการเรียนรู้จากทั้งสองด้านคือกลุ่มโครโมโซมที่มีคุณภาพคำตอบดีและกลุ่มโครโมโซมที่มีคุณภาพคำตอบด้อยทำให้เกิดกระบวนการเปรียบเทียบเพื่อนำทางการค้นหาไปในทิศทางที่นำไปสู่คำตอบที่ดีขึ้น และการเรียนรู้จากกลุ่มคำตอบก็ช่วยลดความเสี่ยงที่จะปรับปรุงค่าของแบบจำลองผิดพลาดเนื่องจากใช้หลักฐานเป็นจำนวนมากในการตัดสินใจ แต่อย่างไรก็ตาม ขั้นตอนวิธีที่งานวิจัยนี้นำเสนอยังมีข้อจำกัดบางประการที่จะต้องปรับปรุงต่อไป เช่น ขั้นตอนการจัดเตรียมฮีลิกยังไม่สมบูรณ์ ฟังก์ชันวัตถุประสงค์ที่เลือกใช้ยังมีข้อจำกัดบางประการอันเนื่องมาจากความไม่สมบูรณ์ของพารามิเตอร์ที่ใช้คำนวณค่าพลังงาน และ การทำนายหลายโครงสร้างยังสามารถปรับปรุงให้ดีขึ้นได้โดยการเพิ่มการพิจารณาในประเด็นอื่น ๆ นอกเหนือจากแค่ค่าพลังงานของโครงสร้าง เช่น ความหลากหลายในเชิงรูปร่าง การปรับปรุงประเด็นต่าง ๆ เหล่านี้อาจช่วยพัฒนาให้ขั้นตอนวิธีที่นำเสนอสามารถทำนายโครงสร้างทุติยภูมิของอาร์เอ็นเอได้ถูกต้องมากยิ่งขึ้น

5.2 งานวิจัยในอนาคต

5.2.1 ปรับปรุงประสิทธิภาพของขั้นตอนวิธีในการทำนายโครงสร้างของอาร์เอ็นเอในกลุ่ม 16S Ribosomal RNA ให้มีความแม่นยำมากยิ่งขึ้น เช่น อาจแบ่งสายลำดับออกเป็นส่วนย่อย ทำการทำนายแต่ละส่วน แล้วค่อยนำทุกส่วนมาประกอบรวมกันเป็นโครงสร้างผลลัพธ์

5.2.2 ปรับปรุงขั้นตอนการจัดเตรียมฮีลิกให้สามารถระบุตำแหน่งของฮีลิกได้มีความแม่นยำมากยิ่งขึ้น

5.2.3 ปรับปรุงฟังก์ชันวัตถุประสงค์ที่ใช้ประเมินคุณภาพของโครงสร้างที่ทำนายได้ให้มีคุณภาพสอดคล้องกับโครงสร้างที่เป็นคำตอบมากยิ่งขึ้น เช่น การใช้หลายฟังก์ชันวัตถุประสงค์

5.2.4 ปรับปรุงวิธีการทำนายหลายโครงสร้างโดยเพิ่มเติมเกณฑ์การพิจารณาอื่น ๆ นอกเหนือจากค่าพลังงานเพียงอย่างเดียว เช่น ความคล้ายคลึงกันของโครงสร้างในแง่ของรูปร่างเพื่อให้โครงสร้างที่ถูกจัดเก็บในอาไคว์มีความหลากหลายมากยิ่งขึ้นและอาจทำให้ได้โครงสร้างที่ตรงกับคำตอบมากยิ่งขึ้นภายใต้ขนาดของอาไคว์ที่ไม่ใหญ่นัก

5.2.5 ปรับปรุงเพิ่มเติมขั้นตอนวิธีให้สามารถรองรับการทำนายโครงสร้างในส่วนของซูโดนอท

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

บรรณานุกรม

[1] Li, F., *Genome-wide analysis of rna secondary structure in eukaryotes.* 2013.

[2] Nussinov, R., et al., *Algorithms for loop matchings.* SIAM Journal on Applied mathematics, 1978. 35(1): p. 68-82.

[3] Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic acids research, 1981. 9(1): p. 133-148.

[4] Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic acids research, 2003. 31(13): p. 3406-3415.

[5] Gruber, A.R., et al., *The vienna RNA websuite.* Nucleic acids research, 2008. 36(suppl 2): p. W70-W74.

[6] Lorenz, R., et al., *ViennaRNA Package* 2.0. Algorithms for Molecular Biology, 2011. 6(1): p. 26.

[7] Anderson, J.W., et al., *Evolving stochastic context-free grammars for RNA secondary structure prediction.* BMC bioinformatics, 2012. 13(1): p. 78.

[8] Knudsen, B. and J. Hein, *Pfold: RNA secondary structure prediction using stochastic context-free grammars.* Nucleic acids research, 2003. 31(13): p. 3423-3428.

[9] Song, D. and Z. Deng. *A BP-SCFG based approach for RNA secondary structure prediction with consecutive bases dependency and their relative positions information*. in *International Symposium on Bioinformatics Research and Applications*. 2007. Springer.

[10] Ding, Y., *Statistical and Bayesian approaches to RNA secondary structure prediction.* Rna, 2006. 12(3): p. 323-331.

[11] McCaskill, J.S., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 1990. 29(6-7): p. 1105-1119.

[12] Wiese, K.C., A.A. Deschenes, and A.G. Hendriks, *RnaPredict—an evolutionary algorithm for RNA secondary structure prediction.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2008. 5(1): p. 25-41.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[13] Grypma, P. and H.H. Tsang. *SARNA-Predict: Using adaptive annealing schedule and inversion mutation operator for RNA secondary structure prediction*. in *Computational Intelligence in Multi-Criteria Decision-Making (MCDM),* 2014 *IEEE Symposium on*. 2014. IEEE.

[14] Lalwani, S., R. Kumar, and N. Gupta, *An efficient two-level swarm intelligence approach for RNA secondary structure prediction with bi-objective minimum free energy scores.* Swarm and Evolutionary Computation, 2016. 27: p. 68-79.

[15] El Fatmi, A., M.A. Bekri, and S. Benhlima. *RNA secondary structure prediction based on genetic algorithm and comparative approach*. in *Optimization and Applications (ICOA),* 2018 *4th International Conference on*. 2018. IEEE.

[16] Legendre, A., E. Angel, and F. Tahi, *Bi-objective integer programming for RNA secondary structure prediction with pseudoknots.* BMC bioinformatics, 2018. 19(1): p. 13.

[17] Mühlenbein, H. and G. Paass. *From recombination of genes to the estimation of distributions I. Binary parameters*. in *International conference on parallel problem solving from nature*. 1996. Springer.

[18] Eiben, A.E. and J.E. Smith, *Introduction to evolutionary computing*. Vol. 53. 2003: Springer.

[19] Armañanzas, R., et al., *A review of estimation of distribution algorithms in bioinformatics.* BioData mining, 2008. 1(1): p. 6.

[20] Hauschild, M. and M. Pelikan, *An introduction and survey of estimation of distribution algorithms.* Swarm and Evolutionary Computation, 2011. 1(3): p. 111-128.

[21] Lozano, J.A., et al., *Towards a new evolutionary computation: advances on estimation of distribution algorithms*. Vol. 192. 2006: Springer.

[22] Inza, I., et al., *Feature subset selection by Bayesian network-based optimization.* Artificial intelligence, 2000. 123(1-2): p. 157-184.

[23] Saeys, Y., et al., *Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction.* Bioinformatics, 2003. 19(suppl_2): p. ii179-ii188.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[24] Santana, R., *Advances in probabilistic graphical models for optimization and learning. Applications in protein modelling.* 2006.

[25] Santana, R., P. Larranaga, and J.A. Lozano. *Protein folding in 2-dimensional lattices with estimation of distribution algorithms*. in *International Symposium on Biological and Medical Data Analysis*. 2004. Springer.

[26] Santana, R., P. Larrañaga, and J.A. Lozano, *Protein folding in simplified models with estimation of distribution algorithms.* IEEE transactions on Evolutionary Computation, 2008. 12(4): p. 418-438.

[27] Santana, R., P. Larrañaga, and J.A. Lozano, *Combining variable neighborhood search and estimation of distribution algorithms in the protein side chain placement problem.* Journal of Heuristics, 2008. 14(5): p. 519-547.

[28] Chen, S.-H., et al., *Guidelines for developing effective estimation of distribution algorithms in solving single machine scheduling problems.* Expert Systems with Applications, 2010. 37(9): p. 6441-6451.

[29] Wang, K., S. Choi, and H. Lu, *A hybrid estimation of distribution algorithm for simulation-based scheduling in a stochastic permutation flowshop.* Computers & Industrial Engineering, 2015. 90: p. 186-196.

[30] Liu, H., L. Gao, and Q. Pan, *A hybrid particle swarm optimization with estimation of distribution algorithm for solving permutation flowshop scheduling problem.* Expert Systems with Applications, 2011. 38(4): p. 4348-4360.

[31] Tzeng, Y.-R., C.-L. Chen, and C.-L. Chen, *A hybrid EDA with ACS for solving permutation flow shop scheduling.* The International Journal of Advanced Manufacturing Technology, 2012. 60(9-12): p. 1139-1147.

[32] Peter, F., *Recent advances in RNA folding.* Journal of biotechnology, 2017.

[33] Andronescu, M., et al., *RNA STRAND: the RNA secondary structure and statistical analysis database.* BMC bioinformatics, 2008. 9(1): p. 340.

[34] Krol, J., et al., *Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design.* Journal of Biological Chemistry, 2004. 279(40): p. 42230-42239.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[35] Echegoyen, C., et al., *The impact of exact probabilistic learning algorithms in edas based on bayesian networks*, in *Linkage in Evolutionary Computation*. 2008, Springer. p. 109-139.

[36] Hauschild, M., et al. *Analyzing probabilistic models in hierarchical BOA on traps and spin glasses*. in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. 2007. ACM.

[37] Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic acids research, 2010: p. gkq1027.

[38] Shabalina, S.A., A.Y. Ogurtsov, and N.A. Spiridonov, *A periodic pattern of mRNA secondary structure created by the genetic code.* Nucleic Acids Research, 2006. 34(8): p. 2428-2437.

[39] Rich, A. and U. RajBhandary, *Transfer RNA: molecular structure, sequence, and properties.* Annual review of biochemistry, 1976. 45(1): p. 805-860.

[40] Dixon, M.T. and D.M. Hillis, *Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis.* Molecular Biology and Evolution, 1993. 10(1): p. 256-267.

[41] Sharma, D., et al., *RNA: Structure, Prediction, and Visualization Tools*, in *Intelligent Communication, Control and Devices*. 2018, Springer. p. 335-345.

[42] Eddy, S.R., *Non–coding RNA genes and the modern RNA world.* Nature Reviews Genetics, 2001. 2(12): p. 919.

[43] Tanner, D.R., et al., *Genetic analysis of the structure and function of transfer messenger RNA pseudoknot* 1. Journal of Biological Chemistry, 2006. 281(15): p. 10561-10566.

[44] Forbes, D.J., T.B. Kornberg, and M.W. Kirschner, *Small nuclear RNA transcription and ribonucleoprotein assembly in early Xenopus development.* The Journal of cell biology, 1983. 97(1): p. 62-72.

[45] Lin, S.-L., J.D. Miller, and S.-Y. Ying, *Intronic microRNA (miRNA).* BioMed Research International, 2006. 2006.

[46] Carthew, R.W. and E.J. Sontheimer, *Origins and mechanisms of miRNAs and siRNAs.* Cell, 2009. 136(4): p. 642-655.

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[47] Liu, T.-T., et al., *A global identification and analysis of small nucleolar RNAs and possible intermediate-sized non-coding RNAs in Oryza sativa.* Molecular plant, 2013. 6(3): p. 830-846.

[48] Gupta, S., et al., *Antisense technology.* International Journal of Pharmaceutical Sciences Review and Research, 2011. 9(2): p. 38-45.

[49] Zwieb, C., et al., *A nomenclature for all signal recognition particle RNAs.* Rna, 2005. 11(1): p. 7-13.

[50] Turner, D.H. and D.H. Mathews, *NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure.* Nucleic acids research, 2009: p. gkp892.

[51] Edwards, A.L., A.D. Garst, and R.T. Batey, *Determining structures of RNA aptamers and riboswitches by X-ray crystallography.* Nucleic Acid and Peptide Aptamers: Methods and Protocols, 2009: p. 135-163.

[52] Bothe, J.R., et al., *Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy.* Nature methods, 2011. 8(11): p. 919-931.

[53] Weeks, K.M., *Advances in RNA structure analysis by chemical probing.* Current opinion in structural biology, 2010. 20(3): p. 295-304.

[54] Wan, Y., et al., *Understanding the transcriptome through RNA structure.* Nature Reviews Genetics, 2011. 12(9): p. 641-655.

[55] Seetin, M.G. and D.H. Mathews, *RNA structure prediction: an overview of methods.* Bacterial Regulatory RNA: Methods and Protocols, 2012: p. 99-122.

[56] Bellaousov, S. and D.H. Mathews, *ProbKnot: fast prediction of RNA secondary structure including pseudoknots.* Rna, 2010. 16(10): p. 1870-1880.

[57] Theimer, C.A., et al., *Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot.* Journal of molecular biology, 1998. 279(3): p. 545-564.

[58] Mathews, D.H. and D.H. Turner, *Experimentally derived nearest-neighbor parameters for the stability of RNA three-and four-way multibranch loops.* Biochemistry, 2002. 41(3): p. 869-880.

[59] Bellaousov, S., et al., *RNAstructure: web servers for RNA secondary structure prediction and analysis.* Nucleic acids research, 2013. 41(W1): p. W471-W474.

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[60] Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC bioinformatics, 2010. 11(1): p. 129.

[61] Do, C.B., D.A. Woods, and S. Batzoglou, *CONTRAfold: RNA secondary structure prediction without physics-based models.* Bioinformatics, 2006. 22(14): p. e90-e98.

[62] Lu, Z.J., J.W. Gloor, and D.H. Mathews, *Improved RNA secondary structure prediction by maximizing expected pair accuracy.* Rna, 2009. 15(10): p. 1805-1813.

[63] Mathews, D.H., *Revolutions in RNA secondary structure prediction.* Journal of molecular biology, 2006. 359(3): p. 526-532.

[64] Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(19): p. 7287-7292.

[65] Wuchty, S., et al., *Complete suboptimal folding of RNA and the stability of secondary structures.* Biopolymers, 1999. 49(2): p. 145-165.

[66] Ding, Y. and C.E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction.* Nucleic acids research, 2003. 31(24): p. 7280-7301.

[67] Lyngsø, R.B. and C.N. Pedersen, *RNA pseudoknot prediction in energy-based models.* Journal of computational biology, 2000. 7(3-4): p. 409-427.

[68] Cao, S. and S.-J. Chen, *Predicting RNA pseudoknot folding thermodynamics.* Nucleic acids research, 2006. 34(9): p. 2634-2652.

[69] Akutsu, T., *Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.* Discrete Applied Mathematics, 2000. 104(1): p. 45-62.

[70] Ruan, J., G.D. Stormo, and W. Zhang, *An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots.* Bioinformatics, 2004. 20(1): p. 58-66.

[71] Doshi, K.J., et al., *Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction.* BMC bioinformatics, 2004. 5(1): p. 105.

[72] Hamada, M., *RNA secondary structure prediction from multi-aligned sequences.* RNA Bioinformatics, 2015: p. 17-38.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[73] Gardner, P.P. and R. Giegerich, *A comprehensive comparison of comparative RNA structure prediction approaches.* BMC bioinformatics, 2004. 5(1): p. 140.

[74] Bernhart, S.H., et al., *RNAalifold: improved consensus structure prediction for RNA alignments.* BMC bioinformatics, 2008. 9(1): p. 474.

[75] Bernhart, S.H. and I.L. Hofacker, *From consensus structure prediction to RNA gene finding.* Briefings in functional genomics & proteomics, 2009. 8(6): p. 461-471.

[76] Sankoff, D., *Simultaneous solution of the RNA folding, alignment and protosequence problems.* SIAM Journal on Applied Mathematics, 1985. 45(5): p. 810-825.

[77] Gorodkin, J., L.J. Heyer, and G.D. Stormo, *Finding the most significant common sequence and structure motifs in a set of RNA sequences.* Nucleic acids research, 1997. 25(18): p. 3724-3732.

[78] Havgaard, J.H., E. Torarinsson, and J. Gorodkin, *Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix.* PLoS Comput Biol, 2007. 3(10): p. e193.

[79] Shapiro, B.A. and K. Zhang, *Comparing multiple RNA secondary structures using tree comparisons.* Computer applications in the biosciences: CABIOS, 1990. 6(4): p. 309-318.

[80] Steffen, P., et al., *RNAshapes: an integrated RNA analysis package based on abstract shapes.* Bioinformatics, 2006. 22(4): p. 500-503.

[81] Höchsmann, M., B. Voss, and R. Giegerich, *Pure multiple RNA secondary structure alignments: a progressive profile approach.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2004. 1(1): p. 53-62.

[82] Pelikan, M., M.W. Hauschild, and F.G. Lobo, *Introduction to estimation of distribution algorithms.* MEDAL Report, 2012(2012003).

[83] Goldberg, D.E., *Genetic Alogorithms in Search.* Optimization & Machine Learning, 1989.

[84] Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by simulated annealing.* science, 1983. 220(4598): p. 671-680.

[85] Juels, A., S. Baluja, and A. Sinclair, *The equilibrium genetic algorithm and the role of crossover.* Unpublished manuscript, 1993.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[86] Baluja, S., *Population-based incremental learning. a method for integrating genetic search based function optimization and competitive learning*. 1994, Carnegie-Mellon Univ Pittsburgh Pa Dept Of Computer Science.

[87] Harik, G.R., F.G. Lobo, and D.E. Goldberg, *The compact genetic algorithm.* IEEE transactions on evolutionary computation, 1999. 3(4): p. 287-297.

[88] De Bonet, J.S., C.L. Isbell Jr, and P.A. Viola. *MIMIC: Finding optima by estimating probability densities*. in *Advances in neural information processing systems*. 1997.

[89] Baluja, S. and S. Davies, *Using Optimal Dependency-Trees for Combinatorial Optimization: Learning the Structure of the Search Space*. 1997, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[90] Pelikan, M. and H. Mühlenbein, *The bivariate marginal distribution algorithm*, in *Advances in Soft Computing*. 1999, Springer. p. 521-535.

[91] Mühlenbein, H. and T. Mahnig, *FDA-A scalable evolutionary algorithm for the optimization of additively decomposed functions.* Evolutionary computation, 1999. 7(4): p. 353-376.

[92] Etxeberria, R. *Global optimization using Bayesian networks*. in *Proc. 2nd Symposium on Artificial Intelligence (CIMAF-99)*. 1999.

[93] Pelikan, M., D.E. Goldberg, and E. Cantú-Paz. *BOA: The Bayesian optimization algorithm*. in *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1*. 1999. Morgan Kaufmann Publishers Inc.

[94] Harik, G., *Linkage learning via probabilistic modeling in the ECGA.* Urbana, 1999. 51(61): p. 801.

[95] Wattanapornprom, W., et al. *Multi-objective combinatorial optimisation with coincidence algorithm*. in *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*. 2009. IEEE.

[96] Nussinov, R. and A.B. Jacobson, *Fast algorithm for predicting the secondary structure of single-stranded RNA.* Proceedings of the National Academy of Sciences, 1980. 77(11): p. 6309-6313.

[97] Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic acids research, 2003. 31(13): p. 3429-3431.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[98] Zuker, M., *Calculating nucleic acid secondary structure*. Current opinion in structural biology, 2000. 10(3): p. 303-310.

[99] Ding, Y., C.Y. Chan, and C.E. Lawrence, *RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.* Rna, 2005. 11(8): p. 1157-1166.

[100] Wiese, K.C. and E. Glen, *A Permutation Based Genetic Algorithm for RNA Secondary Structure Prediction.* HIS, 2002. 87: p. 173-182.

[101] Rivas, E. and S.R. Eddy, *A dynamic programming algorithm for RNA structure prediction including pseudoknots.* Journal of molecular biology, 1999. 285(5): p. 2053-2068.

[102] Ren, J., et al., *HotKnots: heuristic prediction of RNA secondary structures including pseudoknots.* Rna, 2005. 11(10): p. 1494-1504.

[103] Dirks, R.M. and N.A. Pierce, *A partition function algorithm for nucleic acid secondary structure including pseudoknots.* Journal of computational chemistry, 2003. 24(13): p. 1664-1677.

[104] Reeder, J. and R. Giegerich, *Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics.* BMC bioinformatics, 2004. 5(1): p. 104.

[105] Gultyaev, A.P., F. Van Batenburg, and C.W. Pleij, *The computer simulation of RNA folding pathways using a genetic algorithm.* Journal of molecular biology, 1995. 250(1): p. 37-51.

[106] Jabbari, H., A. Condon, and S. Zhao, *Novel and efficient RNA secondary structure prediction using hierarchical folding.* Journal of Computational Biology, 2008. 15(2): p. 139-163.

[107] Wiese, K.C. and A.G. Hendriks. *RNA pseudoknot prediction via an evolutionary algorithm*. in *Evolutionary Computation,* 2009. *CEC'*09. *IEEE Congress on*. 2009. IEEE.

[108] Wiese, K.C. and A. Hendriks, *Comparison of P-RnaPredict and mfold—Algorithms for RNA secondary structure prediction.* Bioinformatics, 2006. 22(8): p. 934-942.

[109] Sato, K., et al., *IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming.* Bioinformatics, 2011. 27(13): p. i85-i93.

2430366845 CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[110] Huang, X. and H. Ali, *High sensitivity RNA pseudoknot prediction.* Nucleic acids research, 2007. 35(2): p. 656-663.

[111] Knudsen, B. and J. Hein, *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.* Bioinformatics, 1999. 15(6): p. 446-454.

[112] Mathews, D.H. and D.H. Turner, *Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.* Journal of molecular biology, 2002. 317(2): p. 191-203.

[113] Xu, Z. and D.H. Mathews, *Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences.* Bioinformatics, 2011. 27(5): p. 626-632.

[114] Harmanci, A.O., G. Sharma, and D.H. Mathews, *Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign.* BMC bioinformatics, 2007. 8(1): p. 130.

[115] Tabaska, J.E., et al., *An RNA folding method capable of identifying pseudoknots and base triples.* Bioinformatics, 1998. 14(8): p. 691-699.

[116] Meyer, I.M. and I. Miklós, *SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework.* PLoS Comput Biol, 2007. 3(8): p. e149.

[117] Hofacker, I.L., M. Fekete, and P.F. Stadler, *Secondary structure prediction for aligned RNA sequences.* Journal of molecular biology, 2002. 319(5): p. 1059-1066.

[118] Witwer, C., I.L. Hofacker, and P.F. Stadler, *Prediction of consensus RNA secondary structures including pseudoknots.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2004. 1(2): p. 66-77.

[119] Touzet, H. and O. Perriquet, *CARNAC: folding families of related RNAs.* Nucleic acids research, 2004. 32(suppl 2): p. W142-W145.

[120] Seetin, M.G. and D.H. Mathews, *TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots.* Bioinformatics, 2012. 28(6): p. 792-798.

[121] Harmanci, A.O., G. Sharma, and D.H. Mathews, *TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences.* BMC bioinformatics, 2011. 12(1): p. 108.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

[122] Doose, G. and D. Metzler, *Bayesian sampling of evolutionarily conserved RNA secondary structures with pseudoknots.* Bioinformatics, 2012. 28(17): p. 2242-2248.

[123] Metzler, D. and M.E. Nebel, *Predicting RNA secondary structures with pseudoknots by MCMC sampling.* Journal of mathematical biology, 2008. 56(1): p. 161-181.

[124] Bindewald, E. and B.A. Shapiro, *RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers.* Rna, 2006. 12(3): p. 342-352.

[125] Deschenes, A., *A genetic algorithm for RNA secondary structure prediction using stacking energy thermodynamic models*. 2005, School of Interactive Arts and Technology-Simon Fraser University.

[126] Montaseri, S., M. Ganjtabesh, and F. Zare-Mirakabad, *Evolutionary algorithm for RNA secondary structure prediction based on simulated SHAPE data.* PloS one, 2016. 11(11): p. e0166965.

[127] Tong, K.-K., et al. *GAknot: RNA secondary structures prediction with pseudoknots using Genetic Algorithm*. in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* 2013 *IEEE Symposium on*. 2013. IEEE.

[128] Li, J., et al., *RGRNA: prediction of RNA secondary structure based on replacement and growth of stems.* Computer methods in biomechanics and biomedical engineering, 2017. 20(12): p. 1261-1272.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60

CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60
2430366845

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวสุภาวดี ศรีคำดี
วัน เดือน ปี เกิด	27 กุมภาพันธ์ 2531
สถานที่เกิด	ชลบุรี
วุฒิการศึกษา	สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ จากมหาวิทยาลัยบูรพา ในปีการศึกษา 2552 และ สำเร็จการศึกษาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ จากมหาวิทยาลัยบูรพา ในปีการศึกษา 2555 หลังจากนั้นได้เข้าศึกษาในหลักสูตรวิศวกรรมศาสตรดุษฎีบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557 โดยได้รับทุนการศึกษา จากคณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
ผลงานตีพิมพ์	Srikamdee, S., Wattanapornprom, W., Chongstitvatana, P., "RNA Secondary Structure Prediction with Coincidence Algorithm," 16th Int Symposium on Communications and Information Technologies (ISCIT 2016), Qingdao, China, 26-28 September 2016.

2430366845
CU iThesis 5771474521 dissertation / recv: 17122561 10:13:21 / seq: 60