การสำรวจงานวิจัยด้านชีวสารสนเทศและการประยุกต์ใช้

A Survey on Bioinformatics and Its Application

ณลักขณา คิดเหมาะ, พยุง มีสัจ และ สุนันทา สดสี
คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Nalakkhana Khitmoh, Payung Meesad and Sunantha Sodsee
Faculty of Information Technology, King Mongkut's University of Technology North Bangkok

Abstract

Bioinformatics is a new discipline considered has roots in biology and information technology and with the advancement of scientific technology can produce massive amounts of data in a short time. As a result, computer science, has entered a critical role in the study and analysis of experiments, including biological database management. To study the structure of biomolecules, it takes long time to study and requires expertise, some situations require fast data processing. So bioinformatics therefore aims to develop new methods to meet the needs of biologists. This article aims to offer an introduction to bioinformatics. The application of computer techniques to analyze data associated with large biomolecules in the benefits derived from bioinformatics applications in various fields.

Keywords: *Bioinformatics, Biological database, Protein sequence, Nucleotide sequence, Bioinformatics tool*

บทคัดย่อ

ชีวสารสนเทศศาสตร์ถือเป็นสาขาวิชาที่ใหม่ มีรากฐานมาจากวิชาชีววิทยาและมีการจัดการข้อมูลด้วยเทคโนโลยีคอมพิวเตอร์ ด้วยความเจริญก้าวหน้าของเทคโนโลยีทางวิทยาศาสตร์ทำให้สามารถผลิตข้อมูลจำนวนมหาศาลได้ในเวลาอันรวดเร็ว ส่งผลให้วิทยาการคอมพิวเตอร์เข้ามามีบทบาทสำคัญอย่างยิ่งในการศึกษาและวิเคราะห์ผลจากการทดลอง รวมทั้งการจัดการฐานข้อมูลทางชีวภาพ ในการศึกษาโครงสร้างของสารชีวโมเลกุลนั้น จะต้องใช้เวลาในการศึกษานานและต้องอาศัยความชำนาญของนักวิจัย ซึ่งบางสถานการณ์ต้องการความรวดเร็วในการประมวลผลข้อมูล ดังนั้นชีวสารสนเทศจึงมีจุดมุ่งหมายในการพัฒนาวิธีการใหม่ เพื่อตอบสนองความต้องการของนักชีววิทยา โดยบทความนี้มีวัตถุประสงค์เพื่อสรุปองค์ความรู้ของชีวสารสนเทศ การประยุกต์ใช้เทคนิคทางคอมพิวเตอร์ในการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับสารชีวโมเลกุลในขนาดใหญ่ รวมถึงประโยชน์ของการนำชีวสารสนเทศไปประยุกต์ใช้ในด้านต่าง ๆ ทั้งนี้เพื่อเป็นแนวทางสำหรับผู้ที่สนใจให้สามารถศึกษาชีวสารสนเทศศาสตร์

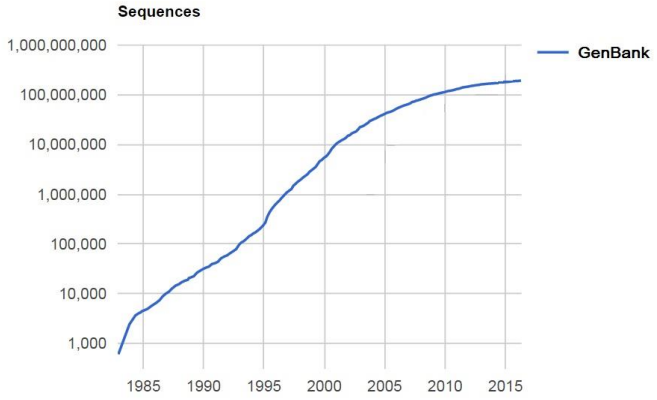
คำสำคัญ: *ชีวสารสนเทศ, ฐานข้อมูลทางชีวภาพ, ลำดับโปรตีน, ลำดับนิวคลีโอไทด์, เครื่องมือชีวสารสนเทศ*

1. บทนำ

ชีวสารสนเทศศาสตร์เป็นสาขาสหวิทยาการที่เติบโตอย่างรวดเร็วโดยได้แรงหนุนจากการพัฒนาที่ไม่หยุดยั้งทั้งในด้านชีววิทยาและวิทยาศาสตร์คอมพิวเตอร์ในช่วงสี่ทศวรรษที่ผ่านมา จากจำนวนของข้อมูลทางชีวภาพที่เพิ่มขึ้นซึ่งถูกสร้างโดยการใช้เทคโนโลยีเอ็นจีเอสหรือ Next generation sequencing ทำให้จำนวนลำดับนิวคลีโอไทด์ทั้งหมดในขณะนี้เพิ่มขึ้นอย่างมหาศาล ข้อมูลขนาดใหญ่นี้ทำให้เกิดการค้นพบรูปแบบต่าง ๆ และมีการตั้งสมมติฐานใหม่ ในขณะเดียวกันข้อมูลเหล่านี้ก็มีความแตกต่างกัน ดังนั้นการจัดการข้อมูลขนาดใหญ่ต้องใช้เทคนิคขั้นสูงที่มีประสิทธิภาพ วิธีการคำนวณที่ถูกต้องและเครื่องมือคอมพิวเตอร์ต่าง ๆ ชีวสารสนเทศจึงเป็นเทคโนโลยีที่มีประสิทธิภาพสำหรับการจัดการ, การค้นหาและการวิเคราะห์ข้อมูลทางชีวภาพ นอกจากนี้ยังเป็นส่วนเติมเต็มวิธีการทดลองแบบดั้งเดิมในการวิจัยด้านชีววิทยา

ในเวลาต่อมาได้มีการค้นพบสารพันธุกรรม ตัวอย่างเช่น ดีเอ็นเอ หรือยีน (Reichhardt, 1999) เดือนกันยายน ปี 2015 ฐานข้อมูลของ GenBank มีลำดับของนิวคลีโอไทด์จำนวน 202,237,081,559 ลำดับ แสดงดังภาพที่ 1 และฐานข้อมูล SWISS-PROT มีลำดับเบสโปรตีนจำนวน 549,646 ลำดับ (Bairoch & Apweiler, 2000) โดยเฉลี่ยฐานข้อมูลเหล่านี้จะเพิ่มขึ้นเป็นสองเท่าทุก 18 เดือน และนอกจากนี้ยังมีข้อมูลมากมายที่ได้จากการวิจัยหลากหลายสาขา เช่น ทางด้านจุล

ชีววิทยา อณูชีววิทยา และการค้นคว้าทางการแพทย์ ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่มีความหลากหลายและมีปริมาณมาก



ภาพที่ 1 อัตราการเจริญเติบโตของฐานข้อมูล GenBank ตั้งแต่ปี 1985 – 2015 (National Center for Biotechnology Information, U.S. National Library of Medicine, 2015)

ด้วยปริมาณข้อมูลที่มากมายดังกล่าวนี้ จึงจำเป็นที่จะต้องมีการพัฒนาเครื่องมือ หรือโปรแกรมประยุกต์ต่าง ๆ ที่ใช้ในการบริหารจัดการข้อมูลที่ซับซ้อนและกำลังเพิ่มขึ้นอย่างรวดเร็วเหล่านี้ (Fleischmann, et al., 1995) โดยมีคุณสมบัติในการจัดเก็บ เปรียบเทียบ วิเคราะห์ คำนวณและประเมินผลข้อมูลที่ได้จากการสืบค้นฐานข้อมูล ซึ่งความรู้ที่ได้จากการศึกษาดังกล่าวสามารถเข้าใจถึงความหมายของรหัสพันธุกรรม รูปแบบและหน้าที่การทำงานของยีน (Parsons & Jhonson, 1993) วิเคราะห์และออกแบบโครงสร้างจากรหัสพันธุกรรม สุดท้ายสามารถนำไปประยุกต์ใช้ในทางการแพทย์ เพื่อให้เกิดการค้นคว้าการค้นพบยารักษาโรคตัวใหม่ การหาสาเหตุของการเกิดโรคเพื่อวินิจฉัยโรค ซึ่งจัดว่ามีความสำคัญเป็นอย่างมากในการพัฒนาให้มีความรวดเร็ว ถูกต้องและแม่นยำ

ในบทความนี้เป็นการนำเสนอหัวข้อที่น่าสนใจในการศึกษาชีวสารสนเทศ เริ่มต้นด้วยการทำความรู้จักกับชีวสารสนเทศ เพื่อให้เข้าใจในความหมายและจุดมุ่งหมายของชีวสารสนเทศซึ่งมี 3 ประการ และฐานข้อมูลทางชีวภาพ ประกอบด้วย ข้อมูลลำดับเบสของโปรตีน และข้อมูลลำดับนิวคลีโอไทด์ จากนั้นจะกล่าวถึงหัวข้อเรื่องการวิเคราะห์ข้อมูลทางชีวภาพโดยได้อธิบายการวิเคราะห์ข้อมูลและยกตัวอย่างเครื่องมือหรือโปรแกรมที่ใช้ในการวิเคราะห์ และส่งท้ายกับประโยชน์ของการนำชีวสารสนเทศไปประยุกต์ใช้งาน ได้แก่ การศึกษาความสัมพันธ์เชิงวิวัฒนาการของสิ่งมีชีวิต, การออกแบบยา เป็นต้น

2. ชีวสารสนเทศและจุดมุ่งหมาย

อณูชีววิทยา (Molecular biology) คือ การศึกษาในระดับโมเลกุลที่เกี่ยวกับโมเลกุลต่าง ๆ ของสิ่งมีชีวิต (Altman, 2001) โดยจะเน้นศึกษาโครงสร้างของโมเลกุล, การทำงานของยีน(gene)และโมเลกุลที่เกี่ยวข้อง, ปฏิสัมพันธ์กันระหว่างดีเอ็นเอ(DNA), อาร์เอ็นเอ(RNA), โปรตีน, โมเลกุลอื่น ๆ ที่เกี่ยวข้อง, ระบบต่าง ๆ ภายในเซลล์รวมถึงกลไกการควบคุมการทำงานของสิ่งเหล่านี้ อณูชีววิทยา (Molecular biology)เป็นสาขาที่มีความผสมผสานระหว่างชีววิทยาและเคมี สาขาพันธุศาสตร์, สาขาชีวเคมี และสาขาจุลชีววิทยา เป็นต้น

สารสนเทศ (Informatics) เป็นศาสตร์ที่ศึกษาเกี่ยวกับสารสนเทศและการคำนวณเพื่อคาดการณ์ในอนาคต หรือก็คือวิทยาศาสตร์และเทคโนโลยีในการประมวลผลข้อมูลโดยใช้คอมพิวเตอร์ (Chen, et al., 1999) ที่ว่าด้วยการรวบรวม, การเก็บรักษา และประมวลผล เพื่อนำมาใช้ประโยชน์ตามจุดประสงค์ ซึ่งการจัดเก็บข้อมูลที่ดี จะต้องมีการกำหนดรูปแบบของข้อมูลให้มีลักษณะง่ายต่อการจัดเก็บ การเรียกใช้งาน และมีรูปแบบเดียวกันอย่างมีระบบ

ดังนั้นคำว่าชีวสารสนเทศ (Bioinformatics) หมายถึง การบูรณาการของหลายสาขาวิชาโดยใช้ความรู้จาก อณูชีววิทยา ชีวเคมี คณิตศาสตร์ประยุกต์,สถิติศาสตร์, สารสนเทศศาสตร์ และวิทยาการคอมพิวเตอร์ ที่ประกอบด้วยการจัดเก็บข้อมูลทางชีววิทยา ประมวลผลข้อมูลทางชีววิทยา (Li & Homer, 2010) ส่วนหนึ่งที่สำคัญคือการพัฒนาโปรแกรมประยุกต์ในการสืบค้น เพื่อใช้ในการหาคำตอบและแก้ปัญหาทางชีววิทยา หรือทำนายความเป็นไปได้ทางชีววิทยาโดยการสร้างแบบจำลอง (Goldberg, 1989) จึงทำให้เกิดศาสตร์ใหม่ เช่น จีโนมิกส์ (Genomics) โปรตีโอมิกส์ (Proteomics) เมตะโบโลมิกส์ (Metabolomics) ฯลฯ

จุดมุ่งหมายของชีวสารสนเทศประกอบด้วย 3 ประการ คือ ประการแรกชีวสารสนเทศจะจัดการข้อมูลให้อยู่ในรูปแบบที่ง่ายต่อการใช้งานข้อมูลที่มีอยู่สำหรับนักวิจัย และการเพิ่มข้อมูลใหม่ที่ได้จากการค้นพบ เช่น ฐานข้อมูลของโปรตีนสำหรับการสร้างรูปแบบ 3 มิติของโครงสร้างโมเลกุล เป็นต้น (Guo, et al., 2010)

ประการที่สองของชีวสารสนเทศคือ การพัฒนาระบบและเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลให้มีความถูกต้อง ตัวอย่างเช่น นักวิจัยต้องการเปรียบเทียบลำดับโปรตีนที่สนใจกับลำดับโปรตีนที่มีอยู่ในฐานข้อมูล ได้แก่ โปรแกรม FASTA (PEARSON & LIPMAN, 1988) และ BLAST (Altschul, et al., 1997) ซึ่งโปรแกรมที่ใช้ในการทำเช่นนี้ต้องมีความสามารถขั้นสูงกว่าการเปรียบเทียบข้อความอย่างง่าย และจำเป็นต้องแสดงให้เห็นความแตกต่างอย่างมีนัยสำคัญของลำดับเบสสองลำดับที่ทำการศึกษา ในการพัฒนาเครื่องมือเหล่านี้ต้องอาศัยความเชี่ยวชาญและความเข้าใจในการคำนวณและทฤษฎีทางชีวภาพ

จุดมุ่งหมายประการที่สามของชีวสารสนเทศเพื่อวิเคราะห์ข้อมูลทางชีวภาพโดยใช้กลไกที่ซับซ้อนและแปลงข้อมูลให้อยู่ในรูปแบบที่มีความหมายทางชีวภาพ (Bhandari, et al., 1996) ในการศึกษาทางชีวภาพแบบดั้งเดิมนั้นการตรวจสอบระบบของแต่ละบุคคลและการเปรียบเทียบจะมีเพียงไม่กี่สาขาที่เกี่ยวข้อง แต่ชีวสารสนเทศศาสตร์ในปัจจุบันนี้ได้รับอนุญาตให้ทำการวิเคราะห์ได้จาก

ทั่วโลกและเปรียบเทียบข้อมูลที่มีอยู่จากหลายสาขา เห็นได้ชัดว่าการพัฒนาวิธีการในการประเมินความแตกต่างระหว่างสายพันธุ์ของสารชีวโมเลกุลและระบุความคล้ายคลึงกันสารชีวโมเลกุลซึ่งมีแหล่งที่มาที่แตกต่างกันเป็นสิ่งที่สำคัญในการจัดการข้อมูลที่มีปริมาณมาก

3. ฐานข้อมูลทางชีวภาพ

ฐานข้อมูลทางชีวภาพเป็นการรวบรวมข้อมูลที่ได้จากการทดลองและการวิจัยทางชีววิทยา ประกอบด้วยข้อมูลที่เกี่ยวข้องกับข้อมูลลำดับเบสของโปรตีนหรือเรียกว่าลำดับกรดอะมิโน และลำดับนิวคลีโอไทด์ (Setubal & Meidanis, 1997) ข้อมูลเหล่านี้เกิดขึ้นจากการร่วมมือกันของนักวิจัยจากหลายประเทศ โดยฐานข้อมูลหลักที่เก็บลำดับเบสและเป็นฐานข้อมูลที่เป็นสาธารณะ ซึ่งนักวิจัยทั่วโลกสามารถดาวน์โหลดข้อมูลเพื่อนำไปใช้เป็นเครื่องมือในการวิเคราะห์ได้ แสดงดังตารางที่ 1 ข้อมูลทั้งหมดจะถูกจัดเก็บอย่างเป็นระบบ และมีการปรับปรุงเนื้อหาเพื่อสะดวกต่อการนำไปใช้งานให้เกิดประโยชน์ ฐานข้อมูลที่ดีนั้นจะต้องมีการเก็บรวบรวมข้อมูลในรูปแบบที่สามารถเข้าถึงได้ง่าย และสามารถนำไปใช้ได้กับระบบหลายผู้ใช้

ลำดับนิวคลีโอไทด์และลำดับโปรตีนเป็นคลังข้อมูลซึ่งเป็นที่สนใจของนักชีววิทยาโมเลกุล การกำเนิดของฐานข้อมูลลำดับโมเลกุลทำให้เกิดการวิเคราะห์ด้วยคอมพิวเตอร์ของลำดับที่มีอยู่ทั้งหมด ฐานข้อมูลลำดับเบสให้บริการ 2 หน้าที่หลักคือ 1) เพื่ออำนวยความสะดวกในการพิจารณาลำดับใหม่ด้วยการเปรียบเทียบกับลำดับเดิมที่มีในฐานข้อมูล 2) เพื่อทำหน้าที่เป็นแหล่งที่มาของข้อมูลสำหรับการทดลองรุ่นใหม่และการทดสอบสมมติฐานเกี่ยวกับการจัดลำดับโมเลกุลและวิวัฒนาการ ปริมาณข้อมูลขนาดใหญ่ที่มีอยู่ตอนนี้จำเป็นต้องมีขั้นตอนวิธีการเพื่อการค้นคืนฐานข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ ซึ่งมีความก้าวหน้าเป็นอย่างมากในการพัฒนาด้านนี้

ตารางที่ 1 แสดงรายชื่อและสถานที่ตั้งของฐานข้อมูลหลักที่ใช้ในการวิจัย

Database	Location	Web link
Protein sequence databases		
- PIR	US	http://pir.georgetown.edu/
- MIPS	Germany	http://mips.gsf.de
- Swiss-Prot	Swiss	http://www.expasy.ch
- TrEMBL	Swiss	http://www.uniprot.org/help/uniprotkb
- NRL_3D	US	http://www.ncifcrf.gov/ NRL-3D
- GenPept	US	http://bioinfo.md.huji.ac.il/databases/genpept
Nucleotide sequence databases		
- EMBL	UK	http://www.ebi.ac.uk/Databases/
- GenBank	US	http://www.ncbi.nlm.nih.gov/Genbank/
- DDBJ	Japan	http://www.ddbj.nig.ac.jp/
- Celera	Celera	http://www.dnalc.org/resources/

3.1 ฐานข้อมูลของโปรตีนและลำดับกรดอะมิโน (Protein sequence databases)

ฐานข้อมูลของโปรตีนเป็นการรวบรวมลำดับที่ได้จากความหลากหลายของแหล่งที่มา ซึ่งลำดับโปรตีนเป็นปัจจัยพื้นฐานของโครงสร้างทางชีวภาพและการทำงานทางชีวภาพ

3.1.1 PIR

PIR เป็นฐานข้อมูลลำดับโปรตีน (McGarvey, et al., 2000) ที่ได้รับการพัฒนาในช่วงต้นปี 1960 ซึ่งตั้งอยู่ที่ National Biomedical Research Foundation (NBRF) และตั้งแต่ปี 1988 ได้รับการเก็บรักษาโดย PIR-International ในสหรัฐอเมริกา และ Martinsried Institute for Protein Sequence (MIPS) ในเยอรมัน ปัจจุบันฐานข้อมูลนี้มี 83,724,331 รายการ (Release: 4.44, 11-Nov-2015)

3.1.2 Swiss-Prot

Swiss-Prot (Stoesser, et al., 2003) เป็นฐานข้อมูลของลำดับกรดอะมิโนที่ได้ก่อตั้งขึ้นในปี 1986 ซึ่งได้รับการดูแลรักษาและรวบรวมข้อมูลโดย SIB (Swiss Institute of Bioinformatics) และ EBI/EMBL ในประเทศสวิสเซอร์แลนด์ ประกอบด้วยคำอธิบายระดับสูงรวมถึงรายละเอียดการทำงานของโปรตีน โครงสร้างของโปรตีน การปรับเปลี่ยนกรดอะมิโน เป็นต้น ซึ่งมีจุดมุ่งหมายที่จะลดความซ้ำซ้อน ฐานข้อมูล Swiss-Prot มีการเชื่อมโยงไปยังแหล่งข้อมูลอื่น ๆ อีกมากมายรวมทั้งฐานข้อมูลลำดับเบสอื่น ๆ อีกด้วย

3.1.3 TrEMBL - Translated EMBL

TrEMBL (José Mateus Silva, et al., 2010) เป็นฐานข้อมูลของลำดับกรดอะมิโนที่ถูกสร้างขึ้นในปี 1996 ซึ่งเป็นส่วนเสริมคำอธิบายประกอบของคอมพิวเตอร์ให้กับฐานข้อมูล Swiss-Prot โดยจะมีการแปลของ coding sequences (CDS) ทั้งหมดในฐานข้อมูลลำดับนิวคลีโอไทด์ของ EMBL, GenBank และฐานข้อมูล DDBJ

3.1.4 GenPept

GenPept (Ray & Pal, 2013) จัดทำขึ้นเพื่อเป็นส่วนเสริมให้กับฐานข้อมูลลำดับนิวคลีโอไทด์ของ GenBank โดยมีการแปล CDS ในรายการของ GenBank ทั้งหมด ซึ่ง GenPept ไม่ได้เป็นการนำเสนออย่างเป็นทางการจาก NCBI แต่จะถูกเก็บรักษาไว้และปรับปรุงข้อมูลให้สอดคล้องกับข้อมูลแต่ละรุ่นของ GenBank

3.1.5 NRL_3D

NRL_3D (Gutmanas, et al., 2014) ได้รับการดูแลรักษาและสร้างโดย PIR ประกอบด้วยลำดับที่สกัดได้จาก Protein DataBank (PDB) รวมถึงโครงสร้างทุติยภูมิ ซึ่ง NRL_3D ทำให้ข้อมูลลำดับใน PDB สามารถใช้ได้สำหรับทั้งการค้นหาแบบข้อความและการค้นหาลำดับเบส นอกจากนี้ยังมีการเชื่อมโยงข้อมูลสำหรับการใช้งานในฐานข้อมูลโปรตีนของ PIR อื่น ๆ

3.2 ฐานข้อมูลลำดับนิวคลีโอไทด์ (Nucleotide sequence databases)

ฐานข้อมูลลำดับนิวคลีโอไทด์ หรือ ฐานข้อมูลรหัสดีเอ็นเอมีอัตราการเติบโตสูงกว่าฐานข้อมูลโปรตีน เพราะส่วนใหญ่ลำดับนิวคลีโอไทด์จะไม่มีการเข้ารหัสเหมือนลำดับโปรตีน และเนื่องจากลำดับนิวคลีโอไทด์เป็นสิ่งที่ได้รับความนิยมในการจัดทำฐานข้อมูล โดยจะมีฐานข้อมูลหลักอยู่ 3 ฐานคือ EMBL, GenBank และ DDBJ

3.2.1 EMBL

EMBL (Metzker, 2010) เป็นฐานข้อมูลลำดับดีเอ็นเอในทวีปยุโรปที่ได้รวบรวมโดย European Bioinformatics Institute (EBI) ซึ่งประกอบด้วยลำดับเบสของดีเอ็นเอที่ได้จากนักวิจัยโดยตรง หรือจากงานวิจัยลำดับจีโนม และงานทางวิทยาศาสตร์ที่ได้จดสิทธิบัตรแล้ว ในปัจจุบันฐานข้อมูล EMBL มีจำนวนลำดับนิวคลีโอไทด์ทั้งสิ้น 1,401,669,271,501 ลำดับ (Release: 125, 01-June-2015)

3.2.2 GenBank

GenBank (Clark, et al., 2015) เป็นฐานข้อมูลลำดับดีเอ็นเอที่จัดการโดย National Center Biotechnology Information (NCBI) ซึ่งจัดเก็บข้อมูลลำดับเบสทั้งที่เป็นรหัสพันธุกรรมของ DNA และ mRNA ข้อมูลทั้งหมดสามารถเผยแพร่ต่อสาธารณะได้ ในปัจจุบันอัตราการขยายตัวของข้อมูลใน GenBank มีจำนวนข้อมูลของลำดับมากถึง 188,372,017 ข้อมูล เป็นจำนวนเบส 202,237,081,559 ลำดับ

3.2.3 DDBJ

DDBJ (MURATA & ISHIBUCHI, 1996) เป็นฐานข้อมูลทางชีวภาพที่ได้ก่อตั้งเมื่อปี 1986 และทำการเก็บรวบรวมข้อมูลดีเอ็นเอ โดยตั้งอยู่ที่ National Institute of Genetics (NIG) ในประเทศญี่ปุ่น นอกจากนี้ยังเป็นสมาชิกของ International Nucleotide Sequence Database Collaboration (INSDC) โดยจะทำการแลกเปลี่ยนข้อมูลกันระหว่างฐานข้อมูล EMBL และฐานข้อมูล GenBank ดังนั้นทั้งสามฐานข้อมูลนี้จะมีการปรับปรุงข้อมูลตลอดเวลา และจะให้ข้อมูลเดียวกันไม่ว่าจะเป็นช่วงเวลาใดก็ตาม

4. เทคนิคการวิเคราะห์ข้อมูลทางชีวภาพ

นอกจากการจัดเก็บข้อมูลอย่างมีประสิทธิภาพแล้วนั้น การวิเคราะห์ข้อมูลทางชีวภาพก็เป็นสิ่งสำคัญเช่นกัน ในหลายปีที่ผ่านมามีการพัฒนาด้านชีวสารสนเทศเป็นไปอย่างอย่างรวดเร็ว ทำให้จำนวนข้อมูลทางชีวภาพมีอัตราการเพิ่มขึ้นเป็นเท่าตัว และการตีความข้อมูลทางชีวภาพเหล่านี้จำเป็นต้องใช้การวิเคราะห์และการคำนวณที่ซับซ้อน (Zhang & Wong, 1997) โดยใช้ศาสตร์ด้านเทคโนโลยีสารสนเทศและวิทยาการคอมพิวเตอร์ในการสร้างเครื่องมือหรือโปรแกรมเพื่อจัดการข้อมูลชีวภาพ ตัวอย่างของการวิเคราะห์ข้อมูลทางชีวภาพ เช่น การทำนายโครงสร้างของโปรตีน, การจำแนกยีนของโรคต่าง ๆ , การจัดกลุ่มข้อมูลตามลักษณะของยีน, การสร้างแบบจำลองทางสถิติของการปฏิสัมพันธ์ระหว่างโปรตีน เป็นต้น จะเห็นได้ว่าหากมีการสร้างเครื่องมือเพื่อทำการวิเคราะห์ข้อมูลที่มีศักยภาพนั้น จะทำให้นักวิจัยสามารถนำผลลัพธ์เหล่านี้ไปใช้ให้เกิดประโยชน์อันสูงสุด

การพัฒนาเครื่องมือสำหรับชีวสารสนเทศนั้นเป็นโปรแกรมที่ถูกออกแบบมาสำหรับการสกัดข้อมูลที่มีความหมายจากข้อมูลอณูชีววิทยา ฐานข้อมูลทางชีวภาพและการดำเนินการกับลำดับเบสของดีเอ็นเอและโปรตีนหรือการวิเคราะห์โครงสร้าง (Memeti & Pllana, 2015) ซึ่งสามารถแยกเป็นกลุ่มของรูปแบบของการวิเคราะห์ข้อมูล และ URL ของแต่ละเครื่องมือที่ใช้ในการวิเคราะห์ โดยแสดงดังตารางที่ 2

ตารางที่ 2 ตัวอย่างของเทคนิคและเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลทางชีวภาพ

Research Areas of Bioinformatics	Tool (Application)	Web Link
Sequence Alignment	BLAST CS-BLAST HMMER FASTA	http://blast.ncbi.nlm.nih.gov/Blast.cgi ftp://toolkit.lmb.unimuenchen.de/csblast/ http://hmmer.janelia.org/ www.ebi.ac.uk/fasta33
Multiple Sequence Alignment	MSAProbs DNA Alignment MultAlin DiAlign	http://msaprobs.sourceforge.net/ http://www.fluxus-engineering.com/align.htm http://multalin.toulouse.inra.fr/multalin/multalin.html http://bibiserv.techfak.uni-bielefeld.de/dialign/
Gene Finding	GenScan GenomeScan GeneMark	genes.mit.edu/GENSCAN.html http://genes.mit.edu/genomescan.html http://exon.biology.gatech.edu/
Protein Domain Analysis	Pfam BLOCKS ProDom	http://pfam.sanger.ac.uk/ http://blocks.fhcrc.org/ http://prodom.prabi.fr/prodom/current/html/home.php
Pattern Identification	Gibbs Sampler AlignACE MEME	http://bayesweb.wadsworth.org/gibbs/gibbs.html http://atlas.med.harvard.edu/ http://meme.sdsc.edu/
Genomic Analysis	SLAM Multiz	http://bio.math.berkeley.edu/slam/ http://www.bx.psu.edu/miller_lab/
Motif finding	MEME/MAST eMOTIF	http://meme.sdsc.edu http://motif.stanford.edu

4.1 Sequence Alignment

การทำ Sequence Alignment เป็นการจัดเรียงความเหมือนของสายลำดับดีเอ็นเอ อาร์เอ็นเอหรือโปรตีนที่มีอยู่ในฐานข้อมูล เพื่อหาช่วงลำดับที่สนใจ (conserved sequence) ซึ่งผลลัพธ์ที่ได้จากการทำ Sequence Alignment สามารถนำไปทำนายลักษณะโครงสร้างของดีเอ็นเอ หรือโปรตีน

ได้ และยังสามารถนำมาใช้ในการแบ่งกลุ่มเพื่อศึกษาความสัมพันธ์ทางวิวัฒนาการของสิ่งมีชีวิตได้อีกด้วย (T.F. & M.S., 1981) การเทียบเรียงของสายลำดับสามารถแบ่งได้เป็น 2 แบบ คือ

4.1.1 การเทียบเรียงแบบแบ่งตามช่วง ได้แก่ การเทียบเรียงแบบโกลบอล (Global Alignment) และการเทียบเรียงแบบโลคอล (Local Alignment)

4.1.2 การเทียบเรียงแบบแบ่งตามจำนวนการเทียบเรียง ได้แก่ การเทียบเรียงคู่ลำดับข้อมูลสองเส้น (Pairwise Alignment) และการเทียบเรียงกลุ่มลำดับข้อมูลที่มากกว่าสองเส้น (Multiple Alignment)

การจัดเรียงสายลำดับเป็นส่วนที่มีความสำคัญอย่างมากในแง่ของการคำนวณข้อมูลชีวภาพขนาดใหญ่ ซึ่งในปัจจุบันมีการให้บริการด้านการจัดเรียงสายลำดับเป็นจำนวนมาก แต่อาจจะมีข้อจำกัดต่างๆ เช่น ความเร็ว ความถูกต้อง เป็นต้น จึงได้มีการนำเสนอกลยุทธ์สำหรับการจัดเรียงสายลำดับยีนด้วยการประมวลผลแบบขนาน และพัฒนาอัลกอริทึมที่ใช้ในการจัดเรียงสายลำดับเพื่อเพิ่มความแม่นยำ และประสิทธิภาพด้านความถูกต้องในการคำนวณ (Yang, et al., 2014)

4.2 Multiple Sequence Alignment

Multiple Sequence Alignment เป็นการเปรียบเทียบความเหมือนกันระหว่างสายลำดับดีเอ็นเอ อาร์เอ็นเอหรือโปรตีนมากกว่าสองเส้นของยีนเดียวกันหรือกลุ่มเดียวกันจากหลายแหล่ง (Zhang & Wong, 1998) ซึ่งอาจจะเป็นลำดับเบสจากสิ่งมีชีวิตชนิดเดียวกันแต่ต่างสายพันธุ์หรือสิ่งมีชีวิตต่างชนิดกัน หรือแม้แต่ต่างอาณาจักรก็สามารถนำมาเปรียบเทียบกันได้ เช่น การจัดเรียงลำดับโดยใช้ การหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization) ซึ่งเป็นการเพิ่มประสิทธิภาพแทนการใช้เทคนิคการเขียนโปรแกรมแบบไดนามิก (Jagadamba, et al., 2011) การเปรียบเทียบลำดับเบสนอกจากทำเพื่อเปรียบเทียบความเหมือนกันแล้วยังใช้เพื่อออกแบบ Primer ได้อีกด้วย

4.3 Gene Finding

Gene Finding หรือ Gene prediction หมายถึงกระบวนการในการระบุ regions ของรหัสดีเอ็นเอ รวมถึงรหัสโปรตีนเช่นเดียวกับอาร์เอ็นเอ และอาจจะรวมการคาดการณ์ขององค์ประกอบการทำงานอื่น ๆ (Fickett, 1996) เช่น ส่วนของลำดับเบสที่มีความสามารถในการเพิ่มหรือลดการแสดงออกที่เฉพาะเจาะจงของยีนภายในสิ่งมีชีวิต ซึ่ง Gene Finding เป็นหนึ่งในขั้นตอนแรกและสำคัญที่สุดในการทำความเข้าใจจีโนมของสายพันธุ์หลังจากที่ได้รับการจัดลำดับ

ในยุคแรก Gene Finding ได้เริ่มจากการพยายามทำการทดลองในเซลล์ของสิ่งมีชีวิต การวิเคราะห์ทางสถิติของอัตราการรวมตัวของความคล้ายคลึงกันของยีนต่าง ๆ สามารถกำหนดได้ในบางโครโมโซมและข้อมูลจากการทดลองดังกล่าวสามารถรวมกันเพื่อสร้างแผนภาพทางพันธุกรรมที่ระบุตำแหน่งของยีนที่ทราบและเชื่อมโยงไปยังยีนที่มีความสัมพันธ์กัน (Afridi & Sheng, 2011)

4.4 Protein Domain Analysis

Protein Domain Analysis เป็นการวิเคราะห์ส่วนหนึ่งของสายโพลีเปปไทด์หรือส่วนหนึ่งของโครงสร้างโปรตีนที่มีการพัฒนาหรือวิวัฒนาการ (WU & GARIBAY, 2002) เพื่อทำหน้าที่เฉพาะอย่างใดอย่างหนึ่งและเป็นอิสระจากส่วนที่เหลือของโครงสร้างโปรตีน

4.5 Pattern Identification

Pattern Identification เป็นการระบุรูปแบบของลำดับนิวคลีโอไทด์หรือลำดับโปรตีนที่สนใจ (V.G. & A.V., 2003) จากการเทียบเรียงดังกล่าวสามารถใช้ในการกำหนดความสัมพันธ์ของวิวัฒนาการและการทำงานได้

4.6 Genomic Analysis

Genomic Analysis คือ การระบุ การวัดหรือการเปรียบเทียบคุณสมบัติของจีโนม เช่น ลำดับดีเอ็นเอ, การเปลี่ยนแปลงทางโครงสร้าง, การแสดงออกของยีนหรือคำอธิบายองค์ประกอบการทำงานในระดับยีน (Tsai, et al., 2004) วิธีการในการวิเคราะห์จีโนม ได้แก่ microarrays ซึ่งมีลักษณะเป็นความสามารถของเทคโนโลยีทางวิทยาศาสตร์ที่สามารถผลิตข้อมูลจำนวนมหาศาลได้ในเวลารวดเร็ว (high throughput method) และใช้ร่วมกับเทคนิคทางชีวสารสนเทศ เช่น การวิเคราะห์จีโนมอย่างมีประสิทธิภาพด้วย Hadoop MapReduce เพื่อเพิ่มความถูกต้องและคุณภาพของการวิเคราะห์ (Liu, et al., 2015)

4.7 Motif finding

Motif finding หรือที่เรียกกันว่า profile analysis คือการค้นหาลำดับของกรดอะมิโน ลำดับนิวคลีโอไทด์หรือโครงสร้างระดับทุติภูมิที่ทำหน้าที่เฉพาะในปฏิกิริยาทางชีวเคมี (Tsai, Yang, Tsai, & Kao, Applying genetic algorithms to finding the optimal order in displaying the microarray data, 2002) ด้วยการเปรียบเทียบความเหมือนกันระหว่างสายลำดับแบบโกลบอลที่พยายามที่จะเทียบเรียงลำดับอนุรักษ์ที่มีความจำเพาะ (Motif) ในชุดของลำดับเบส ซึ่งจะทำการเปรียบเทียบความเหมือนกันของลำดับแบบกลุ่ม จากนั้นลำดับอนุรักษ์ที่มีค่าสูงจะถูกแยกออกและใช้ในการสร้างชุดของตารางเมทริกซ์ (Maiti & Mukherjee, 2015) ซึ่งสามารถนำไปใช้ในการทำนายตำแหน่งของ promoter, ribosome binding site, immunogenicity site เป็นต้น

5. ประโยชน์ของการศึกษาชีวสารสนเทศ

5.1 การศึกษาความสัมพันธ์เชิงวิวัฒนาการของสิ่งมีชีวิต (phylogenetic analysis)

ความสัมพันธ์เชิงวิวัฒนาการของสิ่งมีชีวิตเป็นข้อมูลพื้นฐานที่มีความสำคัญเป็นอย่างมากในการศึกษาชีววิทยา ซึ่งการศึกษาทางอณูชีววิทยาสามารถนำข้อมูลลำดับโปรตีนและลำดับดีเอ็นเอมาเปรียบเทียบกันและคำนวณแบบแผนวิวัฒนาการออกมา จากนั้นจึงนำไปสร้างเป็นแผนภาพเรียกว่า phylogenetic tree (Jafarzadeh & Iranmanesh, 2012) โดยใช้ในการเปรียบเทียบวิวัฒนาการของยีนที่ต้องการศึกษากับยีนในสิ่งมีชีวิตอื่น ๆ เพื่ออธิบายการทำงานของยีน หากยีนที่ได้มีความคล้ายคลึงกันมาก แสดงว่าอาจจะทำหน้าที่อย่างเดียวกัน

งานวิจัยด้านนี้ได้แก่ การวัดอัตราการกลายพันธุ์ (mutation rate) อัตราการวิวัฒนาการ (evolutionary rate) ของสิ่งมีชีวิตจาก sequence หรือการคำนวณเวลาที่สิ่งมีชีวิตใช้ในการวิวัฒนาการจาก sequence ซึ่งเรียกว่า molecular clock ก็เป็นงานวิจัยที่ได้รับความสนใจมากเช่นกัน

5.2 การออกแบบและการค้นหายาใหม่ (Drug Design and Discovery)

การออกแบบยา และการค้นพบตัวยาใหม่โดยการใช้คอมพิวเตอร์และเทคโนโลยีสารสนเทศการวิเคราะห์โครงสร้างและการทำงานของสารชีวมหโมเลกุล (Biological macromolecules) เช่น โปรตีน และกรดนิวคลีอิก เป็นต้น การวิเคราะห์ความสัมพันธ์ระหว่างโครงสร้างส่วนที่ทำหน้าที่ต่างๆ (Tzafriri, et al., 2005) และโครงสร้างส่วนที่เกิดกิจกรรมการทำงานของสารออกฤทธิ์ กลไกระดับโมเลกุลของสาเหตุการเกิดโรค การออกแบบลิแกนด์ (ligand) และการจำลองปฏิกิริยาระหว่างลิแกนด์กับสารชีวโมเลกุล การทำนายคุณสมบัติทางเภสัชของสารหรือตัวยาใหม่ (Salisbury, et al., 2011) รวมทั้งการทำนายรูปโมเลกุลของยา และการออกแบบโมเลกุลตั้งแต่เริ่มแรก เพื่อเพิ่มโอกาสในการค้นพบยาตัวใหม่ที่มีประสิทธิภาพ ซึ่งจะช่วยลดค่าใช้จ่าย และประหยัดเวลา ที่สำคัญคือลดการใช้สัตว์ในการทดลอง

6. สรุป

ชีวสารสนเทศศาสตร์เป็นสาขาหนึ่งของวิทยาศาสตร์ที่ใช้เทคโนโลยีสารสนเทศและวิทยาศาสตร์คอมพิวเตอร์มาประยุกต์กับชีววิทยาโมเลกุล โดยทั่วไปแล้วจะทำการพัฒนาขั้นตอนวิธีและซอฟต์แวร์สำหรับคำนวณข้อมูลทางชีวภาพ จากอัตราการเพิ่มขึ้นของข้อมูลทางชีวภาพในปัจจุบันทำให้วิธีการคำนวณจากคอมพิวเตอร์เป็นสิ่งที่ขาดไม่ได้ในการตรวจสอบทางชีวภาพซึ่งพัฒนามาเพื่อการวิเคราะห์ลำดับทางชีวภาพ ชีวสารสนเทศในขณะนี้ครอบคลุมหลากหลายสาขาวิชารวมทั้งโครงสร้างทางชีววิทยา จีโนม และการศึกษาการแสดงออกของยีน

ในบทความนี้ได้นำเสนอข้อมูลและขอบเขตของการศึกษาชีวสารสนเทศในปัจจุบัน โดยเฉพาะอย่างยิ่งได้กล่าวถึงฐานข้อมูลทางชีวภาพที่มีการใช้งานกันทั่วไปในการวิเคราะห์ข้อมูล ซึ่งการวิเคราะห์ข้อมูลทางชีวภาพนั้นสามารถแบ่งออกเป็น 7 ประเภทของการศึกษาชีวสารสนเทศ ได้แก่ Sequence Alignment, Multiple Sequence Alignment, Gene Finding, Protein Domain

Analysis, Pattern Identification, Genomic Analysis และ Motif finding พร้อมทั้งยกตัวอย่างเครื่องมือที่นิยมใช้ในการวิเคราะห์ข้อมูลแต่ละประเภท และยังมีการนำการศึกษาชีวสารสนเทศไปประยุกต์ใช้ให้เกิดประโยชน์ในด้านต่าง ๆ เช่น การศึกษาความสัมพันธ์เชิงวิวัฒนาการของสิ่งมีชีวิต และการออกแบบและการค้นหายาใหม่ เป็นต้น ดังนั้นจะเห็นได้ว่าชีวสารสนเทศได้เข้ามามีบทบาทและมีความสำคัญอย่างมาก ทำให้เกิดการเปลี่ยนแปลงและการพัฒนาในหลาย ๆ ด้าน โดยสามารถประยุกต์ใช้ในด้านอุตสาหกรรม ด้านการแพทย์ ด้านการเกษตร ด้านสิ่งแวดล้อม และ ด้านพลังงาน เป็นต้น

7. เอกสารอ้างอิง

Afridi, M. I., & Sheng, Y. G. (2011). **Novel gene finding through BLAST search. Phylogenitic analysis through MEGA.** *Computer Science and Network Technology (ICCSNT)* (pp. 115 - 119). Harbin: IEEE.

Altman, R. B. (2001). **Challenges for intelligent systems in biology.** *IEEE Intelligent Systems* (pp. 14-18). IEEE.

Altschul, S. F., Madden, T. L., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research*, 3389-3402.

Bairoch, A., & Apweiler, R. (2000). **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Research*, 45-48.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., & Wheeler, D. L. (2000). **GenBank.** *Nucleic Acids Research*, 15-18.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). **The Protein Data Bank.** *Nucleic Acids Research*, 235-242.

Bhandari, D., Murthy, C. A., & Pal, S. K. (1996). **Genetic algorithm with elitist model and its convergence.** *International Journal of Pattern Recognition and Artificial Intelligence* , 731-747.

Chen, J., Antipov, E., Lemieux, B., Cedeño, W., & David, H. W. (1999). **DNA Computing Implementing Genetic Algorithms.** *Preliminary Proceedings DIMACS Workshop on Evolution as Computation* (pp. 39-51). Piscataway NJ: DIMACS.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). **GenBank.** *Nucleic Acids Research*, D67-D72.

Fickett, J. W. (1996). **Finding genes by computer: the state of the art.** *Trends in genetics : TIG*, 316-320.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . Fitzhugh, W. (1995). **Whole-genome random sequencing and assembly of Haemophilus influenza Rd.** *Science*, 496–512.

Goldberg, D. E. (1989). **Genetic Algorithms in Search, Optimization and Machine Learning.** Boston: Addison-Wesley Longman Publishing Co., Inc.

Guo, F., Wang, L., Yang, Y., & Lin, G. (2010). **Efficient Algorithms for 3D Protein Substructure Identification.** *Bioinformatics and Biomedical Engineering (iCBBE)* (pp. 1 - 4). Chengdu: IEEE.

Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M., Bochet, E., Conroy, M. J., . . . Hendrickx, P. M. (2014). **PDBe: Protein Data Bank in Europe.** *Nucleic Acids Research*, 285-291.

Jafarzadeh, N., & Iranmanesh, A. (2012). **A novel graphical and numerical representation for analyzing DNA sequences based on codons.** *MATCH: Communications in Mathematical and in Computer Chemistry*, (pp. 611-620).

Jagadamba, P. V., Babu, M. S., Rao, A. A., & Rao, P. K. (2011). **An improved algorithm for Multiple Sequence Alignment using Particle Swarm Optimization.** *2011 IEEE 2nd International Conference on Software Engineering and Service Science* (pp. 544 - 547). Beijing: IEEE.

José Mateus Silva, F., Sánchez-Pérez, J. M., Antonio Gómez-Pulido, J., & Vega-Rodríguez, M. A. (2010). **An evolutionary approach for performing multiple sequence alignment.** *IEEE Congress on Evolutionary Computation* (pp. 1-7). Barcelona: IEEE.

Li, H., & Homer, N. (2010). **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform*, 473-483.

Liu, W. Y., Hsiao, H.-I., & Dai, S. Y. (2015). **Genomic analysis with MapReduce.** *Big Data (Big Data)* (pp. 1330 - 1335). Santa Clara, CA: IEEE.

Maiti, A., & Mukherjee, A. (2015). **On the Monte-Carlo Expectation Maximization for Finding Motifs in DNA Sequences.** *IEEE Journal of Biomedical and Health Informatics*, 677 - 686.

McGarvey, P. B., Huang, H., Barker, W. C., Orcutt, B. C., Garavelli, J. S., Srinivasarao, G. Y., Wu, C. H. (2000). **PIR: a new resource for bioinformatics.** *Bioinformatics*, 290-291.

Memeti, S., & Pllana, S. (2015). **Analyzing Large-Scale DNA Sequences on Multi-core Architectures.** *Computational Science and Engineering (CSE)* (pp. 208 - 215). Porto: IEEE.

Metzker, M. L. (2010). **Sequencing technologies — the next generation.** *Nature Reviews Genetics*, 31-46.

MURATA, T., & ISHIBUCHI, H. (1996). **Positive and negative combination effects of crossover and mutation operators in sequencing problems.** *Proceedings of IEEE International Conference on Evolutionary Computation* (pp. 170 - 175). Nagoya: IEEE.

National Center for Biotechnology Information, U.S. National Library of Medicine. (2015, December 2). Retrieved from National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/genbank/statistics/

Parsons, R., & Jhonson, M. E. (1993). **Genetic algorithms for DNA sequence assembly.** *International Conference on Intelligent Systems for Molecular Biology* (pp. 310-318). United States: Menlo Park, Calif. : AAAI Press.

PEARSON, W. R., & LIPMAN, D. J. (1988). **Improved tools for biological sequence comparison.** *Proceedings of the National Academy of Sciences of the United States of America* (pp. 2444-2448). Washington, DC : National Academy of Sciences.

Ray, S. S., & Pal, S. K. (2013). **RNA Secondary Structure Prediction Using Soft Computing.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (pp. 2-17). IEEE.

Reichhardt, T. (1999). **It's sink or swim as a tidal wave of data approaches.** *Nature*, 517-520.

Salisbury, J. P., Morgan, S. W., & Williams, J. C. (2011). **Drug design as a video game: A summer program integrating chemical biology and computer science.** *Integrated STEM Education Conference (ISEC)* (pp. 8B-1 - 8B-3). Ewing, NJ: IEEE.

Setubal, C., & Meidanis, J. (1997). ***Introduction to Computational Molecular Biology 1st Edition.*** Pacific Grove, CA: .

Stoesser, G., Baker, W., Broek, A. v., Garcia-Pastor, M., Kanz, C., Kulikova, T.,Tzou, K. (2003). **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Research*, 17-22.

T.F., S., & M.S., W. (1981). **Identification of common molecular subsequences.** *Journal of molecular biology*, 195-197.

Tsai, H.-K., Yang, J.-M., Tsai, Y.-F., & Kao. (2002). **Applying genetic algorithms to finding the optimal order in displaying the microarray data.** *Genetic and Evolutionary Computation — GECCO* (pp. 610-617). Springer Berlin Heidelberg.

Tsai, H.-K., Yang, J.-M., Tsai, Y.-F., & Kao, C.-Y. (2004). **An evolutionary approach for gene expression patterns.** *IEEE transactions on information technology in biomedicine* (pp. 69-78). IEEE.

Tzafriri, A. R., Lerner, E. I., Flashner-Barak, M., Hinchcliffe, M., Ratner, E., & Parnas, H. (2005). **Mathematical modeling and optimization of drug delivery from intratumorally injected microspheres.** *Clinical Cancer Research*, 826-834.

V.G., L., & A.V., K. (2003). **Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis.** *In silico biology*, 81-87.

WU, A. S., & GARIBAY, I. (2002). **The Proportional Genetic Algorithm: Gene Expression in a Genetic Algorithm.** *Genetic Programming and Evolvable Machines*, 157-192.

Yang, R., Zhao, Y., Su, Y., Pan, C., Duan, H., & Deng, N. (2014). **A high-throughput gene sequence alignment strategy using parallel computing.** *2014 7th International Conference on Biomedical Engineering and Informatics* (pp. 638 - 642). Dalian: IEEE.

Zhang, C., & Wong, A. K. (1997). **A genetic algorithm for multiple molecular sequence alignment.** *Bioinformatics*, 565-581.

Zhang, C., & Wong, A. K. (1997). **Toward efficiesunt multiple molecular sequence alignment: A system of genetic algorithm and dynamic programming.** *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (pp. 918-932). IEEE.

Zhang, C., & Wong, A. K. (1998). **A technique of genetic algorithm and sequence synthesis for multiple molecular sequence alignment.** *1998 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2442 - 2447). San Diego, CA: IEEE.