

แนะนำ Bio-Mirror ของฐานข้อมูลชีวภาพของ เครือข่ายนักวิจัยภาคพื้นเอเชีย-แปซิฟิก

สมนึก คีรีโต และ จิตินันต์ เขียนดวงจันทร์

ฝ่ายสารสนเทศงานวิจัย สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์

พันธุ์ปิติ เปี่ยมสง่า

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

อีเมลล์ {sk,g4265081,pp}@ku.ac.th

ABSTRACT – Kasetsart University Research and Development Institute (KURDI) by cooperation with Asia-Pacific Advanced Network (APAN) establishes and services biological sequence databanks. We provide free of charge for researchers and general person at web site: <http://bio-mirror.ku.ac.th> and ftp site: <ftp://bio-mirror.ku.ac.th/biomirror>. The currently total databanks are about 28 Gigabyte in compressed format and include many subdata that separate 3 widely groups. 1) DNA sequence database include GENBANK, BLAST, DDBJ, EMBL and RICE; 2) Protein sequence database include PROSITE, BLOCKS, PFAM, PIR, SWISSPROT, and TREMBL and 3) Other sub database include ENZYME, REBASE, TAXONOMY, and UNIGENE.

Beside, there are many software engines used for analyze and manage biological sequence database in Bio-Mirror that is useful to researcher and general person. Bio-Mirror helps users in Thailand for high-speed access to increase performance in using network that connects other countries. Moreover Bio-Mirror is useful to biological technology researcher and information technology researcher for research and development in the future.

KEY WORDS – Biological Database, DNA and Protein Sequences

บทคัดย่อ – สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์ โดยความร่วมมือกับเครือข่ายนักวิจัยภาคพื้นเอเชีย-แปซิฟิก (Asia-Pacific Advanced Network: APAN) ได้ดำเนินการติดตั้งและให้บริการระบบข้อมูลทางชีวภาพให้กับนักวิจัยและผู้สนใจทั่วไปโดยไม่คิดค่าบริการใดๆ ในลักษณะของ WWW ที่ <http://bio-mirror.ku.ac.th> และ FTP ที่ <ftp://bio-mirror.ku.ac.th/biomirror/> ระบบดังกล่าวเก็บรวบรวมข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพ ไว้จำนวนมากกว่า 19 กิกะไบต์ (ขนาดของข้อมูลเมื่อถูกบีบอัด) โดยประกอบด้วยข้อมูลย่อยต่างๆ มากมาย โดยแบ่งกว้างๆ ได้สามกลุ่มคือ 1) ฐานข้อมูลลำดับนิวคลีโอไทด์และกรดอะมิโนประกอบด้วย GENBANK, BLAST, DDBJ, EMBL และฐานข้อมูลลำดับนิวคลีโอไทด์ของข้าว (RICE); 2) ฐานข้อมูลลำดับกรดอะมิโนและโปรตีนประกอบด้วย PROSITE, BLOCKS, PFAM, PIR, SWISSPROT, และ TREMBL และ 3) ฐานข้อมูลย่อยอื่นๆ ประกอบด้วย ENZYME, REBASE, TAXONOMY, และ UNIGENE

นอกจากนี้ยังมีโปรแกรมที่ช่วยในการวิเคราะห์และจัดการฐานข้อมูลในระบบข้อมูล Bio-Mirror ในลักษณะต่างๆ ซึ่งจะเป็นประโยชน์ต่อนักวิจัยและผู้สนใจทั่วไปด้วย ระบบข้อมูล Bio-Mirror นี้จะช่วยทำให้ผู้ใช้งานในประเทศไทยสามารถเข้าถึงข้อมูลได้สะดวกและรวดเร็วขึ้น ช่วยเพิ่มประสิทธิภาพในการใช้สายเครือข่ายข้ามประเทศให้มี ประสิทธิภาพสูงขึ้น และเพื่อให้ นักวิจัยทั้งทาง ด้านเทคโนโลยีชีวภาพ และนักวิจัยทางด้านเทคโนโลยีสารสนเทศได้ใช้ประโยชน์ระบบข้อมูลนี้ เพื่อการค้นคว้า วิจัยและพัฒนาให้เกิด ประโยชน์มากขึ้นต่อไป

คำสำคัญ – ระบบข้อมูลทางชีวภาพ, ลำดับเบสของดีเอ็นเอและโปรตีน

1. บทนำ

ในปัจจุบัน การจัดทําระบบฐานข้อมูลที่เป็นศูนย์กลางของข้อมูลและสารสนเทศเพื่อการวิจัยเฉพาะเรื่องหรือเฉพาะสาขานับเป็นโครงสร้างพื้นฐานที่สำคัญมากในการดำเนินงานวิจัย และถือได้ว่าเป็นข้อมูลพื้นฐานที่นักวิชาการและนักวิจัยต้องการเพื่อทำการศึกษาดังข้อมูลเบื้องต้นหรือความเป็นไปได้ของการที่จะทำการวิจัยต่อไปในอนาคต รวมทั้งเป็นแหล่งที่สามารถใช้ในการศึกษาถึงความก้าวหน้าของศาสตร์นั้น ซึ่งจะเห็นได้ว่าในประเทศไทยจำนวนฐานข้อมูลสารสนเทศมิให้บริการไม่เพียงพอและยังขาดข้อมูลเฉพาะเรื่องอีกมาก ดังนั้นจึงได้มีการศึกษาและคิดดั่งระบบข้อมูลทางด้านชีววิทยาขึ้นมา โดยระบบข้อมูลทางด้านชีววิทยานี้ได้เก็บรวบรวมฐานข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพไว้ในเซิร์ฟเวอร์ ที่เรียกว่า Bio-Mirror เซิร์ฟเวอร์นี้ได้ติดตั้งเว็บไซต์สำหรับเข้าถึงข้อมูลได้อย่างรวดเร็วและมีข้อมูลที่เปลี่ยนแปลงใหม่อยู่ตลอดเวลา ซึ่งระบบดังกล่าวนี้ได้รับความร่วมมือในการพัฒนาจากเครือข่ายนักวิจัยภาคพื้นเอเชียและแปซิฟิก หรือ Asia-Pacific Advanced Network (APAN) [1] การพัฒนาข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพนี้ได้มาจากความร่วมมือระหว่างนักวิจัยและพัฒนาจากสหรัฐอเมริกา ญี่ปุ่น และยุโรป โดยมีการแลกเปลี่ยนข้อมูลระหว่างเว็บไซต์กันตลอดเวลา การแลกเปลี่ยนข้อมูลกันได้อย่างรวดเร็วนี้ มีองค์กรและกลุ่มความร่วมมือต่างๆ ที่มีส่วนช่วยในการติดต่อสื่อสารคือ Very High Speed Backbone Service (vBNS) [26], TransPAC (Trans-Pacific Network) [22] และ Asia-Pacific Advanced Network (APAN)

2. จุดมุ่งหมายของระบบข้อมูล Bio-Mirror

การคิดดั่งระบบข้อมูล Bio-Mirror มีจุดมุ่งหมายสำคัญคือ การให้บริการข้อมูลแก่นักวิจัยและผู้สนใจทั่วไปทั้งภายในและภายนอกประเทศ โดยผ่านทางเครือข่ายอินเทอร์เน็ตได้อย่างสะดวกและรวดเร็ว และเพื่อให้ นักวิจัยและผู้สนใจทั่วไปสามารถเข้ามาศึกษาหรือนำข้อมูลไปใช้โดยไม่คิดค่าบริการใดๆ สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์ ได้ให้ความร่วมมือกับเครือข่ายนักวิจัยระหว่างประเทศ (APAN) และได้จัดตั้ง Mirror เว็บไซต์เพื่อให้บริการสำหรับผู้ใช้ในประเทศไทยและประเทศใกล้เคียงได้ที่เว็บไซต์ <http://bio-mirror.ku.ac.th/> และที่เอฟทีพีไซต์ <ftp://bio-mirror.ku.ac.th/biomirror/> เว็บไซต์ที่จัดตั้งขึ้นนี้จะช่วยลดการใช้ bandwidth หรือช่วยเพิ่มอรรถประโยชน์จากสายเครือข่ายข้ามประเทศให้มีประสิทธิภาพสูงขึ้นด้วย

3. ฐานข้อมูลในระบบข้อมูล Bio-Mirror

ฐานข้อมูลลำดับเบสของดีเอ็นเอ หรือบางครั้งเรียกว่าลำดับนิวคลีโอไทด์ และลำดับเบสของโปรตีนหรือที่เรียกว่าลำดับกรดอะมิโนนั้น ในปัจจุบันมีอยู่มากมายมหาศาล นักวิจัยและพัฒนาระหว่างประเทศได้ร่วมกัน จัดเก็บรวบรวมข้อมูลที่เกี่ยวข้องกับข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพไว้อย่างมีระบบ ซึ่งฐานข้อมูลหลักในโลกจะมีอยู่ 3 ฐานคือ Genbank [18] ในสหรัฐอเมริกา, EMBL [7] ในทวีปยุโรป และ DDBJ [5] ในประเทศญี่ปุ่น ซึ่งฐานข้อมูลทั้งสามนี้จะมีการแลกเปลี่ยนข้อมูลกันอยู่ตลอดเวลาทำให้มีข้อมูลที่เรียกว่าเกือบจะเหมือนกัน ฐานข้อมูลเหล่านี้จะมีทั้งฐานข้อมูลทั่วไปและฐานข้อมูลเฉพาะ เช่น ฐานข้อมูลของลำดับนิวคลีโอไทด์ ฐานข้อมูลของลำดับกรดอะมิโนและเว็บไซต์ที่เกี่ยวข้องกับโปรตีน ฐานข้อมูลของโครงสร้างตติยภูมิ เป็นต้น นอกจากนี้ยังได้จัดตั้ง Mirror เว็บไซต์ของระบบข้อมูลทางชีวภาพที่เรียกว่า Bio-Mirror ขึ้นเพื่อให้บริการแก่นักวิจัยและผู้สนใจทั่วไป ระบบดังกล่าวเก็บรวบรวมข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพไว้จำนวนมากกว่า 19 กิกะไบต์ ในลักษณะข้อมูลบีบอัด ถ้านำมาขยายแล้วจะมีข้อมูลมากถึงประมาณ 80 กิกะไบต์ โดยประกอบด้วยข้อมูลย่อยต่างๆ มากมาย แยกได้ตามหมวดหมู่ดังนี้

3.1 ฐานข้อมูลของลำดับนิวคลีโอไทด์

3.1.1 DDBJ

DDBJ [5] เป็นฐานข้อมูลของลำดับนิวคลีโอไทด์ในประเทศญี่ปุ่นซึ่งสนับสนุนโดยรัฐบาลญี่ปุ่น ฐานข้อมูลนี้เริ่มให้บริการในปี ค.ศ. 1986 โดยความร่วมมือของกลุ่มผู้เชี่ยวชาญทางด้านอนุชีววิทยาของประเทศญี่ปุ่น และการสนับสนุนของกระทรวงศึกษา วิทยาศาสตร์ กีฬา และวัฒนธรรม ฐานข้อมูลนี้ได้รับการดูแลโดย DDBJ ร่วมกับองค์กรหลักอื่นๆที่มีฐานข้อมูลลำดับนิวคลีโอไทด์อยู่

3.1.2 EMBL

EMBL [7] เป็นฐานข้อมูลของลำดับนิวคลีโอไทด์ในทวีปยุโรป ซึ่งแหล่งข้อมูลหลักสำหรับข้อมูลลำดับเบสของดีเอ็นเอ และอาร์เอ็นเอ (RNA) ได้มาจากนักวิจัยที่ค้นคว้าด้านนี้โดยตรง โครงการหาลำดับของจีโนม (Genome) และงานอื่นๆที่จัดลิขสิทธิ์แล้ว ฐานข้อมูล EMBL สร้างขึ้นโดยความร่วมมือระหว่างประเทศคือ Genbank ของสหรัฐอเมริกา และ DDBJ ในประเทศญี่ปุ่น แต่ละกลุ่มจะเก็บสะสมข้อมูลลำดับเบสบางส่วนไว้และเผยแพร่ต่อทั่วโลก อนุญาตให้นำมาใช้ และฐานข้อมูลมีการเปลี่ยนแปลงใหม่ทุกวันระหว่างกลุ่มความร่วมมือของฐานข้อมูลนี้

3.1.3 GENBANK

GENBANK [18] เป็นฐานข้อมูลของลำดับนิวคลีโอไทด์ในประเทศสหรัฐอเมริกาซึ่งสนับสนุนโดย National Institute of Health (NIH) [11] ฐานข้อมูลนี้ให้บริการตั้งแต่ปี ค.ศ.1992 และบริหารโดย National Center for Biotechnology Information (NCBI) โดยได้รับความร่วมมือจากองค์กรหลักอื่นๆที่มีฐานข้อมูลลำดับนิวคลีโอไทด์และลำดับกรดอะมิโนอยู่ Genbank ได้จดทะเบียนการค้ากับ National Institute of Health (NIH) แต่ฐานข้อมูลที่มีอยู่เปิดให้บริการโดยไม่คิดค่าบริการใดๆทั้งสิ้น

3.1.4 RICE

RICE [19] เป็นฐานข้อมูลของลำดับนิวคลีโอไทด์ของข้าว

3.2 ฐานข้อมูลของโปรตีนและลำดับกรดอะมิโน

3.2.1 BLOCKS

BLOCKS [3] เป็นระบบสำหรับจำแนกโปรตีนซึ่งได้มาจากลำดับเบสอนุรักษ์ (Conserved Regions) ในกลุ่มของโปรตีนที่อยู่ในฐานข้อมูล PROSITE ฐานข้อมูลบล็อกนี้เป็นกลุ่มของลำดับเบสที่เหมือนกัน โดยได้รับจากลำดับเบสหลายสายที่นำมาเปรียบเทียบความคล้ายกัน ซึ่งการนำลำดับเบสหลายสายมาเปรียบเทียบกันนี้จะช่วยลดสิ่งรบกวน และเพิ่มประสิทธิภาพในการหาความสัมพันธ์ของลำดับเบส ฐานข้อมูลนี้มีจุดมุ่งหมายเพื่ออธิบายโครงสร้างและหน้าที่การทำงานของลำดับเบส

ขั้นตอนในการสร้างฐานข้อมูลบล็อกจะมีอยู่สามขั้นตอน โดยขั้นตอนแรกจะเป็นการหาลำดับเบสอนุรักษ์ในกลุ่มของโปรตีนที่อยู่ในฐานข้อมูล PROSITE และในขั้นตอนที่สองจะนำบล็อกของโปรตีนที่ได้มาจัดเรียงใหม่ให้ดีขึ้นเพื่อที่จะสร้างกลุ่มของบล็อกของโปรตีน ส่วนในขั้นตอนที่สามบล็อกของโปรตีนจะถูกวัดขนาดเทียบกับฐานข้อมูล SWISS-PROT เพื่อที่จะอธิบายระดับของการจับคู่ที่มีโอกาสเกิดขึ้นอีกครั้งหนึ่ง ฐานข้อมูลบล็อกจะถูกสร้างขึ้นเป็นกลุ่มของบล็อกของโปรตีนอย่างสมบูรณ์

3.2.2 PFAM

PFAM [16] เป็นฐานข้อมูลการเปรียบเทียบความคล้ายกันของลำดับเบสหลายสาย และโมเดล Hidden Markov (HMMs) ซึ่งครอบคลุมโดเมนของโปรตีน หรือลำดับเบสอนุรักษ์ของโปรตีนมากมาย ฐานข้อมูลนี้ได้จากความร่วมมือกันระหว่างห้องปฏิบัติการวิจัยของ Erik Sonnhammer (Sanger Centre, Hinxton UK) และ Sean Eddy (Dept. of Genetics, Washington University, St. Louis USA)

3.2.3 PIR

PIR [17] เป็นแหล่งข้อมูลของโปรตีนและฐานข้อมูลลำดับกรดอะมิโนของนานาชาติ ซึ่งได้รับการดูแลรักษาโดย PIR ที่ National Biomedical Research Foundation (NBRF) ในสหรัฐอเมริกา Martinsried Institute for Protein Sequence (MIPS) ในเยอรมัน และ Japan International Protein Information Database (JIPID)

ในญี่ปุ่น ฐานข้อมูลนี้เป็นที่เก็บรวบรวมข้อมูลของโปรตีนที่เกิดขึ้นตามธรรมชาติซึ่งลำดับกรดอะมิโนของมันจะถูกกำหนดขึ้น ข้อมูลในฐานข้อมูลนี้ประกอบด้วยข้อมูลอ้างอิงถึงชื่อ ประเภท และแหล่งที่ได้รับโปรตีน และลักษณะทางชีวภาพของโปรตีน เช่น เครื่องหมายแสดงและขั้นตอนการทำงาน ขอบเขตและความเปลี่ยนแปลงภายในโปรตีน ลักษณะหน้าที่การทำงาน และข้อมูลของลำดับเบสขั้นแรก นอกจากนี้ข้อมูลจำนวนมากในฐานข้อมูลนี้ยังถูกเก็บไว้ในฐานข้อมูลอื่นๆ ด้วยเช่น ฐานข้อมูล SWISS-PROT และฐานข้อมูลอื่นๆ ซึ่งดูแลรักษาโดย NCBI และ EBI

3.2.4 PROSITE

PROSITE [14] เป็นแหล่งข้อมูลซึ่งเก็บโครงร่างของโปรตีน และเว็บไซต์ที่เกี่ยวข้องกับโปรตีน ฐานข้อมูลนี้ได้รับการบริหารและจัดการโดย Dr. Amos Bairoch ที่มหาวิทยาลัยของเมืองเจนีวา ประเทศสวิสเซอร์แลนด์ ซึ่งสามารถได้รับข้อมูลของ PROSITE จากซีดีรอมของ EBI โดยตรง หรือโดยเอฟทีพีจากเซิร์ฟเวอร์ของ NCBI EBI หรือ DDBJ หรือส่งผ่านเซิร์ฟเวอร์ ExPASy [9] บนเว็บไซต์

3.2.5 SWISSPROT

SWISSPROT [15] เป็นฐานข้อมูลของลำดับกรดอะมิโนซึ่งได้จากแหล่งข้อมูลหลายแหล่ง ประกอบด้วยกรดอะมิโนที่แปลมาจากกรดนิวคลีอิก กรดอะมิโนจากฐานข้อมูลลำดับกรดอะมิโนหลักของนานาชาติ ฐานข้อมูลนี้เริ่มต้นใน ปี ค.ศ. 1985 และในปี ค.ศ. 1987 ได้รับการดูแลรักษาโดยมหาวิทยาลัยของเมืองเจนีวา ประเทศสวิสเซอร์แลนด์ และห้องสมุดข้อมูล EBI

ข้อมูลใน SWISS-PROT มาจากแหล่งข้อมูล 3 แหล่งคือ การแปลลำดับนิวคลีโอไทด์จากฐานข้อมูลลำดับนิวคลีโอไทด์ของ EMBL จากกลุ่มของนักวิจัยโดยตรง และจากเอกสารทางวิทยาศาสตร์ ฐานข้อมูลโปรตีนอื่นๆมากมายเช่น PROSITE, BLOCKS และ PRINTS ได้รับข้อมูลมาจากฐานข้อมูล SWISS-PROT

3.2.6 TREMBL

TREMBL [23] เป็นฐานข้อมูลลำดับกรดอะมิโนซึ่งแปลมาจากฐานข้อมูลลำดับนิวคลีโอไทด์ของ EMBL โดยตรง ประกอบด้วยส่วนเสริม

ให้กับฐานข้อมูล SWISS-PROT เป็นคำอธิบายประกอบของคอมพิวเตอร์ ซึ่งรูปแบบจะเหมือนกับรูปแบบของ SWISS-PROT รูปแบบนี้ได้มาจากการแปลลำดับเบสที่รู้ลำดับแล้วทั้งหมดหรือ Coding Sequences (CDS) ในฐานข้อมูลลำดับนิวคลีโอไทด์ของ EMBL ซึ่งฐานข้อมูลของ EMBL ต้องการลำดับเบสที่รู้ลำดับเรียบร้อยแล้วรวมทั้งฐานข้อมูล SWISS-PROT ด้วย

3.3 ฐานข้อมูลอื่นๆ

3.3.1 BLAST

BLAST [2] เป็นฐานข้อมูลลำดับนิวคลีโอไทด์และกรดอะมิโนของ Genbank ซึ่งเก็บอยู่ในรูปของ FASTA Format เป็นฐานข้อมูลหลักที่ใช้ในโปรแกรมสำหรับวิเคราะห์และจัดการข้อมูลในระบบข้อมูล Bio-Mirror

3.3.2 ENZYME

ENZYME [8] เป็นฐานข้อมูลที่เกี่ยวข้องกับการตั้งชื่อวิทยาศาสตร์ของเอนไซม์ ซึ่งได้รับการสนับสนุนและแนะนำเบื้องต้นจากคณะกรรมการการตั้งชื่อทางชีวเคมีและอนุชีววิทยาแห่งสหประชาชาติ หรือ Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) และอธิบายรายละเอียดลักษณะของเอนไซม์แต่ละชนิดพร้อมรหัส EC (Enzyme Commission)

3.3.3 REBASE

REBASE [24] เป็นฐานข้อมูลของเอนไซม์ตัดจำเพาะซึ่งเก็บรายละเอียดของข้อมูลเกี่ยวกับชนิดเอนไซม์ตัดจำเพาะ ชนิดของจุลินทรีย์ที่ผลิตเอนไซม์ ลำดับเบสที่เอนไซม์เลือกตัดจำเพาะ ตำแหน่งที่เอนไซม์ตัดจำเพาะ เอนไซม์ที่หาได้ในเชิงพาณิชย์ และเอกสารอ้างอิงทั้งที่ตีพิมพ์และยังไม่ได้ตีพิมพ์

3.3.4 SRS-DATABANKS

SRS-DATABANKS [20] เป็นฐานข้อมูลที่ได้รับมาจาก EBI [6] เป็นฐานข้อมูลที่เก็บรวบรวมข้อมูลของระบบสืบค้นข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพ ซึ่งระบบสืบค้น SRS เขียนขึ้น เพื่อเลือกข้อมูลย่อยจากฐานข้อมูลลำดับเบสทั้งหมด

3.3.5 TAXONOMY

TAXONOMY [21] เป็นฐานข้อมูลลำดับชั้นของสิ่งมีชีวิตซึ่งเก็บอยู่ในรูปแบบของโครงสร้างต้นไม้ โดยจะเก็บชื่อของสิ่งมีชีวิตทั้งหมดที่เป็นตัวแทนในฐานข้อมูลลำดับเบส และลำดับนิวคลีโอไทด์หรือกรดอะมิโนที่น้อยที่สุดหนึ่งสาย ฐานข้อมูลนี้ส่วนใหญ่ได้รับมาจาก NCBI และ EBI

3.3.6 UNIGENE

UNIGENE [25] เป็นระบบการทดลองเพื่อแบ่งลำดับเบสของ Genbank ให้เป็นเซตที่ไม่ซ้ำซากหรือยาวเกินไปของกลุ่มยีนที่จัดตำแหน่งแล้วอย่างอัตโนมัติ แต่ละกลุ่มของ UniGene เก็บลำดับเบสซึ่งเป็นตัวแทนยีนที่มีลักษณะเฉพาะ และข้อมูลที่เกี่ยวข้องเช่น ชนิดของเนื้อเยื่อในยีนที่สามารถระบุตำแหน่งได้ ฐานข้อมูลนี้ได้รับจาก National Center for Biotechnology Information (NCBI) หรือ Genbank ฐานข้อมูลนี้จะเก็บแยกเป็นไฟล์ด้วยคำหน้าชื่อที่ต่างกันตามชนิดของยีน ไฟล์ในฐานข้อมูลที่น่าหน้าด้วย Hs ย่อมาจาก Homo sapiens ไฟล์ที่น่าหน้าด้วย Mm ย่อมาจาก Mus musculus ไฟล์ที่น่าหน้าด้วย Rn ย่อมาจาก Rattus norvegicus และไฟล์ที่น่าหน้าด้วย Dr ย่อมาจาก Danio rario

ตารางที่ 1. เซิร์ฟเวอร์ของ Bio-Mirror ที่มีอยู่ในปัจจุบัน

ประเทศ	เว็บเซิร์ฟเวอร์และเอฟทีพีเซิร์ฟเวอร์
เกาหลีใต้	http://bio-mirror.kr.apan.net ftp://bio-mirror.kr.apan.net/pub/biomirror
จีน	http://bio-mirror.cn.apan.net ftp://bio-mirror.cn.apan.net
ญี่ปุ่น	http://bio-mirror.jp.apan.net ftp://bio-mirror.jp.apan.net/pub/biomirror
ไทย	http://bio-mirror.ku.ac.th ftp://bio-mirror.ku.ac.th/biomirror
สหรัฐ-อเมริกา	http://bio-mirror.us.apan.net ftp://bio-mirror.us.apan.net/biomirror
สิงคโปร์	http://bio-mirror.sg.apan.net ftp://bio-mirror.sg.apan.net/biomirrors
ออสเตรเลีย	http://bio-mirror.au.apan.net ftp://bio-mirror.au.apan.net/biomirrors

4. ระบบวิเคราะห์และจัดการฐานข้อมูลชีวภาพ

ในทศวรรษที่ผ่านมาจนถึงปัจจุบันนับได้ว่าเป็นยุคของเทคโนโลยีทางวิทยาศาสตร์การแพทย์ ความรู้จากการค้นพบสิ่งใหม่ๆ ก่อให้เกิดประโยชน์นานัปการ การนำคอมพิวเตอร์เข้ามาช่วยในงานทางด้านวิทยาศาสตร์การแพทย์นับว่ามีบทบาทสำคัญมาก ช่วยทำให้เกิดความก้าวหน้าและการค้นพบสิ่งใหม่ๆ ขึ้นมา ในการศึกษาภาคการดำเนินงานของมนุษยศาสตร์ซึ่งมีปริมาณถึง 3 พันล้านนิวคลีโอไทด์ หรือประมาณ 1 แสนยีนซึ่งเป็นเสมือนหนึ่งพิมพ์เขียวของมนุษย์นั้น หากว่านักวิทยา

ศาสตร์มีความสามารถในการวิเคราะห์ขั้นได้นาที่ละ 1 นิวคลีโอไทด์ แม้จะทำงานวันละ 24 ชั่วโมงทุกวัน ก็ยังต้องใช้เวลานานนับ 100 ปีที่จะศึกษาได้ทั้งหมด ดังนั้นจึงได้มีการพัฒนาเครื่องมือและเทคนิคต่างๆ เข้ามาช่วยเพื่อวิเคราะห์และประมวลผลได้อย่างสะดวกและรวดเร็ว โปรแกรมประยุกต์มากมายได้นำมาใช้กับฐานข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพสามารถแยกแยะเป็นกลุ่มดังนี้

1. โปรแกรมแปลลำดับเบส (Translation)

- โปรแกรมแปลลำดับเบสของดีเอ็นเอเป็นโปรตีน (DNA Translation)
- โปรแกรมแปลลำดับเบสของโปรตีนกลับเป็นดีเอ็นเอ (Back Translation)

2. โปรแกรมค้นหาลำดับเบสของดีเอ็นเอและโปรตีนโดยเปรียบเทียบความคล้ายกันของลำดับเบส (Sequence Homology or Similarity Searching)

3. โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีนที่กำลังศึกษา (Sequence Alignment)

- โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีนสองเส้น (Pairwise sequence alignment)
- โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีนมากกว่าสองเส้น (Multiple sequence alignment)

4. โปรแกรมจำแนกชนิดและลักษณะของโปรตีน (Protein Identification and Characterization)

5. โปรแกรมวิเคราะห์โครงสร้างปฐมภูมิ (Primary Structure Analysis)

- โปรแกรมคาดคะเนตำแหน่งของสายโปรตีนที่เกิดการขดเป็นวง
- โปรแกรมคาดคะเนตำแหน่งที่เอนไซม์ตัดจำเพาะ

5. โปรแกรมการหาตำแหน่งของ Transmembrane (Transmembrane regions detection)

ตารางที่ 2. โปรแกรมประยุกต์ที่ใช้กับฐานข้อมูลลำดับเบสของดีเอ็นเอและโปรตีน

1. โปรแกรมแปลลำดับเบส	
โปรแกรมแปลลำดับเบสของดีเอ็นเอเป็นโปรตีน	
DNA Sequence Translation	http://alces.med.umn.edu/cuse.html
Protein Machine	http://www2.ebi.ac.uk/translate/
โปรแกรมแปลลำดับเบสของโปรตีนกลับเป็นดีเอ็นเอ	

Back Translation	http://www.entelechon.com/english/backtranslation.html
2. โปรแกรมค้นหาลำดับเบสของดีเอ็นเอและโปรตีนโดยเปรียบเทียบความคล้ายของลำดับเบส	
FASTA	http://crick.genes.nig.ac.jp/homology/fasta-e.html
BLAST	http://www.ncbi.nlm.nih.gov/blast/blast.cgi
SSEARCH	http://watson.genes.nig.ac.jp/homology/ssearch-e.html
3. โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีนที่กำลังศึกษา	
โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีน 2 เส้น	
ALIGN	http://vega.igh.cnvs.fr/bin/align-guess.cgi
LALIGN	http://vega.igh.cnvs.fr/bin/lalign-guess.cgi
LFasta	http://www2.igh.cnrs.fr/fasta/lfasta-query.html
โปรแกรมเปรียบเทียบความคล้ายกันระหว่างสายดีเอ็นเอหรือโปรตีนมากกว่า 2 เส้น	
CLUSTALW	http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html
MAP	http://dot.imgen.bcm.tmc.edu:9331/seq-seqrch/protein-search.html
MSA	http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
MultAlign	http://www.toulouse.inra.fr/multalin.html
Molsoft	http://molsoft.com/serv/walign.htm
4. โปรแกรมจำแนกชนิดและลักษณะของโปรตีน	
ACompIdent	http://expasy.nhri.org.tw/tools/aacomp/
MultiIdent	http://expasy.nhri.org.tw/tools/multiident/
PeptideMiss	http://expasy.nhri.org.tw/tools/peptide-mass.html
5. โปรแกรมวิเคราะห์โครงสร้างปฐมภูมิ	
โปรแกรมคาดคะเนตำแหน่งของสายโปรตีนที่เกิดการขดเป็นวง	
Coils	http://www.ch.embnet.org
PairCoils	http://nightingale.lcs.mit.edu
MultiCoils	http://nightingale.lcs.mit.edu

โปรแกรมคาดคะเนตำแหน่งที่เอนไซม์ตัดจำเพาะ	
DGPI	http://www.bigfoot.com
NetPicoRNA	http://www.cbs.dtu.dk
6. โปรแกรมการหาตำแหน่งของ Transmembrane (Transmembrane regions detection)	
DAS	http://www.biokemi.su.se
HMMTOP	http://www.enzim.hu
SOSUI	http://azusa.proteome.bio.tuat.ac.jp
Tmpred	http://www.ch.embnet.org

5. ตัวอย่างการประยุกต์ Bio-Mirror

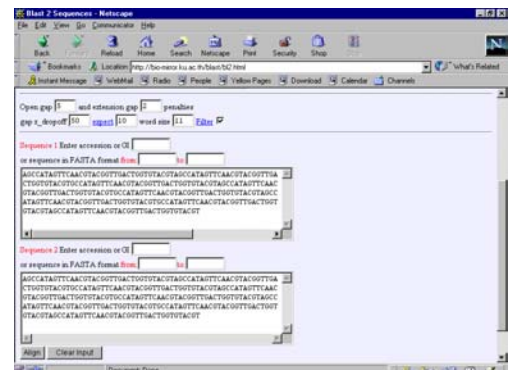
ในทศวรรษที่ผ่านมาจนถึงปัจจุบันนับได้ว่าเทคโนโลยีทางชีวภาพได้ก้าวหน้าไปอย่างรวดเร็ว การค้นพบสิ่งใหม่ๆ ก่อให้เกิดประโยชน์นานัปการ เช่น การวินิจฉัยโรคโดยอาศัยการตรวจความผิดปกติทางโครโมโซม การพัฒนาผลิตภัณฑ์ยาและวัคซีนรักษาโรค ความเข้าใจกลไกของโรคภัยไข้เจ็บ ฯลฯ หนึ่งในบรรดาผลสำเร็จเหล่านั้นที่ได้รับความสนใจจากสังคมโลกคือ เทคโนโลยีเกี่ยวกับดีเอ็นเอที่สามารถประยุกต์ใช้กับงานต่างๆ ได้มากมาย เช่น

- งานทางด้านการพิสูจน์หลักฐาน
 - การพิสูจน์ความเป็นพ่อแม่ ลูก
 - การพิสูจน์หลักฐานในคดีฆาตกรรม
 - การพิสูจน์หลักฐานในโบราณคดี
- งานทางด้านเกษตรกรรม
 - การปรับปรุงพันธุ์พืช พันธุ์สัตว์
 - การตรวจความผิดปกติในพืช และสัตว์
- งานทางด้านวิทยาศาสตร์การแพทย์
 - การตรวจสายพันธุ์ของไวรัส
 - การวินิจฉัยโรคโดยอาศัยการตรวจความผิดปกติของลำดับเบสของดีเอ็นเอ

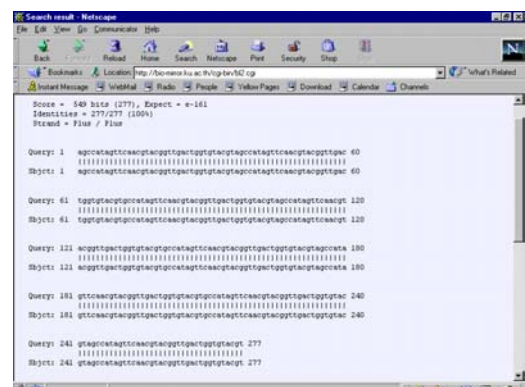
เพื่อให้เห็นถึงการนำข้อมูลของ Bio-Mirror มาประยุกต์ได้อย่างชัดเจน เราจะแสดงตัวอย่างการประยุกต์ดังรูปที่ 1 และรูปที่ 2 ซึ่งเป็นรูปการเปรียบเทียบลำดับเบสของดีเอ็นเอของไวรัสสองสาย โดยสมมติให้ลำดับเบสของดีเอ็นเอสายหนึ่งได้มาจากเชื้อไวรัสที่ได้มาจากคนที่มีความผิดปกติของโรค คล้ายกับคนเป็นโรคไวรัสตับอักเสบ B คือ Sequence 1 ในรูปที่ 1 และให้ลำดับเบสอีกสายหนึ่งคือลำดับเบสของดีเอ็นเอของไวรัสตับอักเสบ B ที่มีอยู่ในฐานข้อมูล Bio-Mirror คือ

Sequence 2 ในรูปที่ 1 เมื่อกดปุ่ม Align เพื่อเปรียบเทียบความคล้ายกันของลำดับเบสทั้ง 2 สายจะได้ผลลัพธ์ดังรูปที่ 2

ในการวินิจฉัยโรคโดยอาศัยการตรวจลำดับเบสของดีเอ็นเอ เราจะนำลำดับเบสสองสายมาเปรียบเทียบความคล้ายกัน หากลำดับเบสทั้งสองสายมีความคล้ายกัน เราก็สามารถวินิจฉัยได้ว่าคนๆนั้นติดเชื้อไวรัสตับอักเสบ B จริง ซึ่งมีความถูกต้องเกือบร้อยเปอร์เซ็นต์ในการตรวจวินิจฉัยโรคด้วยวิธีนี้



รูปที่ 1 การเปรียบเทียบความคล้ายกันระหว่างลำดับเบสของเชื้อไวรัส 2 สายเพื่อตรวจวินิจฉัยโรค



รูปที่ 2 ผลลัพธ์ของการเปรียบเทียบความคล้ายกันระหว่างลำดับเบสของเชื้อไวรัสสองสาย

จากตัวอย่างจะเห็นว่ากรนำข้อมูลลำดับเบสในระบบข้อมูล Bio-Mirror มาประยุกต์ใช้ในงานทางด้านวิทยาศาสตร์การแพทย์นั้นนับว่ามีประโยชน์มาก ยังมีงานวิจัยที่สามารถนำข้อมูลเหล่านี้ไปใช้ประโยชน์ได้อีกมากมาย งานทางด้านเกษตรกรรม การปรับปรุงพันธุ์พืช พันธุ์สัตว์ก็เป็นอีกงานหนึ่งที่ในปัจจุบันมีการศึกษาและวิจัยกันอยู่มาก ข้าวซึ่งเป็นพืชทางเศรษฐกิจที่สำคัญของประเทศนั้น ในปัจจุบันก็ได้มีองค์กรต่างๆ มากมายทั่วโลกได้ร่วมมือกันหาลำดับทางพันธุกรรมของข้าว เพื่อประโยชน์ในการปรับปรุงพันธุ์ สามารถต้านทานศัตรูพืช ทนต่อสภาพแวดล้อม และแพร่พันธุ์ได้มากขึ้น [13] เป็นต้น นอกจากนี้การทำโคลนนิ่งวัวที่สำเร็จไปไม่นานมานี้ ก็ได้มา

จากการวิเคราะห์และศึกษาข้อมูลลำดับเบสของดีเอ็นเอของมันนั่นเอง ซึ่งนับได้ว่าข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพในระบบข้อมูล Bio-Mirror นั้นมีประโยชน์สูงมาก หากมีผู้สนใจและนำข้อมูลเหล่านี้ไปประยุกต์ใช้หรือศึกษาเพื่อนำมาทำการวิจัยต่อไปในอนาคต

6. งานวิจัยที่สามารถทำต่อไปในอนาคตกับระบบข้อมูล

Bio-Mirror

งานวิจัยที่สามารถทำต่อไปกับฐานข้อมูลลำดับเบสของดีเอ็นเอและโปรตีนทางชีวภาพนั้น มีมากมายหลายด้าน แต่ที่จำเป็นมากที่สุดคือ งานทางด้านชีววิทยาซึ่งใช้ทั้งในงานวิจัยทั่วไปและในการเรียนระดับสูง งานวิจัยที่จำเป็นอีกงานหนึ่งก็คือ งานทางด้านวิทยาการคอมพิวเตอร์ เพราะฐานข้อมูลนี้มีขนาดใหญ่และเป็นข้อมูลเฉพาะด้าน ในการวิเคราะห์และจัดการกับฐานข้อมูลนี้เพื่อให้ได้ผลลัพธ์ที่ถูกต้องและรวดเร็วจำเป็นต้องใช้โปรแกรมคอมพิวเตอร์เข้ามาช่วย ซึ่งในปัจจุบันมีโปรแกรมคอมพิวเตอร์สำหรับวิเคราะห์และจัดการกับฐานข้อมูลนี้มากมายดังที่กล่าวมาแล้ว แต่ขั้นตอนวิธีและเทคนิคต่างๆ ที่ใช้ในการวิเคราะห์และจัดการข้อมูลให้ได้อย่างรวดเร็ว และมีความถูกต้องแม่นยำตามความต้องการของผู้ใช้นั้นยังไม่มีดีเพียงพอ

ดังนั้นการออกแบบขั้นตอนวิธีและหาเทคนิคใหม่ๆ มาใช้ในการวิเคราะห์และจัดการข้อมูลจึงเป็นสิ่งจำเป็น นอกจากนั้นฐานข้อมูลที่เพิ่มขึ้นอยู่ทุกวันซึ่งมีทั้งข้อมูลใหม่ที่เป็นข้อมูลชนิดเดิมและชนิดใหม่รวมอยู่ ยังต้องการโปรแกรมใหม่ที่จะรองรับสำหรับวิเคราะห์และจัดการกับข้อมูลนั้นตามความต้องการของผู้ใช้งานอีกด้วย

เอกสารอ้างอิง

- [1] Asia-Pacific Advanced Network (APAN), Available URL: <http://www.apan.net/>
- [2] BLAST, Available URL: <http://www.ncbi.nlm.nih.gov/BLAST/>
- [3] Blocks WWW Server, Available URL: <http://blocks.fhrc.org/>
- [4] Dan Gusfield, "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology", Cambridge University Press, 1997.
- [5] DDBJ Dna Data Bank of Japan, Available URL: <http://ddbj.nig.ac.jp/>
- [6] EBI, Available URL: <http://www.ebi.ac.uk/>
- [7] EMBL Outstation European Bioinformatics Institute, Available URL: <http://www.ebi.ac.uk/embl/>
- [8] Enzyme Nomenclature Database, Available URL: <http://www.expasy.ch/enzyme/>
- [9] Expasy Database, Available URL: <http://www.expasy.ch/>
- [10] FASTA, Available URL: <http://www.ncbi.nlm.nih.gov/BLAST/fasta.html>
- [11] National Institute of Health (NIH), Available URL: <http://www.nih.gov>
- [12] Pieve Baldi and Soren Brunak, "Bioinformatics: The Machine Learning Approach", The MIT Press Cambridge, 1999.
- [13] Prapa Sripichit, Pradit Pongtongkam and Apichart Vanavichit, Breeding of rice (oryza sativa.) variety KDML 105 for photoperiod intensivity through anther culture of F1 Hybrids, Kasetsart Journal, 1994, 28:499-511.
- [14] Prosite Database of protein families and domains, Available URL: <http://www.expasy.ch/prosite/>
- [15] SWISS-PROT annotated protein sequence database, Available URL: http://expasy.nhri.org.tw/sprot/sprot_details.html
- [16] The PFAM Database of protein domains and HMMS, Available URL: <http://pfam.wustl.edu/>
- [17] Protein Information Resource, Available URL: <http://www-nbrf.georgetown.edu/pir/>
- [18] Genbank, Available URL: <http://www.ncbi.nlm.nih.gov/Database/>
- [19] Rice Genome Project, Available URL: <http://www.dna.affrc.go.jp/>
- [20] SRS, Available URL: <http://srs.ebi.ac.uk/>
- [21] Taxonomy, Available URL: <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>
- [22] Trans-Pacific Network (TransPAC), Available URL: <http://www.transpac.org/>
- [23] TREMBL, Available URL: <http://srs.ebi.ac.uk/>
- [24] The Restriction Enzyme Database, Available URL: <http://www.neb.com/rebase/>
- [25] UniGene Resources, Available URL: <http://www.ncbi.nlm.nih.gov/UniGene/>
- [26] Very High Speed Backbone Service (vBNS), Available URL: <http://www.vbns.net/>

Somnuk Keretho

Assistant Professor



Somnuk Keretho received bachelor degree and master degree in electrical engineering from Kasetsart University in 1970 and 1976 respectively and received master degree in computer application and doctoral degree in computer science from Asian Institute of Technology in 1995 and 1982 respectively. He has research experiences in the fields of Artificial Intelligence, Neural Computation, Object-Oriented and Java Technology, Design and Analysis of Algorithms and Management Information Systems. At the present, he is an assistant professor at the Department of Computer Engineering and a deputy director of Kasetsart University and Research Development Institute (KURDI). His current research interests are Management Information Systems and Bioinformatics.

Punpiti Piamsa-nga

Lecturer

Punpiti Piamsa-nga received his B.Eng. and M.Eng. (Electrical Engineering) from Kasetsart University (KU,) Thailand in 1988 and 1991, respectively, and received his D.Sc. (Computer Engineering) from George Washington University in 1999. He has research experiences in the fields of Natural Language Processing, Parallel Computing, and Multimedia. He has been a faculty at the Department of Computer Engineering, KU since 1991. His current research interests are Bioinformatics and Multimedia.