

บทที่ 1

บทนำ

เพื่อแสดงถึงความสำคัญของประเด็นปัญหาและแนวคิดในการแก้ปัญหา ในบทนี้จึงได้นำเสนอหลักการ ทฤษฎี และเหตุผล ที่เกี่ยวข้องกับการจำแนกชนิดโมเลกุลของเอชแอลเอในแต่ละกลุ่ม ซึ่งสามารถนำไปใช้เป็นองค์ความรู้ในการสร้างความสัมพันธ์ระหว่างโมเลกุลได้ต่อไปในอนาคต และในการวิเคราะห์ข้อมูลนั้นจะพิจารณาลักษณะของยีน ซึ่งใช้วิธีซอฟต์แวร์คอมพิวเตอร์-แมชชีนในการจำแนกชนิดข้อมูล ตลอดจนนำเสนอวัตถุประสงค์ของการศึกษาและประโยชน์ที่คาดว่าจะได้รับ อันเป็นแรงจูงใจในการศึกษาวิจัยนี้ นอกเหนือจากนั้น ยังได้นำเสนอขอบเขตและวิธีการวิจัย เพื่อแสดงลำดับขั้นตอนของการทำวิจัยโดยละเอียด ซึ่งถูกใช้เป็นแนวทางในการดำเนินการวิจัยและเพื่อให้การวิจัยบรรลุถึงเป้าหมายที่ตั้งไว้อย่างสมบูรณ์ ท้ายที่สุดจะกล่าวถึงสถานที่และอุปกรณ์ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูลในการวิจัยครั้งนี้

1.1 หลักการ ทฤษฎี และเหตุผล

ในปัจจุบันนี้เทคโนโลยีทางด้านชีววิทยามีความก้าวหน้าขึ้นมาก ทำให้ข้อมูลทางวิชาการต่างๆ ได้ถูกนำไปวิเคราะห์ใช้งานให้เกิดประโยชน์จำนวนมหาศาล ซึ่งข้อมูลดังกล่าวที่มักนิยมนำมาใช้ในการวิเคราะห์สามารถจัดได้เป็น 2 หมวดใหญ่ๆด้วยกัน คือ ข้อมูลสายลำดับดีเอ็นเอ (DNA Sequence) และข้อมูลการแสดงออกของยีน (Gene Expression)

การวิเคราะห์ข้อมูลในสมัยก่อนนั้นจะพิจารณาข้อมูลสายลำดับดีเอ็นเอและการแสดงออกของยีนนั้นจะแยกจากกันอย่างสิ้นเชิง การวิเคราะห์ข้อมูลสายลำดับของยีนนั้นจะเน้นกระบวนการเทียบเรียงข้อมูลลำดับของยีน (Sequence Alignment) เพื่อนำไปใช้ระบุว่าลำดับยีนที่เราสงสัยนั้นคือยีนที่เราสนใจหรือไม่ รวมถึงการจัดกลุ่มของยีน (Sequence Clustering) และการระบุตำแหน่งของยีนในสายลำดับยีนของสิ่งมีชีวิต (Gene Detection) ส่วนการวิเคราะห์ข้อมูลการแสดงออกของยีนจะเน้นที่การจัดกลุ่มของยีน (Gene Clustering) และการทำนายกลุ่มยีน (Gene Classification and Prediction)

การวิเคราะห์ข้อมูลเหล่านี้มีเป้าหมายสำคัญคือการหาความสัมพันธ์หรือการควบคุมกันระหว่างยีน (Gene Regulatory Network Reconstruction) ซึ่งเป็นสิ่งสำคัญที่จะสามารถทำให้นักวิทยาศาสตร์เข้าใจถึงกระบวนการต่างๆ ที่เกิดขึ้นภายในเซลล์ และจะสามารถทำให้สร้างองค์

ความรู้ใหม่ในการพัฒนาเทคโนโลยีชีวภาพให้ก้าวหน้ายิ่งขึ้น เพื่อแก้ไขปัญหาต่างๆของมวลมนุษยชาติเช่น ด้านโรคภัยไข้เจ็บ และด้านเกษตรกรรม ในการเพาะปลูกให้เพิ่มผลผลิตมากขึ้น เป็นต้น การใช้ข้อมูลอย่างใดอย่างหนึ่งมาวิเคราะห์บางครั้งยังไม่สามารถให้ผลการทดลองที่ดีได้เท่าที่ควร ดังนั้นจึงต้องมีการนำข้อมูลทางชีววิทยาอื่นๆ มาทำการวิเคราะห์ประกอบด้วยเพื่อให้เกิดผลที่น่าเชื่อถือมากยิ่งขึ้น

ในงานวิจัยฉบับนี้ จะนำเสนอคอร์เนลฟังก์ชันสำหรับการจำแนกชนิดของข้อมูลสายลำดับโดยวิธีซอฟต์แวร์แมชชีนการเรียนรู้การจำแนกชนิดชีวแมน ลิวโคไซต์ แอนติเจนเพื่อจำแนกชนิดโมเลกุลของแอลกอฮอล์ในแต่ละกลุ่ม ซึ่งสามารถนำไปใช้เป็นองค์ความรู้ในการสร้างความสัมพันธ์ระหว่างกันได้ต่อไปในอนาคต ในการวิเคราะห์ข้อมูลด้วยลักษณะของโมเลกุล ซึ่งมีพื้นฐานแนวคิดมาจากการตัดสินใจ ในการจำแนกหรือจัดกลุ่มของข้อมูลด้วยการแบ่งขอบเขต โดยการประมวลผลของการจำแนกจะพิจารณาจากข้อมูล โดยมองออกเป็นเซตของเวกเตอร์ โดยมีวัตถุประสงค์เพื่อให้ใช้งานได้อย่างยืดหยุ่นกับกลุ่มข้อมูลต่างๆ และเป็นวิธีการที่ใช้เกี่ยวกับการเรียนรู้ด้วยเครื่องที่สามารถใช้กันได้เป็นวงกว้างในหลายๆโดเมน และสามารถแบ่งแยกข้อมูลในมิติที่สูงได้ และครอบคลุมทุกกรณีที่เป็นไปได้ในขอบเขตการพิจารณาที่เป็นมาตรฐาน โดยเฉพาะอย่างยิ่งการเพิ่มคุณลักษณะของข้อมูลใหม่ๆ เป็นต้น หลังจากนั้นจึงนำผลลัพธ์ที่ได้มาวัดค่าความถูกต้องและเปรียบเทียบกับเทคนิคทันสมัยอื่นๆ ต่อไป

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อนำเสนอขั้นตอนวิธีในการจำแนกชนิดของข้อมูลสายลำดับโดยวิธีซอฟต์แวร์แมชชีน

1.2.2 เพื่อนำเสนอคอร์เนลฟังก์ชันสำหรับการจำแนกชนิดของข้อมูลสายลำดับโดยวิธีซอฟต์แวร์แมชชีน

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

จากการศึกษาค้นคว้า คาดว่าจะได้รับประโยชน์ดังนี้

- 1) ได้นำเสนอขั้นตอนวิธีในการจำแนกชนิดของข้อมูลสายลำดับ
- 2) ได้องค์ความรู้ใหม่ที่เกี่ยวข้องกับการจำแนกชนิดของข้อมูลสายลำดับ

1.4 ขอบเขต และ วิธีการวิจัย

1.4.1 ขอบเขต

เพื่อให้บรรลุวัตถุประสงค์ทางการศึกษา จึงกำหนดขอบเขตของการค้นคว้าดังนี้

- 1) พัฒนาขั้นตอนวิธีในการจำแนกชนิดข้อมูลสายลำดับโดยวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีศึกษาการจำแนกชนิดอิวแมน ลิวโคไซต์ แอนติเจนเพื่อจำแนกชนิดของยีนในแต่ละกลุ่ม
- 2) พิสูจน์ความถูกต้องของขั้นตอนวิธีที่ได้ในข้อ 1 โดยใช้วิธีการวัดประสิทธิภาพของข้อมูล
- 3) ข้อมูลที่ใช้เป็นกรณีศึกษาคือ ตัวแบบของกลุ่มยีนจากฐานข้อมูลของข้อมูลสายลำดับเอชแอลเอ

1.4.2 วิธีการวิจัย

เพื่อให้การทำวิจัยสำเร็จตามแผนที่กำหนดจึงออกแบบวิธีการดำเนินการวิจัยดังนี้

- 1) ศึกษาทฤษฎี งานวิจัยที่เกี่ยวข้อง และเครื่องมือที่ใช้
 - 1.1) ศึกษาและค้นคว้าข้อมูลโดยศึกษาถึงดีเอ็นเอของโมเลกุลเอชแอลเอ เพื่อเป็นกลุ่มข้อมูลในการศึกษา
 - 1.2) ศึกษาขั้นตอนวิธีในการจำแนกชนิดของข้อมูลสายลำดับ
 - 1.3) ศึกษาเครื่องมือและทฤษฎีต่างๆในการจำแนกข้อมูลสายลำดับ
 - 1.4) ศึกษางานวิจัยที่เกี่ยวข้อง
- 2) ออกแบบขั้นตอนวิธีการในการจำแนกข้อมูลสายลำดับ
 - 2.1) กำหนดนิยามของรูปแบบข้อมูลสายลำดับในเชิงรูปนัย
 - 2.2) กำหนดนิยามของการจำแนกข้อมูลสายลำดับ
 - 2.3) กำหนดนิยามของแนวทางการแก้ปัญหาการจำแนกข้อมูลสายลำดับในรูปของการเรียนรู้ด้วยเครื่อง
 - 2.4) ออกแบบวิธีการแปลงข้อมูลสายลำดับที่สามารถทำให้เครื่องคอมพิวเตอร์เรียนรู้ได้
 - 2.5) พัฒนาวิธีการจำแนกข้อมูลสายลำดับในรูปของการเรียนรู้ด้วยเครื่องให้มีประสิทธิภาพมากขึ้นได้
- 3) ทดสอบขั้นตอนวิธีที่ได้พัฒนา
 - 3.1) นำตัวแบบของโมเลกุลเอชแอลเอจากฐานข้อมูลเพื่อใช้เป็นกรณีศึกษา
 - 3.2) นำขั้นตอนวิธีที่ได้จากข้อ 2 มาทดสอบการจำแนกชนิดของข้อมูลสายลำดับโดยใช้ตัวแบบที่ได้ในข้อ 3.1

4) วัดความถูกต้องของขั้นตอนวิธี

4.1) วัดความถูกต้องของการจำแนกชนิดข้อมูลสายลำดับโดยวิธีซัพพอร์ต

เวกเตอร์แมชชีน

4.2) เปรียบเทียบความถูกต้องของการจำแนกชนิดข้อมูลสายลำดับกับ
วิธีการหรือเทคนิคทันสมัยอื่นๆ

5) ประเมินผลและจัดทำเอกสารสรุปผลการศึกษา

5.1) วิเคราะห์ผลการทดสอบของขั้นตอนวิธีในการจำแนกชนิดของ
ข้อมูลสายลำดับที่ได้พัฒนาขึ้น

5.2) สรุปผล

1.5 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล

1.5.1 สถานที่

สถานที่ที่ใช้ในการดำเนินการวิจัยประกอบด้วย

- 1) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
- 2) ห้องปฏิบัติการวิจัยชีวสารสนเทศศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

1.5.2 อุปกรณ์ที่ใช้ในการวิจัย

อุปกรณ์ที่ใช้ในการวิจัยนี้ประกอบไปด้วย

- 1) ฮาร์ดแวร์ ได้แก่ เครื่องไมโครคอมพิวเตอร์ ภายใต้ระบบปฏิบัติการไมโครซอฟท์ วินโดวส์ เอ็กซ์พี จำนวน 1 เครื่อง
- 2) ซอฟต์แวร์ ได้แก่ โปรแกรมภาษาอาร์ (R) เพื่อใช้พัฒนาโปรแกรมสำหรับการวิเคราะห์ข้อมูลยีน (สนับสนุน อาร์ เวอร์ชัน 2.9.0)