บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ลักษณะและวิธีการแสดงผลภาพ (Visualization Techniques)

การแสดงผลภาพจากข้อมูล เป็นการสรุปรวมข้อมูลให้แสดงเป็นภาพในรูปแบบต่างๆ เช่น รูปภาพกราฟแท่ง ภาพกราฟเส้น หรือภาพวงกลม เพื่อให้ง่ายในการพิจารณาและทำความเข้าใจ ด้วยเหตุนี้เองทำให้การแสดงผลภาพจากข้อมูล [1][2][3] ถูกพัฒนาและนำเสนอในรูปแบบอื่นๆ มากขึ้น เพื่อให้เหมาะสมกับข้อมูลแต่ละประเภท โดยเฉพาะข้อมูลที่มีความซับซ้อนและมีปริมาณมาก จนไม่สามารถพิจารณาหรือทำความเข้าใจได้โดยง่าย [2] การแสดงผลภาพจากข้อมูลดังกล่าว ได้มีการนำเสนอในหลากหลายวิธี โดยมีลักษณะและวิธีการแสดงผลภาพที่น่าสนใจและเป็นแนวทางในงานวิจัยนี้ ยกตัวอย่างเช่น

2.1.1 การแสดงผลแบบเคออสเกม

การแสดงผลแบบเคออสเกม (Chaos Game Representations) ถูกประยุกต์มาจากทฤษฎีเคออสเกม [1][4] ที่มีอัลกอริธึมในการแสดงผลภาพจากการเกิดขึ้นของจุดโดยการทำซ้ำแบบสุ่ม การแสดงผลแบบเคออสเกมมีหลักการมาจากวิธีการกำหนดจุดเริ่มต้น เป็นเสมือนมุมของภาพและกำหนดจุดที่เกิดขึ้นต่อไปบริเวณตรงกลางระหว่างจุดสองจุดใดๆ ที่มีอยู่แล้วโดยการสุ่ม ซึ่งรายละเอียดจะกล่าวในหัวข้อ 2.2 ต่อไป

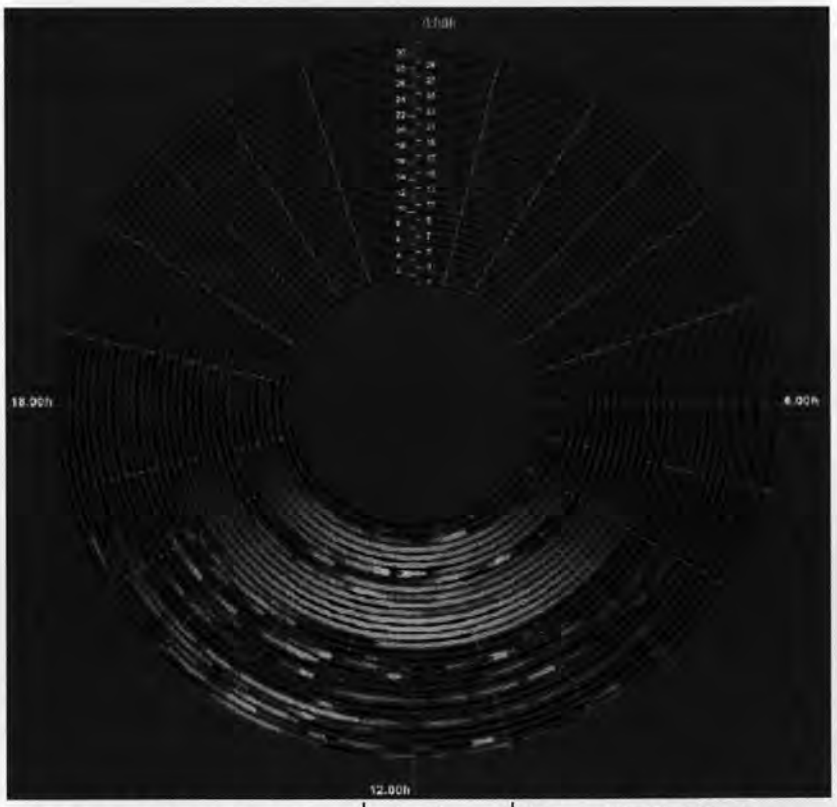
การกำหนดจุดแต่ละจุดจะเกิดโดยการทำซ้ำไปเรื่อยๆ จนได้ภาพที่มีลักษณะเฉพาะเกิดขึ้น การแสดงผลแบบเคออสเกมได้มีการนำไปใช้ในการศึกษารูปแบบของชุดข้อมูลดีเอ็นเอ [1] โดยระบุจุดเริ่มต้นขอบเขต 4 จุด เป็นตัวอักษรที่กำกับด้วยอักขระในข้อมูลดีเอ็นเอ แล้วใช้ข้อมูลของดีเอ็นเอสร้างจุดต่างๆขึ้นมา ต่อมาการแสดงผลแบบเคออสเกมได้ถูกพัฒนาจากการกำหนดจุดมาเป็นการนับความถี่ของอักขระของข้อมูลดีเอ็นเอที่เกิดขึ้นแทน ซึ่งการแสดงผลแบบเคออสเกม ของข้อมูลโครงสร้างของดีเอ็นเอหรือหน่วยพันธุกรรมแบบนับความถี่ เป็นการแสดงผลภาพจากข้อมูลที่มีประสิทธิภาพในการวิเคราะห์โครงสร้างและการจัดกลุ่ม ทำให้สามารถทำความเข้าใจในข้อมูล ได้ง่ายและรวดเร็วขึ้น [4]

2.1.2 การแสดงผลภาพแบบวงแหวน

การแสดงผลภาพแบบวงแหวน (Spiral Representation) เป็นการนำเสนอวิธีการในการวิเคราะห์ข้อมูลอนุกรมเวลา (Time Series) ที่มีปริมาณมากโดยเฉพาะ [3] โดยส่วนมากการแสดงผลภาพแบบวงแหวน ถูกนำมาใช้ในการวิเคราะห์ข้อมูลเชิงวิศวกรรมศาสตร์ วิทยาศาสตร์ หรือแม้แต่สาขาอื่นๆ ที่มีข้อมูลเป็นแบบอนุกรมเวลา และมีขนาดข้อมูลที่ต้องการวิเคราะห์เป็น

จำนวนมาก การแสดงผลภาพแบบวงแหวนอาศัยหลักการสร้างรูปภาพที่มีลักษณะเป็นวงแหวน ซึ่งเกิดจากการนำข้อมูลในแต่ละช่วงเวลามาวางเรียงกันขดเป็นวงอย่างต่อเนื่อง ทำให้มีลักษณะเหมือนเป็นวงแหวนที่ซ้อนกันอยู่หลายวง

การแสดงภาพแบบวงแหวน มีลักษณะการจัดวางข้อมูลแต่ละช่วงในวง ให้เกิดความสัมพันธ์กับทุกๆส่วนของวง ซึ่งนำเสนอเพื่อเปรียบเทียบรายละเอียดของข้อมูลในช่วงเวลาที่ต่างกันในแต่ละคาบของข้อมูล เช่น การเปรียบเทียบปริมาณการใช้ไฟฟ้าในรอบวัน หรือสัปดาห์ ช่วงเวลาที่มีการใช้ไฟฟ้ามากจะปรากฏช่วงของวงแหวนที่มีความเข้มมาก เป็นต้น ซึ่งความเข้มนั้นจะเกิดจากความถี่ที่มีปริมาณของข้อมูลเป็นจำนวนมาก นอกจากนี้การแสดงผลภาพแบบวงแหวนยังสามารถเพิ่มเติมคุณลักษณะพิเศษเพื่อใช้แสดงรายละเอียดต่างๆ ได้มากยิ่งขึ้น เช่น การขยายส่วนเฉพาะจุด (Zooming) เป็นการเลือกเข้าไปดูข้อมูลโดยละเอียดเฉพาะส่วนใดส่วนหนึ่ง โดยเลือกดูข้อมูลเฉพาะช่วงที่สนใจบางส่วนได้ [2] หรือการเน้นและเชื่อมโยงข้อมูล คือ สามารถเลือกเอาข้อมูลบางช่วงที่น่าสนใจหลายๆช่วง แยกออกมาวิเคราะห์และหาความสัมพันธ์กัน วิธีการแสดงผลภาพแบบวงแหวนยังเป็นวิธีการหนึ่ง ที่สามารถแสดงข้อมูลที่มีขนาดใหญ่และมีปริมาณข้อมูลมาก ให้สามารถแสดงผลอยู่ในพื้นที่ที่จำกัดได้ (Information Mural) [5] แสดงตัวอย่างของการแสดงผลภาพแบบวงแหวน ดังรูปที่ 2.1



รูปที่ 2.1 การแสดงผลภาพแบบวงแหวน ซึ่งเป็นข้อมูลที่มีความสัมพันธ์กับช่วงเวลาทุกๆวัน ในเวลา 24 ชั่วโมง (ที่มา: Weber, M., Alexa, M., and Mueller, W.) [3]

หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

2.1.3 การแสดงผลภาพแบบบิตแม็บ

การแสดงผลภาพบิตแม็บ (Bitmap Representation) เป็นวิธีการแสดงผลภาพแบบดิจิทัล ที่มีข้อมูลหรือโครงสร้างเป็นลักษณะรูปสี่เหลี่ยมที่เกิดจากจุดของสี ซึ่งแสดงอยู่บนจอคอมพิวเตอร์ หรือในอุปกรณ์แสดงผลต่างๆ ที่เรารู้จักกันโดยทั่วไป โดยจุดสีที่เกิดขึ้นมาแต่ละจุดเกิดจากค่าที่ประกอบด้วยจำนวนบิตจำนวน 3 ค่า ซึ่งเป็นไปตามรูปแบบการแสดงสีแบบอาร์จีบี (RGB Color Space) กล่าวคือ ประกอบไปด้วยค่าบิตสีจำนวน 3 แบบ คือ บิตที่แสดงค่าสีแดง สีเขียว และสีน้ำเงิน การเพิ่มจำนวนบิตต่อหนึ่งจุดการแสดงสีสามารถทำให้การแสดงสีต่อจุดมีได้มากขึ้น แต่จำเป็นต้องใช้ขนาดของหน่วยความจำในการเก็บข้อมูลต่อจุดของภาพบิตแม็บ หรือภาพดิจิทัลมากขึ้นตามไปด้วย ในทางกลับกันถ้าต้องการแสดงสีของภาพออกมาเป็นเพียงสเกลสีเทา (Grayscale) จะอาศัยจำนวนบิตข้อมูลเพียง 2 บิต หรือหากต้องการแสดงผลข้อมูลเพียงภาพสีขาวดำ จะใช้จำนวนบิตข้อมูลเพียง 1 บิตเท่านั้น

การแสดงผลภาพบิตแม็บบนหน้าจอ มีการแสดงค่าความละเอียดและสีสันจะขึ้นอยู่กับลักษณะความกว้างและความยาวของภาพ ประกอบกับจำนวนบิตและจุดสี (Pixel) ที่ใช้ซึ่งบ่งบอกถึงจำนวนของสีที่สามารถแสดงได้ ในปัจจุบันการแสดงภาพบิตแม็บโดยทั่วไปจะใช้ขนาดของบิตจำนวน 24 บิตต่อหนึ่งจุดสี เพื่อทำให้สามารถแสดงสีให้ได้เพียงพอต่อภาพเสมือนจริง ซึ่งในบางครั้งอาจมากเกินกว่าที่สายตามนุษย์สามารถรับรู้ได้ และทำให้ต้องใช้หน่วยความจำขนาดใหญ่ในการเก็บข้อมูล รวมถึงต้องใช้เวลาในการประมวลผลมากอีกด้วย

การแสดงผลภาพดังกล่าวข้างต้น เป็นเพียงส่วนหนึ่งของการแสดงผลภาพที่เป็นที่รู้จักและถูกนำไปใช้กันอย่างกว้างขวาง อย่างไรก็ตามยังมีการแสดงผลภาพวิธีอื่นๆ อีกเป็นจำนวนมาก ซึ่งถูกนำเสนอในลักษณะต่างๆ ขึ้นอยู่กับรูปแบบของข้อมูลและจุดประสงค์ที่จะนำไปใช้ [2]

2.2 ทฤษฎีเคออสเกม

เคออสเกม (Chaos Game) เป็นอัลกอริธึมที่ใช้ในการสร้างรูปภาพจากจุด ที่เกิดจากการทำซ้ำโดยการสุ่ม [1][4] เคออสเกมถูกนำมาใช้ในการแสดงรูปแบบของข้อมูลดีเอ็นเอ หรือโครงสร้างหน่วยพันธุกรรม ซึ่งมีอัลกอริธึม [1] ดังนี้

อัลกอริธึมของเคออสเกมเริ่มจากการกำหนดจุดเริ่มต้น จากนั้นจะสุ่มกำหนดจุดต่อๆไปเพื่อสร้างเป็นภาพขึ้นมา โดยมีขั้นตอนดังนี้

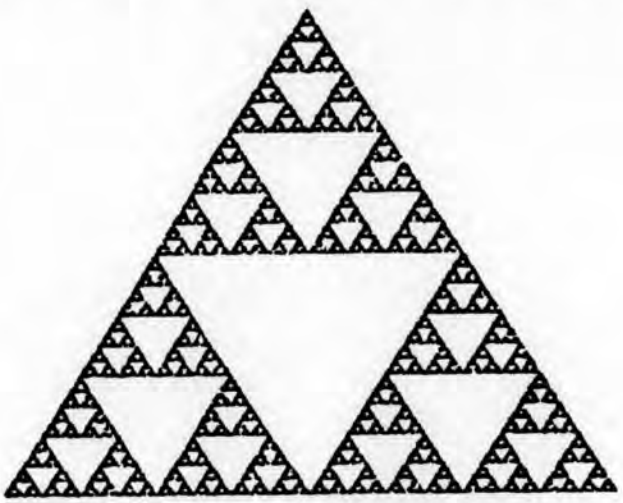
ขั้นตอนที่ 1 กำหนดจุดเริ่มต้น อย่างน้อย 3 จุด ซึ่งจะกำหนดที่ตำแหน่งใดก็ได้ แต่จุดที่กำหนดนั้นต้องไม่เป็นตำแหน่งที่สามารถลากผ่านได้โดยเส้นตรงเส้นเดียวพร้อมกัน ทั้ง 3 จุด

ขั้นตอนที่ 2 กำหนดชื่อของแต่ละจุดเริ่มต้นนั้น เป็นเลขจำนวนนับที่ต่อเนื่อง

ขั้นตอนที่ 3 เลือกจุดตั้งต้น 2 จุด จากจุดเดิมที่เคยกำหนดมาแล้วโดยการสุ่ม จะกำหนดจุดที่เกิดขึ้นมาใหม่จากบริเวณตรงกลางระหว่างจุด 2 จุดที่เลือกนั้น

ขั้นตอนที่ 4 ดำเนินการแบบขั้นตอนที่ 3 ไปเรื่อยๆ โดยยึดหลักการแบบเดิม

เมื่อทำซ้ำจากขั้นตอนข้างต้นเป็นจำนวนหลายพันครั้ง ผลที่ได้ออกมาจากการเลือกจุดแบบสุ่มที่กำหนดจุดเริ่มต้นเป็น 3 จุด เป็นภาพสามเหลี่ยมซ้อนกัน ดังรูปที่ 2.2

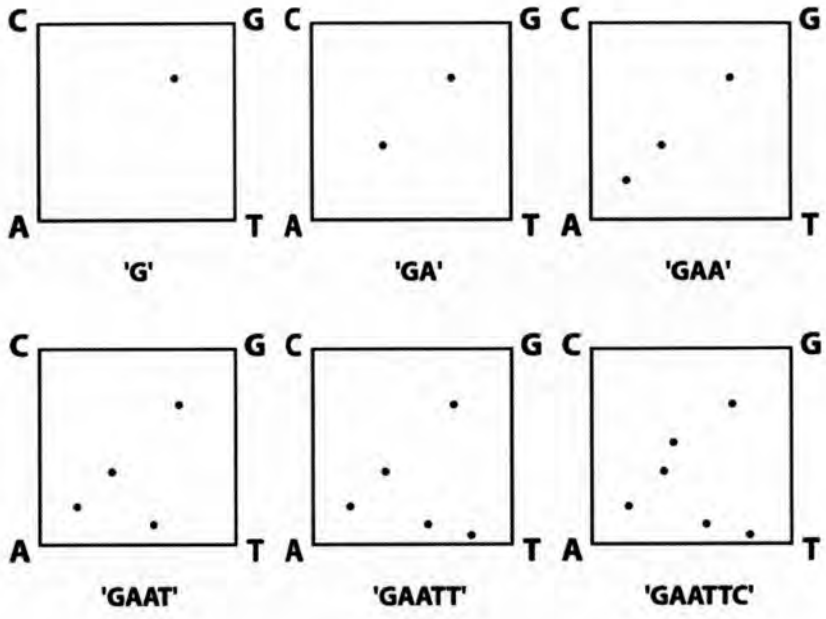


รูปที่ 2.2 ผลของการทำซ้ำจากอัลกอริธึมของเคออสเกม ที่เลือกจุดเริ่มต้น 3 จุด
(ที่มา: Jeffrey, H. J.) [1]

จากรูปที่ 2.2 เป็นผลจากหลักการทางคณิตศาสตร์ที่ได้มีการค้นพบมานานแล้ว และเป็นที่รู้จักกันดีในชื่อของ สามเหลี่ยมเซอร์ปินส์กี (Sierpinski Triangle) นอกจากนี้อัลกอริธึมจากหลักการของเคออสเกมยังสามารถกำหนดจุดเริ่มต้นที่มีจำนวนมากกว่า 3 จุดได้ ซึ่งให้ผลออกมาในลักษณะคล้ายกัน เช่น การกำหนดจุดเริ่มต้นเป็น 5 จุด 6 จุด หรือ 7 จุด จะทำให้เกิดรูปหลายเหลี่ยมตามจำนวนจุดเริ่มต้นที่กำหนด แต่สิ่งที่น่าสนใจในอัลกอริธึมของเคออสเกม คือ กรณีกำหนดจุดเริ่มต้นเป็น 4 จุด ผลที่ได้จะแตกต่างออกไปจากกรณีกำหนดจุดเริ่มต้นเป็น 3 จุด และมากกว่า 4 จุดขึ้นไป ซึ่งผลที่เกิดขึ้นจากการกำหนดจุดเริ่มต้น 4 จุดนั้น จะได้ภาพที่ไม่มีรูปแบบเฉพาะ โดยเกิดเป็นรูปสี่เหลี่ยมที่มีลักษณะไม่แน่นอนขึ้นอยู่กับจำนวนการเลือกจุดสุ่ม

ประโยชน์จากการเกิดรูปแบบภาพที่ไม่แน่นอน ที่เกิดจากการสุ่มโดยอัลกอริธึมของเคออสเกมนั้น ถูกนำไปประยุกต์ใช้ในการแสดงข้อมูลของดีเอ็นเอ เนื่องจากข้อมูลของดีเอ็นเอเป็นข้อมูลที่มีโครงสร้างเฉพาะ ซึ่งสามารถแปลงให้เกิดลักษณะของภาพที่มีลักษณะเฉพาะได้ อีกทั้งข้อมูลพื้นฐานของดีเอ็นเอประกอบไปด้วยตัวอักษร 4 ตัว คือ A C G และ T พอดีกับการนำไปกำหนดเป็นจุดเริ่มต้นที่มี 4 จุดได้

ดังนั้นเมื่อประยุกต์ใช้อัลกอริธึมของเคออสเกมกับข้อมูลของดีเอ็นเอ ยกตัวอย่างข้อมูลเช่น "GATTC" สามารถแสดงรายละเอียดขั้นตอนการกำหนดจุดได้ ดังรูปที่ 2.3



รูปที่ 2.3 ขั้นตอนการกำหนดจุดตามอัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอ "GAATTC"

จากรูปที่ 2.3 แสดงการเกิดจุดขึ้นอย่างต่อเนื่อง โดยมีลักษณะเหตุการณ์ดังต่อไปนี้

รูป 'G' เกิดจุดจากข้อมูล 'G' ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดบริเวณกึ่งกลางของภาพสี่เหลี่ยมกับมุมสี่เหลี่ยมด้าน 'G'

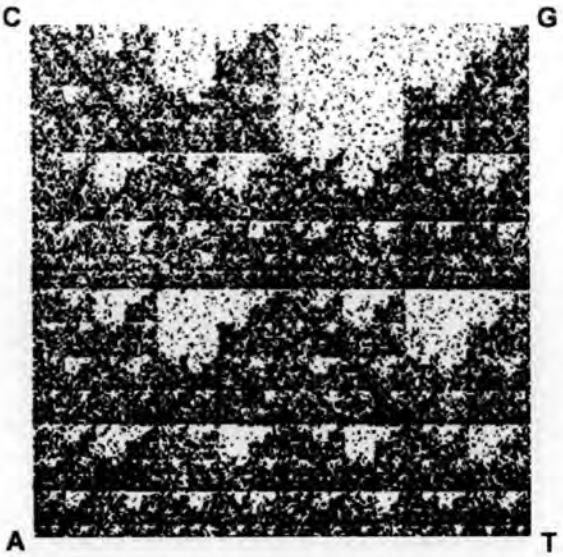
รูป 'GA' เกิดจุดจากข้อมูล 'A' ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดที่สร้างก่อนหน้านี้ ซึ่งคือจุด 'G' กับมุมสี่เหลี่ยมด้าน 'A'

รูป 'GAA' เกิดจุดจากข้อมูล 'A' ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดที่สร้างก่อนหน้านี้ ซึ่งคือจุด 'GA' กับมุมสี่เหลี่ยมด้าน 'A'

รูป 'GAAT', 'GAATT' และ'GAATTC' จะเกิดขึ้นจากเหตุการณ์ลักษณะเดียวกันไปเรื่อยๆ

เมื่อใช้อัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอขนาดความยาว 73,357 ตัวอักษร ซึ่งเป็น ดีเอ็นเอที่มาจาก HUMHBB [1] จะได้ภาพที่เกิดมาจากการกำหนดจุด ดังรูปที่ 2.4

ทฤษฎีเคออสเกมจัดเป็นแนวคิดพื้นฐานที่สำคัญของงานวิจัยเป็นจำนวนมาก สามารถนำไปประยุกต์และพัฒนาต่อเนื่องได้หลากหลาย งานวิจัยนี้จึงนำเอาแนวคิดดังกล่าวมาเป็นแนวทางในการทำวิจัยเช่นกัน เพราะทฤษฎีเคออสเกมเป็นแนวทางที่ทำให้เข้าใจถึงหลักการพื้นฐานของการแสดงผลภาพจากข้อมูลที่มีปริมาณมาก และสามารถนำไปเป็นแนวคิดในการวิจัยต่อเนื่องได้เป็นอย่างดี



รูปที่ 2.4 ภาพการประยุกต์ใช้อัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอขนาดความยาว 73,357 ตัวอักษร (HUMHBB) (ที่มา: Jeffrey, H. J.) [1]

2.3 การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ

การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ (Symbolic Time Series Representations) เป็นวิธีการนำข้อมูลอนุกรมเวลาที่มีปริมาณมาก และมีลักษณะข้อมูลเป็นเลขจำนวนจริง แปลงให้เป็นสัญลักษณ์หรืออักขระ เพื่อที่จะสามารถใช้ข้อมูลลักษณะดังกล่าวมาทำการสรุปรวมข้อมูล แบ่งแยก จำแนก หาลักษณะที่ผิดแปลกของข้อมูล หรือนำไปใช้ในงานต่อเนื่องอื่นๆ [6] จุดเด่นที่สำคัญของการแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ คือ การทำให้ข้อมูลที่เป็นเลขจำนวนจริงให้มาอยู่ในรูปแบบของสัญลักษณ์หรืออักขระ เพราะการประมวลผลกับข้อมูลอนุกรมเวลาโดยตรงในบางกรณี จะยากต่อการใช้อัลกอลริธึมในการจัดการรูปแบบข้อมูล เพราะอัลกอริธึมบางประเภทไม่รองรับการทำงานกับข้อมูลแบบต่อเนื่อง จำเป็นต้องทำให้ข้อมูลนั้นอยู่ในรูปแบบไม่ต่อเนื่องก่อน

การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ มีการวิจัยและนำมาใช้อยู่หลายวิธี อาทิเช่น วิธีสัญลักษณ์นิยม (Symbolizing) วิธีโทเค็น (Tokenizing) หรือวิธีการแจงหน่วย (Quantizing) [7][6] แต่ในการวิจัยนี้ได้ศึกษาและเลือกใช้วิธี การแปลงจากข้อมูลอนุกรมเวลาไปเป็นข้อมูลสัญลักษณ์หรืออักขระที่ได้มาจากการหาค่าเฉลี่ยโดยรวม หรือแซ็ค (Symbolic Aggregate approXimation - SAX) เนื่องจากวิธีการแบบแซ็คเป็นวิธีการที่ไม่ซับซ้อน ประมวลผลเร็ว และข้อมูลจากการแปลงจะไม่เสียคุณสมบัติจากข้อมูลเดิม [2]

สายสัญลักษณ์หรืออักขระของข้อมูลอนุกรมเวลาที่ได้มาจากการหาค่าเฉลี่ยโดยรวม ทำให้ข้อมูลอนุกรมเวลาขนาดความยาว n เปลี่ยนไปเป็นข้อมูลที่ลดขนาดลงเหลือ w ได้ โดยที่ขนาดของ w จะน้อยกว่าหรือเท่ากับ n เสมอ โดยปกติจะน้อยกว่าขนาดของ n มากๆ ซึ่งจะน้อยลงมากเพียงใดขึ้นกับการกำหนดขนาดสัดส่วนจำนวนเฉลี่ย (Piecewise Aggregate Approximation - PAA) ที่เป็นตัวกำหนดช่วงของข้อมูลเพื่อหาค่าเฉลี่ย ทำให้เกิดสายอักขระใหม่ที่มีขนาดเท่ากับ w ส่งผลให้เกิดการลดขนาดของข้อมูลจำนวนมาก ทำให้สามารถพิจารณาข้อมูลโดยรวมทั้งหมดได้ง่ายและข้อมูลไม่เสียรูปแบบจากเดิม [6]

วิธีการแปลงข้อมูลอนุกรมเวลาออกมาเป็นสัญลักษณ์หรืออักขระแบบแซ็ค มีหลักการที่สำคัญอยู่ 2 ประการ คือ การลดขนาดหรือมิติของข้อมูลโดยสัดส่วนจำนวนเฉลี่ย (PAA Dimensionality Reduction) และวิธีการแปลงข้อมูลจำนวนจริงของข้อมูลอนุกรมเวลา (Time Series Data) ไปเป็นข้อมูลอักขระ ซึ่งจะอธิบายรายละเอียดและวิธีการในหัวข้อ 3.2 ต่อไป

2.4 ทฤษฎีการจัดกลุ่มแบบเคมีน

กลุ่ม (Cluster) คือคอลเลคชัน (Collection) ของวัตถุ ซึ่งมีคุณสมบัติคือ วัตถุที่อยู่ในกลุ่มเดียวกันจะคล้ายกัน แต่แตกต่างจากวัตถุในกลุ่มอื่น การจัดกลุ่ม (Clustering) จึงเป็นการจำแนกประเภทของวัตถุที่มีความคล้ายกันให้อยู่ในกลุ่มเดียวกัน และจัดแยกวัตถุที่แตกต่างกันให้อยู่ต่างกลุ่มกัน ซึ่งเป็นการจำแนกโดยที่ไม่ทราบจำนวนกลุ่มและประเภทของกลุ่มล่วงหน้า [8]

การจัดกลุ่มที่ดีจะผลิตกลุ่มที่มีคุณภาพสูง ซึ่งเป็นกลุ่มที่วัตถุภายในกลุ่มเดียวกันมีความคล้ายกันสูง ขณะเดียวกันวัตถุที่อยู่ต่างกลุ่มจะมีความคล้ายกันต่ำ คุณภาพของผลลัพธ์การจัดกลุ่มขึ้นอยู่กับมาตรวัดความคล้ายคลึงที่ใช้และการนำวิธีการไปใช้ให้เกิดผล วิธีการจัดกลุ่มที่ดีควรจะสามารถค้นพบรูปแบบที่ซ่อนอยู่ในข้อมูลบางส่วนหรือทั้งหมด

โดยทั่วไปมาตรวัดค่าความคล้ายหรือไม่คล้ายกันของวัตถุ อยู่ในรูปแบบฟังก์ชันระยะห่าง $d(i,j)$ ซึ่งจัดเก็บในเมทริกซ์ความ(ไม่)คล้าย เปรียบเทียบกับข้อมูลของวัตถุ n ตัว แต่ละตัวจะมีคุณลักษณะจำนวน p คุณลักษณะ ดังรูปที่ 2.5

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & xi_f & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & xn_f & \cdots & x_{np} \end{bmatrix} \qquad \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

รูปที่ 2.5 ตารางเมทริกซ์คุณสมบัติของวัตถุและเมทริกซ์ความไม่คล้าย

นิยามฟังก์ชันระยะห่าง $d(i,j)$ จะมีความแตกต่างกันไปตามประเภทของข้อมูล และวิธีการวิเคราะห์ข้อมูล ทำให้ฟังก์ชันการหาระยะห่างที่ใช้วัดความคล้ายหรือไม่คล้ายระหว่างวัตถุ 2 ตัว มีหลายฟังก์ชัน ซึ่งฟังก์ชันที่ไม่ซับซ้อนและนิยมใช้กันอย่างแพร่หลายคือ การคำนวณระยะทางแบบแมนฮัทตัน (Manhattan Distance) ดังสมการที่ (2.1)

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \quad (2.1)$$

วิธีการจัดกลุ่มแบบเคมีน (*k*-Means Clustering) ใช้หลักการการตัดแบ่งของวัตถุ *n* ตัว ออกเป็นจำนวน *k* กลุ่ม (เมื่อทราบค่า *k*) อัลกอริธึมของเคมีน จะทำการตัดแบ่งของวัตถุเป็น *k* กลุ่ม โดยการแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลางของกลุ่มในการวัดระยะห่างของตัวอย่างในกลุ่มเดียวกัน โดยมีอัลกอริธึมดังรูปที่ 2.6

อัลกอริทึมการจัดกลุ่มด้วยวิธีเคมีน

ข้อมูลเริ่มต้น: จำนวนกลุ่ม *k* และวัตถุที่นำมาจัดกลุ่ม

ผลลัพธ์: ชุดของวัตถุในกลุ่ม จำนวน *k* กลุ่ม

วิธีการ:

(1) ทำการเลือกวัตถุจำนวน *k* ตัว เพื่อเป็นตัวแทนกลุ่มเริ่มต้น

(2) ทำซ้ำ

(2.1) ทำการเลือกวัตถุในกลุ่ม จากระยะทางของวัตถุกับตัวแทนกลุ่ม

(2.2) ทำการเลือกตัวแทนกลุ่มใหม่ จากค่าเฉลี่ยของวัตถุในกลุ่ม

(2.3) ทำซ้ำต่อไป จนกระทั่งสมาชิกในกลุ่มไม่มีการเปลี่ยนแปลง

รูปที่ 2.6 อัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีน [8]

วิธีการจัดกลุ่มแบบเคมีน อาจมีความแตกต่างกันไปบ้าง ในประเด็นของการคัดเลือกค่าเริ่มต้น การคำนวณค่าความคล้ายหรือไม่คล้ายของวัตถุ และวิธีการที่ใช้ในการคำนวณค่าเฉลี่ยของกลุ่ม

ข้อดีของวิธีเคมีน คือ เป็นวิธีการจัดกลุ่มที่มีประสิทธิภาพ เวลาที่ใช้ในการทำงานของอัลกอริธึม คือ $O(tkn)$ โดยที่ *n* คือจำนวนวัตถุ *k* คือจำนวนกลุ่ม และ *t* คือจำนวนรอบที่ทำซ้ำ ซึ่งปกติแล้วค่าของ *k* จะน้อยกว่าค่าของ *n* และ *t* มากๆ ($k << n,t$) แต่วิธีแบบเคมีนมีข้อด้อยที่การทำการจัดกลุ่มข้อมูลจำเป็นต้องกำหนดค่า *k* หรือจำนวนกลุ่มล่วงหน้า สามารถทำการจัดกลุ่มกับวัตถุที่สามารถหาค่าเฉลี่ยได้เท่านั้น และไม่เหมาะกับการจัดกลุ่มที่มีข้อมูลรบกวนหรือข้อมูลผิดปกติเป็นจำนวนมาก

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 **ภาพบิตแม็บของข้อมูลอนุกรมเวลา: เครื่องมือการแสดงผลภาพสำหรับการทำงานกับฐานข้อมูลของข้อมูลอนุกรมเวลาขนาดใหญ่** (Time-series Bitmaps: A Practical Visualization Tool for working with Large Time Series Databases) [2]

งานวิจัยนี้นำเสนอเครื่องมือที่ทำให้ผู้ใช้งานสามารถแยกแยะ จัดการ และบริหารข้อมูลอนุกรมเวลาได้ง่ายขึ้น โดยใช้วิธีการจัดการข้อมูลอนุกรมเวลาให้แสดงออกมาเป็นลักษณะรูปภาพบิตแม็บ ซึ่งนำเอาลักษณะต่างๆของข้อมูลอนุกรมเวลา แสดงออกมาเป็นสีสันในภาพบิตแม็บ ทำให้ผู้ใช้งานสามารถที่จะทำการจัดกลุ่ม หาความแตกต่าง จัดการกับชุดของข้อมูลได้อย่างรวดเร็ว และง่ายขึ้น เครื่องมือดังกล่าวนี้ยังสามารถนำไปใช้ร่วมกับระบบปฏิบัติการที่สนับสนุนการต่อประสานด้วยภาพกับผู้ใช้ (Graphical User Interfaces) เช่น ไมโครซอฟต์วินโดวส์ (Microsoft Windows) อควา (Aqua) และ เอ็กซ์วินโดวส์ (X-windows) เป็นต้น ผู้ใช้เครื่องมือนี้จะสามารถจำแนกข้อมูลอนุกรมเวลาจากสีของรูปภาพบิตแม็บ แทนการพิจารณาข้อมูลดิบที่เป็นเลขอนุกรมจำนวนมาก

งานวิจัยนี้ได้นำเอาแนวคิดจากทฤษฎีของเคออสเกม ที่นำไปใช้ในการแสดงผลภาพจากข้อมูลดีเอ็นเอ มาประยุกต์ใช้ในการแสดงผลภาพให้เข้าใจง่าย และดีขึ้น โดยการนำเอาความถี่ของการเกิดสายอักขระต่างๆในข้อมูลดีเอ็นเอมาใช้แทนการกำหนดจุด และยังพัฒนานำไปประยุกต์ใช้กับข้อมูลอนุกรมเวลาที่ยากแก่การพิจารณาข้อมูลโดยรวม ที่มีปริมาณมากได้ดีอีกด้วย แต่อย่างไรก็ตามเครื่องมือนี้ยังจำเป็นต้องใช้ผู้เชี่ยวชาญในการกำหนดตัวแปรสำคัญ ที่มีผลต่อการแสดงผลภาพของแต่ละชนิดและรูปแบบของข้อมูลที่แตกต่างกันออกไป ซึ่งไม่ตรงกับจุดประสงค์ของผู้ทำการวิจัย เช่น การกำหนดขนาดของช่วงแบ่งข้อมูลในวิธีการแบบแซ็ค หรือการกำหนดระดับขั้นของการแสดงภาพบิตแม็บกับลักษณะของผลลัพธ์ที่ต้องการ อีกทั้งการจำแนกประเภทและการแบ่งกลุ่ม จำเป็นต้องใช้ผู้เชี่ยวชาญที่มีความรู้ทั้งทางด้านวิธีการแบบแซ็ค และทางด้านลักษณะของผลลัพธ์ที่ต้องการ จึงจะสามารถกำหนดตัวแปรที่เหมาะสมและแสดงผลลัพธ์ออกมาได้อย่างมีประสิทธิภาพ

2.5.2 **สัญรูปอัจฉริยะ: การทำเหมืองข้อมูลขนาดย่อม และทำการแสดงผลภาพสู่ระบบปฏิบัติการแบบส่วนต่อประสานด้วยภาพกับผู้ใช้** (Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating Systems) [9]

งานวิจัยนี้นำเสนอแนวคิดที่จะแทนสัญรูป (Icon) มาตรฐาน ที่ปรากฏอยู่ในระบบปฏิบัติการที่สนับสนุนการต่อประสานด้วยภาพกับผู้ใช้ (Graphical User Interfaces) เช่น ไมโครซอฟต์วินโดวส์ (Microsoft Windows) โอเอสเอกซ์ (OS X) หรือลีนุกซ์ (Linux) ด้วยสัญรูปที่

ถูกสร้างขึ้นมาโดยอัตโนมัติ โดยลักษณะของสัญรูปใหม่จะมีความสัมพันธ์กับคุณสมบัติในข้อมูลนั้นๆ ซึ่งงานวิจัยนี้เรียกว่า สัญรูปอัจฉริยะ (Intelligent Icon) นอกจากนี้ยังสามารถทำการจัดกลุ่มข้อมูลจากความเหมือน หรือความแตกต่างของสัญรูปได้อีกด้วย

งานวิจัยนี้มีแนวทางมาจากวิชวลไอดีส์ (VisualIDs) [10] ที่มีแนวความคิดที่ว่าการค้นหาและจดจำภาพทำให้สามารถจดจำและค้นหาได้ดีและมีประสิทธิภาพ กว่าการค้นหาและการจำเป็นประโยค ซึ่งวิชวลไอดีส์จะทำการสร้างภาพสัญรูป (Icon) ที่เกิดมาจากขั้นตอนวิธีแบบแฮช (Hashing) กับชื่อของแฟ้มข้อมูล (File) ด้วยเหตุนี้ทำให้แฟ้มข้อมูลที่มีชื่อคล้ายกันจะได้สัญรูปที่มีลักษณะใกล้เคียงกัน สามารถทำให้ผู้ใช้สามารถเห็นความเหมือนหรือแตกต่างกันของแฟ้มข้อมูลได้ง่ายขึ้น ซึ่งสัญรูปอัจฉริยะนำเอาแนวคิดนี้มาประยุกต์โดยการนำเอาข้อมูลของแฟ้มข้อมูลนั้นมาทำการสร้างสัญรูปแทนที่จะใช้เพียงชื่อของแฟ้มข้อมูล

ถึงแม้ว่างานวิจัยนี้จะนำเอาสัญรูปอัจฉริยะ มาทำการทดลองกับข้อมูลในรูปแบบต่างๆ ซึ่งได้ผลการทดลองที่น่าสนใจ แต่อย่างไรก็ตามผลการทดลองของข้อมูลหน่วยพันธุกรรมและข้อมูลอนุกรมเวลาได้เคยถูกนำเสนอในงานวิจัยอื่นๆมาแล้ว [2][3][5][6] สำหรับข้อมูลเอกสารและข้อมูลที่เป็นวีดีโอเกม ที่นำเสนอในงานวิจัยนี้ไม่ได้ระบุถึงขั้นตอนที่จำเป็นในการนำเอาข้อมูลมาทำให้เป็น สัญรูปอัจฉริยะ เช่นหลักการแปลงข้อมูลตัวอักษรของเอกสารมาเป็นสัญรูป ซึ่งในงานวิจัยได้กล่าวถึงเพียงหลักการพิจารณาคุณลักษณะของเอกสารเพียงเบื้องต้นเท่านั้น อีกทั้งไม่มีหลักการและเหตุผลในการกำหนดรูปแบบ (Template) ที่แตกต่างกันของสัญรูปอัจฉริยะสำหรับข้อมูลแต่ละประเภท