

การพัฒนาชุดคำสั่งการทำเหมืองข้อความสำหรับการระบุปฏิสัมพันธ์ระหว่าง ทรานสคริปชันแฟคเตอร์และดีเอ็นเอในการแสดงออกของไนโตรจีเนส Text Mining Development Tool for Identification of Transcription Factor and DNA Interactions in Nitrogenase Expression

สิทธิรักษ์ สุนทรารักษ์^{1*} เทพปัญญา เจริญรัตน์¹ และภัทรพร คุ่มภัย¹

Sittirak Soontarak^{1*}, Theppanya Charoenrat¹ and Phataraporn Khumchai¹

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาชุดคำสั่งการทำเหมืองข้อความและวิเคราะห์เครือข่ายความสัมพันธ์ของยีนโดยระบุคู่ของทรานสคริปชันแฟคเตอร์และดีเอ็นเอที่มีความสำคัญในการแสดงออกของไนโตรจีเนสในไรโซเบียม ผลของการพัฒนาชุดคำสั่งการทำเหมืองข้อความใช้ข้อมูลที่เกี่ยวข้องกับแบคทีเรียที่มีความสามารถในการตรึงไนโตรเจน จากฐานข้อมูลทางวิชาการ PubMed เป็นข้อมูลนำเข้าจำนวนทั้งสิ้น 18,011 บทคัดย่อ พัฒนาชุดคำสั่งด้วยภาษาคอมพิวเตอร์ Java ให้กลายเป็นข้อความรวมได้จำนวน 935,900 บรรทัด และสกัดเป็นประโยคได้ 264,624 ประโยค เมื่อนำข้อความทั้งหมดไปประมวลผลด้วยชุดคำสั่งต่อ พบว่ามีข้อมูลผลลัพธ์เป็นคำสำคัญที่ใช้บ่งบอกความสัมพันธ์ระหว่างยีนจำนวน 39,524 บรรทัด และเป็นรายชื่อยีนหรือรายชื่อโปรตีนจำนวน 51,193 บรรทัด ซึ่งข้อมูลทั้งสองนี้จะแสดงลำดับตำแหน่งและลำดับประโยคที่พบในข้อความรวมในแต่ละบรรทัด ผลสุดท้ายที่ได้จากการประมวลผลชุดคำสั่งการทำเหมืองข้อความ เป็นข้อมูลความสัมพันธ์ระหว่างคู่ของยีนหรือโปรตีนกับคำสำคัญที่บ่งบอกความสัมพันธ์ของข้อมูลคู่กันจำนวน 187 บรรทัด นอกจากนี้ยังแสดงข้อความเต็มทั้งประโยคที่ค้นพบความสัมพันธ์ และแสดงคำระยะห่างของข้อมูล (token) ซึ่งบ่งบอกระยะห่างระหว่างคู่ของคำ เมื่อนำข้อมูลผลลัพธ์ที่ได้จากการทำเหมืองข้อความมาแสดงเป็นภาพเครือข่ายความสัมพันธ์ของยีนด้วยโปรแกรม Cytoscape พบโหนดจำนวน 119 โหนด และเส้นเชื่อมจำนวน 187 เส้นเชื่อม จากการวิเคราะห์การเป็นจุดเชื่อมต่อศูนย์กลางเครือข่าย และพิจารณาความสัมพันธ์ระหว่างโหนด พบว่าโหนดที่เป็นยีน *nifH* และโหนดที่เป็น

¹ สาขาวิชาเทคโนโลยีชีวภาพ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

* Corresponding author e-mail: sittiraksoontarak@gmail.com

Received: 27 November 2020, Revised: 4 January 2021, Accepted: 4 January 2021

โปรตีน NifA มีความสัมพันธ์กันในรูปแบบของยีนและทรานสคริปชันแฟคเตอร์ที่ทำงานร่วมกันในการแสดงออกของไนโตรจีเนส ซึ่งส่งผลต่อการตรึงไนโตรเจนของไรโซเปียม ซึ่งข้อมูลที่ได้นี้สามารถนำไปประยุกต์ใช้ในงานด้านพันธุวิศวกรรมต่อไป

คำสำคัญ: เหมืองข้อความ เครือข่ายความสัมพันธ์ของยีน การตรึงไนโตรเจน

Abstract

The purpose of this research was to develop a text mining program and to analyze gene networks identifying pairs of transcription factor and DNA that were important in nitrogenase expression in *Rhizobium*. The result of text mining implementation used information related to nitrogen-fixing bacteria by searching from the PubMed database. A total of 18,011 abstracts were input data. Text mining program implemented with Java computer language. The program compiled all abstracts into 935,900 lines of the combined text and extracted to 264,624 sentences. When the combined text compiled with the program, it was found that the output data keywords used to indicate gene relationships were 39,524 lines and the output data gene or protein lists were 51,193 lines. Each line of these two data showed the position and sentence order found in the combined text. The final result obtained by text mining implementation was 187 lines of correlation data between keyword and pairs of gene or protein. It also showed the full text of the relationship found and the token value indicated the distance between word pairs. When showing the result obtained from the text mining used to visualize a gene regulatory network using Cytoscape program, 119 nodes and 187 edges were found. When analyzing the network hub nodes and considering the relationship between nodes, it was found that the *nifH* gene nodes and the NifA protein nodes were related in the gene and the transcription factor that work together in the nitrogenase expression. The results affected the nitrogen fixation of *Rhizobium*. This information can be further applied to genetic engineering works.

Keywords: Text mining, Gene network, Nitrogen fixation

บทนำ

ไรโซเบียมเป็นแบคทีเรียที่มีความสามารถในการตรึงไนโตรเจนบนรากพืชและในลำต้นของพืชน้ำบางชนิด ปฏิสัมพันธ์นี้เรียกว่าภาวะพึ่งพาอาศัย (symbiotic) พืชที่พบความสามารถในการตรึงไนโตรเจน เช่น พืชตระกูลถั่วมีไรโซเบียมอาศัยอยู่ภายในปมของระบบราก การผลิตสารประกอบไนโตรเจนช่วยให้พืชเจริญเติบโต เมื่อพืชตายและย่อยสลาย ไนโตรเจนที่ถูกตรึงไว้ในพืชจะถูกปลดปล่อยออกมาเพื่อเป็นปุ๋ยให้กับดินและให้แก่พืชอื่นต่อไป (Postgate, 1998) จัดเป็นการเพิ่มผลผลิตทางการเกษตร ไรโซเบียมมีเอนไซม์ที่เกี่ยวข้องกับการตรึงไนโตรเจน ได้แก่ ไนโตรจีเนส (nitrogenase) ซึ่งมีความสำคัญอย่างมากในการตรึงไนโตรเจน (Newton, 1993) ทั้งนี้ ไนโตรจีเนสเป็นผลิตภัณฑ์ที่ได้มาจากการแสดงออกของยีน *nikH* *nifD* และ *nifK* โดยการควบคุมการแสดงออกของยีนในระดับการถอดรหัส (transcription) จะเกี่ยวข้องกับการเกิดปฏิสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์ (transcription factor) และดีเอ็นเอ (Sarkar and Reinhold, 2014) ทรานสคริปชันแฟกเตอร์สามารถช่วยส่งเสริมการแสดงออกของยีนให้เพิ่มมากขึ้น หรือมีความสามารถในการยับยั้งการแสดงออกของยีนได้เช่นกัน การเข้าใจกระบวนการเกิดปฏิสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์กับดีเอ็นเอ เช่น ลักษณะการเข้าจับระหว่างทรานสคริปชันแฟกเตอร์กับดีเอ็นเอ จะช่วยให้เข้าใจลักษณะการแสดงออกของยีนได้ดีขึ้น

ปัจจุบันมีการใช้เทคโนโลยีชีวสารสนเทศในการศึกษากระบวนการทางชีววิทยาแบบองค์รวม (systems biology) เช่น การสร้างเครือข่ายควบคุมการแสดงออกของยีน (regulatory network) โดยได้ข้อมูลมาจากการทำเหมืองข้อความ (text mining) และการศึกษาปฏิสัมพันธ์ของโปรตีนที่มีความซับซ้อน จึงมีบทบาทสำคัญในด้านต่าง ๆ ของโครงสร้างและหน้าที่ของเซลล์ ความรู้เหล่านั้นเปิดเผยให้เห็นถึงกลไกของโมเลกุลของกระบวนการทางด้านชีววิทยา (Zhou and He, 2008) ซึ่งวิธีการนี้จะช่วยให้สามารถระบุทรานสคริปชันแฟกเตอร์ที่สามารถพบได้ในไรโซเบียมสายพันธุ์ต่าง ๆ ซึ่งแสดงให้เห็นถึงความสำคัญของทรานสคริปชันแฟกเตอร์เหล่านั้น หรืออาจทำให้พบทรานสคริปชันแฟกเตอร์ที่โดดเด่นอย่างมากในกระบวนการการควบคุมการแสดงออกของยีน งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาชุดคำสั่งการทำเหมืองข้อความสำหรับระบุกลุ่มทรานสคริปชันแฟกเตอร์และยีนที่สำคัญในกระบวนการตรึงไนโตรเจน โดยใช้กระบวนการสร้างเครือข่ายปฏิสัมพันธ์ของยีน ซึ่งองค์ความรู้ที่ได้สามารถนำไปประยุกต์ใช้กับการออกแบบทรานสคริปชันแฟกเตอร์เพื่อเอื้อต่อการแสดงออกของไนโตรจีเนสที่ใช้ในการตรึงไนโตรเจนได้มากขึ้น

วิธีดำเนินการวิจัย

1. การสืบค้นข้อมูลบทความทางวิชาการจากฐานข้อมูล

สืบค้นบทความทางวิชาการจากฐานข้อมูล PubMed ซึ่งเป็นฐานข้อมูลที่ประกอบไปด้วยวรรณกรรมชีวการแพทย์ วารสารวิทยาศาสตร์ชีวภาพ และหนังสือออนไลน์ นักวิจัยทางด้านวิทยาศาสตร์ชีวภาพมักใช้ PubMed เป็นเครื่องมือค้นหาออนไลน์เพื่อเข้าถึงข้อมูลในฐานข้อมูล (Vibert *et al.*, 2009) ในขั้นตอนนี้จะใช้คำสำคัญในการสืบค้นหลายคำ ซึ่งรวมเป็นหลายชุดข้อมูลเพื่อให้ได้ผลลัพธ์ที่เกี่ยวข้องกับการตรึงไนโตรเจน ในโตรจีนัส และไรโซเบียม โดยใช้คำสำคัญที่ใช้สืบค้นทั้งหมดมี 7 ชุดดังนี้ (1) rhizobium nitrogen fixation (2) rhizobia nitrogen fixation (3) bacteria nitrogen fixation (4) bacteria nitrogen fixing (5) bacteria nitrogenase (6) rhizobium nitrogenase และ (7) rhizobia nitrogenase

2. การเขียนชุดคำสั่งสำหรับทำเหมืองข้อความ

การพัฒนาชุดคำสั่งในงานวิจัยนี้ใช้ซอฟต์แวร์ Eclipse Integrated Development Environment (IDE) (Murphy *et al.*, 2006) รุ่น Eclipse Java Oxygen จากเว็บไซต์ <https://www.eclipse.org/downloads/> พัฒนาขึ้นมาทั้งหมด

2.1 แยกประโยคย่อยออกมาจากบทความและกำหนดลำดับที่ให้กับประโยคย่อย โดยใช้สถาปัตยกรรมซอฟต์แวร์สำหรับการจัดการข้อมูลที่ไม่มีโครงสร้าง (unstructured information management architecture: UIMA) ในการวิเคราะห์ข้อมูลที่ไม่มีโครงสร้าง การเข้าถึงแหล่งข้อมูลของโครงสร้าง การจัดเก็บข้อมูล การทำดัชนี และการค้นหาข้อมูลจากเนื้อหาความหมายที่ค้นพบ (Ferrucci and Lally, 2004) และใช้เครื่องมือ OpenNLP ในการสร้างคำอธิบายประกอบอย่างอัตโนมัติ เพื่อตรวจสอบขอบเขตของประโยค จัดการเครื่องหมาย จัดการส่วนของคำ แยกส่วนของวลี และแยกวิเคราะห์ประโยค (Wilcock, 2009)

2.2 ดาวน์โหลดย่อคำสำคัญที่เกี่ยวข้องกับการระบุหาปฏิสัมพันธ์ระหว่างยีนจากเว็บไซต์ <https://www.biomedcentral.com/content/supplementary/1471-2105-10-233-S1.txt> โดยพิจารณาจากคำที่สามารถใช้บ่งบอกความสัมพันธ์ระหว่างยีนและโปรตีน ยีนและยีนโปรตีนและโปรตีนได้

2.3 นำคำสำคัญแต่ละคำไปสืบค้นกับบทความ หากพบว่าคำสำคัญนั้นปรากฏอยู่ในบทความให้แสดงตำแหน่งเริ่มต้นและสิ้นสุดของคำสำคัญ และแสดงลำดับที่ของประโยคย่อยที่พบโดยให้อยู่ในรูปแบบของไฟล์ข้อความ

2.4 กำหนดนิพจน์ทั่วไป (regular expression) สำหรับหารายชื่อยีนหรือโปรตีนโดยมีรูปแบบดังนี้คือ อักษรตัวแรกเป็นพยัญชนะภาษาอังกฤษตัวพิมพ์ใหญ่หรือตัวพิมพ์เล็ก อักษรตัวที่สองและสามเป็นพยัญชนะภาษาอังกฤษตัวพิมพ์เล็กเท่านั้น อักษรตัวที่สี่เป็นพยัญชนะภาษาอังกฤษ

ตัวพิมพ์ใหญ่หรือตัวพิมพ์เล็กเช่นเดียวกับตัวอักษรแรก จากนั้นสามารถมีตัวเลขต่อท้ายข้อความได้จำนวนหนึ่งถึงสามตัวเลข

2.5 นำนิพจน์ทั่วไปของยีนหรือโปรตีนไปสับคั่นกับบทความ หากพบว่านิพจน์ทั่วไปของยีนหรือโปรตีนนั้นปรากฏอยู่ในบทความให้แสดงตำแหน่งเริ่มต้นและสิ้นสุดของนิพจน์ทั่วไปของยีนหรือโปรตีน และแสดงลำดับที่ของประโยคย่อยที่พบ โดยให้อยู่ในรูปแบบของไฟล์ข้อความ

2.6 จัดรูปแบบและแสดงไฟล์ข้อความของรายชื่อคำสำคัญที่ได้จากข้อ 2.3 และรายชื่อยีนหรือโปรตีนที่ได้จากข้อ 2.5 ให้อยู่ในรูปแบบลำดับที่ของประโยคย่อย ตั้งแต่ประโยคย่อยที่หนึ่งว่าพบคำสำคัญ ชื่อยีน หรือชื่อโปรตีนใดบ้าง ทำเช่นนี้จนถึงประโยคย่อยสุดท้าย

2.7 สร้างไฟล์ข้อความให้ข้อมูลอยู่ในรูปแบบของไฟล์ tab-separated value (TSV) โดยข้อมูลคอลัมน์แรกเป็นชื่อยีน คั่นด้วยสัญลักษณ์ tab ตามด้วยข้อมูลคอลัมน์ที่สองเป็นชื่อยีน คั่นด้วยสัญลักษณ์ tab แล้วตามด้วยข้อมูลคอลัมน์ที่สามเป็นคำสำคัญที่บ่งบอกปฏิสัมพันธ์ ทำลักษณะเช่นเดียวกันนี้ไปจนครบจำนวนข้อมูลที่ได้มาจากขั้นตอนข้อที่ 2.6 ซึ่งข้อมูลดังกล่าวในแต่ละแถวนี้มีความสัมพันธ์กัน

3. การแสดงภาพและวิเคราะห์เครือข่ายความสัมพันธ์ระหว่างยีน

แสดงภาพของเครือข่ายความสัมพันธ์ด้วยโปรแกรม Cytoscape v3.6.0 จากนั้นวิเคราะห์เครือข่ายความสัมพันธ์จากโหนดที่เป็นจุดเชื่อมต่อศูนย์กลาง (hub node) และวิเคราะห์จุดเชื่อมต่อความสัมพันธ์แรกสุด (first neighbour analysis) เพื่อหาความสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์และดีเอ็นเอที่ส่งผลต่อการแสดงออกของโนโตรจีเนส

ผลการวิจัย

1. ผลการสืบค้นข้อมูลบทความทางวิชาการจากฐานข้อมูล

จากขั้นตอนการสืบค้นข้อมูลบทความทางวิชาการจากฐานข้อมูลเพื่อให้ได้มาซึ่งบทความที่เกี่ยวข้องกับการตรึงไนโตรเจน ไนโตรจีเนส และไรโซเปียม ด้วยคำค้นหาที่แตกต่างกัน ได้ผลดังตารางที่ 1 โดยจากการสืบค้นได้ผลลัพธ์ออกมาเป็นบทความรวมทั้งสิ้น 18,011 บทความ โดยต่อมาจะนำบทความทั้งหมดนี้มาเรียงต่อกันให้อยู่ในไฟล์ข้อความรวมเป็นไฟล์เดียว เพื่อสะดวกต่อการนำไปเป็นข้อมูลนำเข้าในการเขียนชุดคำสั่งสำหรับการทำเหมืองข้อความของขั้นตอนถัดไป เมื่อนำบทความทั้ง 18,011 บทความมารวมกันไว้เป็นข้อความรวมจะมีจำนวนของบรรทัดรวมทั้งสิ้น 592,645 บรรทัด ซึ่งเป็นข้อมูลขนาดใหญ่

ตารางที่ 1 ชุดของคำที่ใช้สืบค้นในฐานข้อมูล PubMed

คำที่ใช้ในการสืบค้น	จำนวนบทความที่สืบค้นได้
rhizobium nitrogen fixation	1,748
rhizobia nitrogen fixation	846
bacteria nitrogen fixation	7,141
bacteria nitrogen fixing	4,150
bacteria nitrogenase	3,438
rhizobium nitrogenase	500
rhizobia nitrogenase	188
รวม	18,011

2. ผลการเขียนชุดคำสั่งสำหรับทำเหมืองข้อความ

ผลลัพธ์ของชุดคำสั่งนี้ใช้ข้อมูลนำเข้าคือไฟล์ข้อความรวมที่มีบทความ 592,645 บทความที่ได้มาจากการสืบค้นฐานข้อมูลทางวิชาการมาสังกัดเป็นประโยคย่อย ซึ่งได้ผลลัพธ์เป็นไฟล์ข้อความรวมไฟล์ใหม่ที่มีลำดับประโยคและตำแหน่งของประโยคเพิ่มขึ้นมา ทำให้ข้อความรวมใหม่มีจำนวนบทความเพิ่มขึ้นเป็น 935,900 บทความ และสังกัดเป็นประโยคได้ 264,624 ประโยค เช่น จากภาพที่ 1 แสดงข้อความ sentence: 7 ซึ่งหมายถึงข้อความรวมถูกสกัดออกมาเป็นประโยคย่อยลำดับที่ 7 และมีข้อมูลในประโยคแสดงดังภาพ โดยส่วนท้ายของประโยคแสดงข้อความ (449..666) หมายถึงประโยคนี้อักษรตัวแรกเริ่มต้นที่ลำดับ 449 และลงท้ายด้วยตัวอักษรลำดับที่ 666 ของข้อความรวม

```

20 Sentence: 7
21 To analyze microbial communities in environmental samples, this study combined
22 Denaturing Gradient Gel Electrophoresis of amplified 16S rRNA-genes in total
23 genomic DNA extracts from those samples with gene sequencing. [449..666]
24 Sentence: 8
25 The environmental
26 samples studied were oily seawater and soil samples, that had been bioaugmented
27 with natural materials rich in hydrocarbonoclastic bacteria. [667..825]
```

ภาพที่ 1 ผลลัพธ์บางส่วนของชุดคำสั่งในการสกัดประโยคย่อยจากบทความที่เป็นข้อความรวม

ตารางที่ 2 ตัวอย่างตำแหน่งและลำดับของประโยคที่ค้นพบคำสำคัญ

คำสำคัญ (keyword)	ตำแหน่งเริ่มต้น-สิ้นสุด (position)	ลำดับของประโยค (sentence)
bind	15,097-15,101	129
bind	50,852-50,856	431
bind	98,218-98,222	827
binded	11,451,152-11,451,158	100,367
binds	516,729-516,734	4,453
binds	9,386,773-9,386,778	80,732
binding	1,111,019-1,111,026	9,536
binding	8,297,486-8,297,493	71,421
express	732,267-732,274	6,326
express	3,206,236-3,206,243	27,690

จากตารางที่ 2 เป็นผลตัวอย่างบางส่วนของชุดคำสั่งที่ใช้ข้อมูลนำเข้าจากข้อความรวมที่มีการจัดแบ่งลำดับของประโยคไว้เรียบร้อยแล้ว มาประมวลผลข้อมูลด้วยชุดคำสั่งเพื่อให้ได้ผลลัพธ์ออกมาเป็นไฟล์ข้อมูลใหม่ ซึ่งมีจำนวนข้อมูลรวมทั้งสิ้น 39,524 บรรทัด โดยข้อมูลบรรทัดแรกสุดเป็นข้อมูลบ่งบอกถึงความหมายที่อยู่ในสดมภ์ของแต่ละบรรทัด ตัวอย่างในบรรทัดที่สองของตารางที่ 2 การทำงานของชุดคำสั่งนี้ค้นพบคำสำคัญคำว่า bind อยู่ในตำแหน่งที่ 15,097 ถึงตำแหน่งที่ 15,101 ของข้อความรวม และพบคำนี้อยู่ในประโยคที่ 129 ของข้อความรวมเช่นกัน โดยคำสำคัญคำเดียวกันสามารถค้นพบได้หลายที่ในข้อความรวม ซึ่งทราบได้ว่าเป็นคำสำคัญคนละคำกันจากการระบุลำดับที่ของประโยคและตำแหน่งที่พบจากในข้อความรวม เช่น คำสำคัญ bind พบได้จากบรรทัดที่ 2 บรรทัดที่ 3 และบรรทัดที่ 4 ซึ่งพบได้จากประโยคลำดับที่ 129 431 และ 827 ตามลำดับ

ตารางที่ 3 ตัวอย่างตำแหน่งและลำดับของประโยคที่ค้นพบรายชื่อยีนหรือโปรตีน

ยีน (gene)	ตำแหน่งเริ่มต้น-สิ้นสุด (position)	ลำดับของประโยค (sentence)
<i>nifH</i>	1,555-1,559	13
<i>ogyO</i>	1,987-1,991	17
BioE	4,858-4,862	45
<i>nzaT</i>	5,052-5,056	47
<i>hydG</i>	5,854-5,858	53
GmbH	10,402-10,406	84
RpoN	14,235-14,239	124

ตารางที่ 3 (ต่อ)

ยีน (gene)	ตำแหน่งเริ่มต้น-สิ้นสุด (position)	ลำดับของประโยค (sentence)
<i>rpoN</i>	14,431-14,435	125
NtrB	15,475-15,479	131
HydA	22,383-22,387	186

จากตารางที่ 3 เป็นผลตัวอย่างบางส่วนจากการเขียนชุดคำสั่งสำหรับการกำหนดตำแหน่งและลำดับของประโยคที่ค้นพบรายชื่อยีนหรือโปรตีน ซึ่งได้ข้อมูลผลลัพธ์รวมทั้งสิ้น 51,193 บรรทัด โดยข้อมูลบรรทัดแรกสุดเป็นข้อมูลบ่งบอกถึงความหมายที่อยู่ในสดมภ์ของแต่ละบรรทัด ตัวอย่างในบรรทัดที่สองของตารางที่ 3 การทำงานของชุดคำสั่งนี้ค้นพบรายชื่อยีน *nifH* อยู่ในตำแหน่งที่ 1,555 ถึงตำแหน่งที่ 1,559 ของข้อความรวม และพบยีนนี้อยู่ในประโยคที่ 13 ของข้อความรวม ความแตกต่างระหว่างรายชื่อยีนและรายชื่อโปรตีนคือ รายชื่อยีนจะขึ้นต้นด้วยตัวอักษรอังกฤษพิมพ์เล็ก ส่วนรายชื่อโปรตีนจะขึ้นต้นด้วยตัวอักษรอังกฤษพิมพ์ใหญ่ เช่น RpoN และ NtrB

ตารางที่ 4 ตัวอย่างรายชื่อคำสำคัญที่ค้นพบในแต่ละประโยค

ลำดับของประโยค (sentence)	คำสำคัญ (keyword)
122	N/A
123	N/A
124	regulation transcript transcription
125	N/A
126	N/A
127	transcript
128	N/A
129	bind binding regulate regulated transcript
130	express expression transcript
131	bind binding regulate regulated regulation

หมายเหตุ: - N/A คือ ไม่พบข้อมูล

จากตารางที่ 4 เป็นผลตัวอย่างบางส่วนของชุดคำสั่งที่เขียนขึ้นมาเพื่อให้ทราบว่าแต่ละประโยคที่มีในข้อความรวม สามารถพบคำสำคัญที่บ่งบอกถึงความสัมพันธ์ระหว่างคู่ของยีนหรือโปรตีนได้บ้าง เช่น ผลในตารางที่ 4 บรรทัดที่สี่ แสดงถึงประโยคในข้อความรวมลำดับที่ 124 และพบคำสำคัญดังนี้ regulation transcript และ transcription

ตารางที่ 5 ตัวอย่างรายชื่อยีนที่ค้นพบในแต่ละประโยค

ลำดับของประโยค (sentence)	ยีน (gene)
24,954	N/A
24,955	<i>nifH</i>
24,956	N/A
24,957	N/A
24,958	N/A
24,959	<i>nifH</i>
24,960	<i>nifH nifD nifK</i>
24,961	<i>nifH nifD</i>
24,962	<i>nifD nifD</i>
24,963	<i>nifH nifH</i>

หมายเหตุ: - N/A คือ ไม่พบข้อมูล

จากตารางที่ 5 ผลลัพธ์คล้ายกับข้อมูลในตารางที่ 4 แต่ผลลัพธ์นี้จะเป็นข้อมูลของรายชื่อยีนหรือรายชื่อโปรตีนแทนข้อมูลของคำสำคัญ โดยผลบางส่วนที่ได้เป็นไปตามตารางที่ 5 นอกจากนี้ ข้อมูลที่อยู่ในสดมภ์คำสำคัญของตารางที่ 4 และข้อมูลในสดมภ์ยีนของตารางที่ 5 อาจพบว่ามี ความสัมพันธ์กันได้หากเป็นข้อมูลที่อยู่ในลำดับประโยคเดียวกัน

ตารางที่ 6 ตัวอย่างรายชื่อยีนหรือโปรตีนและคำสำคัญที่มีความสัมพันธ์กันพร้อมแสดงประโยคที่ค้นพบ

ลำดับของ ประโยค (sentence)	ยีน (gene) 1	คำสำคัญ (keyword)	ยีน (gene) 2	ระยะห่าง ของข้อมูล (token)	ข้อมูลทั้งหมดที่พบ ในประโยค (full-sentence)
206,311	NifA	binding	<i>nifH</i>	10	We now report that the introduction of half turns of the DNA helix into the DNA separating the upstream NifA binding site from the downstream promoter element of the <i>nifH</i> promoter decreases NifA-mediated activation

ตารางที่ 6 (ต่อ)

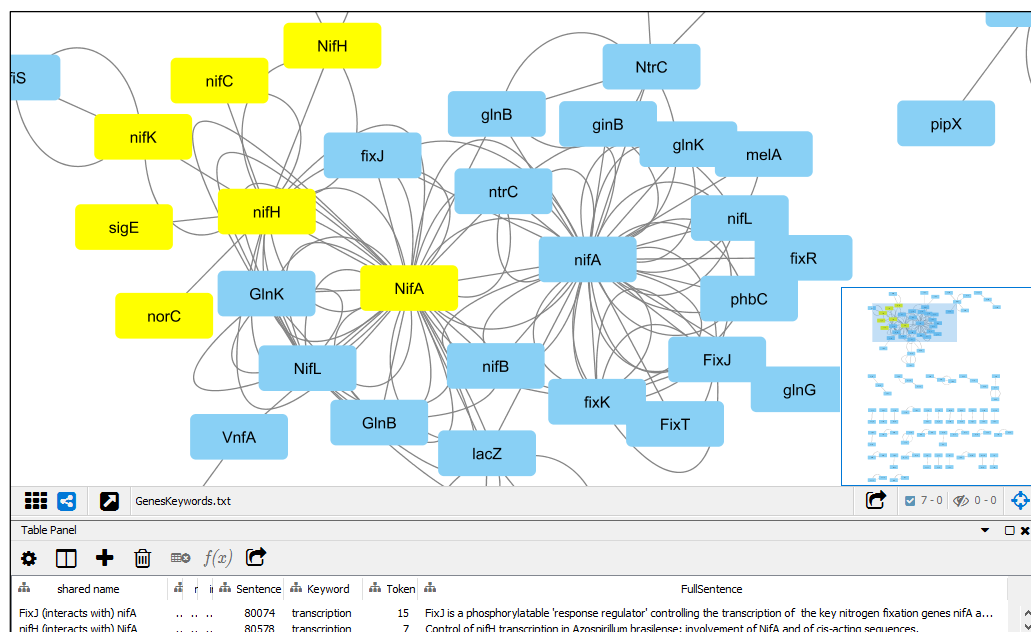
ลำดับของประโยค (sentence)	ยีน (gene) 1	คำสำคัญ (keyword)	ยีน (gene) 2	ระยะห่างของข้อมูล (token)	ข้อมูลทั้งหมดที่พบในประโยค (full-sentence)
252,678	NifA	transcription	<i>nifH</i>	10	to a greater extent than does the introduction of full helical turns. The central domain of <i>Rhizobium meliloti</i> NifA is sufficient to activate transcription from the <i>R. meliloti nifH</i> promoter.

จากตารางที่ 6 เป็นผลตัวอย่างบางส่วนของกระบวนการเขียนชุดคำสั่งสำหรับแสดงความสัมพันธ์ระหว่างรายชื่อยีนหรือรายชื่อโปรตีนที่มีความสัมพันธ์กันด้วยคำสำคัญใดคำสำคัญหนึ่ง ในส่วนค่าของระยะห่างของข้อมูลมีเงื่อนไขว่าข้อมูลแรกที่พบจะต้องเป็นรายชื่อยีนหรือโปรตีนก่อน จากนั้นตามด้วยข้อมูลคำสำคัญเป็นข้อมูลที่สองซึ่งเป็นข้อมูลระหว่างกลางของข้อมูลแรกและข้อมูลที่สามที่เป็นข้อมูลรายชื่อยีนหรือรายชื่อโปรตีน แล้วจึงคำนวณออกมาเป็นค่าระยะห่างของข้อมูล หากค่านี้มีค่าน้อยหมายความว่าระยะห่างของคู่อายชื่อยีนหรือรายชื่อโปรตีนมีระยะห่างน้อย หากค่านี้มีค่ามากหมายความว่าระยะห่างของคู่อายชื่อยีนหรือรายชื่อโปรตีนมีระยะห่างมาก ซึ่งอาจนำมาช่วยพิจารณาความสัมพันธ์ของรายชื่อยีนหรือโปรตีนดังกล่าวได้ เพราะถ้าค่าระยะห่างของข้อมูลสูงมาก เช่นค่าเท่ากับ 100 แสดงว่าคู่อายชื่อยีนหรือโปรตีนมีระยะห่างของทั้งสองคำอยู่ถึง 100 ช่องว่าง โดยมีคำสำคัญอยู่ระหว่างกลางของสองคำนี้ในประโยคเพียงประโยคเดียว อาจเลือกที่จะไม่พิจารณาความสัมพันธ์ของข้อความดังกล่าว ข้อมูลในสดมภ์ข้อมูลทั้งหมดที่พบในประโยค สามารถอ่านข้อความได้ทั้งประโยคเพื่อช่วยพิจารณาความสัมพันธ์ของข้อมูลรายชื่อยีนหรือโปรตีน และยังสามารถอ้างอิงกลับไปหาประโยคข้อความรวมได้ เพื่อหาต่อไปว่ามาจากบทคัดย่อใด ตัวอย่างของข้อมูลผลลัพธ์ที่ได้จากโปรแกรมทำเหมืองข้อความบางส่วนจากตารางที่ 6 ข้อมูลในสดมภ์ลำดับของประโยค ข้อมูลแรกคือ 206,311 หมายถึงประโยคลำดับที่ 206,311 ในข้อความรวม พบรายชื่อยีนโปรตีน NifA และยีน *nifH* มีความสัมพันธ์กันด้วยคำสำคัญ binding โดยมีระยะห่างเป็นช่องว่างระหว่าง NifA คั่นด้วย expressed แล้วตามด้วย *nifH* อยู่จำนวน 10 ช่องว่างของคำ และมีข้อมูลที่เป็นข้อความทั้งประโยคคือ “We now report that the introduction of half turns of the DNA helix

into the DNA separating the upstream NifA binding site from the downstream promoter element of the *nifH* promoter decreases NifA-mediated activation to a greater extent than does the introduction of full helical turns.” เมื่อดูความหมายของข้อความในประโยคแล้วพบว่า โปรตีน NifA มีความสัมพันธ์เกี่ยวข้องกับการแสดงออกของยีน *nifH* ซึ่งเป็นประโยชน์ในการพิจารณาความสัมพันธ์ของคูยีนดังกล่าว

3. ผลการวิเคราะห์เครือข่ายความสัมพันธ์ระหว่างยีน

จากข้อมูลผลลัพธ์ที่ได้จากการประมวลผลด้วยชุดคำสั่งของโปรแกรมการทำเหมืองข้อความได้เป็นข้อมูลความสัมพันธ์ระหว่างรายชื่อยีนและโปรตีน นำมาใช้เป็นข้อมูลนำเข้าสำหรับโปรแกรม Cytoscape เพื่อสร้างเป็นเครือข่ายความสัมพันธ์ระหว่างรายชื่อยีนและโปรตีน แสดงให้เห็นเป็นภาพเครือข่ายความสัมพันธ์ได้ดังภาพที่ 2 โดยกรอบสี่เหลี่ยมคือโหนด (node) แต่ละโหนดแทนข้อมูลที่เป็นรายชื่อยีนหรือโปรตีน และเส้นเชื่อม (edge) ของแต่ละโหนดแทนความสัมพันธ์ของโหนดดังกล่าว ผลของขั้นตอนนี้พบโหนดทั้งหมด 119 โหนด เป็นรายชื่อโปรตีน 47 โหนด รายชื่อยีน 72 โหนด และเส้นเชื่อมทั้งหมด 187 เส้นเชื่อม จากเครือข่ายความสัมพันธ์ที่แสดงสามารถเลือกโหนดหรือเส้นเชื่อมเพื่อดูรายละเอียดของข้อมูลโหนดนั้นได้ โดยจะแสดงข้อมูลเหมือนกับข้อมูลผลลัพธ์ของโปรแกรมการทำเหมืองข้อความ คือแสดงข้อมูลลำดับของประโยคที่พบในข้อความรวม ค่าระยะห่างของข้อมูล ซึ่งเป็นระยะห่างของรายชื่อยีนหรือโปรตีน รวมถึงแสดงข้อความทั้งประโยคที่เลือกอีกด้วย เพื่อการค้นหาข้อมูลกลับไปยังข้อความในบทความย่อ นำไปถึงวารสารวิชาการต้นทางได้ ผลที่ได้จากการวิเคราะห์เครือข่ายความสัมพันธ์นี้พบว่าที่โหนด *nifA* และ *nifH* เป็นรายชื่อยีนที่พบเส้นเชื่อมความสัมพันธ์จำนวนมาก และที่โหนด NifA ซึ่งเป็นรายชื่อโปรตีนที่พบเส้นเชื่อมความสัมพันธ์จำนวนมากเช่นกัน



ภาพที่ 2 เครือข่ายความสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์และยีนที่ได้จากการทำเหมืองข้อมูล โดยแสดงออกมาเป็นภาพเครือข่ายด้วยโปรแกรม Cytoscape

การอภิปรายผลการวิจัย

ในงานวิจัยนี้เลือกใช้การทำให้เหมือนข้อความสำหรับการระบุปฏิสัมพันธ์ระหว่างทรานสคริปต์
 ชันแฟกเตอร์และดีเอ็นเอในการแสดงออกของไนโตรจีเนส เนื่องจากการทำให้เหมือนข้อความเป็
 เทคโนโลยีที่มีประสิทธิภาพสำหรับการกลั่นกรองข้อมูลสำคัญให้ออกมาได้อย่างรวดเร็วจาก
 วรรณกรรมจำนวนมหาศาล (Przybyla *et al.*, 2016) การทำให้เหมือนข้อความสามารถนิยามได้ว่าเป็น
 การค้นพบข้อมูลใหม่ ๆ ที่ยังไม่เคยรู้จักมาก่อน โดยแยกข้อมูลออกจากข้อมูลที่เป็นลายลักษณ์อักษร
 โดยอัตโนมัติ การสกัดข้อมูลสามารถสกัดปฏิสัมพันธ์ระหว่างยีนกับยีน หรือปฏิสัมพันธ์ระหว่างโปรตีน
 กับโปรตีนและวิถีทางชีวภาพ โดยถูกนำมาใช้อย่างแพร่หลายคือการประมวลผลภาษาธรรมชาติ
 (natural language processing) ซึ่งเป็นเทคโนโลยีที่รวบรวมไวยากรณ์และความหมายของภาษา
 (Ananiadou *et al.*, 2006) งานวิจัยนี้ได้ใช้ระเบียบวิธีการทำให้เหมือนข้อความตามด้วยการสร้าง
 เครือข่ายความสัมพันธ์ คล้ายกับงานวิจัยของ Song and Chen (2009) ซึ่งทำเหมือนข้อมูลชีว
 การแพทย์โดยอัตโนมัติเพื่อสกัดข้อมูลการควบคุมของยีน (จำนวนสายพันธุ์ พีโนไทป์ ความสัมพันธ์
 ของการควบคุมยีน) แต่ต่างกันที่การพัฒนาชุดคำสั่งการทำให้เหมือนข้อความ จึงได้ข้อมูลที่สกัดออกมา
 ได้ต่างกันออกไป โดยงานวิจัยนี้ได้ผลลัพธ์จากการประมวลผลชุดคำสั่งเป็นข้อมูลความสัมพันธ์
 ระหว่างยีนกับยีน โปรตีนกับโปรตีน หรือยีนกับโปรตีน ซึ่งจะได้ผลที่สอดคล้องกับหัวข้องานวิจัย

มากกว่า โดยต้องการปฏิสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์และดีเอ็นเอ ในส่วนของค่าระยะห่างของข้อมูลที่ใช้สำหรับพิจารณาจำนวนของการเว้นวรรคระหว่างคำในประโยค มีค่าสูงสุดอยู่ที่ 47 จากความสัมพันธ์ระหว่างรายชื่อยีนหรือโปรตีนกับรายชื่อคำสำคัญที่ค้นพบในแต่ละประโยคจำนวน 187 บรรทัด ซึ่งยอมรับได้ว่าที่ค่าระยะห่างของข้อมูล 47 นี้เป็นระยะห่างระหว่างคำในประโยคที่ไม่ได้ห่างมากเกินไป หากพบว่าค่าระยะห่างของข้อมูลมีค่ามากกว่า 100 ไปขึ้นจะไม่พิจารณา เนื่องจากที่ระยะห่างดังกล่าวอาจไม่พบความสัมพันธ์ระหว่างคำ จากขั้นตอนการแสดงผลด้วยเครือข่ายความสัมพันธ์ของยีนหรือโปรตีน พบว่าโหนด *NifA* และ *nifH* เป็นรายชื่อโปรตีนและยีนที่พบเส้นเชื่อมความสัมพันธ์จำนวนมาก ซึ่งสอดคล้องกับงานวิจัยของ Ng *et al.*, (2016) ที่กล่าวว่าเครือข่ายความสัมพันธ์ของโปรตีนทรานสคริปชันแฟกเตอร์สามารถวิเคราะห์หาโหนดเพื่อนบ้านแรกสุดที่เชื่อมต่อกับโหนดหรือทรานสคริปชันแฟกเตอร์ที่สนใจ หากมีโหนดเพื่อนบ้านแรกสุดเชื่อมต่อกันแสดงถึงการมีปฏิสัมพันธ์ที่มากเช่นกัน จึงนำผลที่ได้วิเคราะห์ต่อไปด้วยการค้นหาคำกลับไปยังบทความทางวิชาการและทราบถึงยีน *nifH* ที่สามารถแสดงออกเป็นเอ็นไซม์ไนโตรจีเนส โดยมีโปรตีน *NifA* เป็นทรานสคริปชันแฟกเตอร์ที่ช่วยในการแสดงออกของยีน *nifH*

สรุปผลการวิจัย

ผลจากการสืบค้นฐานข้อมูลได้ผลลัพธ์เป็นบทความที่เกี่ยวข้องกับแบคทีเรียที่มีความสามารถในการตรึงไนโตรเจนจำนวน 18,011 บทความ ได้ผลจากการเขียนชุดคำสั่งสำหรับการทำเหมืองข้อความแบ่งเป็นประโยคทั้งหมดจำนวน 264,660 ประโยค ได้ตำแหน่งและลำดับของประโยคที่ค้นพบคำสำคัญจำนวน 39,523 บรรทัด ได้ตำแหน่งและลำดับของประโยคที่ค้นพบรายชื่อยีนหรือโปรตีนจำนวน 51,192 บรรทัด ได้ความสัมพันธ์ระหว่างรายชื่อยีนหรือโปรตีนกับรายชื่อคำสำคัญที่ค้นพบในแต่ละประโยคจำนวน 187 บรรทัด ได้ค่าระยะห่างของข้อมูลสำหรับพิจารณาจำนวนของการเว้นวรรคระหว่างคำในประโยค เพื่อหาความสัมพันธ์ระหว่างยีนอย่างคร่าว ๆ นำผลที่ได้จากการวิเคราะห์เครือข่ายความสัมพันธ์ระหว่างทรานสคริปชันแฟกเตอร์และยีนที่ได้จากการทำเหมืองข้อความมาแสดงเป็นภาพเครือข่าย พบโหนดจำนวน 119 โหนด เป็นรายชื่อโปรตีน 47 โหนด รายชื่อยีน 72 โหนด และเส้นเชื่อมจำนวน 187 เส้นเชื่อม จากการวิเคราะห์การเป็นจุดเชื่อมต่อศูนย์กลางเครือข่ายและพิจารณาความสัมพันธ์ระหว่างโหนด พบว่าโหนดที่เป็นยีน *nifH* และโหนดที่เป็นโปรตีน *NifA* มีความสัมพันธ์กันในรูปแบบของยีนและทรานสคริปชันแฟกเตอร์ที่ทำงานร่วมกันในการแสดงออกของไนโตรจีเนส ซึ่งส่งผลต่อการตรึงไนโตรเจนของไรโซเบียม โดยศึกษากลับไปยังบทความทางวิชาการต้นทางเพื่อตรวจสอบความถูกต้องของข้อมูล ซึ่งข้อมูลที่ได้นี้สามารถนำไปประยุกต์ใช้ในงานด้านพันธุวิศวกรรมเพื่อใช้ประโยชน์ทางด้านการเกษตร เช่น เพิ่มผลผลิตของพืชด้วยธาตุไนโตรเจน หรือประโยชน์ทางด้านอุตสาหกรรม เช่น เพิ่มมูลค่าทางการค้าจากการผลิตปุ๋ยไนโตรเจน

ข้อเสนอแนะ

ผลลัพธ์ที่ได้จากการทำเหมืองข้อความและการวิเคราะห์เครือข่ายความสัมพันธ์ ควรศึกษาเพิ่มเติมในส่วนของการมีปฏิสัมพันธ์ระดับโมเลกุลที่ลึกลงไปถึงความจำเพาะเจาะจงของตำแหน่งและชนิดกรดอะมิโนที่เข้าจับกับบริเวณใดบริเวณหนึ่งของโครงสร้างดีเอ็นเอ ซึ่งอาจนำไปสู่งานทางด้านพันธุวิศวกรรมที่ต้องการออกแบบทรานสคริปชันแฟกเตอร์สำหรับช่วยส่งเสริมการแสดงออกของยีนที่เกี่ยวข้องกับเอนไซม์ไนโตรจีเนส นอกจากนั้นขั้นตอนต่าง ๆ ในงานวิจัยนี้เป็นเพียงการพัฒนาระเบียบวิธีการทางด้านชีวสารสนเทศ เพื่อศึกษาเพียงการแสดงออกของไนโตรจีเนส หากต้องการศึกษาในเรื่องอื่น ๆ ของสาขาทางด้านชีววิทยา ยังสามารถนำระเบียบวิธีการทางด้านชีวสารสนเทศนี้ ตั้งแต่ขั้นตอนการสืบค้นฐานข้อมูล การพัฒนาชุดคำสั่งโปรแกรมการทำเหมืองข้อความ จนถึงการสร้างเครือข่ายความสัมพันธ์ นำไปประยุกต์หรือใช้เป็นแนวทางการศึกษาข้อมูลที่น่าสนใจได้

กิตติกรรมประกาศ

ขอขอบคุณสาขาวิชาเทคโนโลยีชีวภาพ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ อาจารย์และผู้มีส่วนเกี่ยวข้องทุกท่าน ที่สนับสนุนและให้ความอนุเคราะห์ในการทำวิจัยครั้งนี้

เอกสารอ้างอิง

- Ananiadou, S., Kell, D.B. and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12), 571-579.
- Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4), 327-348.
- Murphy, G.C., Kersten, M. and Findlater, L. (2006). How are Java software developers using the Eclipse IDE. *Institute of Electrical and Electronics Engineers Software*, 23(4), 76-83.
- Newton, W.E. (1993). Nitrogenases: distribution, composition, structure and function. In Palacios, R., Mora, J., Newton, W.E. (Eds.). *New Horizons in Nitrogen Fixation*, pp. 5-18. Springer: Dordrecht.

- Ng, F.S.L., Ruau, D., Wernisch, L. and Gottgens, B. (2016). A graphical model approach visualizes regulatory relationships between genome-wide transcription factor binding profiles. *Briefings in Bioinformatics*, 19(1), doi: <http://doi.org/10.1093/bib/bbw102>.
- Postgate, J. (1998). *Nitrogen Fixation*. (3rd ed). Cambridge: Cambridge University Press.
- Przybyla, P., Shardlow, M., Aubin, S., Bossy, R., Castilho, R.E., Piperidis, S., McNaught, J. and Ananiadou, S. (2016). Text mining resources for the life sciences. *Database*, 2016(2016), doi: <https://doi.org/10.1093/database/baw145>.
- Sarkar, A. and Reinhold-Hurek, B. (2014). Transcriptional profiling of nitrogen fixation and the role of NifA in the diazotrophic endophyte *Azoarcus* sp. strain BH72. *PLOS ONE*, 9(2), doi: <https://doi.org/10.1371/journal.pone.0086527>.
- Song, Y.L. and Chen, S.S. (2009). Text mining biomedical literature for constructing gene regulatory networks. *Interdisciplinary Sciences Computational Life Sciences*, 1(3), 179-186.
- Vibert, N., Ros, C., Bigot, L.L., Ramond, M., Gatefin, J. and Rouet, J. (2009). Effects of domain knowledge on reference search with the PubMed database: An experimental study. *Journal of the American Society for Information Science and Technology*, 60(7), 1423-1447.
- Wilcock, G. (2009). *Introduction to Linguistic Annotation and Text Analytics*. California: Morgan and Claypool Publishers.
- Zhou, D. and He, Y. (2008). Extracting interactions between proteins from the literature. *Journal of Biomedical Informatics*, 41(2), 393-407.