The Journal of Applied Science **Vol. 14 (No. 1): 68-77 [2015]**
วารสารวิทยาศาสตร์ประยุกต์
ISSN 1513-7805 Printed in Thailand

Review Article

การระบุความสัมพันธ์ของโรคโดยใช้ดัชนีความสัมพันธ์

Identifying disease relationships using association indices

อภิชาต ศุรธณี

Apichat Suratanee

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok

E-mail: apichat.s@sci.kmutnb.ac.th

บทคัดย่อ

การเข้าใจถึงความสัมพันธ์ระหว่างโรคถือเป็นสิ่งสำคัญที่จะช่วยให้พัฒนาความรู้ทางชีววิทยาระบบ รวมถึงช่วยพัฒนาการวินิจฉัยและการรักษาโรคทางการแพทย์ได้ มีการศึกษามากมายที่พยายามศึกษายีนที่มีความสำคัญต่อโรคเพื่อให้เข้าใจกลไกการทำงานของโรค อย่างไรก็ดี ความรู้จากยีนที่สำคัญต่อโรคเพียงอย่างเดียวอาจไม่เพียงพอต่อการทำความเข้าใจกระบวนการทำงานของโรคอย่างเป็นระบบได้อย่างชัดเจน ปัจจุบันมีการพัฒนาวิธีการทางการคำนวณเพื่อใช้วัดความสัมพันธ์ระหว่างโรค ซึ่งวิธีการเหล่านี้ได้รับการยอมรับอย่างกว้างขวางเพื่อสนับสนุนให้เกิดความเข้าใจในชีววิทยาระบบของโรคมากยิ่งขึ้น ในบทความนี้จึงได้ทำการทบทวนเกี่ยวกับตัววัดความสัมพันธ์ เรียกว่า ดัชนีความสัมพันธ์ เพื่อหาความสัมพันธ์ระหว่างโรคในมนุษย์ โดยเน้นจำเพาะไปที่วิธีการทางโครงข่าย

คำสำคัญ: ดัชนีความสัมพันธ์ของโรค, ความสัมพันธ์ระหว่างโรคและยีน, วิธีการทางโครงข่าย

Abstract

Understanding disease relationships is important to enhance knowledge in systems biology as well as disease diagnosis and treatment. Several studies attempt to find genes essential for diseases. However, knowledge of essential genes alone is not enough to complete understanding of disease mechanisms. Several computational methods have been developed to find relationships between diseases. These methods are widely used to support studying in systems biology. This paper reviews disease relationship metrics, named disease association indices, for identifying relationships between diseases in human. In particular, the measurements are based on network-based approaches.

Keywords: disease association index, disease gene relationship, network-based approach

1. บทนำ

ปัจจุบันวิธีการทางการคำนวณได้ถูกพัฒนาขึ้นเพื่อช่วยให้การเข้าใจกระบวนการเกิดโรค รวมถึงพัฒนาวิธีการวินิจฉัยโรคให้มีความแม่นยำมากยิ่งขึ้น งานวิจัยส่วนใหญ่ได้พยายามพัฒนาวิธีการเพื่อบ่งชี้ยีนที่สำคัญต่อโรค ซึ่งถือเป็นข้อมูลสำคัญที่ช่วยลดความเสี่ยงที่จะเกิดโรคนั้นๆได้ หากพิจารณาโครงข่ายปฏิสัมพันธ์ของสิ่งมีชีวิต โดยเฉพาะในมนุษย์จะพบว่ายีนในระบบมีจำนวนมาก และหากต้องการศึกษากระบวนการทำงานของโรคที่มีความซับซ้อนจะพบกับว่ามียีนที่คาดว่าจะสำคัญต่อโรคในปริมาณที่มากเกินกว่าจะทำการทดลองได้ในห้องปฏิบัติการ ด้วยเหตุนี้วิธีการทางการคำนวณจึงมีส่วนสำคัญในการช่วยลดจำนวนของยีนที่คาดว่าน่าจะสำคัญต่อโรคนั้นให้เหลือจำนวนน้อย พอที่จะสามารถทำการทดลองในห้องปฏิบัติการได้ ถึงแม้ว่าในปัจจุบันเทคโนโลยี high-throughput จะเข้ามามีบทบาทเป็นอย่างมากในห้องปฏิบัติการ โดยสามารถทำการทดลองได้ในปริมาณมากด้วยเงื่อนไขในการทดลองที่หลากหลาย และช่วยตอบโจทย์ปัญหาเรื่องจำนวนของยีนที่กล่าวมาข้างต้นได้เป็นอย่างดี แต่กระนั้น ในความเป็นจริงจะพบว่า ต้นทุนในการออกแบบการทดลองรวมถึงความซับซ้อนทางวิธีการยังคงเป็นปัญหาหลักที่ทำให้การทดลองในบางประเทศซึ่งมีต้นทุนในการวิจัยต่ำไม่สามารถทำได้

ดังนั้นวิธีการทางการคำนวณจึงยังคงมีบทบาทสำคัญเป็นอย่างมากสำหรับชีววิทยาระบบวิธีการส่วนใหญ่ได้ถูกพัฒนาขึ้น บนพื้นฐานของข้อมูลทางชีววิทยาที่หลากหลายไม่ว่าจะเป็นจากภาพทางการทดลองเพื่อศึกษาฟีโนไทป์ (Suratanee และคณะ, 2010) จากการแสดงออกของยีนในการทดลอง (Wu และคณะ, 2012; Wong และคณะ, 2014) หรือ จากการสืบค้นข้อความในบทความที่ถูกตีพิมพ์ (Wiegers และคณะ, 2009) ข้อมูลทางชีววิทยาอย่างหนึ่งที่ได้รับความนิยมเป็นอย่างมากในการนำมาใช้ในทางการคำนวณคือข้อมูลปฏิสัมพันธ์ของยีนหรือโปรตีน หลากหลายงานวิจัยประสบความสำเร็จในการพัฒนาวิธีการคำนวณบนโครงข่ายปฏิสัมพันธ์เพื่อหายีนที่สำคัญต่อโรค เช่นการระบุยีนที่สำคัญต่อโรคลำไส้อักเสบด้วยวิธีการทางโครงข่าย (Suratanee และ Plaimas, 2014) อย่างไรก็ดี การระบุได้ถึงยีนที่สำคัญต่อโรคนั้น ยังไม่เพียงพอต่อการเข้าใจถึงชีววิทยาระบบของโรคนั้นๆได้อย่างลึกซึ้ง เนื่องจากความซับซ้อนในโครงข่ายนั้นมีสูง ดังนั้นจึงได้มีผู้พัฒนาวิธีการทางการคำนวณเพื่อการหาความสัมพันธ์ระหว่างโรคขึ้น ซึ่งช่วยพัฒนาฐานความรู้ในระดับกลุ่มของโรค และนำไปสู่การพัฒนาการวินิจฉัยและการรักษาโรคต่อไป ในบทความนี้จะทบทวนตัววัดความสัมพันธ์ระหว่างโรคเพื่อการทำนายความสัมพันธ์ของยีนโดยจำเพาะไปที่วิธีการทางโครงข่าย

2. ข้อมูลแสดงความสัมพันธ์ของยีน โปรตีน และ โรคในมนุษย์

ความสัมพันธ์ของยีนสามารถถูกแสดงออกในรูปแบบของปฏิสัมพันธ์กันของยีนในโครงข่าย ลักษณะของโครงข่ายที่นิยมใช้กันอาจจำแนกออกได้เป็น 3 ลักษณะ คือ 1) Protein-protein interaction 2) Gene ontology (GO) 3) Disease-gene association

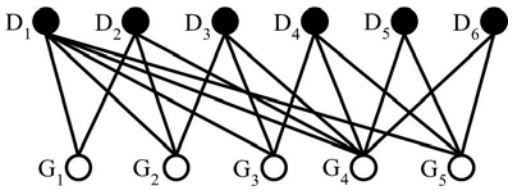
กำหนดให้ $G(V,E)$ แทนกราฟ ซึ่งประกอบด้วย V แทนเซตของโหนด และ $E \subseteq V \times V$ แทนเซตของเส้นเชื่อม ในที่นี้หาก $G(V,E)$ แทนโครงข่ายปฏิสัมพันธ์ของโปรตีน โหนดของโครงข่ายจะหมายถึงโปรตีน และ เส้นเชื่อมของโครงข่ายหมายถึงปฏิสัมพันธ์ระหว่างโปรตีนนั่นเอง โดยโครงข่ายปฏิสัมพันธ์นี้จะมีลักษณะเป็นกราฟไม่ระบุทิศทาง (undirected graph) เพราะเส้นเชื่อมนั้นแสดงเพียงปฏิสัมพันธ์ระหว่างโปรตีน ยกเว้นบางงานวิจัย (Suratanee และคณะ, 2014) ที่ได้มีการกำหนดสถานะของเส้นเชื่อมไว้แตกต่างกันเช่น กำหนดให้เป็นสถานะ activation และ inhibition ข้อมูลปฏิสัมพันธ์ระหว่างโปรตีนปัจจุบันได้มาจากข้อมูลทางการทดลอง เช่น yeast two-hybrid screening (Stelzl และคณะ, 2005) และ affinity capture mass spectrometry (Krogan และคณะ, 2006). เป็นต้น ข้อมูลเหล่านี้ได้ถูกรวบรวมและเก็บไว้เป็นฐานข้อมูล เช่น HPRD (Prasad และคณะ, 2009), BIND (Liu และคณะ, 2007), STRING (Franceschini และคณะ, 2013) เป็นต้น โดยมีการแจกจ่ายให้ใช้โดยไม่เสียค่าใช้จ่ายเพื่อการศึกษา โดยพบว่าฐานข้อมูลส่วนใหญ่นี้จะมีขนาดใหญ่ กล่าวคือจะมีจำนวนโหนดมากกว่า 10,000 โหนด และเส้นเชื่อมมากกว่า 60,000 เส้น

สำหรับ gene ontology (The Gene Ontology Consortium, 2008) ยีนจะถูกอธิบายในรูปแบบของ GO term เพื่อบ่งบอกคุณสมบัติทางชีววิทยา เทอมต่างๆเหล่านี้จะถูกจัดการให้อยู่ใน 3 โดเมนประกอบด้วย cellular component, molecular function, และ biological process โดยข้อมูลเหล่านี้ถูกเก็บเข้าในฐานข้อมูล Gene Ontology Database (http://www.geneontology.org) ซึ่งประกอบไปด้วยข้อมูลของมนุษย์และสิ่งมีชีวิตอื่นๆ ด้วย

ในส่วนของ disease-gene association ซึ่งถูกแทนด้วยกราฟจะประกอบไปด้วย โรคในมนุษย์ที่ทราบแน่ชัดแล้ว และกลุ่มของยีนที่เกี่ยวข้องกับโรค โดยโรคที่มีการศึกษามาก มักจะมีจำนวนยีนที่เกี่ยวข้องมาก ในทางกลับกันยีนที่มีการศึกษามากหรือมีหน้าที่ที่หลากหลาย มักจะเกี่ยวข้องกับโรคเป็นจำนวนมากเช่นกัน ฐานข้อมูลที่เก็บรวบรวมเกี่ยวกับโรคและยีนที่เกี่ยวข้องกับโรคและเป็นที่นิยมใช้กันอย่างแพร่หลาย มีหลากหลายฐานข้อมูล เช่น Online Mendelian Inheritance in Man (OMIM) (Hamosh และคณะ, 2005), Comparative Toxicogenomics Database (CTD) (Davis และคณะ, 2011) และ Functional Disease Ontology annotations (FunDO) (Osborne และคณะ, 2009) เป็นต้น

3. ลักษณะของโครงข่าย

หากกล่าวถึงโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนอาจมองเป็นกราฟในลักษณะกราฟไม่ระบุทิศทาง แต่หากกล่าวถึงโครงข่ายของ Gene ontology หรือ Disease-gene association (Sun และคณะ, 2014) จำเป็นจะต้องพิจารณากราฟให้มีลักษณะเป็น Bi-partite graph กล่าวคือ กราฟที่มีชนิดของโหนดสองชนิด และเส้นเชื่อมจะเกิดขึ้นระหว่างโหนดที่มีชนิดต่างกัน เช่น โหนดชนิดแรกแทนโรค และโหนดชนิดที่สองแทนยีน เป็นต้น



รูปที่ 1. กราฟที่มีลักษณะแบบ Bi-partile network

จากรูปที่ 1. หากกำหนดให้ D_i แทนโรคที่ i เมื่อ $i \in \{1,2,3,4,5,6\}$ และ G_j แทนยีนที่ j เมื่อ $j \in \{1,2,3,4,5\}$ กล่าวคือจากรูปดังกล่าวมีโรคทั้งหมด 6 โรค และมียีนทั้งหมด 5 ยีน สำหรับโรคที่ 1 (D_1) พบว่ามีความเกี่ยวข้องกับยีนทั้งหมด 5 ยีน (G_1 ถึง G_5) ในขณะที่ โรคที่ 2 (D_2) พบว่ามีความเกี่ยวข้องกับยีนเพียง 3 ยีนเท่านั้น คือ G_1, G_2 และ G_4

4. การวัดความสัมพันธ์ระหว่างโรค

การวัดความสัมพันธ์ระหว่างโรคสามารถทำได้บนโครงข่าย Bi-partile ที่กล่าวมาข้างต้นหรือทำบนกราฟไม่ระบุทิศทางก็ได้ โดยงานวิจัยส่วนใหญ่จะทำการออกแบบตัววัดคะแนนความเหมือน (similarity score) หรือ ดัชนีความสัมพันธ์ (association index) (Bassและคณะ, 2013) ที่แตกต่างกันออกไป

หากกำหนดให้ โหนดในโครงข่ายมีสองชนิดคือ ชนิด X และ Y หากโหนดชนิด X แทนโรค และชนิด Y แทนยีน การหาความสัมพันธ์ระหว่างโรคนั้น คือการหาความสัมพันธ์ระหว่างโหนดสองโหนด เช่น โหนด D_1 และ โหนด D_2 ซึ่งเป็นโหนดชนิด X ว่ามีการเชื่อมโยงไปยังยีน หรือ โหนดชนิด Y ว่ามีร่วมกันเป็นจำนวนเท่าใด นั่นคือ $\left|N(D_1) \cap N(D_2)\right|$ เมื่อ $\left|N(D_1)\right|$ แทนดีกรีของโหนด D_1 กล่าวคือจำนวนยีนที่ โรค D_1 มีปฏิสัมพันธ์ด้วย และ $\left|N(D_2)\right|$ แทนดีกรีของโหนด D_2 กล่าวคือจำนวนยีนที่ โรค D_2 มีปฏิสัมพันธ์ด้วย

ปัจจุบันได้มีผู้พัฒนาตัววัดหรือดัชนีความสัมพันธ์ไว้มากมาย ในที่นี้ขอสรุปเฉพาะที่นิยมถูกนำมาใช้กันมากดังต่อไปนี้

4.1 Jaccard index

Jaccard index ถูกใช้อย่างแพร่หลายในหลากหลายงานวิจัย รวมทั้งการหาความสัมพันธ์ระหว่างโรคด้วย เนื่องด้วยสูตรที่ง่ายต่อการทำความเข้าใจ โดย Jaccard index จะทำการคำนวณสัดส่วนของยีนที่ถูกใช้ร่วมกันระหว่างโรคสองโรคหารด้วยจำนวนยีนทั้งหมดที่เชื่อมต่อไปยังโรคทั้งสองนั้น ซึ่ง Jaccard index สามารถแสดงได้ดังต่อไปนี้

$$\text{Jaccard}(D_1, D_2) = \frac{\left|N(D_1) \cap N(D_2)\right|}{\left|N(D_1) \cup N(D_2)\right|}$$

4.2 Simpson index

Simpson index มีลักษณะคล้ายกับ Jaccard index เพียงแต่มีการปรับเปลี่ยนตัวหารให้เป็นดีกรีที่น้อยที่สุดของโหนดของโรค โดยสูตรการคำนวณ Simpson index สามารถแสดงได้ดังต่อไปนี้

$$\text{Simpson}(D_1, D_2) = \frac{\left|N(D_1) \cap N(D_2)\right|}{\min(\left|N(D_1)\right|, \left|N(D_2)\right|)}$$

4.3 Geometric index

สำหรับ Geometric index นั้นจะทำการพิจารณาสัดส่วนของโหนดร่วมระหว่างโรคสองโรคในรูปแบบของผลคูณ ซึ่งสามารถแสดงได้ดังต่อไปนี้

$$\text{Geometric}(D_1, D_2) = \frac{\left|N(D_1) \cap N(D_2)\right|^2}{\left|N(D_1)\right| \cdot \left|N(D_2)\right|}$$

4.4 Cosine index

Cosine index มีลักษณะที่คล้ายคลึงกับทั้ง Geometric index คือพิจารณาสัดส่วนในรูปแบบของผลคูณ โดยสามารถแสดงได้ดังต่อไปนี้

$$\text{Cosine}(D_1, D_2) = \frac{\left|N(D_1) \cap N(D_2)\right|}{\sqrt{\left|N(D_1)\right| \cdot \left|N(D_2)\right|}}$$

4.5 Pearson correlation coefficient

Pearson correlation coefficient (PCC) พิจารณาความสัมพันธ์กัน (correlation) ระหว่างคุณลักษณะของโรค D_1 และโรค D_2 ซึ่งจะพิจารณาถึง จำนวนของยีนทั้งหมดในระบบ (n_Y) เข้ามาเกี่ยวข้องด้วย โดย PCC สามารถคำนวณได้ดังต่อไปนี้

$$\text{PCC}(D_1, D_2) = \frac{\left|N(D_1) \cap N(D_2)\right| \cdot n_Y - \left|N(D_1)\right| \cdot \left|N(D_2)\right|}{\sqrt{\left|N(D_1)\right| \cdot \left|N(D_2)\right| \cdot (n_Y - \left|N(D_1)\right|) \cdot (n_Y - \left|N(D_2)\right|)}}$$

4.6 Hypergeometric index

Hypergeometric index พิจารณาการใช้ค่า Logarithm เพื่อแปลงค่าความน่าจะเป็นของการมีส่วนร่วมกันของปฏิสัมพันธ์ระหว่างโรคสองโรค ซึ่ง Hypergeometric index สามารถคำนวณได้ดังต่อไปนี้

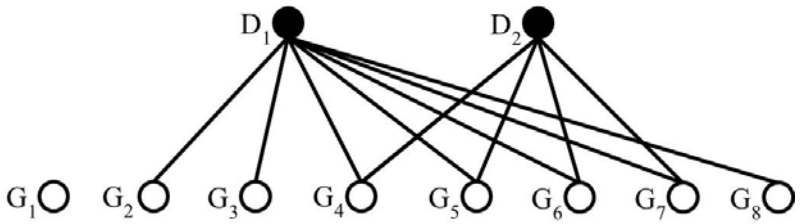
$$\text{Hypergeometric}(D_1, D_2) = -\log_{10} \sum_{i=|N(D_1)\cap N(D_2)|}^{\min(|N(D_1)|,|N(D_2)|)} \frac{\binom{|N(D_1)|}{i} \cdot \binom{n_Y - |N(D_1)|}{|N(D_2)| - i}}{\binom{n_Y}{|N(D_2)|}}$$

จากดัชนีความสัมพันธ์ทั้ง 6 ที่กล่าวมาข้างต้นนั้น ไม่อาจกล่าวได้อย่างแน่ชัดว่าดัชนีใดดีที่สุดหรือแย่ที่สุด ทั้งนี้ขึ้นกับลักษณะของปัญหา ในบางกรณีจะพบว่าหากใช้ตัววัดดัชนีตัวเดียวกัน วัดกราฟที่มีลักษณะแตกต่างกันอาจคำนวณได้ค่าเดียวกันได้ โดยตัวอย่างของการใช้ดัชนีความสัมพันธ์จะแสดงไว้ในหัวข้อถัดไป

5. ตัวอย่างการใช้ดัชนีความสัมพันธ์

ค่าที่คำนวณได้จากดัชนีความสัมพันธ์จะบ่งบอกถึงความสัมพันธ์ระหว่างโรคที่กำลังพิจารณา กล่าวคือ หากค่าดัชนีมีค่ามากจะหมายถึงว่าโรคสองโรคนั้นมีความเกี่ยวพันกันสูง ในทางตรงกันข้ามหากค่าดัชนีที่คำนวณได้มีค่าต่ำ นั่นบ่งบอกว่าโรคสองโรคนั้นอาจมีความเกี่ยวพันกันต่ำหรือไม่มีความเกี่ยวพันกันเลย โดยตัววัดแต่ละตัวจะมีช่วงของค่าดัชนีที่แตกต่างกัน ส่วนใหญ่จะมีค่าดัชนีอยู่ในช่วง 0 ถึง 1 โดยค่า 1 แทนความสัมพันธ์ที่สูง และ ค่า 0 แทนความสัมพันธ์ที่ต่ำ ยกเว้น PCC ซึ่งมีช่วงค่าดัชนีอยู่ในช่วง -1 ถึง 1 โดยค่า 1 บ่งบอกถึงโรคสองโรคที่มีความสัมพันธ์กันโดยสมบูรณ์ ส่วนค่า 0 หมายถึงจำนวนของยีนที่ร่วมกันของทั้งสองโรคนั้นมีค่าเท่ากับจำนวนของยีนที่ร่วมกันที่สามารถเป็นไปได้ของทั้งสองโรค แต่หากค่าที่คำนวณได้มีค่าเป็น -1 นั้นหมายถึงโรคทั้งสองไม่มีส่วนสัมพันธ์กันเลยเมื่อพิจารณายีนทั้งระบบ

หากพิจารณาดัชนีทั้ง 6 จากกรณีศึกษาดังรูปที่ 2 โรค D_1 เชื่อมโยงกับยีน G_2 G_3 G_4 G_5 G_6 G_7 และ G_8 ในขณะที่โรค D_2 เชื่อมโยงกับยีน G_4 G_5 G_6 และ G_7



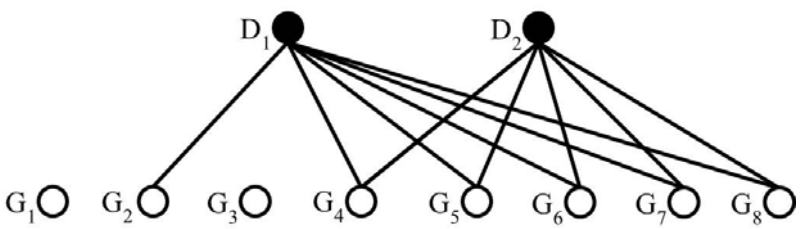
รูปที่ 2. ตัวอย่างโครงข่ายของโรคสองโรคที่มียีนร่วมกันจำนวน 4 ยีน และมียีนที่ต่างกัน 3 ยีน

ตามโครงข่ายในรูปที่ 2 สามารถคำนวณค่าดัชนีได้ผลการคำนวณแสดงไว้ดังตารางที่ 1 ซึ่งพบว่า Simpson index ให้ค่าดัชนีที่มากที่สุด ในขณะที่ Hypergeometric index แสดงผลค่าดัชนีค่อนข้างต่ำ

ตารางที่ 1. แสดงค่าดัชนีที่แตกต่างกันที่คำนวณได้ตามโครงข่ายในรูปที่ 2

ตัววัด	**Jaccard**	**Simpson**	**Geometric**	**Cosine**	**PCC**	**Hypergeometric**
ค่าดัชนี	0.5714	1.0000	0.5714	0.7559	0.3779	0.3010

หากพิจารณาโครงข่ายที่มียีนร่วมกันระหว่างโรคมากขึ้น ดังรูปที่ 3 ซึ่งโรค D_1 และ D_2 มียีนร่วมกันถึง 5 ยีนคือ ยีน G_4 G_5 G_6 G_7 และ G_8 ขณะเดียวกันมียีนที่ต่างกันเพียงยีนเดียวคือยีน G_2 โดยผลการคำนวณค่าดัชนีที่ได้แสดงไว้ในตารางที่ 2



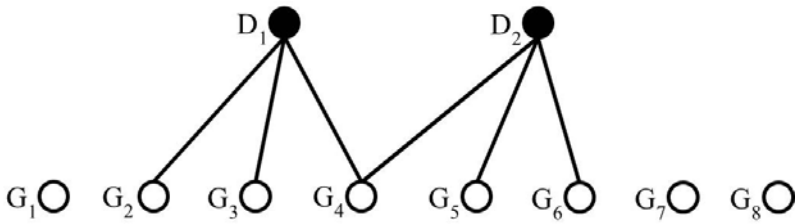
รูปที่ 3. ตัวอย่างโครงข่ายของโรคสองโรคที่มียีนร่วมกันจำนวน 5 ยีน และมียีนที่ต่างกันเพียงยีนเดียว

จากตารางที่ 2 พบว่าค่าดัชนีที่คำนวณได้มีค่าสูงกว่าผลในตารางที่ 1 ทั้งนี้เนื่องจากมียีนร่วมกันระหว่างโรคทั้งสองเป็นจำนวน 5 ยีน และมียีนที่ต่างกันจำนวนน้อย นอกจากนั้นพบว่า Simpson index ยังคงให้ค่าเป็น 1 ซึ่งมีค่าเท่ากับกับค่าที่คำนวณได้ในโครงข่ายในรูปที่ 2

ตารางที่ 2. แสดงค่าดัชนีที่แตกต่างกันที่คำนวณได้ตามโครงข่ายในรูปที่ 3

ตัววัด	Jaccard	Simpson	Geometric	Cosine	PCC	Hypergeometric
ค่าดัชนี	0.8333	1.0000	0.8333	0.9129	0.7454	0.9700

พิจารณากรณีที่มียีนร่วมกันเพียงยีนเดียวและมียีนที่ต่างกันโรคละสองยีนดังรูปที่ 4 ผลจากการคำนวณค่าดัชนีได้ถูกแสดงไว้ในตารางที่ 3



รูปที่ 4. ตัวอย่างโครงข่ายของโรคสองโรคที่มียีนร่วมกันเพียง 1 ยีน และมียีนที่ต่างกันโรคละ 2 ยีน

จากค่าดัชนีที่คำนวณได้สำหรับโครงข่ายในรูปที่ 4 พบว่าค่าดัชนีทั้งหมดอยู่ในเกณฑ์ต่ำเมื่อเปรียบเทียบกับโครงข่ายทั้งสองโครงข่ายก่อนหน้า นอกจากนั้นพบว่าค่า PCC ยังแสดงค่าลบ ซึ่งแสดงให้เห็นถึงความไม่สัมพันธ์กันระหว่างโรคทั้งสอง

ตารางที่ 3. แสดงค่าดัชนีที่แตกต่างกันที่คำนวณได้ตามโครงข่ายในรูปที่ 4

ตัววัด	Jaccard	Simpson	Geometric	Cosine	PCC	Hypergeometric
ค่าดัชนี	0.1667	0.3333	0.0833	0.2887	-0.2582	0.0322

6. บทสรุป

วิธีการคำนวณที่ใช้พื้นฐานข้อมูลจากการทดลองได้ถูกพัฒนามาอย่างต่อเนื่อง นอกจากการศึกษาความสัมพันธ์กันระหว่างยีนหรือโปรตีนโดยใช้โครงข่ายที่ถูกพัฒนากันอย่างกว้างขวางแล้วนั้น ยังมีการศึกษาในระดับความสัมพันธ์ระหว่างโรคซึ่งช่วยให้ข้อมูลภาพกว้างที่มีประโยชน์ต่อการเข้าใจกลไกการทำงานของโรคที่ซับซ้อนที่มีลักษณะของโรคใกล้เคียงกัน และเป็นประโยชน์ต่อการวินิจฉัยโรค โดยวิธีการส่วนใหญ่จะใช้ข้อมูลแสดงความสัมพันธ์กันระหว่างยีนและโรค ข้อมูลเหล่านี้ได้ถูกนำมาพัฒนาเป็นโครงข่ายของโรคได้ การสร้างโครงข่ายแสดงความสัมพันธ์ของโรคและยีนที่เกี่ยวข้องนั้นอาจแสดงความสัมพันธ์ดังกล่าวได้ในหลายลักษณะ เช่น

การใช้ข้อมูลโดยตรงจากฐานข้อมูล Disease-gene association หรืออาจพิจารณาในเชิงหน้าที่ โดยใช้ GO term หรืออาจใช้ข้อมูลปฏิสัมพันธ์จากโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน การพิจารณาเซตของยีนที่สำคัญต่อโรคหนึ่งๆนั้นเปรียบเสมือนเป็นคุณลักษณะของโรคนั้นๆ เพื่อใช้ในการเปรียบเทียบกับโรคอื่นๆ หากโรคสองโรคที่ทำการเปรียบเทียบมียีนร่วมกันเป็นจำนวนมากและมียีนที่ต่างกันเป็นจำนวนน้อย นั่นแสดงถึงความสัมพันธ์ระหว่างโรคทั้งสองนั้นมีสูง โดยดัชนีความสัมพันธ์ได้ถูกพัฒนาเพื่อเป็นตัววัดที่ใช้วัดความสัมพันธ์ดังกล่าว ซึ่งถือเป็นค่าที่มีความสำคัญและสามารถใช้เป็นข้อมูลในการวิเคราะห์ความสัมพันธ์ระหว่างโรคต่อไป

งานวิจัยที่เกี่ยวข้องกับการระบุความสัมพันธ์ระหว่างโรคนั้นมีความสำคัญและยังคงมีการพัฒนาตัววัดต่างๆ ขึ้นอย่างต่อเนื่องในปัจจุบัน การมีข้อมูลความสัมพันธ์ของโรคมีประโยชน์ในการศึกษาทางชีววิทยาระบบ ชีวเวชศาสตร์ และทางการแพทย์เป็นอย่างมาก และยังเป็นความท้าทายอย่างหนึ่งสำหรับนักคณิตศาสตร์ คอมพิวเตอร์และชีววิทยาเชิงคำนวณที่จะค้นหาตัววัดความสัมพันธ์ที่น่าสนใจต่อไป

7. เอกสารอ้างอิง

Bass J.I.F, Diallo A., Nelson J., Soto J.M., Myers C.L., Walhout A.J., (2013) Using networks to measure similarity between genes: association index selection. Nat. Methods, 10(12): 1169-1176.

Davis A.P., King B.L., Mockus S., Murphy C.G., Saraceni-Richards C. et al., (2011) The comparative toxicogenomics database: update 2011. Nucleic Acids Res., 39(Database issue): D1067-72.

Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M. et al., (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res., 41(Database issue): D808-815.

Hamosh A., Scott A.F., Amberger J.S., Bocchini C.A. and McKusick V.A., (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res., 33(Database issue): D514-D517.

Krogan N.J., Cagney G., Yu H., Zhong G., Guo X. et al., (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature, 440(7084): 637-643.

Liu T., Lin Y., Wen X., Jorissen R.N. and Gilson M.K., (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res., 35(Database issue): D198-201.

Osborne J.D., Flatow J., Holko M., Lin S.M., Kibbe W.A. et al., (2009) Annotating the human genome with Disease Ontology. BMC Genomics, 10(Suppl 1): S6.

Prasad T.S., Kandasamy K. and Pandey A., (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol. Biol., 577: 67-79.

Stelzl U., Worm U., Lalowski M., Haenig C., Brembeck F.H. et al., (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell, 122(6): 957-968.

Sun K., Gonçalves J.P., Larminie C. and Przulj N., (2014) Predicting disease associations via biological network analysis. BMC Bioinformatics, 15: 304.

Suratanee A. and Plaimas K., (2014) Identification of inflammatory bowel disease-related proteins using a reverse k-nearest neighbor search. J. Bioinform. Comput. Biol., 12(4): 1450017.

Suratanee A., Rebhan I., Matula P., Kumar A., Kaderali L., et al., (2010) Detecting host factors involved in virus infection by observing the clustering of infected cells in siRNA screening images. Bioinformatics, 26(18): i653-i658.

Suratanee A., Schaefer M.H., Betts M.J., Soons Z., Mannsperger H., et al., (2014) Characterizing protein interactions employing a genome-wide siRNA cellular phenotyping screen. PLoS Comput. Biol., 10(9): e1003814.

The Gene Ontology Consortium, (2008) The Gene Ontology project in 2008. Nucl. Acids Res., 36(suppl 1): D440-D444.

Wiegers T.C., Davis A.P., Cohen K.B., Hirschman L. and Mattingly C.J., (2009) Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). BMC Bioinformatics, 10: 326.

Wong D.C., Sweetman C. and Ford C.M., (2014) Annotation of gene function in citrus using gene expression information and co-expression networks. BMC Plant Biol., 14: 186.

Wu C., Zhu J. And Zhang X., (2012) Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics, 13: 182.