การค้นหาหน้าที่และความสัมพันธ์ของกลุ่มยีนด้วย Gene Ontology

จุฑารัตน์ มณีวัฒนาพฤกษ์

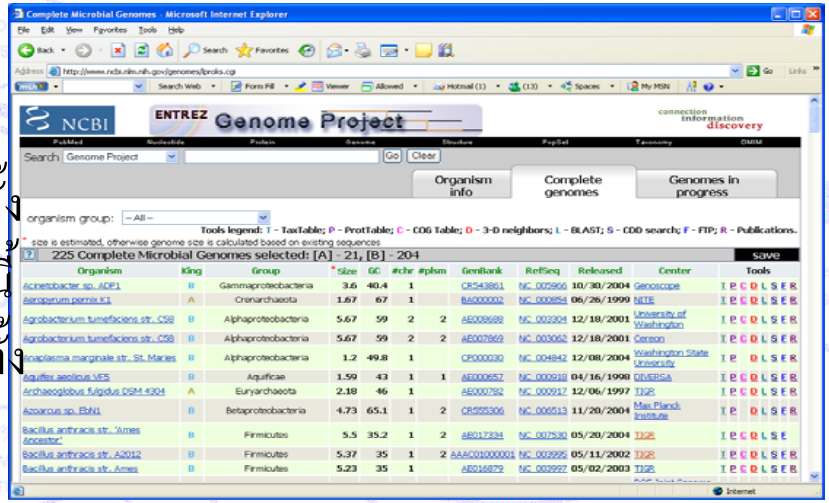
ดร. นพดล คีรีเพ็ชร

ฝ่ายวิจัยและพัฒนาเทคโนโลยีคอมพิวเตอร์เพื่อการคำนวณ

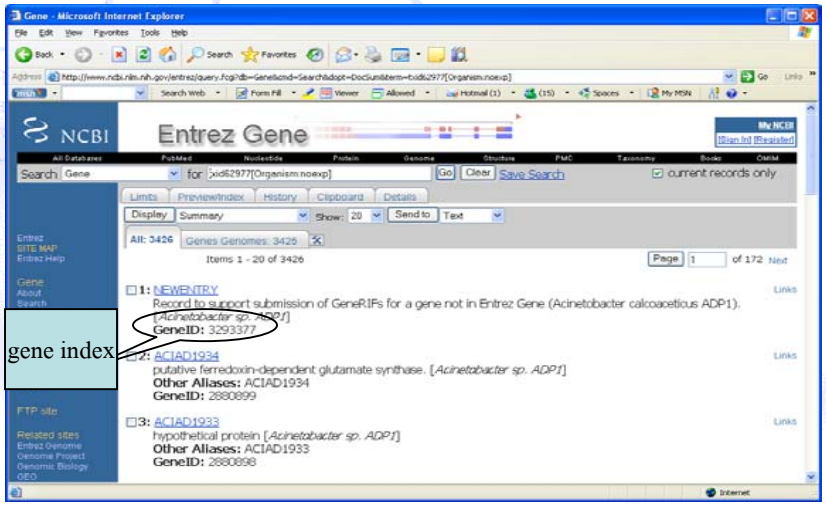
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

ข้อมูลของ Microbial Genomes

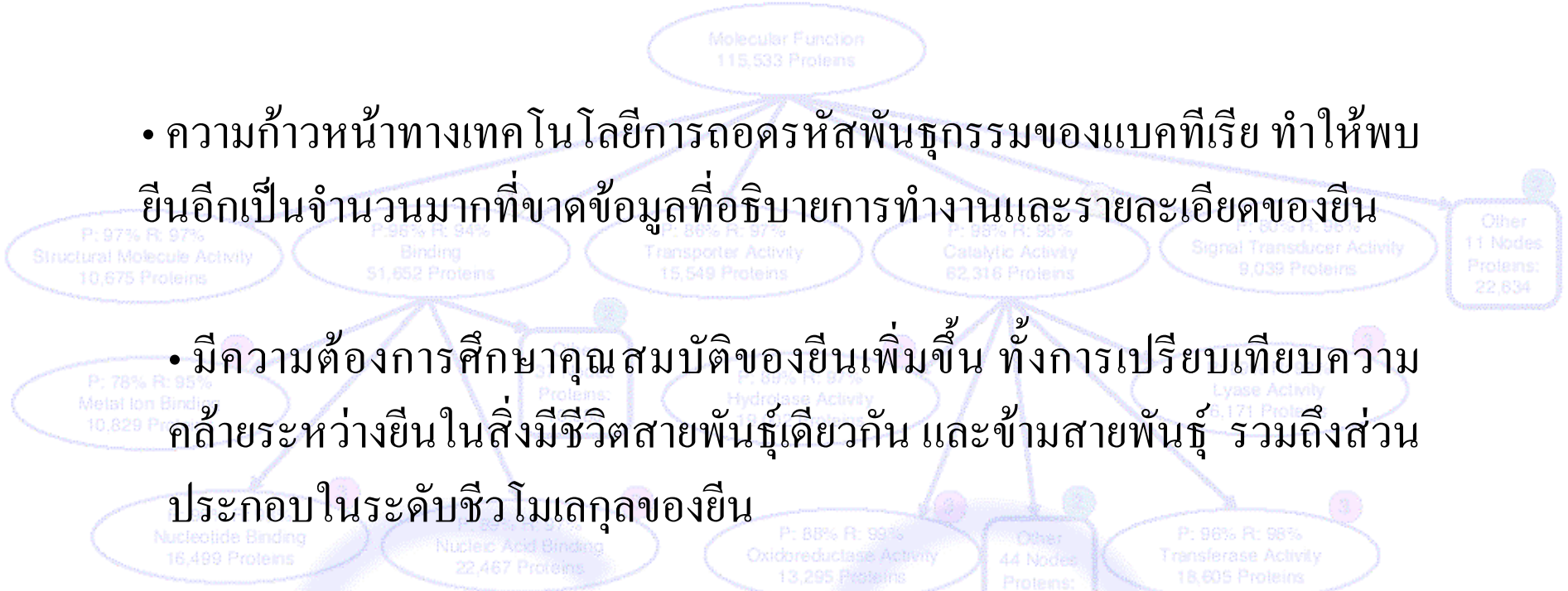
ข้อมูลพันธุกรรมของสิ่งมีชีวิตหลายชนิด ได้รับการรวบรวมไว้ที่Genbank (http://www.ncbi.nlm.nih.gov) รวมทั้งแบคทีเรีย ซึ่งเป็นสิ่งมีชีวิตที่ใช้สำหรับการศึกษาในงานนี้ โดยปัจจุบันพบว่ามีข้อมูลสายพันธุกรรมของแบคทีเรียทั้งสิ้น 225 สายพันธุ์



ระบบของ Genbank มีการจัดเก็บข้อมูลรายละเอียดของยีนและส่วนประกอบที่เกี่ยวข้องกับแบคทีเรียเหล่านั้นไว้อย่างสมบูรณ์ สามารถสืบค้นและดึงข้อมูลมาใช้ได้ด้วยการใช้ gene index (gi) ตามมาตรฐาน Genebank



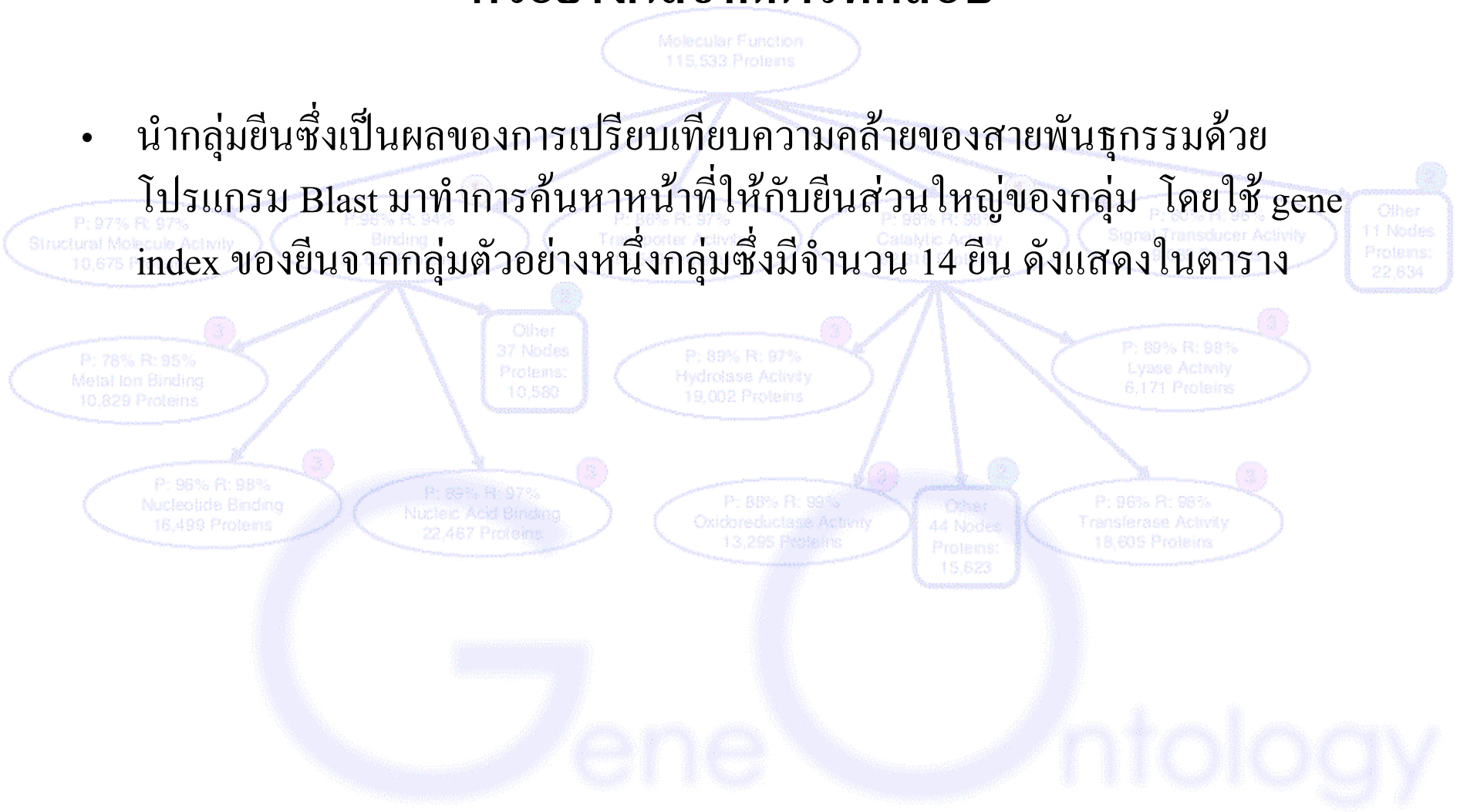
แนวคิดในการพัฒนางาน



- ความก้าวหน้าทางเทคโนโลยีการถอดรหัสพันธุกรรมของแบคทีเรีย ทำให้พบยีนอีกเป็นจำนวนมากที่ขาดข้อมูลที่อธิบายการทำงานและรายละเอียดของยีน
- มีความต้องการศึกษาคุณสมบัติของยีนเพิ่มขึ้น ทั้งการเปรียบเทียบความคล้ายระหว่างยีนในสิ่งมีชีวิตสายพันธุ์เดียวกัน และข้ามสายพันธุ์ รวมถึงส่วนประกอบในระดับชีวโมเลกุลของยีน

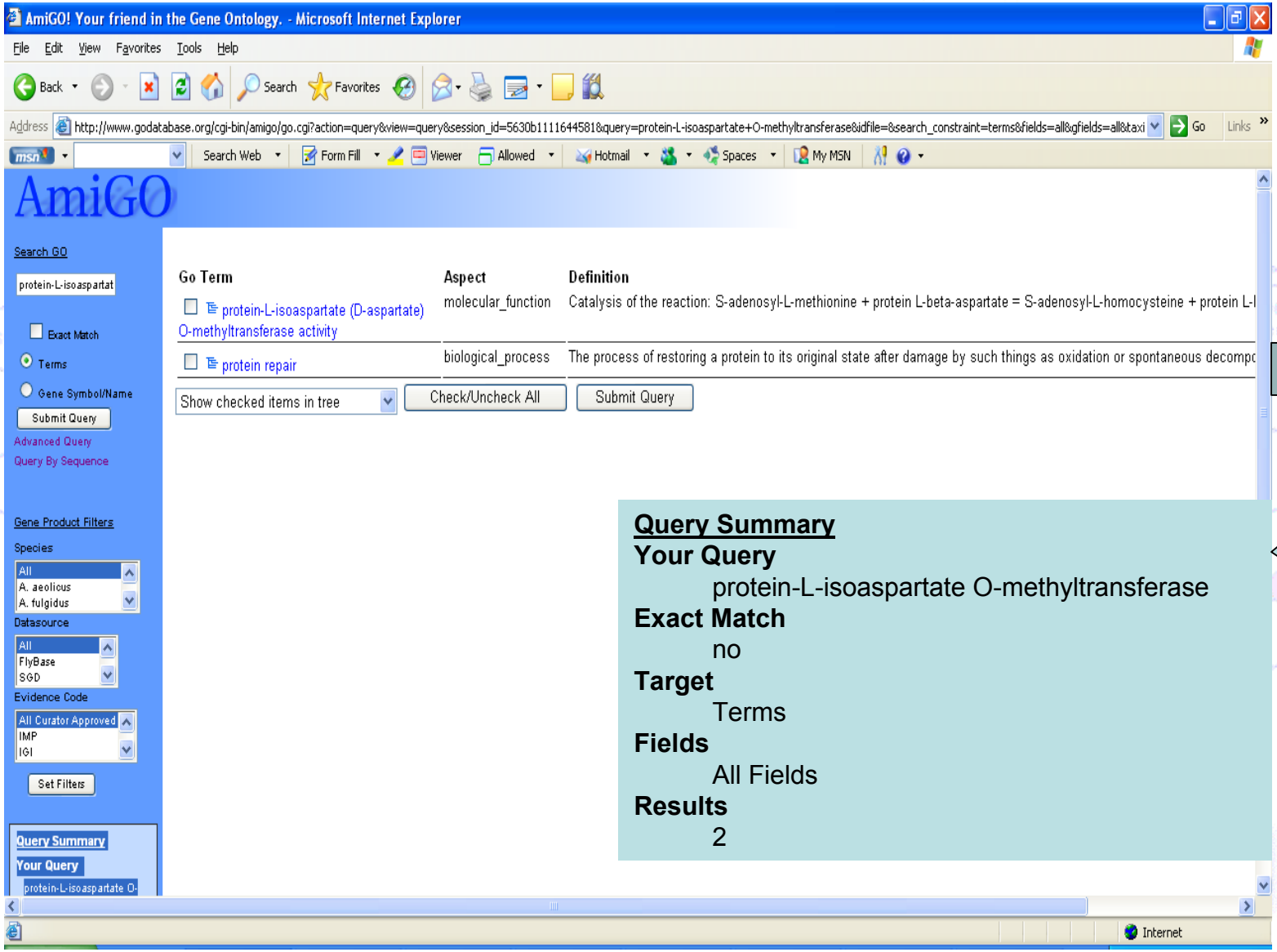
ตัวอย่างผลจากการทดสอบ

- นำกลุ่มยีนซึ่งเป็นผลของการเปรียบเทียบความคล้ายของสายพันธุกรรมด้วยโปรแกรม Blast มาทำการค้นหาหน้าที่ให้กับยีนส่วนใหญ่ของกลุ่ม โดยใช้ gene index ของยีนจากกลุ่มตัวอย่างหนึ่งกลุ่มซึ่งมีจำนวน 14 ยีน ดังแสดงในตาราง



gi	Accession	Organism	gene	synonym	product
11498123	NC_000917	Archaeoglobus_fulgidus		AF0512	chloroplast inner envelope membrane protein
11499903	NC_000917	Archaeoglobus_fulgidus		AF2322	L-isoaspartyl protein carboxyl methyltransferase (pcm-2)
14601138	NC_000854	Aeropyrum_pernix		APE1011	hypothetical protein-L-isoaspartate O-methyltransferase
146012				83	hypothetical protein
156066				57	hypothetical protein
156070				39	putative protein
15888383	NC_003062	Agrobacterium_tumefaciens_C58_Cereon	AGR_C_1920	AGR_C_1920p	
15888938	NC_003062	Agrobacterium_tumefaciens_C58_Cereon	AGR_C_2998	AGR_C_2998p	
15889008	NC_003062	Agrobacterium_tumefaciens_C58_Cereon	AGR_C_3127	AGR_C_3127p	
15889027	NC_003062	Agrobacterium_tumefaciens_C58_Cereon	AGR_C_3159	AGR_C_3159p	
17934948	NC_003304	Agrobacterium_tumefaciens_C58_UWash		Atu1041	methyltransferase
17935521	NC_003304	Agrobacterium_tumefaciens_C58_UWash		Atu1625	methyltransferase
17935595	NC_003304	Agrobacterium_tumefaciens_C58_UWash		Atu1701	L-isoaspartyl protein carboxyl methyltransferase
17935615	NC_003304	Agrobacterium_tumefaciens_C58_UWash		Atu1721	protein-L-isoaspartate O-methyltransferase

- นำข้อมูล gene product ของแต่ละยีน ไปทำการค้นหา Go term บนระบบของ GO



Query Summary

Your Query
protein-L-isoaspartate O-methyltransferase

Exact Match
no

Target
Terms

Fields
All Fields

Results
2

การวิเคราะห์ผล

ตัวอย่างยีนจำนวน 14 ยีน มี gene product 10 ยีน พบข้อมูล GO 9 ตัว

GO ID	Go term	จำนวนยีนที่สัมพันธ์กับ GO term	จำนวนยีนที่ไม่สัมพันธ์กับ GO term	gi	Accession
GO:0008150	biological_process unknown	3	7		
GO:0005525	GTP binding	3	7		
GO:0005554	molecular_function	3	7	14601138	NC_000918
GO:0006355	regulation of transcription, DNA-dependent	3	7	14601233	NC_000918
GO:0008168	methyltransferase activity	3	7	15606625	NC_000854
GO:0008372	cellular_component unknown	3	7		
GO:0016020	membrane	3	7		
GO:0004719	protein-L-isoaspartate (D-aspartate) O-methyltransferase activity	4	6	17934948	NC_003304
GO:0030091	protein repair	4	6	17935521	NC_003304
				17935595	NC_003304
				17935615	NC_003304

Fisher's exact test

A	B	A+B
C	D	C+D
A+C	B+D	N

2x2 contingency table

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{A!\,B!\,C!\,D!\,N!}$$

P (outcome)

สรุปผลการค้นหา Go term ที่ตรงกับยีน

GO ID	Go term	P-value
GO:0008150	biological_process unknown	0.6758
GO:0005525	GTP binding	0.6758
GO:0005554	molecular_function	0.6758
GO:0006355	regulation of transcription, DNA-dependent	0.6758
GO:0008168	methyltransferase activity	0.6758
GO:0008372	cellular_component unknown	0.1111
GO:0016020	membrane	0.1111
GO:0004719	protein-L-isoaspartate (D-aspartate) O-methyltransferase activity	0.5
GO:0030091	protein repair	0.6612

Cutt off P-Value = 0.529

ข้อสรุปและงานที่ดำเนินการต่อ

- ข้อจำกัดของการวิเคราะห์ผล

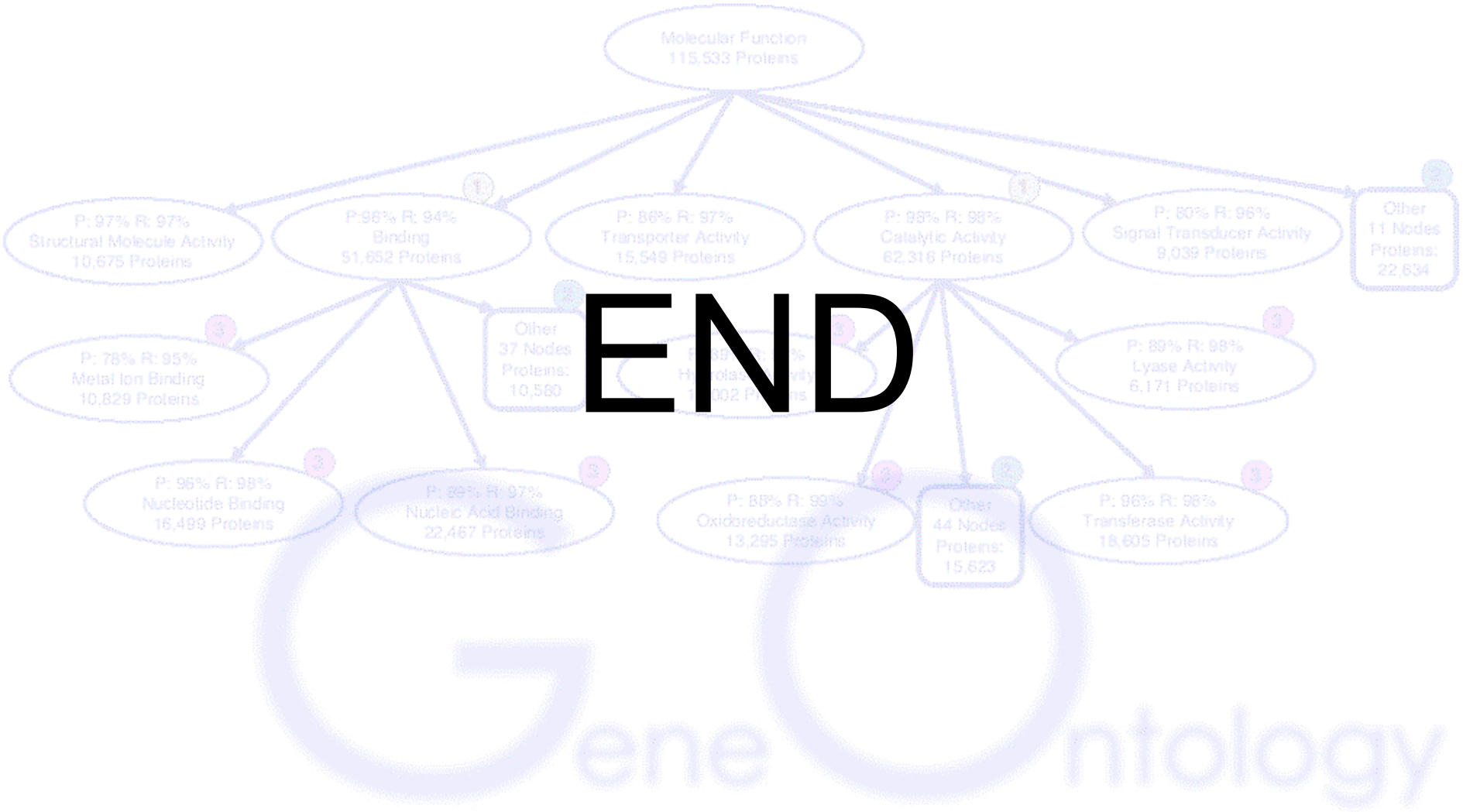
  การใช้ Fisher's exact test เหมาะกับปริมาณข้อมูลน้อยๆ ใช้การเปรียบเทียบครั้งละสองกลุ่ม ถ้าข้อมูลมากๆ ต้องเสียเวลาวนวิเคราะห์หลายรอบ และทำให้ผลมีค่าผิดพลาดได้ จึงควรทดลองใช้การวิเคราะห์อื่นๆ ทดสอบด้วย

- งานส่วนที่ต้องดำเนินการต่อจากการทดลองข้อมูลข้อมูลตัวอย่าง

  การสร้าง Engine ของระบบที่สามารถทำการคัดเลือก GO term โดยไม่ต้องพึ่งระบบของ GO และส่วนของการวิเคราะห์ผลให้ทำงานอัตโนมัติ

เอกสารอ้างอิง

[1] Gene Ontology consortium
[2] SGD (Saccaromyces cerevisiae orbudding yeast) : http://www.yeastgenome.org/help/yeastGeneNomenclature.shtml
[3] FlyBase (Drosophila) : http://flybase.bio.indiana.edu/
[4] MGI (Mouse) : http://www.informatics.jax.org/
[5] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*, 25:25-29.
[6] Zhang B, Schmoyer D, Kirov S, Snoddy J. (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *Bioinformatics*, 5(1):16.
[7] Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. (2002). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4:R7.
[8] Al-Shahrour,F., Diaz-Uriarte,R.andDopazo,J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20, 578–580.
[9] Zhong S, Tian L, Li C, Storch FK, and Wong WH. (2004). Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework. *Proc IEEE Computational Systems Bioinformatics*, 425-435.
[10] Blast: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html


Molecular Function
115,533 Proteins
P: 97% R: 97%
Structural Molecule Activity
10,675 Proteins
P:96% R: 94%
Binding
51,652 Proteins
P: 86% R: 97%
Transporter Activity
15,549 Proteins
P: 98% R: 98%
Catalytic Activity
62,316 Proteins
P: 80% R: 96%
Signal Transducer Activity
9,039 Proteins
Other
11 Nodes
Proteins:
22,634
P: 78% R: 95%
Metal Ion Binding
10,829 Proteins
Other
37 Nodes
Proteins:
10,580
P: 89% R: 98%
Lyase Activity
6,171 Proteins
P: 96% R: 98%
Nucleotide Binding
16,499 Proteins
P: 89% R: 97%
Nucleic Acid Binding
22,467 Proteins
P: 88% R: 99%
Oxidoreductase Activity
13,295 Proteins
Other
44 Nodes
Proteins:
15,623
P: 96% R: 98%
Transferase Activity
18,605 Proteins

END

- นำผลรวมของจำนวนยีนที่ตรง และไม่ตรงกับแต่ละ Go term มาทำการวิเคราะห์ค่าทางสถิติเพื่อระบุ Go term ที่เป็นคำนิยามสำหรับกลุ่มยีน โดยการใช้ Fisher's exact test สำหรับข้อมูลที่เป็นความสัมพันธ์ระหว่างสอง Go term

  โดยสัดส่วนระหว่างผลรวมการพบยีนใน GO term ที่ 1 กับผลรวมการพบยีนใน GO term ที่ 1 มากกว่าสัดส่วนระหว่างผลรวมการพบยีนใน GO term ที่ 2 กับผลรวมการพบยีนใน GO term ที่ 2

  ในกรณีที่พบว่ามี Go term ที่เป็นไปได้มีมากกว่าสองคำ ต้องใช้การคำนวณ P-Value เพื่อตัดสินใจเลือก Go term โดยเลือก Go term ที่มีค่า Fisher's exact test ใกล้ P-Value โดยทำการคิด P-value ดังนี้ P-Value = ผลรวมของค่า Fisher's exaction test ของแต่ละ Go term / จำนวน Go term

- ในกรณีที่พบว่ามี Go term ที่เป็นไปได้มีมากกว่าสองคำ ต้องใช้การคำนวณ P-Value เพื่อตัดสินใจเลือก Go term โดยเลือก Go term ที่มีค่า Fisher's exaction test ใกล้ P-Value โดยทำการคิด P-value ดังนี้ P-Value = ผลรวมของค่า Fisher's exaction test ของแต่ละ Go term / จำนวน Go term