

โปรแกรมวิเคราะห์ฮอมอโลยีของโปรตีนโดยใช้แผนภูมิ 2 มิติของกรดอะมิโน

นาย วิฑูร วิริยพิพัฒน์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2542

ISBN 974-333-158-1

๕ ๑๕๘ ๖๑๑๔๓

PROTEIN HOMOLOGY ANALYSIS USING 2-DIMENSIONAL AMINO
ACID PATTERN PROGRAM

Mr. Witoon Wiriyapiat

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computational Science

Department of Mathematics

Graduate School

Chulalongkorn University

Academic year 1999

ISBN 974-333-158-1

หัวข้อวิทยานิพนธ์ โปรแกรมวิเคราะห์สอมอโลยีของโปรตีนโดยใช้แผนภูมิ 2 มิติของกรดอะมิโน

โดย นายวิฑูร วิริยพิพัฒน์

ภาควิชา คณิตศาสตร์

อาจารย์ที่ปรึกษา อ.ดร.รัฐ พิชญางกูร

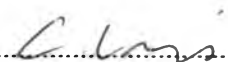
อาจารย์ที่ปรึกษาร่วม รองศาสตราจารย์.ดร.เดวิด รุฟไฟโล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

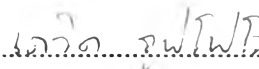



.....คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร.สุชาดา กีระนันท์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ศาสตราจารย์ ดร.ชิตชนก เหลือสินทรัพย์)

.....อาจารย์ที่ปรึกษา
(อ.ดร.รัฐ พิชญางกูร)

.....อาจารย์ที่ปรึกษาร่วม
(รองศาสตราจารย์ ดร.เดวิด รุฟไฟโล)

.....กรรมการ
(อ.ดร.เลอสร ฐนสุกาญจน์)

วิทูร วิริยพิพัฒน์ : โปรแกรมวิเคราะห์ฮอมอโลยีของโปรตีนโดยใช้แผนภูมิ 2 มิติของกรดอะมิโน (Protein Homology Analysis Using 2-Dimensional Amino Acid Pattern Program) อ. ที่ปรึกษา อ.ดร.รัฐ พิษณุางกูร อ. ที่ปรึกษาร่วม รองศาสตราจารย์ ดร.เดวิด รุฟโฟโล , 101 หน้า ISBN 974-333-158-1.

ความก้าวหน้าทางพันธุวิศวกรรม และชีวเคมี ทำให้ทราบถึงลำดับของกรดอะมิโนของโปรตีนมากมาย การทดสอบฮอมอโลยีของโปรตีนจากลำดับของกรดอะมิโนยังมีข้อจำกัดอยู่มาก วิธีการที่นิยมใช้และเป็นมาตรฐานในปัจจุบันคือ วิธี alignment ลำดับของกรดอะมิโน แต่วิธีการนี้ยังประสบปัญหาอันเกิดจาก การเว้นช่วง (gap) และการมี sequence identity ต่ำ ซึ่งทำให้ไม่สามารถทดสอบความคล้ายกันของลำดับของกรดอะมิโนได้

การวิเคราะห์ลำดับของกรดอะมิโนโดยปรับข้อมูลให้อยู่ในรูปแผนภูมิ 2 มิติเป็นเทคนิคที่มีผู้นำมาใช้ได้อย่างได้ผลในการทำนายโครงสร้างทุติยภูมิและความคล้ายของโปรตีน โครงการนี้ได้นำแนวคิดนี้มาผนวกกับข้อมูลทางเคมี และชีวเคมีของกรดอะมิโนแต่ละตัว สร้างแผนภูมิ 2 มิติใหม่เพื่อใช้ในการเปรียบเทียบฮอมอโลยีหรือความคล้ายของโปรตีน ด้วยแนวคิดที่ว่าโปรตีนที่ฮอมอโลกัสนั้นจะให้แผนภูมิ 2 มิติที่เหมือนกันหรือคล้ายกัน แล้วใช้ neural networks ช่วยในการพิจารณาความเหมือนกันหรือคล้ายกันของแผนภูมิ 2 มิติใหม่ที่สร้างขึ้น พบว่าเมื่อนำโปรตีนที่มี sequence identity ต่างๆกัน ตั้งแต่ร้อยละ 35 ถึง ร้อยละ 97 มาสร้างแผนภูมิ 2 มิติ ผลที่ได้คือ โปรตีนที่ฮอมอโลกัสนั้นก็จะให้แผนภูมิ 2 มิติที่เหมือนหรือคล้ายกัน และได้สร้างโครงข่ายประสาทเทียมพิจารณาความเหมือนหรือคล้ายกันของแผนภูมิ 2 มิติขนาดเล็ก ที่เกิดจากลำดับของกรดอะมิโน 18 ตัว จากโปรตีน 5 กลุ่ม พบว่าโครงข่ายประสาทเทียมสามารถพิจารณาเปรียบเทียบได้ถูกต้องประมาณร้อยละ 86 ถึง ร้อยละ 92

ภาควิชาคณิตศาสตร์..... ลายมือชื่อนิสิต *AA กิตติพันธ์*
สาขาวิชาวิทยาการคอมพิวเตอร์..... ลายมือชื่ออาจารย์ที่ปรึกษา *→*
ปีการศึกษา2540..... ลายมือชื่ออาจารย์ที่ปรึกษาร่วม *เดวิด รุฟโฟโล*

4072388523 : Major Computational Sci

Key word : Protein / Homology / Amino Acid / Neural Network /

WITON WIRIYAPIPAT : PROTEIN HOMOLOGY ANALYSIS USING 2-DIMENSIONAL AMINO ACID PATTERN PROGRAM. THESIS ADVISOR : DR. RATH PICHYANGKURA, Ph.D., THESIS COADVISOR : ASSOC. PROF. DAVID RUFFOLO, Ph.D. 101 pp. ISBN 974-333-158-1

Presently, an enormous number of protein sequences has been generated by advanced techniques in molecular biology and genetic engineering. However, protein homology searching by comparing amino acid sequences is complicated. The standard and most popular method is linear alignment. This method has many limitations, for example, gaps present in the protein sequence and low sequence identity could interfere with the protein alignment.

A two-dimensional representation of proteins has been successfully used in classification of secondary structure as well as similarity of proteins. This research uses this method to generate 2-dimensional patterns and incorporate data on the chemical/biochemical characteristics of each amino acid are incorporated. Computer neural networks were used to identify the similarity of the amino acid patterns. When proteins with a sequence identity of 35%-97% were used to generate the 2-dimensional patterns, similar patterns of amino acids were found. Neural networks were built to identify the similarity category of a 18-amino acid window from 5 groups of homologous proteins. The neural networks we built can identify the similarity category to an accuracy of approximately 86-92%.

ภาควิชาคณิตศาสตร์..... ลายมือชื่อนิติกร
สาขาวิชาวิทยาการคอมพิวเตอร์..... ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา2540..... ลายมือชื่ออาจารย์ที่ปรึกษาร่วมDavid Ruffolo.....



กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของอ.ดร.รัฐ พิชญานุกร อาจารย์ที่ปรึกษาวิทยานิพนธ์และ รองศาสตราจารย์ ดร. เดวิด รุฟโฟโล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่างๆ ในการทำวิจัยมาด้วยดีตลอด

เนื่องจากงานวิจัยนี้ ได้รับทุน Graduate Consortium Scholarship Computational Science & Engineering Program (GREC01-41-010) จากสำนักงานวิทยาศาสตร์และเทคโนโลยีแห่งชาติ จึงขอกราบขอบพระคุณสำนักงานวิทยาศาสตร์และเทคโนโลยีแห่งชาติมา ณ ที่นี้ด้วย

ขอขอบพระคุณที่ท่านคณะกรรมการสอบวิทยานิพนธ์ ศาสตราจารย์.ดร.ชิตชนก เหลือสินทรัพย์ และ อ.ดร.เลอสรร ธนสุกาญจน์ ที่กรุณาให้คำแนะนำ และแก้ไขวิทยานิพนธ์ฉบับนี้จนเสร็จสมบูรณ์

และขอขอบคุณนายวรวัฒน์ วรศิลป์ และนายจุมพล ตันประยูร ผู้ให้คำแนะนำในการเขียนโปรแกรม Visual C++ และ การใช้ โปรแกรม SNNS และเพื่อนๆที่ให้กำลังใจ

ท้ายนี้ ผู้วิจัยใคร่ขอขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงิน และให้กำลังใจแก่ผู้วิจัย เสมอมาจนสำเร็จการศึกษา

สารบัญ

	หน้าที่
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญตาราง.....	ฌ
สารบัญภาพ.....	ญ
บทที่	
1 บทนำ.....	1
2 ปรีทัศน์วรรณกรรม.....	4
ความรู้เบื้องต้นเกี่ยวกับโปรตีน.....	4
ประเภทของโปรตีน.....	4
กรดอะมิโนและพันธะเปปไทด์.....	6
โครงสร้างของโปรตีน.....	12
ไฮโมโลยีของโปรตีน.....	17
ฮอมอโลยี และ ความคล้าย.....	17
ความสัมพันธ์ระหว่าง โครงสร้างสามมิติ และ ฮอโมโลยี.....	17
เครื่องมือในการหาไฮโมโลยี.....	17
Hydrophobic Cluster Analysis(HCA).....	20
Local VS Global Alignment.....	21
โครงข่ายประสาทเทียม.....	22
ความรู้เกี่ยวกับระบบประสาท.....	22
แบบจำลองโครงข่ายประสาทเทียม.....	23
Single layer network ของ s neuron.....	24

บทที่	หน้าที่
McCulloch and Pitte Model (MCP).....	25
Pattern Classification.....	27
The Backpropagation	27
3 วัสดุอุปกรณ์และวิธีการ.....	31
วัสดุอุปกรณ์.....	31
วิธีการ	34
การสร้างแผนภูมิ 2 มิติของกรดอะมิโน.....	34
การสร้างโครงข่ายประสาทเทียม.....	40
4 ผลการทดลอง.....	47
แผนภูมิ 2 มิติของกรดอะมิโน.....	47
โครงข่ายประสาทเทียม.....	57
5 อภิปรายผลการทดลอง.....	64
แรงจูงใจและประโยชน์ในการศึกษา.....	64
อภิปรายผลการทดลอง.....	64
ปัญหาในการทดลอง.....	66
แนวทางการวิจัยในอนาคต.....	66
6 สรุป.....	67
รายการอ้างอิง.....	69
ภาคผนวก ก	72
ภาคผนวก ข.....	94
ประวัติผู้วิจัย.....	101

สารบัญตาราง

	หน้าที่
ตารางที่ 2.1 ตารางแสดง สูตรโครงสร้าง น้ำหนักโมเลกุล ปริมาตร และพื้นที่ผิวของกรดอะมิโนแต่ละชนิด.....	8
ตารางที่ 3.1 ตารางแสดงเครื่องคอมพิวเตอร์ และระบบปฏิบัติการ ที่ใช้ SNNS.....	33
ตารางที่ 3.2 ตารางแสดง ค่า hydrophobicity สัมพันธ์ เทียบกับ glycine.....	38
ตารางที่ 3.3 ตารางแสดงการแบ่งกลุ่มของกรดอะมิโนในการรวม cluster.....	41
ตารางที่ 4.1 ตารางแสดงลำดับของอะมิโนที่นำมาทดสอบแยกตาม sequence identity.....	48
ตารางที่ 4.2 ตารางแสดงผลรวมของความผิดพลาดจำแนกตามใน hidden layer ใน test set ที่ 1 และ test set ที่ 2.....	62
ตารางที่ 4.3 ตารางแสดงเปอร์เซ็นต์ความถูกต้องเทียบกับจำนวน hidden Layer ใน test set ที่ 1 และ test set ที่ 2.....	63

สารบัญภาพ

	หน้าที่
รูปที่ 2.1	รูปแสดงโครงสร้างของกรดอะมิโน..... 7
รูปที่ 2.2	รูปแสดงพันธะเปปไทด์..... 11
รูปที่ 2.3	รูปแสดงระดับโครงสร้างของโปรตีน..... 12
รูปที่ 2.4	รูปแสดงโครงสร้างโปรตีนแบบเกลียวแอลฟา..... 13
รูปที่ 2.5	รูปแสดงโครงสร้างโปรตีนแบบแผ่นพับ..... 14
รูปที่ 2.6	รูปแสดงการศึกษาโปรตีน..... 15
รูปที่ 2.7	รูปแสดง Dynamic programming methods ของ Needleman..... 18
รูปที่ 2.8	รูปแสดงวิธีการ BLAST..... 19
รูปที่ 2.9	รูปแสดงแผนภูมิ 2 มิติจากวิธีการ HCA..... 20
รูปที่ 2.10	รูปแสดง Local VS Global Alignment..... 21
รูปที่ 2.11	รูปแสดงระบบประสาท..... 22
รูปที่ 2.12	รูปแสดงโครงข่ายประสาทเทียมแบบ Multiple-Input..... 23
รูปที่ 2.13	รูปแสดง Layer ของโครงข่ายประสาทเทียมแบบ Single-Layer..... 24
รูปที่ 2.14	รูปแสดง Layer ของโครงข่ายประสาทเทียมแบบ Single-Layer ที่ได้มีการปรับแต่งแล้ว..... 25
รูปที่ 2.15	รูปแสดงแบบจำลองแบบ McCulloch and Pitte..... 26
รูปที่ 2.16	รูปแสดงการสร้างขอบเขตเพื่อจำแนกข้อมูลเป็นกลุ่ม..... 27
รูปที่ 2.17	รูปแสดง Feedforward network..... 27
รูปที่ 2.18	รูปแสดง Back-propagation ใน 3-Layer network..... 29
รูปที่ 3.1	รูปแสดง องค์ประกอบของ SNNS..... 32
รูปที่ 3.2	รูปแสดงอัลฟาฮีลิกที่มีขนาดกรดอะมิโน 3.6 ตัวต่อรอบ..... 35
รูปที่ 3.3	รูปแสดงการสร้าง 2-D α -Helical pattern..... 35

รูปที่ 3.4	รูปแสดงโครงสร้างของโปรตีนที่พยายามนำส่วนที่เป็น hydrophobic ไว้ด้านใน และนำส่วนที่เป็น hydrophilic ออกมาด้านนอก.....	37
รูปที่ 3.5	รูปแสดงรูปแบบของแผนภูมิ 2 มิติของกรดอะมิโนที่เกิดจากลำดับของกรดอะมิโนขนาด 18 ที่เป็นชนิดเดียวกัน.....	44
รูปที่ 3.6	รูปแสดงรูปแบบของแผนภูมิ 2 มิติของกรดอะมิโนที่เกิดจากลำดับของกรดอะมิโนขนาด 18 ต่างชนิดกัน.....	44
รูปที่ 4.1	รูปแสดงโปรแกรมสร้างแผนภูมิ 2 มิติของกรดอะมิโน.....	47
รูปที่ 4.2	แผนภูมิ 2 มิติของกรดอะมิโน ของ Phosphotransferase ของ E. coli กับ S. faecalis.....	51
รูปที่ 4.3	แผนภูมิ 2 มิติของกรดอะมิโน ของ HIV-1 Protease กับ HIV-2 Protease.....	52
รูปที่ 4.4	แผนภูมิ 2 มิติของกรดอะมิโน ของ FK-506 Binding Protein ของ H. sapiens กับ S. cerevisiae.....	53
รูปที่ 4.5	แผนภูมิ 2 มิติของกรดอะมิโน ของ Chey ของ S. typhimurium กับ E. coli.....	54
รูปที่ 4.6	แผนภูมิ 2 มิติของกรดอะมิโน ที่มีช่องว่างในบริเวณที่ไม่ใช่ cluster.....	55
รูปที่ 4.7	แผนภูมิ 2 มิติของกรดอะมิโน ที่มีช่องว่างในบริเวณที่เป็น cluster.....	56
รูปที่ 4.8	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบเมื่อจำนวน hidden layer มีขนาด 30 สำหรับ test set ที่ 1.....	54

รูปที่ 4.9	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 40 สำหรับ test set ที่ 1.....	55
รูปที่ 4.10	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 50 สำหรับ test set ที่ 1.....	55
รูปที่ 4.11	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 60 สำหรับ test set ที่ 1.....	56
รูปที่ 4.12	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 70 สำหรับ test set ที่ 1.....	56
รูปที่ 4.13	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 30 สำหรับ test set ที่ 2.....	57
รูปที่ 4.14	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 40 สำหรับ test set ที่ 2.....	57
รูปที่ 4.15	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 50 สำหรับ test set ที่ 2.....	58
รูปที่ 4.16	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 60 สำหรับ test set ที่ 2.....	58

รูปที่ 4.17	กราฟระหว่างผลต่างกำลังสองกับรูปแบบที่ใช้ทดสอบ เมื่อจำนวน hidden layer มีขนาด 70 สำหรับ test set ที่ 2.....	59
รูปที่ 4.18	กราฟระหว่างจำนวน hidden layer กับเปอร์เซ็นต์ ความถูกต้องใน test set ชุดที่ 1 และ test set ชุดที่ 2.....	60