ความสามารถในการแยกข้อมูลโดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สันสำหรับโครงข่ายประสาทแบบสไปกิง

นายอดิศักดิ์ การบรรจง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

CAPABILITY OF DATA CLASSIFICATION USING EUCLIDEAN DISTANCE AND PEARSON CORRELATION FOR SPIKING NEURAL NETWORKS

Mr.Adisak Karnbanjong

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

Department of Mathematics
Faculty of Science
Chulalongkorn University

Academic Year 2007

หัวข้อวิทยานิพนธ์	ความสามารถในการแยกข้อมูลโดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สันสำหรับโครงข่ายประสาทแบบสไปกิง
โดย	นาย อดิศักดิ์ การบรรจง
สาขาวิชา	วิทยาการคณนา
อาจารย์ที่ปรึกษา	ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ หารหนองบัว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ สุชาดา ศิริพันธุ์)

..... อาจารย์ที่ปรึกษา
(ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พีรวัฒน์ วัฒนพงศ์)

อดิศักดิ์ การบรรจง : ความสามารถในการแยกข้อมูลโดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สันสำหรับโครงข่ายประสาทแบบสไปกิง (CAPABILITY OF DATA CLASSIFICATION USING EUCLIDEAN DISTANCE AND PEARSON CORRELATION FOR SPIKING NEURAL NETWORKS) อ.ที่ปรึกษา : ศ.ดร.ชิดชนก เหลือสินทรัพย์, 46 หน้า.

ลำดับดีเอ็นเอที่สกัดมาจากเซลล์ของสิ่งมีชีวิตโดยเครื่องอ่านลำดับดีเอ็นเอ อาจให้ลำดับดีเอ็นเอที่ไม่สมบูรณ์ นั่นคือ มีลำดับดีเอ็นเอบางลำดับเป็นสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือ ตัวอย่างเช่นสัญลักษณ์ N ที่ปรากฏในลำดับดีเอ็นเอ อาจหมายถึง A หรือ C หรือ G หรือ T งานวิจัยนี้ได้แปลงปัญหาดังกล่าวเป็นปัญหาการรู้จำลำดับนิวคลีโอไทด์ที่ชัดเจนก่อนหน้าสัญลักษณ์ที่คลุมเครือ และใช้โครงข่ายประสาทแบบสไปกิง ซึ่งเป็นโครงข่ายประสาทเทียมที่มีกระบวนการทำงานคล้ายกับระบบประสาทจริงๆ ในสมองของมนุษย์ มาแก้ปัญหานี้ งานวิจัยนี้ได้เสนอวิธีเข้ารหัสข้อมูลจากรูปแบบลำดับนิวคลีโอไทด์เป็นรูปแบบลำดับเวลาที่เกิดสไปค์ เพื่อใช้เป็นข้อมูลนำเข้าสำหรับโครงข่ายประสาทแบบสไปกิง นอกจากนี้เรายังได้ศึกษาวิธีการแบ่งกลุ่มข้อมูล โดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สัน จากการทดลองพบว่าโดยเฉลี่ยการแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบสหสัมพันธ์เพียร์สันได้ให้ความถูกต้องในการทำนายมากกว่าการแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบยุคลิด นอกจากนี้เรายังพบอีกว่าจำนวนขั้นเวลาที่ใช้ขยายสัญญาณสไปค์เท่ากับ 3 ได้ให้ความถูกต้องในการทำนายมากกว่าการขยายสัญญาณสไปค์ด้วยจำนวนขั้นเท่ากับ 1 และ 5 อีกด้วย

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต อดิศักดิ์ การบรรจง
สาขาวิชา วิทยาการคณนา ลายมือชื่ออาจารย์ที่ปรึกษา C.Lursin
ปีการศึกษา 2550

4772553323 : MAJOR COMPUTATIONAL SCIENCE

KEY WORD: spiking neural networks / SpikeProp / distance measure / ambiguous nucleotide

ADISAK KARNBANJONG : CAPABILITY OF DATA CLASSIFICATION USING EUCLIDEAN DISTANCE AND PEARSON CORRELATION FOR SPIKING NEURAL NETWORKS. THESIS ADVISOR : PROF. CHIDCHANOK LURSINSAP, Ph.D., 46 pp.

DNA sequence obtained from cell organism by a DNA sequencer may give incomplete DNA sequence. That is, some order of the sequence is an ambiguous nucleotide symbol. For example, N symbol that appears in the DNA sequence may be A or C or G or T. This research has transformed the problem of recogning an ambiguous symbol at the end of a given nucleotide sequence and study the feasibility of applying spiking neurons, which is a type of neural network having similar functions to that of actual human neurons, to solve this problem. This research has proposed encoding method that encodes nucleotide sequence pattern to be the spike train pattern. The spike train pattern is used to be an input data for the spiking neural networks. In addition, we have studied the partitioning method using Euclidean distance and Pearson correlation. From the experiment, data classification using Pearson correlation is more accurate than using Euclidean distance. Moreover, the number of expanded time step of 3 is more accurate than 1 and 5.

Department Mathematics Student's signature.....
Field of study Computational Science Advisor's signature.....
Academic year 2007

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ ด้วยความช่วยเหลืออย่างดียิ่งของ ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลา เสนอแนะความคิดเห็นและให้คำแนะนำตลอดระยะเวลาการทำวิจัย รวมทั้งตรวจแก้ไขวิทยานิพนธ์ฉบับนี้อย่างละเอียด ซึ่งผู้วิจัยขอกราบขอบพระคุณไว้ ณ ที่นี้ด้วย

ขอบขอบคุณ รองศาสตราจารย์ สุชาดา ศิริพันธุ์ ที่ให้คำแนะนำ และดูแลเป็นอย่างดีตลอดระยะเวลาการทำวิจัย

ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.พีรวัฒน์ วัฒนพงศ์ ที่ให้คำแนะนำ และความรู้เป็นอย่างดี

ขอขอบคุณศูนย์วิจัย AVIC ที่เอื้อเฟื้อสถานที่ และคอมพิวเตอร์สมรรถภาพสูง ในการศึกษาและทดลองงานวิจัยนี้ และขอขอบคุณเพื่อนๆ พี่ๆ และน้องๆ ที่ให้คำแนะนำ มิตรภาพและความรู้ที่เป็นประโยชน์

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา และน้องสาว ที่ได้ให้กำลังใจและให้การสนับสนุนเสมอ

สารบัญ

หน้า

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ

ภาพประกอบ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ลำดับดีเอ็นเอมีความสำคัญต่อการทำนายดีเอ็นเอที่ไม่รู้เบส ลำดับดีเอ็นเออาจพิจารณาให้อยู่รูปของสัญญาณที่ป้อนเข้ามาตามเวลาต่างๆ การจัดการกับสัญญาณตามเวลา สามารถจัดการได้สะดวกโดยใช้แบบจำลองของเซลล์ประสาทชนิดสไปกิง งานวิจัยที่เกี่ยวกับการประยุกต์เซลล์ประสาทแบบสไปกิงที่ผ่านมา ได้กระทำกับกลุ่มข้อมูลที่มีขนาดเล็กมาก เช่น ข้อมูล XOR และข้อมูลที่ไม่ได้มาตามลำดับเวลา อย่างไรก็ตาม การประยุกต์เซลล์ประสาทแบบสไปกิง กับข้อมูลที่เข้ามาตามลำดับเวลา ได้มีการศึกษาน้อยมากในช่วงที่ผ่านมา งานวิจัยนี้จึงเป็นการศึกษาการประยุกต์ใช้เซลล์ประสาทแบบสไปกิงกับปัญหาการทำนายสัญลักษณ์ดีเอ็นเอในรูปของสัญญาณที่เข้ามาตามลำดับเวลา และศึกษาการแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สัน

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาความสามารถในการแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สัน สำหรับการทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือในลำดับดีเอ็นเอโดยใช้โครงข่ายประสาทแบบสไปกิง

1.3 ขอบเขตของการวิจัย

1.3.1 ลำดับดีเอ็นเอที่ใช้ในการวิจัยนี้ คือ แบคทีเรียอีโคไล (E.coli) 4 สายพันธุ์ คือ CFT073, K12, O157: H7 EDL933, และ O157: H7 สายพันธุ์ย่อย RIMD 0509952 ที่ได้มาจากฐานข้อมูลลำดับนิวคลีโอไทด์ EMBL ของสถาบันสารสนเทศแห่งยุโรป (European Bioinformatics Institute, EBI)

1.3.2 ศึกษาการทำนายสัญลักษณ์ที่คลุมเครือ 1 ตัว

1.3.3 ใช้รูปแบบลำดับเวลาที่เกิดสไปค์ (spike train pattern) เป็นข้อมูลนำเข้าของโครงข่ายประสาทแบบสไปกิง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้วิธีการแบ่งกลุ่มที่เหมาะสมสำหรับการทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือในลำดับดีเอ็นเอ โดยใช้โครงข่ายประสาทแบบสไปกิง

1.4.2 ได้โครงข่ายประสาทประสาทแบบสไปกิงที่สามารถทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเคลือได้

1.5 วิธีดำเนินการวิจัย

1.5.1 ศึกษาเอกสารต่างๆ ที่เกี่ยวข้อง

1.5.2 ออกแบบและพัฒนาโปรแกรมคอมพิวเตอร์เพื่อใช้ทดลอง

1.5.3 ทดสอบความถูกต้องของโปรแกรม

1.5.4 แก้ไขข้อผิดพลาด

1.5.5 ทดลองและเก็บผลการทดลอง

1.5.6 วิเคราะห์และสรุปผลการวิจัย

บทที่ 2
งานวิจัยที่เกี่ยวข้อง

2.1 งานวิจัยที่เกี่ยวข้องกับการทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือ

Kitiporn Plaimas, Chidchanok Lursinsap, และ Apichat Suratanee [1,2] เสนอวิธีการหาคุณลักษณะที่สำคัญจากข้อมูลลำดับนิวคลีโอไทด์ เพื่อใช้เป็นข้อมูลนำเข้าสำหรับโครงข่ายประสาทเทียม ในทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือในลำดับดีเอ็นเอ ซึ่งโครงข่ายประสาทเทียมที่ประกอบด้วยเซลล์ประสาทแบบซิกมอยด์เชื่อมโยงกันด้วยโครงสร้างโครงข่ายแบบแพร่ไปข้างหน้าหลายชั้น (Multilayer feed forward network) การหาคุณลักษณะที่สำคัญนั้นอาศัยตำแหน่งความสัมพันธ์ของลำดับนิวคลีโอไทด์ก่อนหน้าสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือ ทำให้ได้รูปแบบข้อมูลนำเข้าดังนี้

$$V_S = \left(\prod_{i,j \in P_A; j>i} \frac{1}{r_{A_i,A_j}} \right) \left(\prod_{i,j \in P_T; j>i} \frac{1}{r_{T_i,T_j}} \right) \left(\prod_{i,j \in P_C; j>i} \frac{1}{r_{C_i,C_j}} \right) \left(\prod_{i,j \in P_G; j>i} \frac{1}{r_{G_i,G_j}} \right)$$

โดยที่ V_S เวกเตอร์คุณลักษณะของลำดับนิวคลีโอไทด์ S

P_A, P_T, P_C และ P_G คือเซตของตำแหน่งนิวคลีโอไทด์ A, T, C และ G ในลำดับนิวคลีโอไทด์ S ตามลำดับ

$\prod$ คือการเรียงประกอบต่อกันของตำแหน่งความสัมพันธ์

นอกจากนี้ยังได้แบ่งชุดข้อมูลเป็นกลุ่มย่อยๆ ตามคุณลักษณะ แล้วให้โครงข่ายเรียนรู้ชุดข้อมูลในแต่ละกลุ่มอย่างเป็นอิสระต่อกัน และจากการทดสอบกับชุดข้อมูลทดสอบ ได้ค่าความถูกต้องเฉลี่ยเท่ากับ 93.34 %

2.2 งานวิจัยที่เกี่ยวข้องกับการแบ่งกลุ่มข้อมูลโดยใช้โครงข่ายประสาทแบบสไปกิง

Sander M. Bohte, Joost N. Nok, และ Han La Poutre [3] เสนอขั้นตอนวิธี SpikeProp ซึ่งเป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน (supervised learning algorithm) ของโครงข่ายประสาทแบบสไปกิง ขั้นตอนวิธีนี้คล้ายกับขั้นตอนวิธี error backpropagation สำหรับโครงข่ายประสาทเทียมแบบซิกมอยด์ ข้อจำกัดของ SpikeProp คือสามารถจัดการกับเวลาที่เกิดสไปค์เพียงครั้งเดียวเท่านั้นต่อหนึ่งนิวรอน เวลาที่เกิดสไปค์หลังจากนั้นจะไม่สนใจและไม่นำมาคำนวณด้วย ขั้นตอนวิธี SpikeProp ใช้วิธีการปรับค่าน้ำหนักแบบ Gradient Descent ดังสมการที่

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \qquad (2.1)$$

โดยที่ w_{ij} คือค่าน้ำหนักของการเชื่อมระหว่างเซลล์ i กับ j

E คือฟังก์ชันความคลาดเคลื่อน (Error function)

Δw_{ij} คือขนาดของการเปลี่ยนค่าน้ำหนัก

η คืออัตราการเรียนรู้ (Learning rate)

$\frac{\partial E}{\partial w_{ij}}$ คือความคลาดเคลื่อนเกรเดียนท์ (Error gradient)

และในงานวิจัยนั้นได้นำขั้นตอนวิธี SpikeProp ไปแก้ปัญหาแบ่งกลุ่มข้อมูล XOR, Iris, Wisconsin breast-cancer และ Statlog Landsat

Jianguo Xin, และ Mark J. Embrechts [4] เสนอวิธีการปรับค่าน้ำหนักโดยการเพิ่มค่าโมเมนตัม (momentum term) ซึ่งเป็นขนาดของการเปลี่ยนค่าน้ำหนักในรอบก่อนหน้า ในสมการการปรับค่าน้ำหนัก และมีการปรับอัตราการเรียนรู้ในแต่ละรอบของการปรับค่าน้ำหนัก ดังสมการที่ 2.2

$$\Delta w_{ij}^{(t)} = -\eta_t \frac{\partial E}{\partial w_{ij}^{(t)}} + \alpha \Delta w_{ij}^{(t-1)} \qquad (2.2)$$

โดยที่ η_t คือ อัตราการเรียนรู้ในรอบที่ t

α คือ ค่าคงที่โมเมนตัม

$\alpha \Delta w_{ij}^{(t-1)}$ คือ ค่าโมเมนตัมในรอบที่ $t-1$

และนำไปใช้แก้ปัญหาการแบ่งกลุ่มข้อมูล Iris

Benjamin Schrauwen, และ Jan Van Campenhout [5] เสนอกฎการเรียนรู้โดยการปรับพารามิเตอร์อื่นๆ ได้แก่ เวลาหน่วง (synaptic delays), ค่าคงที่ทางเวลา (time constants) และ thresholds ของนิวรอน โดยใช้วิธีการปรับพารามิเตอร์แบบ Gradient Descent ดังสมการที่ 2.3, 2.4 และ 2.5 ตามลำดับ

$$\Delta d_{ij} = -\eta_d \frac{\partial E}{\partial d_{ij}} \qquad (2.3)$$

$$\Delta\tau_{ij} = -\eta_{\tau}\frac{\partial E}{\partial \tau_{ij}} \quad (2.4)$$

$$\Delta\theta_{j} = -\eta_{\theta}\frac{\partial E}{\partial \theta_{j}} \quad (2.5)$$

โดยที่ η_d , η_τ และ η_θ คือ อัตราการเรียนรู้สำหรับเวลาหน่วง, ค่าคงที่ทางเวลา และ threshold ของนิวรอน ตามลำดับ

Δd_{ij} , $\Delta\tau_{ij}$ และ $\Delta\theta_j$ คือ ขนาดของการปรับเวลาหน่วง, ค่าคงที่ทางเวลา และ threshold ของนิวรอน ตามลำดับ

d_{ij} คือ เวลาหน่วงระหว่างนิวรอน i กับ j

τ_{ij} คือ ค่าคงที่ทางเวลาระหว่างนิวรอน i กับ j

θ_j คือ threshold ของนิวรอน j

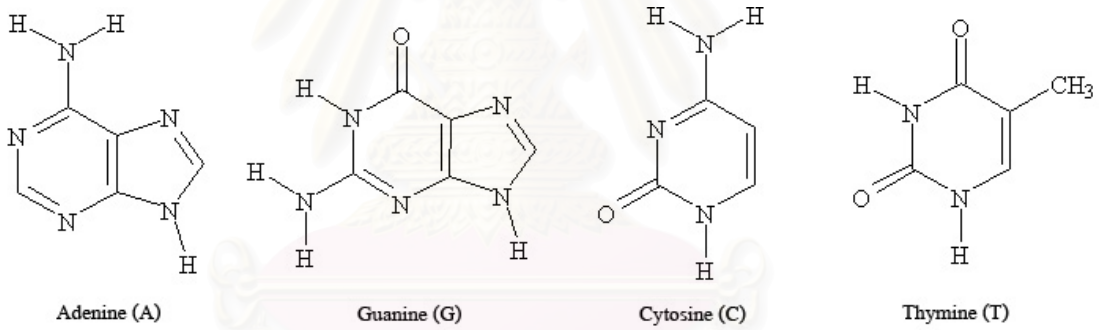
และนำไปใช้แก้ปัญหาแบ่งกลุ่มข้อมูล XOR

บทที่ 3
ความรู้พื้นฐาน

3.1 ความรู้เบื้องต้นเกี่ยวกับดีเอ็นเอ

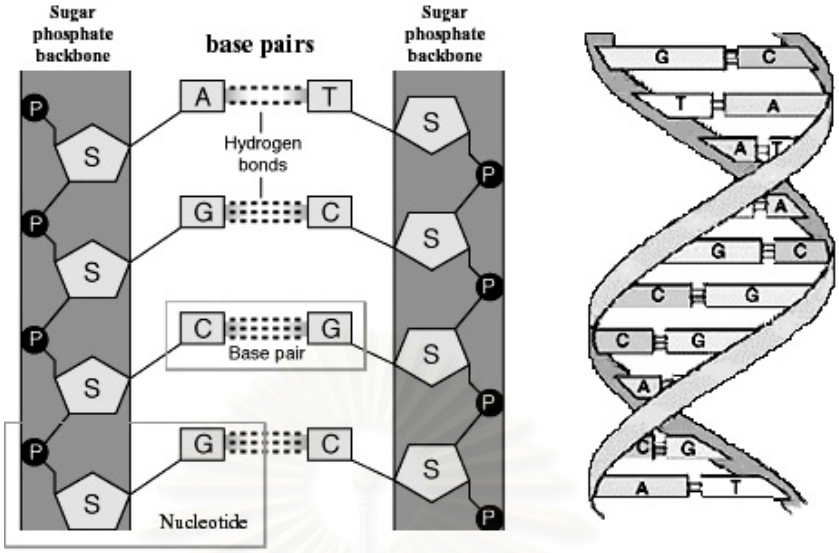
3.1.1 องค์ประกอบและโครงสร้างของดีเอ็นเอ

ดีเอ็นเอ หรือ กรดดีออกซิไรโบนิวคลีอิก (DeoxyriboNucleic Acid, DNA) เป็นสารพันธุกรรมชนิดหนึ่งที่อยู่ในเซลล์ของสิ่งมีชีวิต ประกอบด้วย หน่วยย่อยที่เรียกว่า นิวคลีโอไทด์ (nucleotide) ซึ่งประกอบด้วย 3 ส่วน คือ น้ำตาล หมู่ฟอสเฟต และเบส 4 ชนิด คือ อะดีนีน (Adenine หรือ A), ไซโทซีน (Cytosine หรือ C), กัวนีน (Guanine หรือ G) และไธมีน (Thymine หรือ T) นอกจากนี้ยังจำแนกเบสตามโครงสร้างทางเคมีออกเป็น 2 กลุ่ม คือ พิวรีน (purine) และ พิริมิดีน (pyrimidine) โดยที่ พิวรีนมีโครงสร้างหลักที่ประกอบด้วย วงแหวน 2 วง ได้แก่ เบส A และ G ส่วนพิริมิดีนมีโครงสร้างหลักที่ประกอบด้วย วงแหวน 1 วง ได้แก่ เบส C และ T



รูปที่ 3.1 โครงสร้างทางเคมีของเบส อะดีนีน, ไซโทซีน, กัวนีน และไธมีน [2]

ในปี ค.ศ.1953 นักวิทยาศาสตร์ชาวอเมริกันชื่อ เจมส์ วัตสัน (James Watson) และนักวิทยาศาสตร์ชาวอังกฤษชื่อ ฟรานซิส คลิก (Francis Crick) พบโครงสร้างดีเอ็นเอ มีลักษณะเป็นเกลียวคู่บิดพันกันคล้ายบันไดเวียนขวา (right handed double helix) ซึ่งเกลียวคู่นี้เป็นพอลิเมอร์นิวคลีโอไทด์ 2 สาย ที่มีทิศทางสวนทางกัน โดยมีน้ำตาล และหมู่ฟอสเฟตเป็นแกนของเกลียว (DNA backbone) และมีเบสที่จับคู่กันอยู่ภายในเกลียว นั่นคือ เบส A จับคู่กับ T และเบส G จับคู่กับ C ดังรูปที่ 3.2



รูปที่ 3.2 โครงสร้างดีเอ็นเอ [2]

3.1.2 การหาลำดับดีเอ็นเอ (DNA Sequencing)

การหาลำดับดีเอ็นเอ คือการหาลำดับการเรียงลำดับเบส A, C, G และ T ของดีเอ็นเอที่อยู่ในเซลล์ของสิ่งมีชีวิต วิธีการหาลำดับดีเอ็นเอมี 2 วิธี คือ การหาลำดับเบสโดยวิธีทางเคมี (Chemical degradation method) ซึ่งถูกพัฒนาขึ้นในปี 1977 โดย Maxam และ Gilbert และการหาลำดับเบสโดยวิธีการใช้เอนไซม์ (Enzymatic method) ซึ่งถูกพัฒนาขึ้นในปี 1977 โดย Sanger ซึ่งทั้งสองวิธีใช้เวลานานในการหาลำดับดีเอ็นเอทั้งหมด เทคโนโลยีทางด้านคอมพิวเตอร์ได้เข้ามามีบทบาทในการหาลำดับเบสของดีเอ็นเอได้รวดเร็วมากขึ้น ซึ่งเรียกเครื่องคอมพิวเตอร์ที่ใช้หาลำดับเบสของดีเอ็นเอว่า เครื่องอ่านลำดับดีเอ็นเอ (DNA sequencer) และในปัจจุบันได้มีการพัฒนาเครื่องอ่านลำดับเบสของดีเอ็นเอขึ้นมามากมาย ทำให้สามารถอ่านลำดับเบสได้ครั้งละมากๆ และใช้เวลาน้อยลง

ตารางที่ 3.1 สัญลักษณ์นิวคลีโอไทด์ที่ได้จากเครื่องอ่านลำดับดีเอ็นเอ (DNA sequencer)

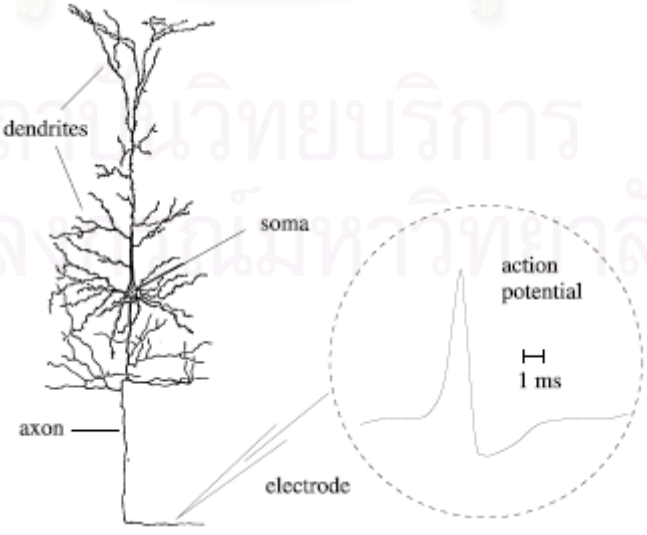
สัญลักษณ์	ความหมาย	คำอธิบาย
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A หรือ G	puRine

Y	C หรือ T	pYrimidine
M	A หรือ C	aMino
K	G หรือ T	Keto
S	C หรือ G	Strong interactions (3 h bonds)
W	A หรือ T	Weak interactions (2 h bonds)
H	A หรือ C หรือ T	H follows G in alphabet
B	C หรือ G หรือ T	B follows A in alphabet
V	A หรือ C หรือ G	V follows T in alphabet
D	A หรือ G หรือ T	D follows C in alphabet
N	A หรือ C หรือ G หรือ T	Any base

3.2 โครงข่ายประสาทแบบสไปกิง (Spiking Neural Networks, SNNs)

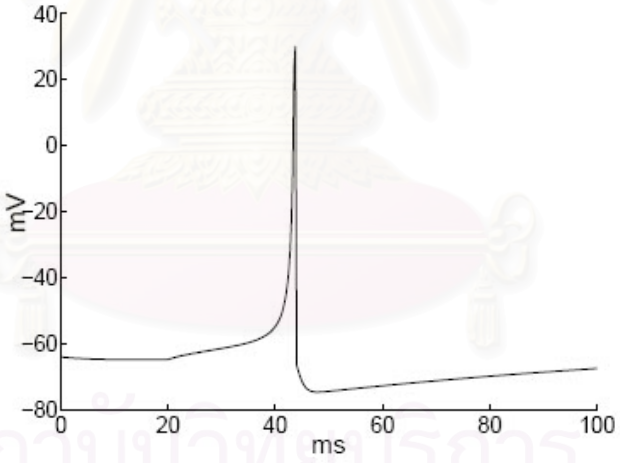
3.2.1 เซลล์ประสาททางชีววิทยา (Biological neuron)

ระบบประสาท (Nervous System) ในสมองของมนุษย์ ประกอบด้วยเซลล์ประสาท (nerve cell) หรือ นิวรอน (neuron) ประมาณ 10^{11} เซลล์ มีการเชื่อมต่อกันในรูปแบบที่ซับซ้อน เซลล์ประสาทหนึ่งมีการเชื่อมต่อกับเซลล์ประสาทอื่นๆ ประมาณ 10,000 เซลล์ ส่วนประกอบหลักของเซลล์ประสาทมี 3 ส่วน คือ ตัวเซลล์ (cell body หรือ soma), แอกซอน (axon) และเดนไดรต์ (dendrite) ดังรูปที่ 3.3



รูปที่ 3.3 องค์ประกอบของเซลล์ประสาท [9]

การสื่อสารระหว่างเซลล์ประสาทที่ส่งสัญญาณกับเซลล์ประสาทที่รับสัญญาณ เซลล์ประสาทที่ส่งสัญญาณ เรียกว่า เซลล์ประสาทก่อนไซแนปส์ (Pre-Synaptic Neuron) และเซลล์ประสาทที่รับสัญญาณ เรียกว่า เซลล์ประสาทหลังไซแนปส์ (Post-Synaptic Neuron) เซลล์ประสาททุกเซลล์จะมีความต่างศักย์ไฟฟ้าระหว่างภายในและภายนอกเซลล์ หรือ ความต่างศักย์ของเยื่อหุ้มเซลล์ หรือ เมมเบรนโพเทนเชียล (Membrane Potential, MP) ในสภาวะที่เซลล์ไม่มีสัญญาณใดๆ มากระตุ้น เมมเบรนโพเทนเชียลของเซลล์มีค่าประมาณ -70 mV เรียกว่า ความต่างศักย์ไฟฟ้าขณะพัก (Resting Potential, RP) เมื่อใดก็ตามที่มีสัญญาณมากระตุ้นทำให้เมมเบรนโพเทนเชียลของเซลล์มีการเปลี่ยนแปลง และถ้าสัญญาณที่มากระตุ้นนั้นแรงพอ ทำให้เมมเบรนโพเทนเชียลมีการเปลี่ยนแปลงเกินระดับที่ทำให้เกิดการกระตุ้น (threshold voltage) ซึ่งมีค่าประมาณ -55 mV เมมเบรนโพเทนเชียลของเซลล์ประสาทนั้นจะมีการเปลี่ยนแปลงอย่างรวดเร็ว เมมเบรนโพเทนเชียลของเซลล์ประสาทที่เปลี่ยนแปลงอย่างรวดเร็ว เรียกว่า แอกชันโพเทนเชียล (action potential) หรือ สไปค์ (spike) แอกชันโพเทนเชียลจะไหลไปตามแอกซอนเพื่อส่งสัญญาณไปให้เซลล์ประสาทอื่นๆ



รูปที่ 3.4 ลักษณะของแอกชันโพเทนเชียล [11]

จุดประสานประสาท (Synapse) คือ ช่องว่างระหว่างปลายแอกซอนของเซลล์ประสาทก่อนไซแนปส์ กับเดนไดรต์ของเซลล์ประสาทหลังไซแนปส์ เมื่อมีแอกชันโพเทนเชียลมาที่จุดประสานประสาทนี้ ทำให้มีการถ่ายโอนสารสื่อประสาทจากปลายแอกซอนของเซลล์ก่อนไซแนปส์ ไปยังเดนไดรต์ของเซลล์ประสาทหลังไซแนปส์ ทำให้มีการเปลี่ยนแปลงทางเคมีขึ้น เรียกว่า Post Synaptic Potential (PSP)

3.2.2 เซลล์ประสาทเทียม (Artificial neuron)

เซลล์ประสาทเทียม คือแบบจำลองที่สร้างขึ้นมา เพื่อเลียนแบบการทำงานของเซลล์ประสาทจริง (Biological neuron หรือ Real neuron) Wolgang Maass [12] ได้แบ่งแบบจำลองของเซลล์ประสาทเทียมออกเป็น 3 รุ่น ตามลักษณะการทำงาน คือ McCulloch-Pitts neuron model, Sigmoid neuron model และ Spiking neuron model แบบจำลองรุ่นที่ 1 ใช้ฟังก์ชันกระตุ้นแบบ Threshold function ซึ่งให้ผลลัพธ์เป็น 0 และ 1 ผลลัพธ์ที่ได้นี้เป็นลักษณะที่สำคัญของแบบจำลองรุ่นนี้ ส่วนแบบจำลองรุ่นที่ 2 ได้ประยุกต์ฟังก์ชันกระตุ้นให้ได้ผลลัพธ์ที่ต่อเนื่อง ฟังก์ชันกระตุ้นในแบบจำลองนี้ ได้แก่ Sigmoid function, Linear Saturated function เป็นต้น แบบจำลองรุ่นที่ 1 และ 2 เรียกได้ว่าเป็นแบบจำลองเซลล์ประสาทแบบเดิม (Traditional neuron model) โครงข่ายประสาทเทียมที่ใช้แบบจำลองเซลล์ประสาทแบบเดิมนี้ได้รับความนิยมและใช้งานกันอย่างแพร่หลายในปัจจุบัน เนื่องจากความสามารถในการแก้ปัญหาในหลายๆ ด้าน ได้แก่ การจำแนกรูปแบบ (Pattern Classification), การประมาณฟังก์ชัน (Function Approximation) เป็นต้น

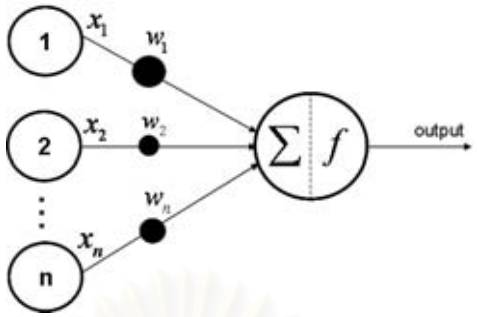
แบบจำลองเซลล์ประสาทแบบสไปกิง (Spiking neuron model) ถือได้ว่าเป็นแบบจำลองรุ่นที่ 3 และเป็นแบบจำลองที่มีความคล้ายคลึงกับเซลล์ประสาทจริง เนื่องจากแบบจำลองนี้ใช้ Spike (เปรียบเสมือนกระแสประสาทของเซลล์ประสาทจริง) ในการสื่อสารระหว่างเซลล์ประสาทเทียม นอกจากมีประสิทธิภาพมากกว่าแบบจำลองเซลล์ประสาทแบบเดิมแล้ว ยังสามารถจำลองการทำงานแบบจำลองเซลล์ประสาทแบบเดิมได้อีกด้วย

ตารางที่ 3.2 ข้อมูลในการสื่อสารของแบบจำลองของเซลล์ประสาท

แบบจำลอง (Model)	ข้อมูลที่สำคัญในการสื่อสาร (Information)
Traditional neuron	จำนวนจริง
Spiking neuron	สไปค์

หลักการทำงานของเซลล์ประสาทเทียม

ลักษณะการทำงานของเซลล์ประสาทเทียมแบบเดิม (Traditional neuron) มีดังนี้ เซลล์ที่รับสัญญาณจะรวมสัญญาณ (x_i) ที่ถูกปรับค่าน้ำหนัก (w_i) จากเซลล์อื่นๆ (i) แล้วนำผลรวมที่ได้ผ่านฟังก์ชันกระตุ้น (f) จะได้ผลลัพธ์ของเซลล์นั้นออกมา ดังรูปที่ 3.5



รูปที่ 3.5 ลักษณะการทำงานของเซลล์ประสาทเทียมแบบเดิม

และสามารถเขียนได้ด้วยสมการ 3.1

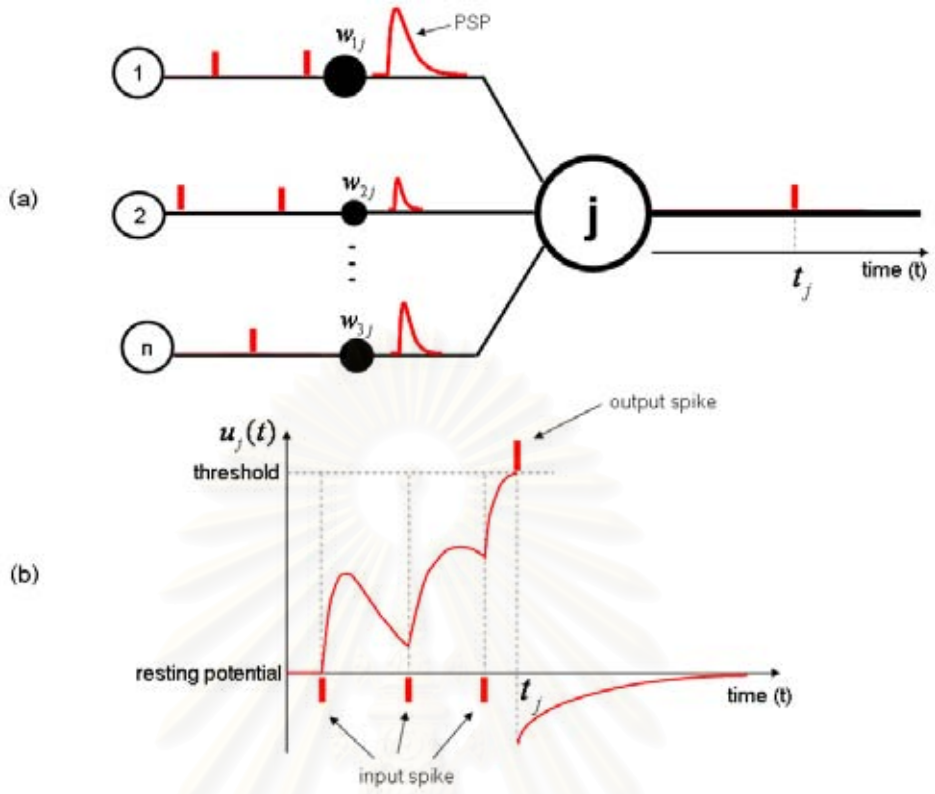
$$output = f(\sum_{i=1}^{n} x_i w_i) \qquad (3.1)$$

โดยที่ x_i คือ ข้อมูลที่ส่งมาจากเซลล์ประสาท i

w_i คือ ค่าน้ำหนักของการเชื่อมต่อ ที่สัมพันธ์กับเซลล์ประสาท i

f คือ ฟังก์ชันกระตุ้น

ส่วนเซลล์ประสาทแบบสไปกิงมีลักษณะการทำงานดังนี้ สถานะของเซลล์ประสาทแบบสไปกิง (เปรียบเสมือน Membrane Potential (MP) ของเซลล์ประสาทจริง) ในภาวะที่ไม่มีสัญญาณใดๆ มากระตุ้น สถานะของเซลล์มีค่าอยู่ระดับปกติ (เปรียบเสมือนระดับ Resting Potential ของเซลล์ประสาทจริง) เมื่อเซลล์ได้รับสัญญาณกระตุ้นจากเซลล์อื่นๆ เซลล์นั้นจะรวมผลการตอบสนองต่อสิ่งกระตุ้นที่ถูกปรับค่าน้ำหนัก (เปรียบเสมือน Post Synaptic Potential (PSP) ของเซลล์ประสาทจริง) และเมื่อใดก็ตามที่สถานะของเซลล์มีค่าเกินระดับ Threshold จะทำให้เซลล์นั้นส่งสัญญาณกระตุ้นออกไป หลังจากที่เซลล์ส่งสัญญาณออกไป จะทำให้สถานะของเซลล์ลดระดับลงต่ำกว่าระดับปกติ และจะค่อยๆ เพิ่มขึ้น จนอยู่ในระดับปกติ ดังแสดงดังรูปที่ 3.6



รูปที่ 3.6 ลักษณะการทำงานของเซลล์ประสาทแบบสไปกิง

3.2.3 เซลล์ประสาทแบบสไปกิง (Spiking neuron)

Hodgkin และ Huxley ได้ศึกษาถึงพฤติกรรมของ Membrane Potential (MP) ของเซลล์ประสาทของปลาหมึกยักษ์ และได้สร้างแบบจำลองขึ้น เรียกว่า Hodgkin-Huxley Model ซึ่งเป็นแบบจำลองที่เหมือนเซลล์ประสาทจริง แต่มีความซับซ้อนมาก และยากต่อการนำไปใช้งาน นักวิจัยหลายคนได้สร้างแบบจำลองขึ้นมา โดยการลดรูป Hodgkin-Huxley Model ให้สามารถนำไปใช้งานได้ง่ายขึ้น ได้แก่ FitzHugh-Nagumo Model, Integrate-and-Fire Model (I&F) เป็นต้น นอกจากนี้ยังมีแบบจำลองอีกแบบหนึ่ง คือ Spike Response Model (SRM) ซึ่งถูกเสนอโดย Wulfram Gerstner [9,10] เป็นแบบจำลองที่เข้าใจได้ง่าย และนำไปประยุกต์ใช้กันอย่างแพร่หลายในงานหลายด้าน ได้แก่ การจำแนกรูปแบบ (Pattern Classification), การประมาณค่าฟังก์ชัน (Function Approximation) และการจัดกลุ่มข้อมูล (Data Clustering) เป็นต้น

3.2.4 แบบจำลองการตอบสนองต่อสไปค์ (Spike Response Model)

สถานะของเซลล์ประสาท j ณ เวลา t ใดๆ กำหนดโดยตัวแปร $u_j(t)$ ในภาวะที่ไม่มีสัญญาณใดๆ มากระตุ้น สถานะของเซลล์มีค่าอยู่ระดับปกติ เมื่อเซลล์ได้รับสัญญาณสไปค์ (สัญญาณกระตุ้น) จากเซลล์อื่นๆ เซลล์ j จะรวมผลการตอบสนองต่อสไปค์ที่ถูกปรับค่าน้ำหนัก เมื่อใดก็ตามที่ u_j มีค่าเกินระดับ threshold (θ) ทำให้เซลล์นั้นส่งสัญญาณออกไป หลังจากนั้นสถานะของเซลล์ลดระดับลงต่ำกว่าระดับปกติทันที และจะค่อยๆ เพิ่มขึ้น จนอยู่ในระดับปกติ และถ้ามีสัญญาณมากระตุ้นอีก การทำงานจะเหมือนเดิม ซึ่งสามารถเขียนลำดับเวลาที่เซลล์ j ส่งสัญญาณ หรือ ลำดับเวลาที่เกิดสไปค์ของเซลล์ j ได้ดังนี้

$$F_j = \left\{ t_j^{(1)}, t_j^{(2)}, ..., t_j^{(n)} \right\} \tag{3.2}$$

โดยที่ $t_j^{(f)}$ คือ เวลาที่เซลล์ประสาท j ส่งสัญญาณสไปค์ลำดับที่ f

n คือ จำนวนครั้งที่นิวรอนส่งสัญญาณสไปค์ในช่วงเวลาหนึ่ง

จากพฤติกรรมดังกล่าว สถานะของเซลล์ j ขึ้นอยู่กับอิทธิพลของ 2 สิ่ง คือ เซลล์ j ได้รับการกระตุ้นจากเซลล์อื่นๆ และผลจากการที่เซลล์ j ส่งสัญญาณสไปค์ออกไป ทำให้สถานะของเซลล์ j เขียนได้ดังนี้

$$u_j(t) = \sum_{t_j^{(f)} \in F_j} \eta(t - t_j^{(f)}) + \sum_{i \in \Gamma_j} \sum_{t_i^{(g)} \in F_i} w_{ji} \varepsilon(t - t_i^{(g)}) \tag{3.3}$$

โดยที่ ε คือ ฟังก์ชันการตอบสนองต่อสไปค์ (Spike Response Function) นิยามโดย

$$\varepsilon_{ij}(s) = \left[\exp\left(-\frac{s}{\tau_m} \right) - \exp\left(-\frac{s}{\tau_s} \right) \right] H(s) \tag{3.4}$$

η คือ ฟังก์ชันที่อธิบายผลจากการที่เซลล์ได้ส่งสัญญาณออกไป (Refractoriness function) นิยามโดย

$$\eta(s) = -\theta \exp\left(-\frac{s}{\tau_r} \right) H(s) \tag{3.5}$$

H คือ Heviside Step function นิยามได้ดังนี้

$$H(s) = \begin{cases} 1 & , s > 0 \\ 0 & , s \le 0 \end{cases} \tag{3.6}$$

w_{ji} คือ ค่าน้ำหนักของการเชื่อมต่อระหว่างเซลล์ i กับ j

Γ_j คือ เซตของเซลล์ที่ส่งสัญญาณมาให้เซลล์ j

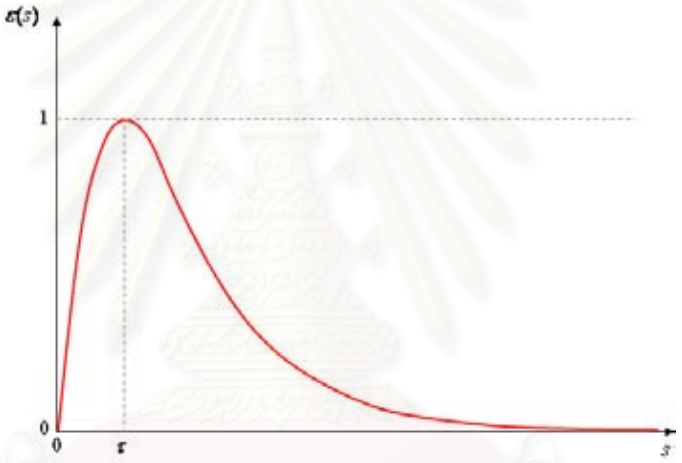
τ_m, τ_s และ τ_r คือ ค่าคงที่ทางเวลา (time constants)

นอกจาก ฟังก์ชันการตอบสนองต่อสไปค์ (Spike Response Function) ที่ได้นิยามไว้ในสมการ 3.4 แล้ว ยังมีอีกรูปแบบหนึ่งที่มีการใช้กันอย่างแพร่หลาย และใช้เป็นรูปแบบฟังก์ชันการตอบสนองต่อสไปค์ในงานวิจัยนี้ด้วยเช่นกัน คือ

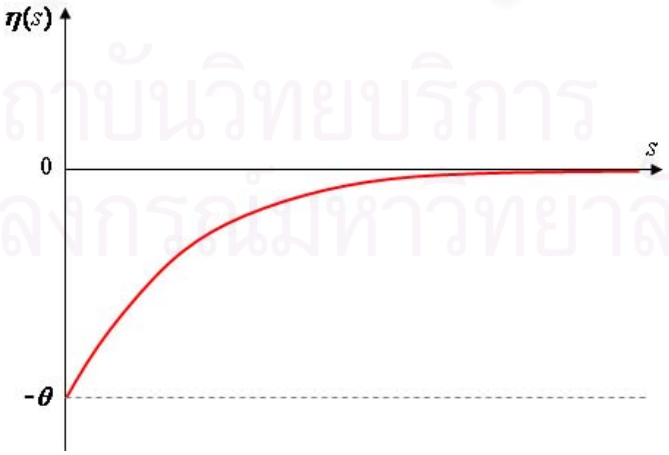
$$\varepsilon_{ij}(s) = \exp\left(-\frac{s}{\tau}\right) H(s) \qquad (3.7)$$

โดยที่ τ คือ ค่าคงที่ทางเวลา (time constant)

ลักษณะของฟังก์ชันการตอบสนองต่อสไปค์ที่ได้นิยามในสมการที่ 3.7 แสดงได้ดังรูปที่ 3.7



รูปที่ 3.7 ลักษณะของฟังก์ชันการตอบสนองต่อสไปค์



รูปที่ 3.8 ลักษณะของ Refractoriness function

3.2.5 ขั้นตอนวิธีการเรียนรู้ (Learning Algorithms)

ในหัวข้อนี้ กล่าวถึง ขั้นตอนวิธี SpikeProp ที่สามารถจัดการกับเวลาที่เกิดสไปค์ได้หลายครั้งต่อหนึ่งนิวรอน [6, 7] และกฎการเรียนรู้แบบ RProp [8] สำหรับการปรับค่าน้ำหนัก

ในงานวิจัยนี้ ใช้โครงข่ายประสาทแบบสไปกิงแบบชั้นเดียว การเชื่อมต่อระหว่างนิวรอนในชั้นนำเข้าและชั้นนำออกมีการเชื่อมต่อแบบหลายเส้น ซึ่งได้อธิบายไว้ในหัวข้อที่ 4.4.1 และเนื่องจากงานวิจัยนี้สนใจสัญญาณสไปค์ที่เกิดขึ้นครั้งแรกของนิวรอนในชั้นนำออกเท่านั้น ดังนั้น สถานะของนิวรอน j ในชั้นนำออก นิยามได้ดังนี้

$$u_j(t) = \sum_{i \in \Gamma_j} \sum_{t_i^{(g)} \in F_i} \sum_k w_{ij}^k \varepsilon(t - t_i^{(g)} - d_{ij}^k) \tag{3.8}$$

โดยที่ w_{ji}^k คือ ค่าน้ำหนักที่ k ของการเชื่อมต่อระหว่างเซลล์ i กับ j

ε คือ ฟังก์ชันการตอบสนองต่อสไปค์ ตามสมการที่ 3.7

$t_i^{(g)}$ คือ เวลาที่นิวรอน i ส่งสัญญาณสไปค์ลำดับที่ g

d_{ij}^k คือ เวลาหน่วงของการเชื่อมโยงที่ k ระหว่างนิวรอน i กับ j

Γ_j คือ เซตของนิวรอนที่ส่งสัญญาณให้นิวรอน j

F_i คือ เซตของเวลาที่เกิดสไปค์ของนิวรอน i

และกำหนดฟังก์ชันความคลาดเคลื่อน (Error function) ดังนี้

$$E = \frac{1}{2} \sum_{p=1}^{\mathrm{P}} \sum_{j \in O} (t_j^{(1)} - \hat{t}_j^{(1)})^2 \tag{3.9}$$

โดยที่ P คือ จำนวนรูปแบบทั้งหมด

O คือ เซตของนิวรอนในชั้นนำออก

$t_j^{(1)}$ คือ เวลาที่เกิดสไปค์จริงของนิวรอน j ลำดับที่ 1 ของรูปแบบ p

$\hat{t}_j^{(1)}$ คือ เวลาที่เกิดสไปค์เป้าหมายของนิวรอน j ลำดับที่ 1 ของรูปแบบ p

การหากฎการเรียนรู้โดยวิธี Gradient Descent สำหรับเซลล์ประสาทแบบสไปกิงที่ส่งสัญญาณหลายครั้ง

จากวิธีปรับค่าน้ำหนักแบบ Gradient Descent

$$\Delta w_{ij}^k = -\eta \frac{\partial E}{\partial w_{ij}^k} \tag{3.10}$$

จากกฎลูกโซ่ จะได้ว่า

$$\frac{\partial E}{\partial w_{ij}^{k}} = \sum_{p=1}^{\mathrm{P}} \frac{\partial E}{\partial t_j^{(1)}} \frac{\partial t_j^{(1)}}{\partial w_{ij}^{k}} \tag{3.11}$$

พจน์ที่ 1 ของสมการที่ 3.11 สามารถหาได้ดังนี้

$$\frac{\partial E}{\partial t_j^{(1)}} = t_j^{(1)} - \hat{t}_j^{(1)} \tag{3.12}$$

จากการศึกษาของ Sander M. Bohte [3], พจน์ที่ 2 ของสมการ 3.11 สามารถหาได้ดังนี้

$$\frac{\partial t_j^{(1)}}{\partial w_{ij}^{k}} = \frac{-\dfrac{\partial u_j(t_j^{(1)})}{\partial w_{ij}^{k}}}{\dfrac{\partial u_j(t_j^{(1)})}{\partial t_j^{(1)}}} = \frac{-\sum\limits_{t_i^g \in F_i} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)}{\sum\limits_{i \in \Gamma_j} \sum\limits_{t_i^g \in F_i} \sum\limits_{k} w_{ij}^{k} \dfrac{\partial}{\partial t_j^{(1)}} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)} \tag{3.13}$$

แทนสมการ 3.12 และ 3.13 ในสมการ 3.11

$$\frac{\partial E}{\partial w_{ij}^{k}} = \sum_{p=1}^{\mathrm{P}} \left(t_j^{(1)} - \hat{t}_j^{(1)} \right) \left(\frac{-\sum\limits_{t_i^g \in F_i} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)}{\sum\limits_{i \in \Gamma_j} \sum\limits_{t_i^g \in F_i} \sum\limits_{k} w_{ij}^{k} \dfrac{\partial}{\partial t_j^{(1)}} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)} \right) \tag{3.14}$$

ดังนั้น

$$\Delta w_{ij}^{k} = -\eta \sum_{p=1}^{\mathrm{P}} \left(t_j^{(1)} - \hat{t}_j^{(1)} \right) \left(\frac{-\sum\limits_{t_i^{(g)} \in F_i} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)}{\sum\limits_{i \in \Gamma_j} \sum\limits_{t_i^g \in F_i} \sum\limits_{k} w_{ij}^{k} \dfrac{\partial}{\partial t_j^{(1)}} \varepsilon(t_j^{(1)} - t_i^{(g)} - d^k)} \right) \tag{3.15}$$

กฎการเรียนรู้แบบ RProp

RProp หรือ Resilient Propagation เป็นวิธีการปรับค่าน้ำหนักที่ถูกพัฒนาขึ้นเพื่อเพิ่มความเร็วในการเรียนรู้ของโครงข่ายประสาทเทียม วิธีการปรับค่าน้ำหนักได้อาศัยทิศทางของความคลาดเคลื่อนเกรเดียนท์ และนำมาประยุกต์ใช้กับขั้นตอนวิธี SpikeProp เพื่อสอนโครงข่ายประสาทแบบสไปกิง [8] ดังสมการที่ 3.16 และ 3.17

$$\Delta_{ij}^{k}(epoch)=\begin{cases}\eta^{+}\Delta_{ij}^{k}(epoch-1) & ,if\ \frac{\partial E}{\partial w_{ij}^{k}}(epoch-1)\cdot\frac{\partial E}{\partial w_{ij}^{k}}(epoch)>0\\ \eta^{-}\Delta_{ij}^{k}(epoch-1) & ,if\ \frac{\partial E}{\partial w_{ij}^{k}}(epoch-1)\cdot\frac{\partial E}{\partial w_{ij}^{k}}(epoch)<0\\ \Delta_{ij}^{k}(epoch-1) & ,otherwise\end{cases} \quad (3.16)$$

และ

$$\Delta w_{ij}^{k}(epoch)=\begin{cases}-\Delta_{ij}^{k}(epoch) & ,if\ \frac{\partial E}{\partial w_{ij}^{k}}(epoch)>0\\ +\Delta_{ij}^{k}(epoch) & ,if\ \frac{\partial E}{\partial w_{ij}^{k}}(epoch)<0\\ 0 & ,otherwise\end{cases} \quad (3.17)$$

โดยที่

$\frac{\partial E}{\partial w_{ij}^{k}}(epoch)$ คือ ความคลาดเคลื่อนเกรเดียนท์ในรอบที่ $epoch$

η^{+} และ η^{-} คือ อัตราการเรียนรู้ โดยทั่วไปมีค่าดังนี้ $0<\eta^{-}<1<\eta^{+}$

ขั้นตอนการเรียนรู้ (Learning Algorithm)

1 กำหนด $\frac{\partial E}{\partial w_{ij}^{k}}(0)=0$ และ $\Delta_{ij}^{k}(0)=0.07$ สำหรับทุก i, j และ k

2 กำหนด $\eta^{+}=1.2$ และ $\eta^{-}=0.5$

3 กำหนดความคลาดเคลื่อนที่ยอมรับได้ (Tolerant) และจำนวนรอบมากที่สุดในการเรียนรู้ (MaxLoop)

4 สุ่มค่าน้ำหนัก w_{ij}^{k} สำหรับทุก i, j และ k

5 กำหนด เวลาหน่วง d_{ij}^{k} สำหรับทุก i, j และ k

6 คำนวณ output spike time สำหรับทุกๆ pattern

7 คำนวณความคลาดเคลื่อน (Error) จากสมการ 3.9

8 กำหนด $epoch = 1$

9 ทำข้อ 9.1- 9.7 จนกระทั่ง Error<Tolerant หรือ Epoch>MaxLoop

9.1 คำนวณ $\frac{\partial E}{\partial w_{ij}^k}(epoch)$ จากสมการ 3.14

9.2 คำนวณ $\Delta_{ij}^k(epoch)$ จากสมการ 3.17

9.3 ปรับค่าน้ำหนักตามสมการ 3.16

9.4 คำนวณ output spike time สำหรับทุก pattern

9.5 คำนวณความคลาดเคลื่อน (Error) จากสมการ 3.9

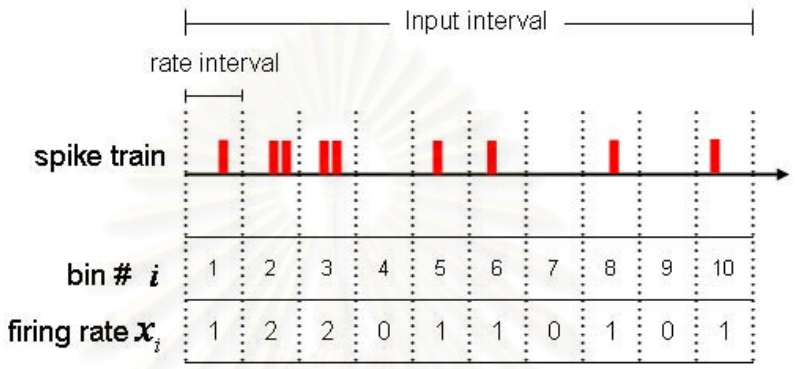
9.6 ปรับปรุงค่า $\Delta_{ij}^k(epoch) = \Delta_{ij}^k(epoch-1)$

และ $\frac{\partial E}{\partial w_{ij}^k}(epoch) = \frac{\partial E}{\partial w_{ij}^k}(epoch-1)$

9.7 ปรับปรุงค่า $epoch = epoch + 1$

3.2.6 การวัดระยะทาง

การวัดระยะทางหรือการวัดความคล้ายกันระหว่าง 2 spike trains สามารถวัดได้โดยอาศัยอัตราการกระตุ้น (Firing rate) ซึ่งเป็นความถี่ของการเกิดสไปค์ในแต่ละช่วงย่อย (bin) ที่มีขนาดเท่ากันของ spike train ดังรูปที่ 3.9



รูปที่ 3.9 ตัวอย่างการหาอัตราการกระตุ้นของ spike train

อัตราการกระตุ้นของ spike train อาจมองเป็นจุดหรือเวกเตอร์ที่อยู่ในระนาบ N มิติ โดยที่ N คือจำนวนช่วงย่อย ดังนั้นการหาระยะทางระหว่าง 2 spike trains สามารถทำได้โดยการวัดระยะทางระหว่างจุด 2 จุด หรือเวกเตอร์ 2 เวกเตอร์ ในระนาบ N มิติ

ในงานวิจัยนี้ศึกษาวิธีการแบ่งกลุ่มข้อมูล spike trains โดยใช้วิธีการวัดระยะทาง 2 แบบ คือ การวัดระยะทางแบบ Euclidean และการวัดระยะทางเชิงมุมแบบ Cosine ซึ่งมีรายละเอียดของแต่ละวิธีดังนี้

การวัดระยะทางแบบ Euclidean

กำหนดให้ $x=(x_1,x_2,...,x_m)$ และ $y=(y_1,y_2,...,y_m)$ คือจุดที่อยู่ในระนาบ m มิติ ระยะทางแบบ Euclidean เขียนแทนด้วย $d_{ED}(x,y)$ นิยามได้ดังนี้

$$d_{ED}(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_m-y_m)^2} \qquad (3.18)$$

การวัดระยะทางเชิงมุมแบบ Cosine

กำหนดให้ $\vec{x}=(x_1,x_2,...,x_m)$ และ $\vec{y}=(y_1,y_2,...,y_m)$ คือเวกเตอร์ที่อยู่ในระนาบ m มิติ

ระยะทางแบบ Cosine เขียนแทนด้วย $d_{\cos ine}(\vec{x},\vec{y})$ นิยามได้ดังนี้

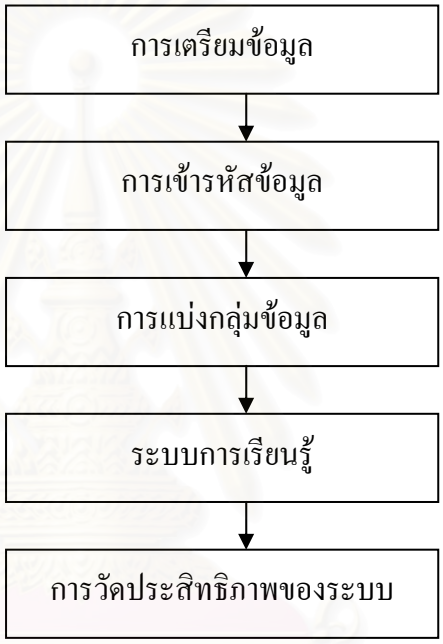
$$d_{\cos ine}(\vec{x},\vec{y}) = \cos^{-1}\left(\frac{\sum_{i=1}^{m} x_i y_i}{\|\vec{x}\| \cdot \|\vec{y}\|}\right) \tag{3.19}$$

โดยที่ $\|\vec{x}\| = \sqrt{\sum_{i=1}^{m} x_i}$ (3.20)

$$\|\vec{y}\| = \sqrt{\sum_{i=1}^{m} y_i} \tag{3.21}$$

บทที่ 4
การประยุกต์ใช้โครงข่ายประสาทแบบสไปกิง

การนำโครงข่ายประสาทแบบสไปกิงไปประยุกต์ใช้ในการแก้ปัญหาการทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือในลำดับดีเอ็นเอ สามารถแบ่งการดำเนินการออกเป็น 5 ขั้นตอน คือ การเตรียมข้อมูล, การเข้ารหัสข้อมูล, การแบ่งกลุ่มข้อมูล, ระบบการเรียนรู้ และการวัดประสิทธิภาพของระบบ ดังรูปที่ 4.1



รูปที่ 4.1 ขั้นตอนการดำเนินการ

ขั้นตอนแรก กล่าวถึงการสร้างชุดข้อมูลเพื่อใช้ในการทดลองจากลำดับดีเอ็นเอ และการแบ่งชุดข้อมูลเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ ขั้นตอนที่ 2 กล่าวถึงขั้นตอนวิธีการเข้ารหัสข้อมูลเพื่อใช้เป็นรูปแบบข้อมูลนำเข้าสำหรับสอนและทดสอบประสิทธิภาพของโครงข่ายประสาทแบบสไปกิง ขั้นตอนที่ 3 กล่าวถึงแนวคิดและวิธีการแบ่งชุดข้อมูลออกเป็นกลุ่มย่อยๆ ขั้นตอนที่ 4 กล่าวถึงโครงสร้างโครงข่ายที่ใช้สอนและทดสอบชุดข้อมูล และโครงสร้างระบบรู้จำ และขั้นตอนสุดท้าย กล่าวถึงขั้นตอนการทดสอบระบบ การวิเคราะห์ผลลัพธ์ที่ได้จากระบบ และการวัดความถูกต้องในการทดสอบ

กำหนดปัญหา

กำหนดลำดับนิวคลีโอไทด์ ซึ่งเขียนแทนด้วย $(A+C+G+T)^{n}N$

ให้หาสัญลักษณ์นิวคลีโอไทด์ที่แท้จริงของ N โดยที่ n คือ จำนวนเต็มบวกที่มากกว่า 0 และ $(A+C+G+T)^{n}$ คือ ลำดับนิวคลีโอไทด์ที่ประกอบด้วย A, C, G, T เรียงติดกัน และมีความยาว n ตัว

แนวทางการแก้ไขปัญหา

เราได้แปลงปัญหาดังกล่าวเป็นปัญหาการรู้จำลำดับนิวคลีโอไทด์ ที่ทราบสัญลักษณ์นิวคลีโอไทด์ที่ชัดเจนว่าเป็น A หรือ C หรือ G หรือ T และสมมติสัญลักษณ์นิวคลีโอไทด์ในตำแหน่งขวาสุดเป็นสัญลักษณ์ที่คลุมเครือ แล้วใช้โครงข่ายประสาทแบบสไปกิงเรียนรู้จำลำดับนิวคลีโอไทด์ก่อนหน้าสัญลักษณ์ที่คลุมเครือนั้น

ตัวอย่างเช่น กำหนด ลำดับนิวคลีโอไทด์ ดังนี้ ACTTGCCAGAA

สมมติ สัญลักษณ์นิวคลีโอไทด์ในตำแหน่งขวาสุดเป็นสัญลักษณ์ที่คลุมเครือ แทนด้วย N นั่นคือ ACTTGCCAGAN

แล้วใช้โครงข่ายประสาทเทียมเรียนรู้จำลำดับนิวคลีโอไทด์ โดยที่ ข้อมูลนำเข้า (input data) คือ ลำดับนิวคลีโอไทด์ก่อนหน้าตัวที่คลุมเครือ นั่นคือ ACTTGCCAGA และข้อมูลที่ต้องการ (desired data) คือ สัญลักษณ์นิวคลีโอไทด์ที่ถูกสมมติว่าคลุมเครือ นั่นคือ N=A

4.1 การเตรียมข้อมูล

ชุดข้อมูล (Data set) ที่ใช้ในงานวิจัยนี้ ได้จากการสุ่มลำดับนิวคลีโอไทด์ที่มีความยาว 15, 20 และ 25 ตัว จากลำดับดีเอ็นเอของแบคทีเรียอีโคไล (E. coli) ทั้ง 4 สายพันธุ์ ในบริเวณตำแหน่งเดียวกัน บริเวณละ 1,000 ตัว โดยเลือกแบบสุ่มมา 80% จากลำดับนิวคลีโอไทด์ที่เป็นไปได้ทั้งหมด เพื่อใช้เป็นชุดข้อมูลตัวอย่าง จากนั้นทำการตัดลำดับนิวคลีโอไทด์ที่ซ้ำกันออก และตัดลำดับนิวคลีโอไทด์ที่มีลำดับก่อนหน้าตัวที่ถูกสมมติว่าคลุมเครือเหมือนกัน แต่ตัวที่ถูกสมมติว่าคลุมเครือต่างกันออกไป หลังจากนั้นจึงทำการแบ่งชุดข้อมูลดังกล่าวออกเป็น 2 ส่วน คือชุดข้อมูลสอน (training data set) และชุดข้อมูลทดสอบ (test data set) ด้วยอัตราส่วนชุดข้อมูลสอนต่อชุดข้อมูลทดสอบ คือ 85/15

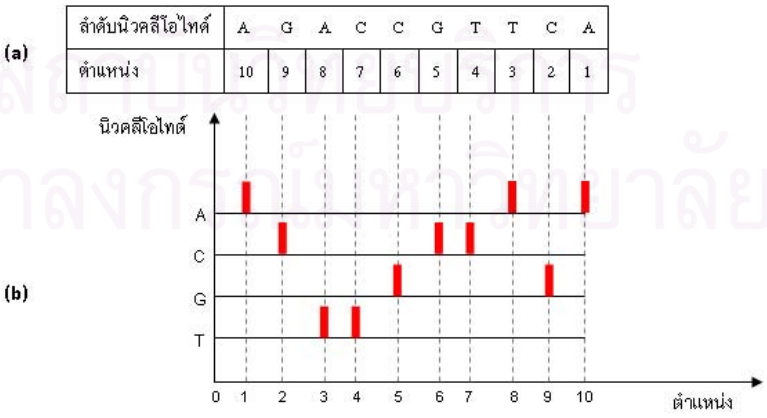
4.2 การเข้ารหัสข้อมูล

ในหัวข้อนี้ กล่าวถึงวิธีการสร้างรูปแบบข้อมูลนำเข้า หรือการเข้ารหัสข้อมูล จากรูปแบบข้อมูลลำดับนิวคลีโอไทด์ ขั้นตอนการเข้ารหัสประกอบด้วย 2 ขั้นตอนหลัก คือ การแทนลำดับนิวคลีโอไทด์ และการกำหนดลำดับเวลาที่เกิดสไปค์ เพื่อใช้เป็นข้อมูลนำเข้าสำหรับสอนและทดสอบโครงข่ายประสาทแบบสไปกิง ดังรูปที่ 4.2

รูปที่ 4.2 ขั้นตอนการเข้ารหัสข้อมูล

การแทนลำดับนิวคลีโอไทด์

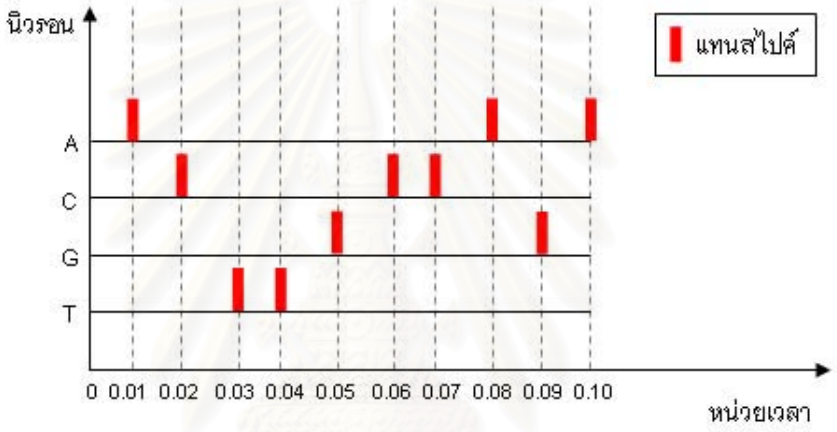
ลำดับนิวคลีโอไทด์จะถูกแทนด้วยลำดับตำแหน่งของการเกิดสัญลักษณ์นิวคลีโอไทด์แต่ละตัว โดยนับจากขวาไปซ้ายของลำดับ ดังรูปที่ 4.3



รูปที่ 4.3 (a) ลำดับตำแหน่งของลำดับนิวคลีโอไทด์ (b) ลำดับตำแหน่งของการเกิดสัญลักษณ์นิวคลีโอไทด์แต่ละตัว

การกำหนดลำดับเวลาที่เกิดสไปค์

การจำลองการทำงานของเซลล์ประสาทในงานวิจัยนี้ ไม่ได้ใช้เวลาจริงในการคำนวณ แต่เป็นเวลาที่เราสมมติขึ้นมาเพื่อจำลองการทำงานของเซลล์ประสาทแทน โดยกำหนดความละเอียดหรือขนาดของขั้นเวลา (time step size) เป็น 0.01 ดังนั้น เราจึงกำหนดลำดับตำแหน่งของการเกิดสัญลักษณ์นิวคลีโอไทด์ เป็นลำดับขั้นเวลาที่เกิดนิวคลีโอไทด์ หรือลำดับเวลาที่เกิดสไปค์ ดังรูปที่ 4.4 นอกจากนี้ เรายังได้ทำการศึกษาการกำหนดลำดับเวลาที่เกิดสไปค์โดยการขยายขั้นเวลาการเกิดสไปค์ อีกด้วย



รูปที่ 4.4 ลำดับขั้นเวลาที่เกิดนิวคลีโอไทด์

ขั้นตอนวิธี การเข้ารหัสข้อมูล

กำหนดให้ P_A, P_C, P_G, P_T แทน ลำดับตำแหน่งของสัญลักษณ์ A, C, G, T ตามลำดับ

T_A, T_C, T_G, T_T แทน ลำดับเวลาที่เกิดสัญลักษณ์ A, C, G, T ตามลำดับ

w_{ij} แทน ค่าน้ำหนักของการเชื่อมโยงระหว่างเซลล์ i กับ j

d_{ij} แทน เวลาที่ถูกหน่วงระหว่างการส่งสัญญาณจากเซลล์ i ไป j

ข้อมูลเข้า ลำดับนิวคลีโอไทด์ แทนด้วย S

ขนาดของขั้นเวลา (time step size) แทนด้วย dt

ข้อมูลออก ชุดลำดับเวลาที่เกิดสไปค์ (Spike trains) แทนด้วย ST = {ST_A, ST_C, ST_G, ST_T}

โดยที่ ST_A, ST_C, ST_G, ST_T คือลำดับเวลาที่เกิดสไปค์ (Spike train)

ของสัญลักษณ์นิวคลีโอไทด์ A, C, G, T ตามลำดับ

1 คำนวณหา P_A, P_C, P_G และ P_T ของ S โดยเริ่มนับจากขวาไปซ้าย

2 คำนวณหา $T_A = dt \times P_A$

$T_C = dt \times P_C$

$T_G = dt \times P_G$

$T_T = dt \times P_T$

3 กำหนดลำดับเวลาที่เกิดสไปค์ (ST_A, ST_C, ST_G, ST_T) โดยการขยายขั้นเวลาของการเกิดสไปค์ ดังนี้

$ST_A = m \times T_A$

$ST_C = m \times T_C$

$ST_G = m \times T_G$

$ST_T = m \times T_T$

โดยที่ m คือ จำนวนของขั้นเวลาที่ขยาย

ตัวอย่างที่ 4.1

กำหนด ลำดับนิวคลีโอไทด์ S = ACTTGCCAGA และ dt = 0.01

ลำดับตำแหน่งของสัญลักษณ์ A, C, G, T ของ S โดยนับจากขวาไปซ้าย และลำดับเวลาที่เกิดสัญลักษณ์ A, C, G, T คือ

$P_A = \{1, 3, 10\}$ $T_A = \{0.01, 0.03, 0.10\}$

$P_C = \{4, 5, 9\}$ $T_C = \{0.04, 0.05, 0.09\}$

$P_G = \{2, 6\}$ $T_G = \{0.02, 0.06\}$

$P_T = \{7, 8\}$ $T_T = \{0.07, 0.08\}$

กำหนดลำดับเวลาที่เกิดสไปค์ โดยการขยายขั้นเวลาของการเกิดสไปค์

สมมติ m=5 คือจำนวนขั้นเวลาที่ขยาย จะได้ลำดับเวลาที่เกิดสไปค์ ดังนี้

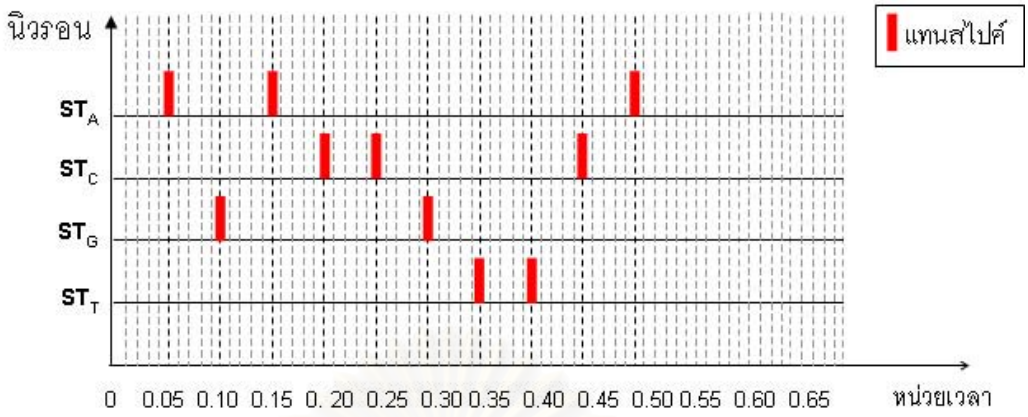
$ST_A = \{0.05, 0.15, 0.50\}$

$ST_C = \{0.20, 0.25, 0.45\}$

$ST_G = \{0.10, 0.30\}$

$ST_T = \{0.35, 0.40\}$

และแสดงได้ดังรูปที่ 4.5



รูปที่ 4.5 ลำดับเวลาที่เกิดสไปค์จากการขยายขั้นเวลาที่เกิดสไปค์

4.3 การแบ่งกลุ่มข้อมูล

ข้อมูลด้านชีวสารสนเทศเป็นข้อมูลด้านชีววิทยาที่มีขนาดใหญ่ การเรียนรู้ข้อมูลชุดสอนที่มีขนาดใหญ่โดยใช้เพียงหนึ่งโครงข่ายเพื่อการเรียนรู้จำนั้น ทำได้ค่อนข้างยากและใช้เวลานาน ดังนั้น การแบ่งชุดข้อมูลออกเป็นกลุ่มย่อยๆ จึงเป็นสิ่งจำเป็น

จากการทำงานของเซลล์ประสาทแบบสไปกิง spike train ที่ไหลเข้ามาในนิวรอนจะทำให้สถานะของนิวรอนนั้นเปลี่ยนแปลงไปตามเวลา สถานะของนิวรอนจะมีลักษณะแตกต่างกันขึ้นอยู่กับลักษณะของ spike train นิวรอนที่ได้รับ spike train คล้ายกัน อาจทำให้เกิดการกระตุ้นที่เวลาใกล้เคียงกัน นิวรอนที่ได้รับ spike train แตกต่างกันอาจทำให้เกิดการกระตุ้นที่เวลาต่างกัน ดังนั้นการแบ่งกลุ่มข้อมูล spike train จะอาศัยความคล้ายกันของ spike train เป็นแนวคิดในการแบ่งกลุ่มข้อมูล และการวัดความคล้ายกันของ spike train สามารถวัดได้โดยอาศัยอัตราการกระตุ้นของ spike train และวิธีการวัดระยะทาง ที่กล่าวไว้ในหัวข้อ 3.2.6

จากวิธีการเข้ารหัสที่กล่าวในหัวข้อ 4.2 ทำให้ได้ลำดับเวลาที่เกิดสไปค์ 4 ลำดับ (4 spike trains) คือ ลำดับเวลาที่เกิดสไปค์ของ A, C, G, T ดังนั้น การวัดระยะทางระหว่าง 2 รูปแบบต้องใช้วิธีการวัดระยะทางระหว่างรูปแบบที่มีหลาย spike trains และในงานวิจัยนี้เราศึกษาวิธีการวัดระยะทาง 2 วิธี ดังนั้นเราจึงเสนอวิธีการวัดระยะทาง 2 วิธี ดังนี้

กำหนดให้

$(x_A^{(1)}, x_A^{(2)}, ..., x_A^{(m)})$, $(x_C^{(1)}, x_C^{(2)}, ..., x_C^{(m)})$, $(x_G^{(1)}, x_G^{(2)}, ..., x_G^{(m)})$, $(x_T^{(1)}, x_T^{(2)}, ..., x_T^{(m)})$ แทนอัตราการกระตุ้นของ spike train ของ A, C, G, T ตามลำดับ ของรูปแบบที่ 1

และ

$(y_A^{(1)}, y_A^{(2)}, ..., y_A^{(m)})$, $(y_C^{(1)}, y_C^{(2)}, ..., y_C^{(m)})$, $(y_G^{(1)}, y_G^{(2)}, ..., y_G^{(m)})$, $(y_T^{(1)}, y_T^{(2)}, ..., y_T^{(m)})$ แทนอัตราการกระตุ้นของ spike train ของ A, C, G, T ตามลำดับ ของรูปแบบที่ 2

การวัดระยะทางแบบ Euclidean

ระยะทางระหว่าง 2 รูปแบบ นิยามได้ดังนี้

$$\begin{aligned} d(x,y) &= \sqrt{(x_A^{(1)} - y_A^{(1)}) + (x_A^{(2)} - y_A^{(2)}) + ... + (x_A^{(m)} - y_A^{(m)})} \\ &+ \sqrt{(x_C^{(1)} - y_C^{(1)}) + (x_C^{(2)} - y_C^{(2)}) + ... + (x_C^{(m)} - y_C^{(m)})} \\ &+ \sqrt{(x_G^{(1)} - y_G^{(1)}) + (x_G^{(2)} - y_G^{(2)}) + ... + (x_G^{(m)} - y_G^{(m)})} \\ &+ \sqrt{(x_T^{(1)} - y_T^{(1)}) + (x_T^{(2)} - y_T^{(2)}) + ... + (x_T^{(m)} - y_T^{(m)})} \end{aligned} \quad (4.1)$$

การวัดระยะทางเชิงมุมแบบ Cosine

ให้ $\vec{x}$, $\vec{y}$ แทนเวกเตอร์ที่เกิดจากการต่อกันของอัตราการกระตุ้นของ spike train ของ A, C, G, T เขียนได้ดังนี้

$$\vec{x} = (x_A^{(1)}, x_A^{(2)}, ..., x_A^{(m)}, x_C^{(1)}, x_C^{(2)}, ..., x_C^{(m)}, x_G^{(1)}, x_G^{(2)}, ..., x_G^{(m)}, x_T^{(1)}, x_T^{(2)}, ..., x_T^{(m)})$$

$$\vec{y} = (y_A^{(1)}, y_A^{(2)}, ..., y_A^{(m)}, y_C^{(1)}, y_C^{(2)}, ..., y_C^{(m)}, y_G^{(1)}, y_G^{(2)}, ..., y_G^{(m)}, y_T^{(1)}, y_T^{(2)}, ..., y_T^{(m)})$$

ดังนั้น ระยะทางเชิงมุมระหว่าง 2 รูปแบบ นิยามได้ดังนี้

$$d(\vec{x}, \vec{y}) = \cos^{-1}\left(\frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|}\right) \quad (4.2)$$

โดยที่ $\vec{x} \cdot \vec{y} = \sum_{i=1}^{m} x_A^{(i)} y_A^{(i)} + \sum_{i=1}^{m} x_C^{(i)} y_C^{(i)} + \sum_{i=1}^{m} x_G^{(i)} y_G^{(i)} + \sum_{i=1}^{m} x_T^{(i)} y_T^{(i)}$

$$\|\vec{x}\| = \sqrt{\sum_{i=1}^{m}\left(x_A^{(i)}\right)^2 + \sum_{i=1}^{m}\left(x_C^{(i)}\right)^2 + \sum_{i=1}^{m}\left(x_G^{(i)}\right)^2 + \sum_{i=1}^{m}\left(x_T^{(i)}\right)^2}$$

$$\|\vec{y}\| = \sqrt{\sum_{i=1}^{m}\left(y_A^{(i)}\right)^2 + \sum_{i=1}^{m}\left(y_C^{(i)}\right)^2 + \sum_{i=1}^{m}\left(y_G^{(i)}\right)^2 + \sum_{i=1}^{m}\left(y_T^{(i)}\right)^2}$$

และ m คือ จำนวนช่วงที่ใช้หาอัตราการกระตุ้นของ spike train

ขั้นตอนวิธี การแบ่งกลุ่มข้อมูลที่มีค่าเป้าหมายเป็น A และไม่เป็น A

กำหนดให้ G{g} แทนกลุ่มของรูปแบบกลุ่มที่ g

A แทนเซตของรูปแบบที่มีค่าเป้าหมายเป็น A

A' แทนเซตของรูปแบบที่มีค่าเป้าหมายไม่เป็น A

ข้อมูลเข้า ชุดข้อมูล (data set)

ข้อมูลออก กลุ่มของชุดข้อมูล

1 แบ่งชุดข้อมูลเป็น 2 กลุ่มคือ กลุ่ม A และ A'

2 หาอัตราการกระตุ้นสำหรับทุกๆ รูปแบบใน A และ A'

3 กำหนด g = 0

4 ทำขั้นตอนที่ 4.1 ถึง 4.10 จนกระทั่ง A หรือ A' ว่าง

4.1 กำหนด g = g + 1

4.2 หาระยะทางระหว่างรูปแบบทุกคู่ ที่อยู่ใน A

4.3 เลือก 1 รูปแบบ ที่อยู่ใน A และมีความแปรปรวน (variance) ของระยะทางระหว่างรูปแบบนั้นกับรูปแบบอื่นๆ ที่อยู่ใน A น้อยที่สุด

4.4 หาส่วนเบี่ยงเบนมาตรฐานและค่าเฉลี่ยของระยะทางระหว่างรูปแบบที่เลือกจากข้อ 4.3 กับรูปแบบอื่นๆ ที่อยู่ใน A

4.5 ย้ายรูปแบบที่เลือกจากข้อ 4.3 จาก A ไป G{g}

4.6 ย้ายรูปแบบใน A ที่มีระยะทางระหว่างรูปแบบนั้นกับรูปแบบที่ได้จากข้อ 4.3 น้อยกว่า ผลต่างระหว่างค่าเฉลี่ยกับส่วนเบี่ยงเบนมาตรฐานที่ได้จากข้อ 4.4 จาก A ไป G{g}

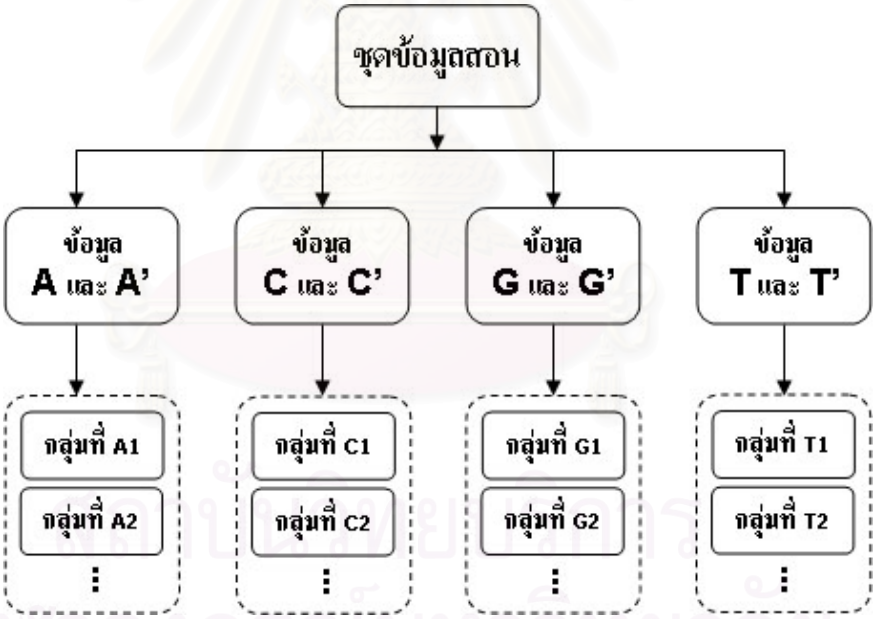
4.7 เลือก 1 รูปแบบที่อยู่ใน A' ที่มีระยะทางระหว่างรูปแบบนั้นกับรูปแบบที่ได้จากข้อ 4.3 มากที่สุด

4.8 คำนวณหาส่วนเบี่ยงเบนมาตรฐานและค่าเฉลี่ยของระยะทางระหว่างรูปแบบที่เลือกจากข้อ 4.7 กับรูปแบบอื่นๆ ที่อยู่ใน A'

4.9 ย้ายรูปแบบที่เลือกจากข้อ 4.7 จาก A' ไป G{g}

4.10 ย้ายรูปแบบใน A' ที่มีระยะทางระหว่างรูปแบบนั้นกับรูปแบบที่ได้จากข้อ 4.7 น้อยกว่า ผลต่างระหว่างค่าเฉลี่ยกับส่วนเบี่ยงเบนมาตรฐานที่ได้จากข้อ 4.8 จาก A' ไป G{g}

ขั้นตอนวิธีการแบ่งกลุ่มข้อมูลข้างต้น เป็นการแบ่งกลุ่มข้อมูลสำหรับโครงข่ายสำหรับ A ส่วนการแบ่งกลุ่มข้อมูลสำหรับโครงข่ายสำหรับ C, G, T ให้ทำในทำนองเดียวกันและสุดท้ายจะได้กลุ่มข้อมูลย่อยสำหรับโครงข่ายสำหรับ A, C, G, T ดังรูปที่ 4.6



รูปที่ 4.6 การแบ่งกลุ่มข้อมูล

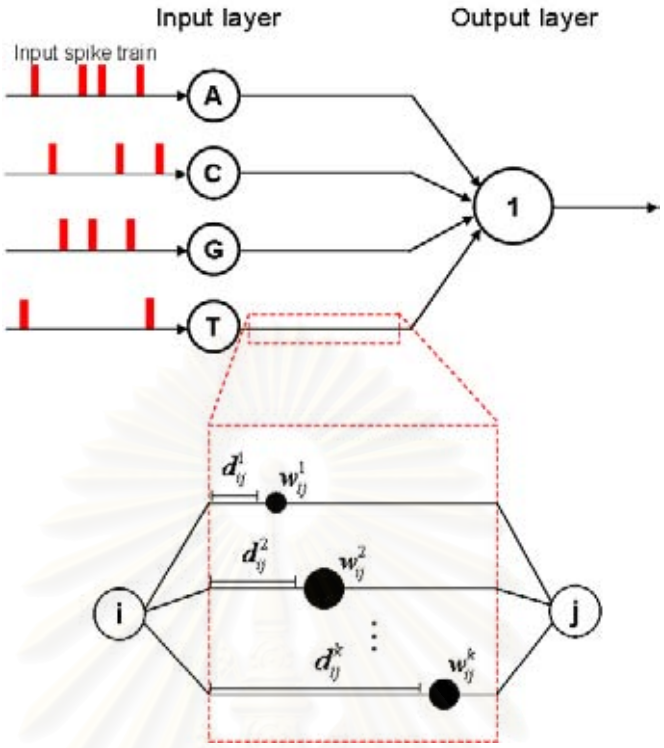
4.4 ระบบการเรียนรู้

การเรียนรู้ข้อมูลที่อยู่ในชุดสอนที่มีขนาดใหญ่โดยใช้เพียงหนึ่งโครงข่ายเพื่อการรู้จำนั้น ทำได้ค่อนข้างยากและใช้เวลานาน ดังนั้น เราจึงแบ่งชุดข้อมูลดังกล่าวออกเป็นชุดข้อมูลกลุ่มย่อยๆ แล้วให้ชุดข้อมูลในแต่ละกลุ่ม เป็นชุดข้อมูลตัวอย่างสำหรับสอนและทดสอบประสิทธิภาพโครงข่ายประสาท ด้วยโครงสร้างโครงข่ายที่ได้ออกแบบไว้ในหัวข้อ 4.2.1 และกฎการเรียนรู้ที่ได้กล่าวไว้แล้วในหัวข้อ 3.2.5 อย่างเป็นอิสระต่อกัน

4.4.1 โครงสร้างโครงข่าย (Network Architecture)

โครงสร้างโครงข่ายประสาทที่ใช้ในการเรียนรู้เป็นโครงสร้างแบบแพร่ไปข้างหน้าแบบชั้นเดียว (Single Layer Feed Forward Neural Network) นั่นคือมีชั้นนำเข้า (input layer) และชั้นนำออก (output layer) โดยที่ชั้นนำเข้ามีนิวรอน 4 โหนด และชั้นนำออกมีนิวรอน 1 โหนด

การเชื่อมโยงระหว่างนิวรอนในชั้นนำเข้าและชั้นนำออก เป็นการเชื่อมโยงแบบสมบูรณ์ (full connection) ดังรูปที่ 4.7 ด้านบน และการเชื่อมโยงระหว่างนิวรอน i กับ j มีการเชื่อมโยงแบบหลายเส้น ดังรูป 4.7 ด้านล่าง ซึ่งเป็นส่วนขยายการเชื่อมต่อระหว่างนิวรอนของชั้นนำเข้า (i) และนิวรอนของชั้นนำออก (j) ของรูปด้านบน โดยที่ d_{ij}^{k} คือเวลาหน่วงของเส้นเชื่อมที่ k ระหว่างนิวรอน i กับ j และ w_{ij}^{k} คือค่าน้ำหนักของเส้นเชื่อมที่ k ระหว่างนิวรอน i กับ j ซึ่งค่าของค่าน้ำหนักที่ต่างกันแสดงด้วยวงกลมที่มีขนาดต่างกัน

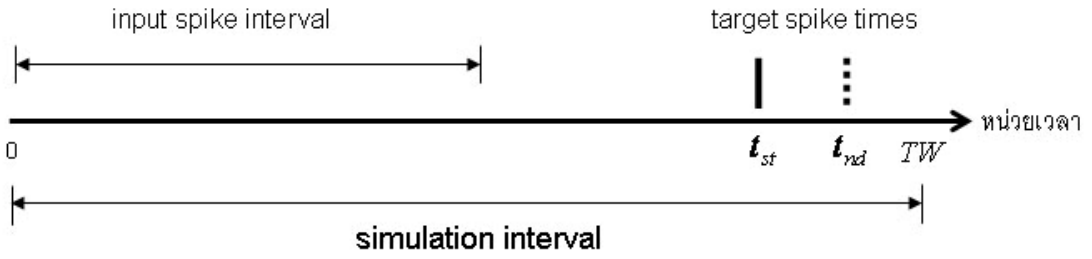


รูปที่ 4.7 โครงสร้างของโครงข่าย

4.4.2 โครงสร้างระบบการรู้จำ

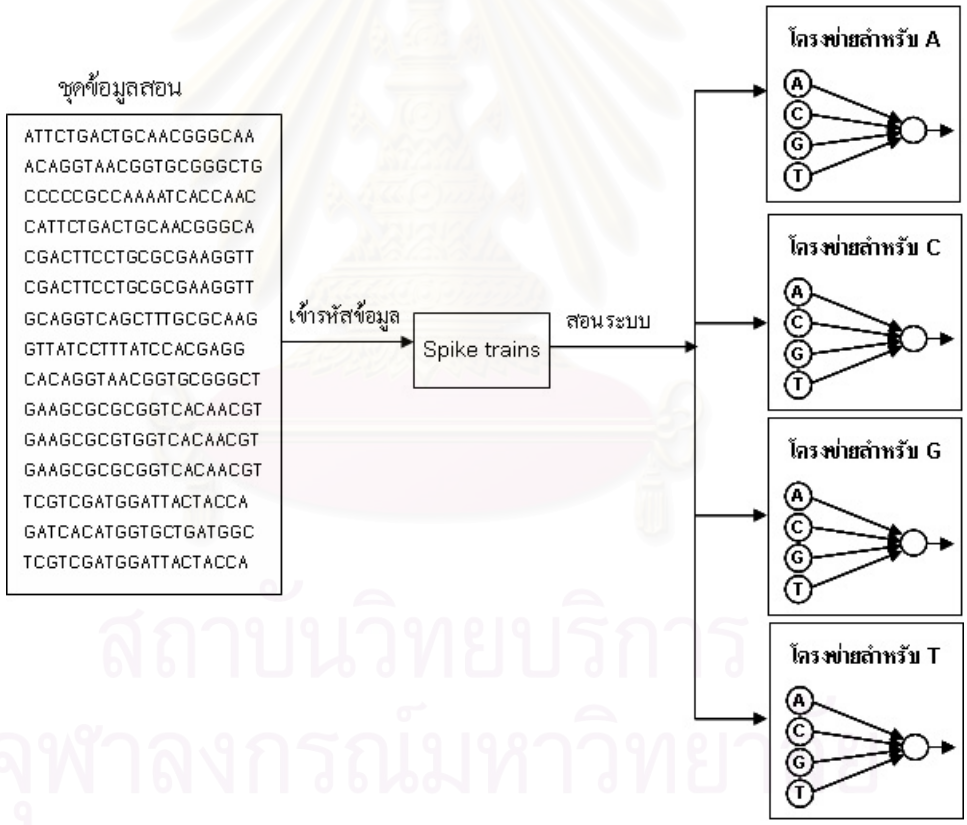
ระบบโครงข่ายประสาทที่ใช้เรียนรู้ชุดข้อมูล ประกอบด้วยโครงข่าย 4 โครงข่าย คือ โครงข่ายสำหรับ A, C, G, T ทำหน้าที่เรียนรู้จำเพื่อแบ่งกลุ่มรูปแบบข้อมูลเป็น 2 กลุ่ม คือ กลุ่มที่มีค่าเป้าหมายเป็นและไม่เป็น A, C, G, T ตามลำดับ

การจำลองการทำงานด้วยโครงข่ายประสาทแบบสไปกิง ต้องกำหนดช่วงเวลาสำหรับการจำลองการทำงานของโครงข่าย (simulation interval, [0,TW]), ช่วงเวลาที่เกิดสไปค์นำเข้า (input spike interval) และเวลาที่เกิดสไปค์เป้าหมาย (target spike time) การกำหนดเวลาที่เกิดสไปค์เป้าหมายนั้น ไม่มีหลักการกำหนดเวลาที่แน่นอน แต่ควรเป็นเวลาหลังจากช่วงเวลาที่เกิดสไปค์นำเข้าแล้ว ในงานวิจัยนี้ ได้ใช้โครงข่ายตามที่ได้กล่าวมาแล้วในหัวข้อ 4.4.1 เพื่อแบ่งกลุ่มข้อมูลลำดับเวลาที่เกิดสไปค์ ออกเป็น 2 กลุ่ม โดยใช้นิวรอนในชั้นข้อมูลออก 1 โหนด เพื่อเป็นโหนดสำหรับตัดสินว่าข้อมูลอยู่ในกลุ่มที่ 1 หรือกลุ่มที่ 2 โดยกำหนดเวลาที่เกิดสไปค์เป้าหมายต่างกัน เขียนแทนด้วยสัญลักษณ์ t_{st} และ t_{nd} ตามลำดับ การกำหนดเวลาที่เกิดสไปค์เป้าหมายนี้ควรกำหนดให้โครงข่ายสามารถเรียนรู้ข้อมูลเพื่อแบ่งกลุ่มข้อมูลได้ ดังรูปที่ 4.8



รูปที่ 4.8 ช่วงเวลาที่ใช้จำลองการทำงานของเซลล์ประสาทแบบสไปกิง

ชุดข้อมูลสอนซึ่งมีรูปแบบเป็นลำดับนิวคลีโอไทด์ จะถูกเข้ารหัสเป็นรูปแบบลำดับเวลาที่เกิดสไปค์ (spike trains pattern) แล้วใช้รูปแบบลำดับเวลาที่เกิดสไปค์นั้นเป็นข้อมูลนำเข้าสำหรับสอนโครงข่ายทั้ง 4 โครงข่าย คือ โครงข่ายสำหรับ A, C, G, T ดังรูปที่ 4.9



รูปที่ 4.9 ระบบโครงข่ายสำหรับการเรียนรู้จำชุดข้อมูล

4.5 การวัดประสิทธิภาพของระบบ

ระบบโครงข่ายประสาทแบบสไปกิง ที่ได้เรียนรู้ชุดข้อมูลสอนเป็นอย่างดีแล้ว จะถูกทดสอบประสิทธิภาพด้วยชุดข้อมูลทดสอบที่โครงข่ายไม่เคยเห็นมาก่อน ข้อมูลทดสอบจะนำไปทดสอบกับกลุ่มที่มีระยะทางใกล้กับข้อมูลนั้นมากที่สุด ของโครงข่ายสำหรับ A, C, G, T และการวิเคราะห์ผลลัพธ์ที่ได้จากโครงข่ายสำหรับ A, C, G, T นั้น จะอาศัยเวลาที่เกิดสไปค์ที่เร็วที่สุดของโครงข่ายทั้ง 4 คือผลของการทำนาย ตัวอย่างเช่น โครงข่าย A และ C เกิดสไปค์เร็วที่สุดพร้อมกัน ผลการทำนาย คือสัญลักษณ์ M เป็นต้น ผลการวิเคราะห์จากผลลัพธ์ที่เป็นไปได้ทั้งหมดจากทั้ง 4 โครงข่าย ได้แสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 การวิเคราะห์สัญลักษณ์นิวคลีโอไทด์ จากผลลัพธ์ของโครงข่ายสำหรับ A, C, G, T

โครงข่ายที่เกิดสไปค์เร็วที่สุด	สัญลักษณ์ที่ถูกวิเคราะห์
A	A
C	C
G	G
T	T
A และ G	R
C และ T	Y
A และ C	M
G และ T	K
C และ G	S
A และ T	W
A, C และ T	H
C, G และ T	B
A, C และ G	V
A, G และ T	D
A, C, G และ T	N

การวัดประสิทธิภาพของระบบนั้น เราวัดโดยเป็นเปอร์เซ็นต์ความถูกต้องของการทำนาย ดังนี้

$$\text{เปอร์เซ็นต์ความถูกต้อง} = n/N \times 100\ \% \qquad (4.3)$$

โดยที่ n คือ จำนวนรูปแบบที่ทำนายถูกต้อง

N คือ จำนวนรูปแบบทั้งหมด

ขั้นตอนวิธี การทดสอบประสิทธิภาพของระบบโครงข่าย

ข้อมูลเข้า : ชุดข้อมูลทดสอบ
โครงข่ายประสาทแบบสไปกิงที่ได้เรียนรู้ชุดข้อมูลสอน

ข้อมูลออก : เปอร์เซ็นต์ความถูกต้องในการทำนาย

ขั้นตอนการทดสอบ :

1 สำหรับแต่ละรูปแบบในชุดข้อมูลทดสอบ ทำขั้นตอน 1.1-1.3

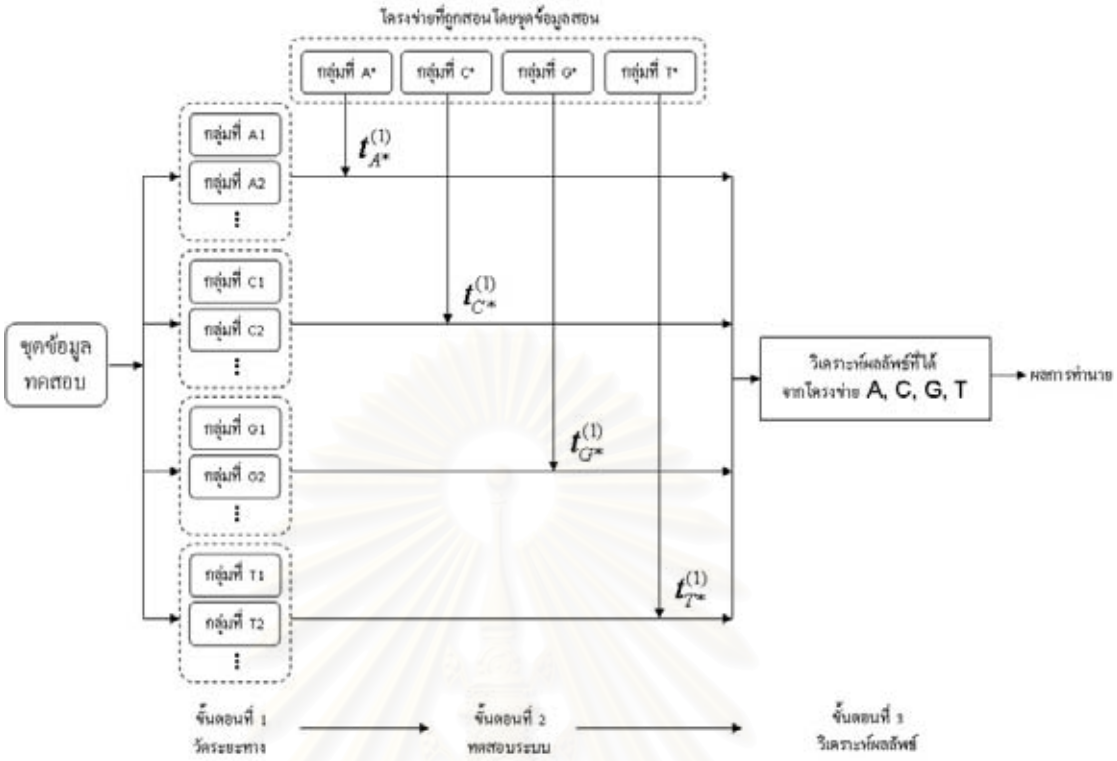
1.1 เลือกกลุ่มเพื่อทำการทดสอบ โดยวัดระยะทางระหว่างรูปแบบทดสอบกับจุดศูนย์กลางทั้ง 2 จุด ทุกกลุ่มย่อย ของกลุ่ม A, C, G, T ที่มีระยะทางน้อยที่สุดสมมติให้เป็นกลุ่ม A*, C*, G*, T* ตามลำดับ

1.2 ทดสอบโครงข่ายสำหรับ A, C, G, T ในกลุ่ม A*, C*, G*, T* จะได้เวลาที่เกิดสไปค์ลำดับที่ 1 เขียนแทนด้วย $t_{A^*}^{(1)}$, $t_{C^*}^{(1)}$, $t_{G^*}^{(1)}$ และ $t_{T^*}^{(1)}$ ตามลำดับ

1.3 วิเคราะห์ผลลัพธ์ที่ได้จากโครงข่ายสำหรับ A, C, G, T โดยใช้ตารางที่ 4.3

2 คำนวณเปอร์เซ็นต์ความถูกต้องในการทำนายจากสมการ 4.3

และเพื่อความเข้าใจในขั้นตอนวิธีการทดสอบ ขั้นตอนดังกล่าวแสดงได้ดังรูปที่ 4.10



รูปที่ 4.10 การทดสอบประสิทธิภาพของระบบ

บทที่ 5
การทดลองและผลการทดลอง

5.1 การทดลอง

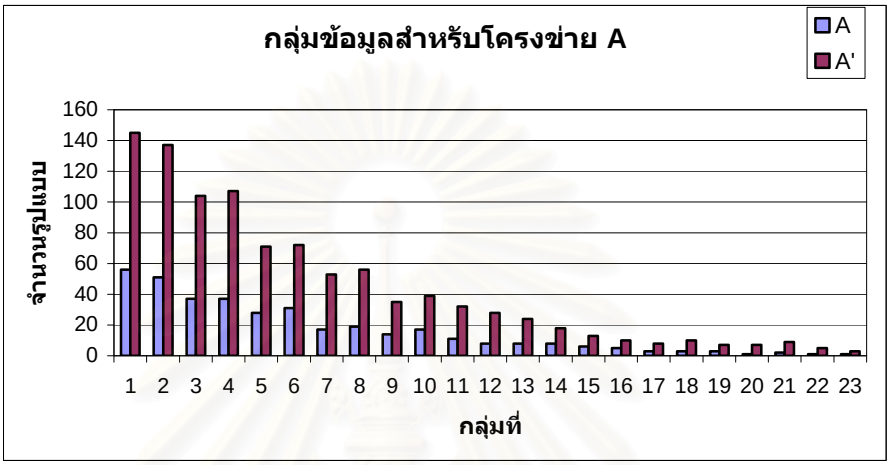
การทดลองใช้ข้อมูลจีโนมของอีโคไล (Escherichia coli หรือ E. coli) 4 สายพันธุ์ คือ CFT073, K12, O157: H7 EDL933, และ O157:H7 สายพันธุ์ย่อย RIMD 0509952 (accession number : AE014075, U00096, AE005174 และ BA000007) ที่ได้มาจากฐานข้อมูลลำดับนิวคลีโอไทด์ EMBL ของสถาบันสารสนเทศแห่งยุโรป (European Bioinformatics Institute, EBI) ซึ่งแต่ละสายพันธุ์มีความยาวประมาณ 4-5 ล้านเบส ในการทดลองได้เลือกลำดับนิวคลีโอไทด์ในบริเวณตำแหน่งเดียวกันของลำดับดีเอ็นเอของทั้ง 4 สายพันธุ์ บริเวณละ 1,000 เบส แล้วทำการสุ่มลำดับนิวคลีโอไทด์ที่มีความยาว 15, 20 และ 25 เบส ที่ไม่มีสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือปรากฏอยู่ในลำดับนั้น โดยสุ่มเลือกมา 80% จากที่เป็นไปได้ทั้งหมดของแต่ละสายพันธุ์ จากนั้นทำการตัดรูปแบบลำดับนิวคลีโอไทด์ที่ซ้ำกันออก โดยเก็บไว้เพียง 1 รูปแบบ และทำการตัดรูปแบบลำดับนิวคลีโอไทด์ที่มีลำดับนิวคลีโอไทด์ก่อนหน้าตัวที่พิจารณาว่าคลุมเครือที่เหมือนกัน แต่ตัวที่คลุมเครือต่างกัน ออกไป จะได้ชุดข้อมูล (Data Set) เพื่อนำไปใช้ในการทดลองต่อไป และชุดข้อมูลถูกแบ่งด้วยอัตราส่วนของชุดข้อมูลสอน (Training data set) กับชุดข้อมูลทดสอบ (Test data set) คือ 85/15 รายละเอียดของชุดข้อมูลที่ใช้ในการทดลองแสดงไว้ในตารางที่ 5.1

ตารางที่ 5.1 รายละเอียดของชุดข้อมูล

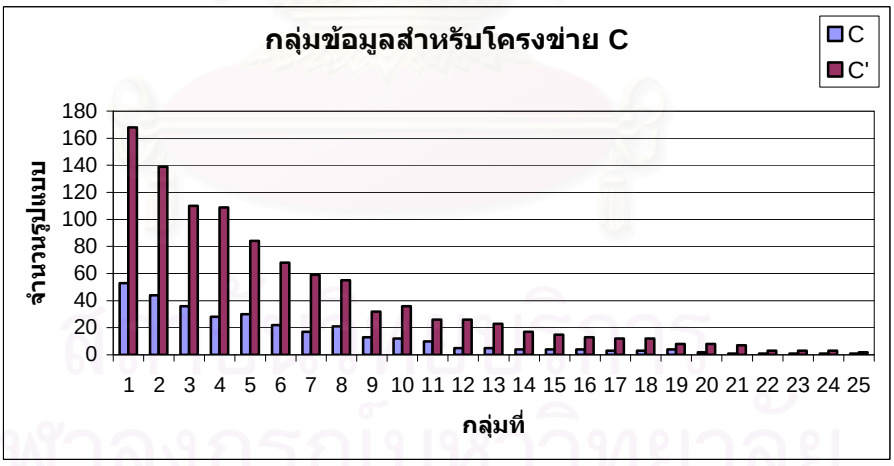
ชุดข้อมูล	ความยาวลำดับนิวคลีโอไทด์ (รวมสัญลักษณ์เป้าหมาย)	จำนวนข้อมูลสอน	จำนวนข้อมูลทดสอบ
1	16	1374	242
2	21	1448	256
3	26	1522	269

การแบ่งกลุ่มข้อมูลโดยใช้วิธีการวัดระยะทางแบบ Euclidean และวิธีการวัดระยะทางเชิงมุมแบบ Cosine ที่ได้กล่าวมาแล้วในหัวข้อ 4.3 ทำให้ได้ข้อมูลสำหรับแต่ละโครงข่ายมีจำนวนกลุ่มและจำนวนรูปแบบแตกต่างกันไป

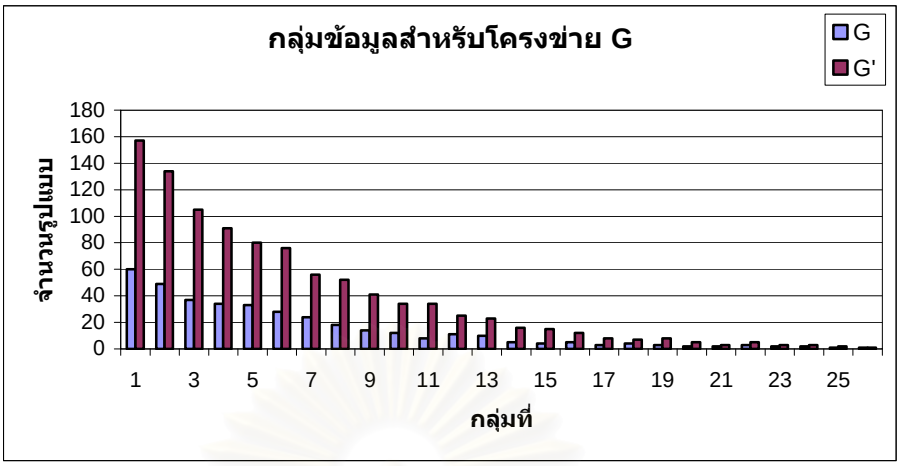
รูปที่ 5.1-5.4 แสดงจำนวนรูปแบบที่มีค่าเป้าหมายต่างกันในแต่ละกลุ่ม ของกลุ่มข้อมูลสำหรับโครงสร้าง A, C, G, T ที่ได้จากการแบ่งกลุ่มข้อมูลชุดที่ 1 โดยวิธีการวัดระยะทางแบบ Euclidean และเห็นได้ว่ากลุ่มข้อมูลสำหรับแต่ละโครงข่าย ข้อมูลในกลุ่มแรกๆ จะมีจำนวนรูปแบบมาก จำนวนรูปแบบจะลดลงเรื่อยๆ และทำให้กลุ่มปลายๆ มีจำนวนรูปแบบน้อย



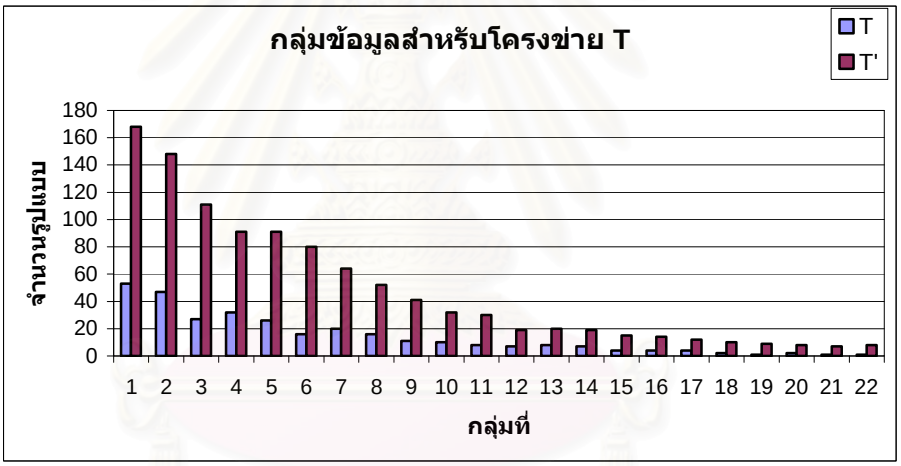
รูปที่ 5.1 กลุ่มข้อมูลสำหรับโครงข่าย A



รูปที่ 5.2 กลุ่มข้อมูลสำหรับโครงข่าย C



รูปที่ 5.3 กลุ่มข้อมูลสำหรับโครงข่าย G



รูปที่ 5.4 กลุ่มข้อมูลสำหรับโครงข่าย T

ข้อมูลที่ได้จากการแบ่งกลุ่ม จะนำมาสอนโครงข่ายประสาท โดยมีการกำหนดพารามิเตอร์สำหรับแต่ละชุดข้อมูลแตกต่างกัน เนื่องจากความยาวของลำดับนิวคลีโอไทด์ และจำนวนขั้นเวลาที่ใช้ขยายสัญญาณแตกต่างกัน รายละเอียดต่างๆ แสดงไว้ในตารางที่ 5.2 – 5.4

ตารางที่ 5.2 พารามิเตอร์ในการทดลองกับข้อมูลชุดที่ 1

จำนวนขั้นเวลาที่ใช้ขยายสัญญาณ	1	3	5
เวลาที่เกิดสไปค์เป้าหมายที่ 1 (t_{st})	0.44	0.74	1.04
เวลาที่เกิดสไปค์เป้าหมายที่ 2 (t_{nd})	0.55	0.85	1.15
เวลาหน่วงระหว่างคู่นิวรอน	0, 0.01, 0.02, …, 0.54	0, 0.01, 0.02, …, 0.84	0, 0.01, 0.02, …, 1.14
จำนวนเส้นเชื่อมระหว่างคู่นิวรอน	55	85	115
ช่วงเวลาการจำลองการทำงาน	[0, 0.65]	[0, 0.95]	[0, 1.25]

ตารางที่ 5.3 พารามิเตอร์ในการทดลองกับข้อมูลชุดที่ 2

จำนวนขั้นเวลาที่ใช้ขยายสัญญาณ	1	3	5
เวลาที่เกิดสไปค์เป้าหมายที่ 1 (t_{st})	0.49	0.89	1.29
เวลาที่เกิดสไปค์เป้าหมายที่ 2 (t_{nd})	0.60	1.00	1.40
เวลาหน่วงระหว่างคู่นิวรอน	0, 0.01, 0.02, …, 0.59	0, 0.01, 0.02, …, 0.99	0, 0.01, 0.02, …, 1.39
จำนวนเส้นเชื่อมระหว่างคู่นิวรอน	60	100	140
ช่วงเวลาการจำลองการทำงาน	[0, 0.70]	[0, 1.10]	[0, 1.50]

ตารางที่ 5.4 พารามิเตอร์ในการทดลองกับข้อมูลชุดที่ 3

จำนวนขั้นเวลาที่ใช้ขยายสัญญาณ	1	3	5
เวลาที่เกิดสไปค์เป้าหมายที่ 1 (t_{st})	0.54	1.04	1.54
เวลาที่เกิดสไปค์เป้าหมายที่ 2 (t_{nd})	0.65	1.15	1.65
เวลาหน่วงระหว่างคู่นิวรอน	0, 0.01, 0.02, …, 0.64	0, 0.01, 0.02, …, 1.14	0, 0.01, 0.02, …, 1.64
จำนวนเส้นเชื่อมระหว่างคู่นิวรอน	65	115	165
ช่วงเวลาการจำลองการทำงาน	[0, 0.75]	[0, 1.25]	[0, 1.75]

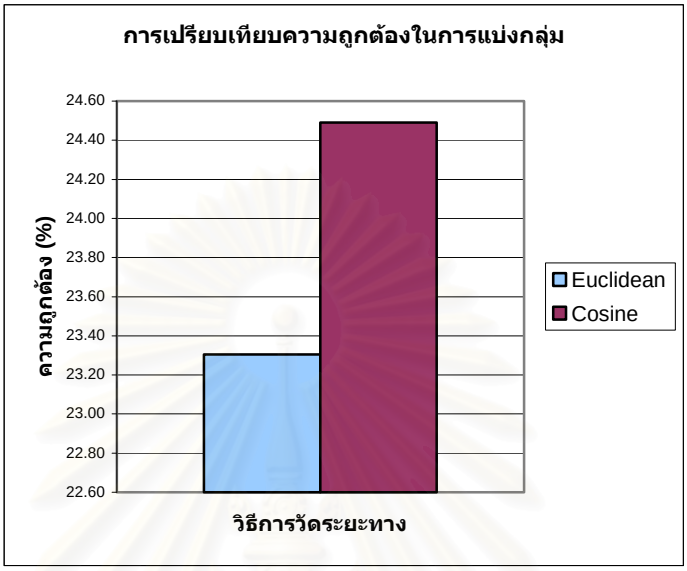
5.2 ผลการทดลอง

จากผลการทดลองในตารางที่ 5.5 เราจะเห็นว่าโดยส่วนใหญ่การแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบ Cosine ได้ให้ความถูกต้องในการทำนายมากกว่าการแบ่งกลุ่มข้อมูลวัดระยะแบบ Euclidean

ตารางที่ 5.5 เปอร์เซ็นต์ความถูกต้องระหว่างการวัดระยะแบบ Euclidean กับแบบ Cosine

ชุดข้อมูล	จำนวนขั้นเวลาที่ใช้ขยายสัญญาณ	วิธีการวัดระยะ	
		Euclidean	Cosine
1	1	22.73	24.38
	3	25.62	26.86
	5	20.25	26.86
2	1	22.27	25.39
	3	26.95	28.13
	5	25.00	21.88
3	1	21.56	21.56
	3	24.54	21.93
	5	20.82	23.42
	ความถูกต้องเฉลี่ย (%)	23.30	24.49

และจากรูปที่ 5.5 จะเห็นว่าโดยเฉลี่ยแล้วการแบ่งกลุ่มข้อมูลโดยใช้ระยะแบบ Cosine ได้ให้ความถูกต้องในการทำนายมากกว่าการแบ่งกลุ่มข้อมูลวัดระยะแบบ Euclidean

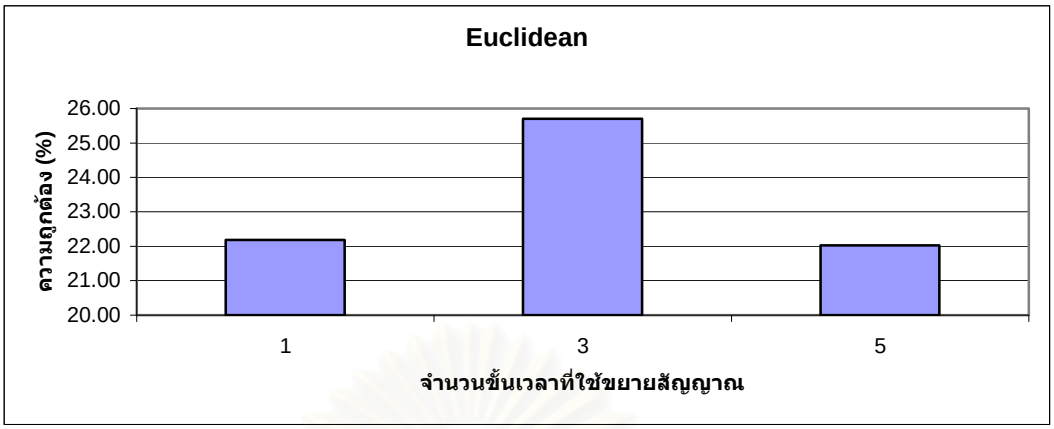


รูปที่ 5.5 การเปรียบเทียบความถูกต้องในการแบ่งกลุ่มข้อมูลโดยใช้วิธีการวัดระยะทางแบบ Euclidean และ Cosine

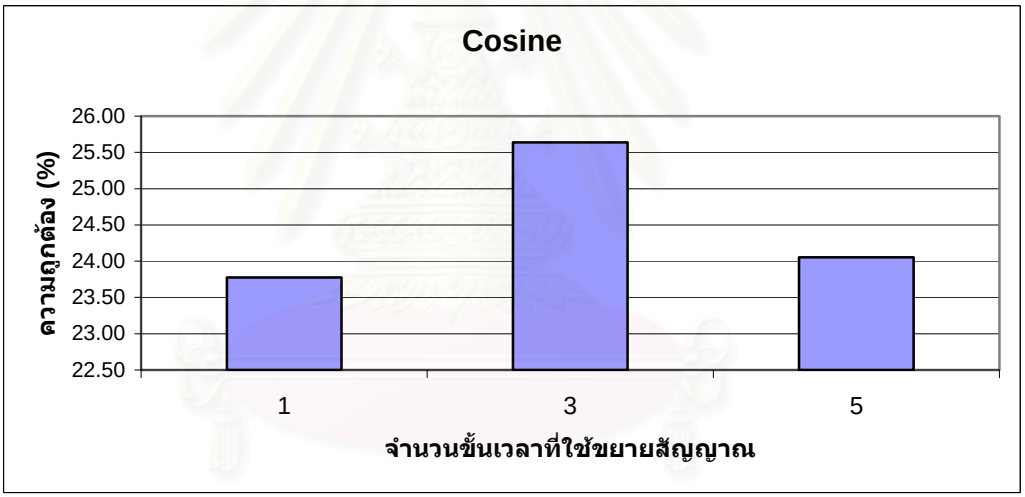
จากตารางที่ 5.5 นำมาเขียนใหม่ได้ตารางที่ 5.6 จะเห็นได้ว่าจำนวนขั้นเวลาที่ขยายสัญญาณสไปค์เท่ากับ 3 ได้ให้ความถูกต้องในการทำนายมากกว่าจำนวนขั้นเวลาที่ขยายสัญญาณสไปค์เท่ากับ 1 และ 5 ทั้งการวัดระยะแบบ Euclidean และ Cosine ดังรูปที่ 5.6 และ 5.7

ตารางที่ 5.6 เปอร์เซ็นต์ความถูกต้องในการขยายจำนวนขั้นเวลาของสัญญาณ

วิธีการวัดระยะ	Euclidean			Cosine		
จำนวนขั้นเวลาที่ใช้ขยายสัญญาณ	1	3	5	1	3	5
ชุดข้อมูล 1	22.73	25.62	20.25	24.38	26.86	26.86
ชุดข้อมูล 2	22.27	26.95	25.00	25.39	28.13	21.88
ชุดข้อมูล 3	21.56	24.54	20.82	21.56	21.93	23.42
ค่าเฉลี่ย	22.19	25.70	22.02	23.78	25.64	24.05



รูปที่ 5.6 การเปรียบเทียบความถูกต้องของจำนวนขั้นเวลาที่ใช้ขยายสัญญาณโดยใช้วิธีการวัดระยะแบบ Euclidean



รูปที่ 5.7 การเปรียบเทียบความถูกต้องของจำนวนขั้นเวลาที่ใช้ขยายสัญญาณโดยใช้วิธีการวัดระยะแบบ Cosine

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

งานวิจัยนี้ศึกษาการใช้ระยะแบบยุคลิดและสหสัมพันธ์เพียร์สัน ในการทำนายสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือในลำดับดีเอ็นเอ โดยใช้โครงข่ายประสาทแบบสไปกิง โดยใช้โครงสร้างโครงข่ายแบบโครงข่ายแบบแพร่ไปข้างหน้าชั้นเดียว และใช้ขั้นตอนวิธี SpikeProp ที่สามารถจัดการกับเวลาที่เกิดสไปค์ได้หลายครั้งต่อหนึ่งนิวรอน และมีการปรับค่าน้ำหนักแบบ RProp ปัญหาการทำนายดังกล่าว ถูกแปลงไปเป็นปัญหาการรู้จำลำดับนิวลีโอไทด์ก่อนหน้าสัญลักษณ์นิวคลีโอไทด์ที่คลุมเครือ งานวิจัยนี้ได้เสนอการเข้ารหัสข้อมูลจากรูปแบบลำดับนิวคลีโอไทด์เป็นรูปแบบชุดลำดับเวลาที่เกิดสไปค์ ซึ่งเป็นรูปแบบข้อมูลนำเข้าสำหรับโครงข่ายประสาทแบบสไปกิง โดยอาศัยลำดับของตำแหน่งที่เกิดนิวคลีโอไทด์แต่ละตัว จากผลการทดลองจะเห็นว่า การทำนายสัญลักษณ์ที่คลุมเครือโดยใช้ระยะแบบเพียร์สันในการแบ่งกลุ่มได้ให้ความถูกต้องเฉลี่ยมากกว่าใช้ระยะแบบยุคลิด และนอกจากนี้เรายังพบว่าจำนวนขั้นเวลาที่ใช้ขยายสัญญาณมีผลต่อความถูกต้องในการทำนายด้วย

6.2 ข้อเสนอแนะ

จากการทดลอง จะเห็นได้ว่าความถูกต้องในการทำนายค่อนข้างต่ำ มีสาเหตุมาจากวิธีวัดระยะทางเพื่อใช้ในการแบ่งกลุ่มข้อมูลที่ได้เสนอไว้แล้วนั้นไม่เหมาะสม ดังนั้นเราควรศึกษาวิธีการวัดระยะทางแบบอื่น ที่สามารถวัดระยะทาง แล้วทำให้ข้อมูลที่อยู่ในกลุ่มเดียวกันมีระยะทางใกล้ศูนย์ และข้อมูลที่อยู่ต่างกลุ่มกันควรมีระยะทางต่างกันมากๆ ถ้าเราสามารถหาวิธีวัดระยะทางที่มีคุณสมบัติแบบนี้ได้ เราก็จะสามารถแบ่งกลุ่มข้อมูลเหล่านี้ได้ถูกต้อง

นอกจากนี้ อาจนำโครงสร้างโครงข่ายประสาทเทียมแบบแพร่ไปข้างหน้าหลายชั้น และขั้นตอนวิธีการเรียนรู้สำหรับโครงสร้างโครงข่ายแบบหลายชั้น [6,7] มาประยุกต์ใช้ อาจทำให้โครงข่ายประสาทแบบสไปกิงสามารถเรียนรู้ชุดข้อมูลที่มีขนาดใหญ่ได้มากขึ้น และอาจให้ผลการทำนายที่ถูกต้องมากขึ้น

รายการอ้างอิง

[1] Plaimas, K., Lursinsap, C., and Suratanee, A. High performance of artificial neural network of resolving ambiguous nucleotide problem. Proceeding of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) Denver, Colorado, USA. April 4-8, 2005.

[2] Plaimas, K. A technique for predicting an ambiguous nucleotide symbol in a DNA sequence. Master's Thesis, Department of Mathematics, Faculty of Science, Chulalongkorn University, 2004.

[3] Bohte, S. M., Nok, J. N., and Poutre H. L. Error-backpropagation in temporally encoded networks of spiking neurons. Neurocomputing 48 (2002): 17-37.

[4] Xin, J., and Embrechts, M. J. Supervised learning with spiking neural networks Proceedings of International Joint Conference on Neural Networks (IJCNN) (2001): 1772-1777.

[5] Schrauwen, B., and Campenhout, J. V. Extending SpikeProp. Proceedings of the International Joint Conference on Neural Networks 1 (July 2004): 471-475.

[6] Booij, O., and Nguyen, H. T. A gradient descent rule for spiking neurons emitting multiple spikes. Information Processing Letters 95 (September 2005): 552-558.

[7] Booij, O. Temporal pattern classification using spiking neural networks. Master's Thesis, Intelligent Sensory Information Systems, University of Amsterdam, 2004.

[8] McKennoch, S., Liu, D., and Bushnell, L. G. Fast modifications of the SpikeProp algorithm. International Joint Conference on Neural Networks Vancouver, BC, Canada July 16-21, 2006.

[9] Maass, W., and Bishop, C. M. Pulsed neural networks. Cambridge: MIT Press, 1999.

[10] Gerstner, W., and Kistler, W. M. Spiking neuron models: Single neurons, populations, plasticity. Cambridge, United Kingdom: Cambridge University Press, 2002.

[11] Berredo, R. C. de. A review of spiking neuron models and applications. Master's Thesis, Post-Graduation Program in Electrical Engineering, Universidade Federal de Minas Gerais, 2005.

[12] Maass, W. Networks of spiking neurons: The third generation of neural network models. Neural Networks 10 (1997): 1659-1671.

ประวัติผู้เขียนวิทยานิพนธ์

นายอดิศักดิ์ การบรรจง เกิดวันที่ 19 สิงหาคม พ.ศ. 2522 ที่จังหวัดมหาสารคาม สำเร็จการศึกษาระดับปริญญาตรีวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น ในปีการศึกษา 2544 ปีการศึกษา 2545-2546 ได้เข้าทำงานตำแหน่งอาจารย์อัตราจ้าง แผนกคณิตศาสตร์ สถาบันเทคโนโลยีราชมงคล วิทยาเขตขอนแก่น และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2547