บทที่ 3
วิธีการทำนายยีนด้วยวิธี Fickett และ Glimmer

จากโครงการจีโนมต่าง ๆ (Genome sequencing projects) ทำให้มีข้อมูลลำดับเบสจำนวนมหาศาล ซึ่งได้เปิดทางให้นักวิทยาศาสตร์ได้ค้นหายีนทั้งหมดในสิ่งมีชีวิตต่าง ๆ อย่างกว้างขวาง ซึ่งระบบชีวสารสนเทศ โดยเฉพาะการทำนายยีน (Gene Prediction) มีความจำเป็นต่อการสืบค้นหายีนได้อย่างรวดเร็ว นอกจากนี้ยังสามารถนำมาประยุกต์ใช้ในการวิเคราะห์บทบาทหน้าที่ อธิบายโครงสร้าง ตำแหน่งที่สำคัญต่าง ๆ และการแสดงออกของยีนต่อไป นักวิจัยทางด้าน ชีวสารสนเทศศาสตร์ให้ความสนใจในเรื่องของการทำนายยีน เพื่อระบุตำแหน่งและหน้าที่ของยีนจากจีโนมของสิ่งมีชีวิตต่าง ๆ รวมทั้งมนุษย์อย่างแพร่หลาย ทำให้มีโปรแกรมซึ่งเป็นผลจากงานวิจัย ที่มีความสามารถในการทำนายยีน และมีการเสนอทฤษฎีและวิธีการทำนายยีนออกมาเป็นจำนวนมาก ยกตัวอย่างเช่น โปรแกรมชื่อ Genscan (Chris Burge และคณะ, 1997), Genemark (Mark Borodovsky และคณะ, 1993), GeneParser (Snyder และคณะ, 1993), Genie (D. Kulp และคณะ, 1996) และ Glimmer (S. Salzberg และคณะ, 1998) เป็นต้น โดยในงานวิจัยนี้ได้นำวิธีการทำนายยีนของ Fickett J.W. (1982) และผลจากโปรแกรมทำนายยีน Glimmer version 3.0 มาทำการทดลอง

3.1 วิธีการทำนายยีนของ Fickett J.W. (1982)

วิธีการเพื่อระบุหายีนหรือส่วนที่แสดงออกเป็นโปรตีน (protein coding region) ในสายลำดับดีเอ็นเอ (DNA sequence) โดยศึกษาคุณสมบัติต่าง ๆ ของ protein coding region โดยให้ความสำคัญในการศึกษาความสัมพันธ์ของจำนวนเบสในตำแหน่งต่าง ๆ ของสายลำดับดีเอ็นเอ ซึ่งมีวิธีการดังนี้

1) หาค่า A_i, C_i, G_i and T_i, i = 1, 2, 3 โดยที่

A_1 = จำนวนของเบส A ในตำแหน่งที่ 1, 4, 7, 10, ... ของลำดับเบส

A_2 = จำนวนของเบส A ในตำแหน่งที่ 2, 5, 8, 11, ... ของลำดับเบส

A_3 = จำนวนของเบส A ในตำแหน่งที่ 3, 6, 9, 12, ... ของลำดับเบส

2) คำนวณหาค่า A-, T-, C-, G-Position โดย

$$\text{A-Position} = \frac{\text{MAX}(A_1,A_2,A_3)}{\text{MIN}(A_1,A_2,A_3)+1}$$

3) คำนวณหาค่า A-,T-,C-,G-Content

A content คือ ค่าร้อยละของจำนวนเบส A ต่อจำนวนเบสทั้งหมด

4) หาค่า Probability of Coding แทนด้วย p_1, p_2, p_3, …, p_8 โดยดูจากค่า Position Parameter และ Content Parameter เช่นหากค่า T-Position เท่ากับ 1.15 จะได้ค่า Probability of Coding เท่ากับ 0.09 แต่หาก T-Position เท่ากับ 1.73 จะได้ค่า Probability of Coding เท่ากับ 0.91 ดังรูป 3.1

5) กำหนดค่า weight แทนด้วย w_1, w_2, w_3, ... ,w_8 ดังรูป 3.2

6) คำนวณหาค่า TESTCODE โดย

$$TESTCODE = p_1w_1 + p_2w_2 + p_3w_3 + ... + p_8w_8$$

7) นำค่า TESTCODE ที่ได้ ไปทำนายผลโดยหาก TESTCODE มากกว่า 0.95 จะทำนายว่าลำดับเบสนี้เป็น Coding แต่ถ้ามีค่าน้อยกว่า 0.74 จะทำนายว่าลำดับเบสนี้เป็น Noncoding และหากมีค่าระหว่าง 0.74 ถึง 0.95 จะทำนายว่า No opinion คือยังไม่สามารถหาข้อสรุปที่ชัดเจนได้ ดังรูป 3.3

Position Parameter		Probability of Coding			
0.0 to 1.1	A: .22	C: .23	G: .08	T: .09
1.1 1.2	.20	.30	.08	.09
1.2 1.3	.34	.33	.16	.20
1.3 1.4	.45	.51	.27	.54
1.4 1.5	.68	.48	.48	.44
1.5 1.6	.58	.66	.53	.69
1.6 1.7	.93	.81	.64	.68
1.7 1.8	.84	.70	.74	.91
1.8 1.9	.68	.70	.88	.97
1.9 2.0+	.94	.80	.90	.97

Content Parameter		Probability of Coding			
.00 to .17	A: .21	C: .31	G: .29	T: .58
.17 .19	.81	.39	.33	.51
.19 .21	.65	.44	.41	.69
.21 .23	.67	.43	.41	.56
.23 .25	.49	.59	.73	.75
.25 .27	.62	.59	.64	.55
.27 .29	.55	.64	.64	.40
.29 .31	.44	.51	.47	.39
.31 .33	.49	.64	.54	.24
.33 .99	.28	.82	.40	.28

รูป 3.1 ค่า Probability of Coding ของ Position และ Content พารามิเตอร์

(แหล่งที่มา; Fickett J.W.,1982)

	Position	Content
A	.26	.11
C	.18	.12
G	.31	.15
T	.33	.14

รูป 3.2 ค่า weight ที่กำหนดให้แต่ละพารามิเตอร์

(แหล่งที่มา; Fickett J.W.,1982)

TESTCODE Indicator	Probability of Coding	Prediction
0.32 to 0.43	0.00	Noncoding
0.43 0.53	0.04	Noncoding
0.53 0.64	0.07	Noncoding
0.64 0.74	0.29	Noncoding
0.74 0.84	0.40	No Opinion
0.84 0.95	0.77	No Opinion
0.95 1.05	0.92	Coding
1.05 1.16	0.98	Coding
1.16 1.26	1.00	Coding
1.26 1.37	1.00	Coding

รูป 3.3 ค่า Probability of Coding ของ Position และ Content พารามิเตอร์

(แหล่งที่มา; Fickett J.W.,1982)

3.2 วิธีการทำนายยีนของ GLIMMER

GLIMMER (Gene Locator and Interpolated Markov ModelER) เป็นระบบสำหรับการค้นหายีนใน microbial DNA โดยใช้ interpolated Markov models (IMMs) เป็นหลักการสำคัญในการระบุหา coding region และแยกออกจาก noncoding DNA ซึ่ง IMMs เป็นการรวม Markov models จาก 1^{st} จนถึง 8^{th}–order แล้วกำหนด weight ตามความสามารถในการทำนายของแต่ละโมเดลซึ่ง Markov models ที่นำมาใช้คือ 3-periodic nonhomogenous Markov models โดย Steven L. Salzberg และคณะ ได้เสนอ Glimmer version 1.0 ในปีค.ศ.1997 และพัฒนาเป็น Glimmer version 2.0 ในปีค.ศ.1997 ซึ่งในปัจจุบันได้พัฒนามาถึง Glimmer version 3.0 โดย GLIMMER มีกระบวนการทำงานหลัก 2 ขั้นตอน ในขั้นตอนแรก โปรแกรมจะทำการสร้าง probability model

ของ coding sequences ซึ่งเรียกว่า interpolated context model (ICM) จากเซตของ training sequences ขั้นที่สอง โปรแกรมจะใช้ ICM ที่ได้จากขั้นตอนแรกมาระบุหายีน โดยทำการค้นหา open reading frames (ORFs) ทั้งหมดที่มีความยาวมากกว่าค่า threshold ที่กำหนดไว้ และให้คะแนน ORFs เหล่านั้นในแต่ละสายของ reading frames ทั้ง 6 สายจาก 6-possible reading frames จากนั้นเลือก ORFs ที่มีคะแนนสูงกว่าค่า threshold ที่ได้ออกแบบไว้ใน reading frame ที่ถูกต้อง แล้วนำมาตรวจสอบความเหลื่อมทับกัน (overlap) ถ้ามี ORFs 2 ส่วนมีความเหลื่อมทับกันใน reading frame ที่แตกต่างกัน overlapping region จะถูกแยกออกมาให้คะแนน และนำคะแนนที่ได้จาก reading framesทั้ง 6 สายมาเปรียบเทียบเพื่อหาว่า frame ไหนมีคะแนนสูงสุด frame นั้นน่าจะเป็นยีน

(S. Salzberg และคณะ, 1998)