การประยุกต์ใช้ดีซิชันทรีอินดักชันในการศึกษาพันธุศาสตร์ประชากร

An Application of Decision Tree Induction on Population Genetics Study

จีรยุทธ ไชยจารุวณิช[1] ดาวรุ่ง กังวานพงศ์*

ศิรินทิพย์ รุ่งไกรลาศ[1] เมธวี ศรีคำมูล[2] จตุพล คำปวนสาย[2]

[1]ภาควิชาวิทยาการคอมพิวเตอร์ [2]ภาควิชาชีววิทยา คณะวิทยาศาสตร์

มหาวิทยาลัยเชียงใหม่ เชียงใหม่ 50200

[*]Corresponding author: Tel: +66-053-943409 Fax: +66-053-943433

Email: jeerayut@science.cmu.ac.th

ABSTRACT - This paper employs the decision tree induction technique for data classification using microsatellite genetic data to provide a phylogenetic tree of four hill tribes of northern Thailand (Lisu, Akha, Karen, and Hmong). The obtained results would guide population geneticist a possible hypothesis of such tribute race evolution.

KEY WORDS - Data Classification, Decision Tree Induction Technique, Genetic Marker, Phylogenetics

บทคัดย่อ - บทความนี้นำเทคนิค Decision Tree Induction สำหรับจัดหมวดข้อมูลมาใช้ในการสืบหาความใกล้เคียงเชิงพันธุกรรมของชาวเขา 4 เผ่า (ลีซอ อีก้อ กะเหรี่ยง และม้ง) โดยวิเคราะห์จากข้อมูลความแปรผันทางพันธุกรรมของ Microsatellites ผลจากการทดลองทำให้ได้สมมุติฐานเบื้องต้นที่จะแนะให้มีการตรวจสอบข้อเท็จจริงในเชิงพันธุศาสตร์ประชากรต่อไป

คำสำคัญ - การจัดหมวดข้อมูล เทคนิคดิซิชันทรีอินดักชัน ข้อมูลพันธุกรรม พันธุศาสตร์ประชากร

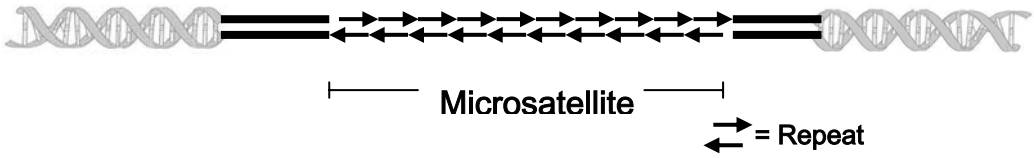
1. บทนำ

ข้อมูลต่างๆในปัจจุบันเริ่มมีจำนวนมากขึ้น การจัดหมวดข้อมูล (data classification) [1] ถือเป็นสิ่งสำคัญที่ทำให้เราสามารถวิเคราะห์เพื่อค้นพบความสัมพันธ์ของข้อมูลในการแบ่งกลุ่มออกเป็นหมวดๆ (class) ตามคุณสมบัติเฉพาะที่เราสนใจ เทคนิคหนึ่งในการจัดหมวดข้อมูล คือ การสร้างต้นไม้การตัดสินใจ (decision tree induction) [4]

decision tree สามารถช่วยจัดหมวดของข้อมูลได้ โดยพิจารณาคุณลักษณะ (attribute) ของข้อมูลว่า คุณลักษณะใดที่ทำให้ข้อมูลถูกจัดให้อยู่ในหมวดเดียวกันได้ดีที่สุด หรืออีกนัยหนึ่งคือหมวดของข้อมูลมีการกระจายอย่างไรเมื่อใช้คุณลักษณะนั้นๆ เป็นตัวแบ่งแยก decision tree ยังสามารถจัดหมวดให้กับข้อมูลใหม่ได้ ตามคุณลักษณะของข้อมูลนั้นๆ

ในบทความฉบับนี้เราจะนำเทคนิค decision tree induction มาใช้ในการจัดหมวดข้อมูลสารพันธุกรรมประเภท microsatellite ที่ตรวจพบเฉพาะในโครโมโซม Y ที่พบในเพศชาย [5] เพื่อใช้ในการสืบหาความใกล้เคียง หรือความแปรผันทางพันธุกรรมของชาวเขา 4 เผ่า (ลีซอ อีก้อ กะเหรี่ยง และม้ง) ผลจากการวิเคราะห์จะทำให้ได้สมมุติฐานเบื้องต้นที่จะแนะให้มีการตรวจสอบข้อเท็จจริงในทางพันธุศาสตร์ประชากรต่อไป

microsatellite เป็นสารพันธุกรรมที่แทรกอยู่โดยทั่วไปในสายของดีเอ็นเอ มีลักษณะเป็นลำดับเบสที่มีการเรียงตัวซ้ำๆ กันเป็นชุด แต่ละชุดมักจะประกอบด้วยคู่เบสจำนวน 2-6 คู่เบส (ดูภาพจำลองในรูปที่ 1) จำนวนชุดของคู่เบสที่ซ้ำกัน ณ ตำแหน่งเดียวกันในสายดีเอ็นเออาจแตกต่างกันได้ในแต่ละบุคคล ความแตกต่างกันดังกล่าวอาจเกิดขึ้นเนื่องจากการกลายพันธุ์ (mutation) และ/หรือ การผสมผสานระหว่างชนเผ่าที่ต่างเชื้อสายกัน



รูปที่ 1 แสดงตัวอย่างจำลองภาพสารพันธุกรรม microsatellite ในโครโมโซม Y

จากคุณลักษณะดังกล่าวของ microsatellite ทำให้เราสามารถศึกษาโครงสร้างพันธุกรรมของชนเผ่าต่างๆได้ โดยอาศัยสมมุติฐานทางพันธุศาสตร์ประชากรที่ว่า บุคคลที่มีสารพันธุกรรม microsatellite ในตำแหน่งต่างๆ เหมือนกันมาก มักจะมีความสัมพันธ์ใกล้ชิดกันทางเชื้อสายมาก ฉะนั้นในประชากรที่มีสารพันธุกรรม microsatellite ใกล้เคียงกัน ก็ย่อมมีความเป็นไปได้ที่จะมีบรรพบุรุษร่วมกัน

อย่างไรก็ตาม ในปัจจุบันประชากรเผ่าพันธุ์ต่างๆ มีการติดต่อสื่อสารกันมากขึ้น ทำให้มีการผสมผสานกันระหว่างประชากรมากขึ้น การศึกษาโครงสร้างพันธุกรรมของชนเผ่าหนึ่งๆ ตามสมมุติฐานข้างต้นจึงไม่ใช่เรื่องง่ายนัก บทความนี้เลือกศึกษากลุ่มประชากรชาวเขาสี่เผ่า ที่อาศัยอยู่บริเวณภาคเหนือตอนบนของประเทศไทย คือ ลีซอ อีก้อ กะเหรี่ยง และ ม้ง เนื่องจากสังคมของชาวเขาเหล่านี้ยังคงเข้มงวดกับการแต่งงานเฉพาะกับคนเผ่าเดียวกัน การสืบทอดเชื้อสายจึงเกิดขึ้นภายในเผ่าเท่านั้น ดังนั้นข้อมูลพันธุกรรมที่ตรวจพบจึงสะท้อนให้เห็นถึงความสัมพันธ์ระหว่างประชากรทั้ง 4 เผ่าได้ดีนั่นเอง

ในการศึกษาครั้งนี้เราจะนำข้อมูลสารพันธุกรรม microsatellite ของชาวเขามาใช้ในการจัดหมวด โดยแต่ละหมวดหมายถึงชนเผ่าต่างๆ และคุณลักษณะของชาวเขาแต่ละคน บ่งบอกได้ด้วยจำนวนชุดของคู่เบสที่ซ้ำกันของ microsatellite ในตำแหน่งต่างๆ (ดูรายละเอียดในตารางที่ 1 ในภาคผนวก) decision tree ที่สร้างได้จะแสดงให้เห็นว่า microsatellite ตำแหน่งใดบ้างที่ช่วยในการแยกกลุ่มของประชากรได้ดีที่สุด ซึ่งเราสามารถนำข้อสรุปดังกล่าวไปใช้ช่วยในการตัดสินใจในการจำแนกความแตกต่างทางพันธุกรรมระหว่างกลุ่มประชากรต่อไป

ในส่วนที่ 2 ของบทความจะอธิบายถึงวิธีการสร้าง decision tree ในส่วนที่ 3 จะเสนอผลของการจัดหมวดข้อมูลพันธุกรรม microsatellites ด้วยเทคนิค decision tree induction และ จะสรุปบทความในส่วนสุดท้าย

2. การสร้าง decision tree

แนวคิดหลักของการสร้าง decision tree คือการเลือกคุณลักษณะ (attribute) ของข้อมูลที่เมื่อแบ่งข้อมูลออกเป็นกลุ่มๆ ตามค่าต่างๆ ของคุณลักษณะดังกล่าวแล้วทำให้ข้อมูลยังจับกลุ่มเป็นหมวดๆ อยู่ไม่แตกกระจายออกจากกัน นั่นคือ ทำให้ entropy [4] ของการแบ่งข้อมูลมีค่าน้อยที่สุด คุณลักษณะที่เลือกจะแทนเป็นโหนดรากของ decision tree และต้นไม้ย่อย ต่างๆ สามารถสร้างได้ในลักษณะเดียวกันจากข้อมูลกลุ่มย่อยที่มีค่าเฉพาะหนึ่งๆ ของคุณลักษณะดังกล่าว ซึ่งจะได้ decision tree เป็นผลลัพธ์สุดท้าย decision tree สามารถสร้างได้ด้วยอัลกอริทึมอย่างง่ายต่อไปนี้

Algorithm DecisionTree (ข้อมูล, กลุ่มคุณลักษณะ)

Input : ข้อมูลที่จะนำมาสร้าง decision tree
: กลุ่มคุณลักษณะของข้อมูล
Output : โครงสร้างของ Decision Tree

1. สร้างโหนดใหม่ N
2. ถ้า ข้อมูลทุกตัวอยู่ในหมวด C เดียวกัน ให้ส่งค่าการคำนวณเป็นต้นไม้โหนดเดี่ยว N ที่มีหมวด C กำหนดไว้
3. ถ้า ข้อมูลมีคุณลักษณะที่กำหนดเหลือเพียง 1 ตัว ให้ส่งค่าการคำนวณเป็นต้นไม้โหนดเดี่ยว N ที่มีหมวดของข้อมูลส่วนใหญ่กำหนดไว้
4. มิเช่นนั้น
5. คำนวณค่า entropy ของการแบ่งข้อมูลออกตามค่าต่าง ๆ ของแต่ละคุณลักษณะ
6. เลือกคุณลักษณะ A ที่ให้ค่า entropy น้อยที่สุดและกำหนดให้ เป็นคุณลักษณะของโหนด N
7. สำหรับทุก ๆ ค่า v ที่เป็นไปได้ของคุณลักษณะ A
8. หาข้อมูลย่อยที่มีค่าของคุณลักษณะ A เท่ากับ v

9. สร้างกิ่งต้นไม้ใหม่จากโหนด N สำหรับค่า v ให้มีต้นไม้ย่อยจากการคำนวณย่อย

DecisionTree (ข้อมูลย่อย, กลุ่มคุณลักษณะยกเว้น A)

10. ให้ส่งค่าการคำนวณเป็นต้นไม้โหนด N เป็นราก

อัลกอริทึมนี้จะสร้าง decision tree ในแนวลึกก่อน แบบ top-down recursive การทำ recursive จะหยุดเมื่อ

- ข้อมูลย่อยที่ได้มีหมวดข้อมูลเดียวกันหมด (ขั้นตอนที่ 2)
- ข้อมูลย่อยมีคุณลักษณะที่กำหนด เหลือเพียง 1 ตัว (ขั้นตอนที่ 3)

ในการเลือกคุณลักษณะที่แบ่งแยกข้อมูลออกเป็นกลุ่ม โดยรักษาหมวดข้อมูลไม่ให้กระจายออกจากกัน ตามระบุในขั้นตอนที่ 5 และ 6 ของอัลกอริทึมสามารถทำได้ โดยเลือกคุณลักษณะที่มี entropy น้อยที่สุด ซึ่ง entropy ของคุณลักษณะ A ใด ๆ คำนวณได้ดังนี้

กำหนดให้ S เป็นเซตของข้อมูล entropy ของ A มีค่าเท่ากับ

$$Entropy(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{|S|} I(s_{1j},\ldots,s_{mj})$$

โดยที่ s_j แทนเซตข้อมูลย่อยของ S ที่มีค่าคุณลักษณะ A เท่ากับ A_j

s_{ij} แทนจำนวนข้อมูลในเซตข้อมูลย่อย s_j ที่อยู่ในหมวดข้อมูล i

m แทนจำนวนของหมวดข้อมูลที่แตกต่างกัน

C_i แทนหมวดข้อมูลที่ i (i =1,…,m)

s_i แทนเซตข้อมูลย่อยของ S ในหมวดข้อมูล C_i

$$I(s_{1j},....,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij})$$

$p_{ij} = s_{ij}/|S_j|$ แทนความน่าจะเป็นที่ข้อมูลใน s_j จะอยู่ในหมวดข้อมูล C_i

อัลกอริทึมนี้จะทำการสร้าง decision tree ในลักษณะ recursive โดยข้อมูลจะถูกแบ่งออกเป็นข้อมูลย่อยเรื่อย ๆ ตามขั้นตอนที่ 7 - 9 และให้ส่งค่าการคำนวณเป็นโครงสร้างของ Decision Treeในท้ายที่สุดในขั้นตอนที่ 10

3. ผลการจัดหมวดข้อมูลพันธุกรรมด้วยเทคนิคดิซิชันทรีอินดักชัน

ข้อมูลที่ใช้ในการสร้าง decision tree คือข้อมูลความยาวของ microsatellites 15 ตำแหน่ง ในโครโมโซม Y ของชาวเขา 51 คน ดูรายละเอียดในตารางที่ 1 ในภาคผนวก

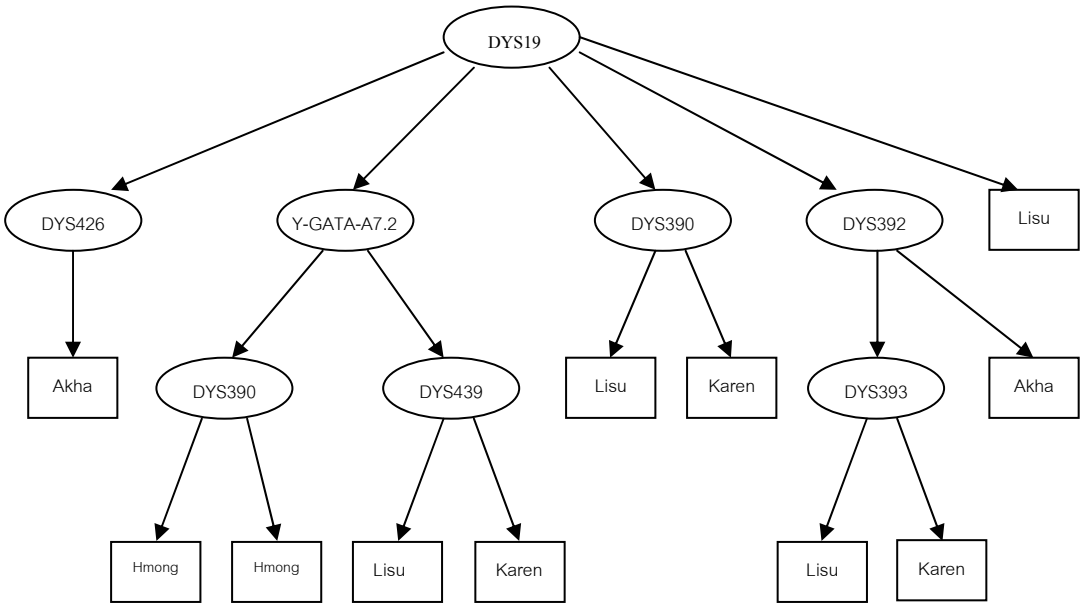
ผลการคำนวณทำให้ได้ decision tree ดังรูปที่ 2 ซึ่งสามารถสรุปได้ว่า

- สารพันธุกรรม microsatellite ตำแหน่งที่ DYS19 ใช้แยกชาวเขาทั้ง 4 เผ่าออกจากกันได้ดีที่สุด
- กลุ่มชาวเขาเผ่าอีก้อมีความใกล้เคียงเชิงพันธุกรรมกับชาวเขาเผ่าอื่นน้อย
- กลุ่มชาวเขาเผ่าลีซอมีความใกล้เคียงเชิงพันธุกรรมกับกลุ่มของชาวเขาเผ่ากะเหรี่ยง โดยมีความแตกต่างกันของพันธุกรรม microsatellite ตำแหน่งที่ DYS339 DYS439 และ DYS390
- กลุ่มชาวเขาเผ่าม้งมีความแตกต่างเชิงพันธุกรรมกับเผ่าลีซอและกะเหรี่ยงด้วย microsatellite ตำแหน่งที่ Y-GATA-A7.2

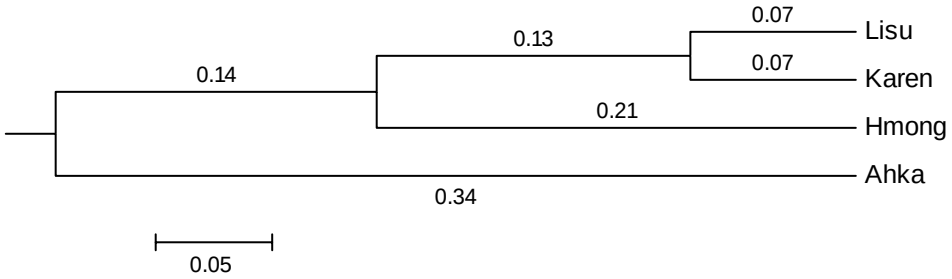
รูปที่ 3 แสดงความสัมพันธ์ทางพันธุกรรม (phylogenetic tree) ระหว่างชาวเขาทั้ง 4 เผ่าที่ได้จากข้อสรุปข้างต้น ซึ่งตรงกับการคำนวณหา phylogenetic tree ด้วยโปรแกรม Mega 2.1 [2] ที่นิยามความแตกต่างระหว่างกลุ่มประชากรด้วยสูตร Fst และ Nei [3]

4. สรุป

จากผลการทดลองจะเห็นได้ว่าการจัดหมวดข้อมูลโดยวิธี decision tree induction สามารถนำมาประยุกต์ใช้ในการสืบหาความใกล้เคียงของกลุ่มชาติพันธุ์มนุษย์ อย่างไรก็ตาม ผลที่ได้จากการทดลองยังเป็นเพียงสมมุติฐานเบื้องต้นที่ได้จากการวิเคราะห์ตามค่าข้อมูลที่พิจารณา ยังมีความจำเป็นที่จะต้องตรวจสอบข้อเท็จจริงในเชิงพันธุศาสตร์ประชากรต่อไป



รูปที่ 2 แสดง decision tree ที่บ่งชี้ตำแหน่งของพันธุกรรม microsatellite ที่แบ่งแยกความแตกต่างของชาวเขา 4 เผ่า



รูปที่ 3 แสดงความสัมพันธ์ทางพันธุกรรมระหว่างชาวเขา 4 เผ่า โดยตัวเลขที่กำกับแสดงถึงความแตกต่างระหว่างเผ่า

เอกสารอ้างอิง

[1] J.Han and M.Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, San Francisco, 2001

[2] S.Kumar et al., Mega 2.1: Molecular Evolutionary Genetics Analysis version 2.1, www. megasoftware.net

[3] M.Nei and S.Kumar, “Molecular Evolution and Phylogenetics”, Oxford University Press, New York, 2000

[4] J.R.Quinlan, “Induction of decision tree”, Machine Learning, Vol. 1, pp.81-106, 1986

[5] M.,Srikummool, D.Kangwanpong, N.Singh, and M.Seielstad, "Y-Chromosomal Variation in Uxorilocal and Patrilocal Populations in Thailand", In: L.Jin, M.Seielstad, and C.Xiao eds. Genetic, Linguistic and Archaeological Perspectives on Human Diversity in Southeast Asia, World Scientific Press, Singapore, pp.69-82, 2001

จีรยุทธ ไชยจารุวณิช สำเร็จการศึกษา BSc. (Computer Science) ปี 1995, MSc. (Computer Science) ปี 1996 จาก University of Grenoble I ประเทศฝรั่งเศส และ PhD. (Computer Science) ปี 2000 จาก University of Evry Val d'Essonne ประเทศฝรั่งเศส ปัจจุบันเป็นอาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

ดาวรุ่ง กังวานพงศ์ สำเร็จการศึกษา วท.บ. (ชีววิทยา) มหาวิทยาลัยเชียงใหม่ ปี พ.ศ. 2515, วท.ม. (พยาธิชีววิทยา) มหาวิทยาลัยมหิดล ปี พ.ศ. 2519 และ Dr.rer.nat (สาขามนุษย์พันธุศาสตร์) มหาวิทยาลัย Johannes-Gutenberg เมือง Mainz ประเทศสหพันธ์สาธารณรัฐเยอรมัน ปี พ.ศ. 2527 ปัจจุบันดำรงตำแหน่งรองศาสตราจารย์ประจำภาควิชาชีววิทยา คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ กรอบงานวิจัยด้านการประยุกต์ใช้ความรู้พันธุศาสตร์พื้นฐาน ในการติดตามผลทางชีวภาพเพื่อเตือนภัย ป้องกันและดูแลรักษา ประชากรมนุษย์ที่สัมผัสมลพิษในระดับที่เกิดความเสี่ยงต่อการเป็นมะเร็ง

ภาคผนวก

ตารางที่ 1 แสดงข้อมูลความยาวของสารพันธุกรรม microsatellites 15 ตำแหน่ง ในโครโมโซม Y ของชาวเขา 51 คน

sample	DYS 390	DYS 391	DYS 392	DYS 393	DYS 389ii	DYS 426	DYS 436	DYS 437	DYS 439	Y-gata-A7.1	Y-gata-A7.2	Y-gata-A10	DYS 389i	DYS 19	DYS 388	class
1	214	162	171	127	117	92	133	185	248	172	178	166	153	199	128	Lisu
2	210	162	174	119	117	92	133	185	244	168	182	166	149	191	122	Lisu
3	214	166	171	127	121	92	133	185	248	176	182	166	153	203	128	Lisu
4	206	162	168	119	117	92	133	189	244	168	182	170	149	191	128	Lisu
5	210	162	171	119	121	92	133	189	244	168	182	166	149	195	128	Lisu
6	210	162	174	119	117	92	133	185	244	168	182	166	149	191	122	Lisu
7	214	162	171	119	117	92	133	185	252	172	182	162	149	203	128	Lisu
8	210	162	171	119	121	92	133	189	244	168	182	166	149	195	128	Lisu
9	210	162	174	119	117	92	133	185	244	168	182	166	149	191	122	Lisu
10	214	162	171	127	121	92	133	185	248	176	182	166	153	203	128	Lisu
11	210	162	171	119	121	92	133	189	244	168	182	166	149	195	128	Lisu
12	210	158	174	123	117	95	136	185	244	180	182	162	157	187	128	Ahka
13	214	162	174	127	117	92	133	185	248	172	182	170	153	199	128	Ahka
14	214	158	174	123	117	95	136	185	244	180	182	162	157	187	128	Ahka
15	214	158	174	123	117	95	136	185	244	176	182	166	157	187	128	Ahka
16	214	162	174	127	117	92	133	185	248	172	182	170	153	199	128	Ahka
17	214	158	174	123	117	95	136	185	244	176	182	162	157	187	128	Ahka

sample	DYS 390	DYS 391	DYS 392	DYS 393	DYS 389ii	DYS 426	DYS 436	DYS 437	DYS 439	Y-gata-A7.1	Y-gata-A7.2	Y-gata-A10	DYS 389i	DYS 19	DYS 388	class
18	214	158	174	123	117	95	136	185	244	172	182	162	157	187	128	Ahka
19	214	162	174	127	117	92	133	185	248	172	182	170	153	199	128	Ahka
20	214	158	174	123	117	95	136	185	244	176	182	162	157	187	128	Ahka
21	214	158	174	123	117	95	136	185	244	176	182	162	157	187	128	Ahka
22	214	158	174	123	117	95	136	185	244	180	182	162	157	187	128	Ahka
23	214	162	174	127	117	92	133	185	248	172	182	170	157	199	128	Ahka
24	214	158	174	123	117	95	136	185	244	176	182	162	157	187	128	Ahka
25	214	158	174	123	117	95	136	185	244	180	182	162	157	187	128	Ahka
26	210	162	174	119	117	92	133	189	248	172	182	166	149	191	122	Karen
27	214	158	171	127	117	92	133	185	252	172	178	174	153	187	128	Karen
28	214	162	171	123	117	92	133	189	244	172	178	166	149	199	128	Karen
29	210	162	174	119	117	92	133	189	248	168	182	166	149	191	122	Karen
30	222	166	171	123	117	92	133	185	248	172	182	162	153	195	128	Karen
31	218	162	174	127	117	92	133	185	244	172	182	166	157	195	128	Karen
32	210	162	174	123	117	92	133	185	244	172	178	162	157	183	131	Karen
33	218	162	174	131	117	92	133	185	244	172	178	166	157	195	128	Karen
34	218	162	174	127	117	92	133	185	244	172	182	166	161	195	128	Karen
35	210	162	174	123	117	92	133	185	244	172	178	162	157	187	131	Karen
36	210	162	174	119	121	92	133	189	244	168	178	170	149	191	122	Karen
37	210	162	174	119	117	92	133	189	248	168	182	162	149	191	122	Karen
38	218	162	174	127	117	92	133	185	244	172	182	166	157	195	128	Karen
39	210	162	174	119	117	92	133	189	248	168	182	166	149	191	122	Karen
40	210	162	174	119	113	92	133	189	248	168	182	166	149	191	122	Karen
41	214	162	174	119	113	92	133	189	248	168	182	170	149	191	122	Karen
42	210	162	174	119	117	92	133	189	248	168	186	166	149	191	122	Karen
43	210	162	174	123	117	92	133	185	244	172	178	162	157	191	131	Karen
44	210	162	174	119	117	92	133	189	248	168	186	166	149	191	122	Karen
45	214	162	174	119	117	92	133	185	248	180	186	162	149	191	122	Hmong
46	214	162	165	131	117	92	133	185	244	176	186	162	153	195	131	Hmong
47	214	162	174	119	117	92	133	185	248	168	186	166	149	191	122	Hmong
48	214	166	165	123	125	92	133	189	244	172	186	162	149	191	134	Hmong
49	206	162	165	119	125	92	133	185	244	172	186	162	149	191	134	Hmong
50	206	162	165	119	117	92	133	185	252	172	186	166	153	191	128	Hmong
51	206	162	165	127	125	92	133	189	244	172	186	162	149	191	134	Hmong

แหล่งข้อมูล: Genetics and Molecular Biology Unit, ภาควิชาชีววิทยา, มหาวิทยาลัยเชียงใหม่, จังหวัดเชียงใหม่ ประเทศไทย