บทที่ 5

การทดลอง

บทนี้จะกล่าวถึงผลการทดลองที่ได้จากโปรแกรม MOMSA เปรียบเทียบกับโปรแกรมในการแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับ โดยชุดข้อมูลที่ใช้ทดลองได้จากฐานข้อมูล BAliBASE

5.1 ข้อมูลที่ใช้ในการทดลอง

งานวิจัยนี้ทำการทดลองกับ 9 ชุดข้อมูลในฐานข้อมูล BAliBASE ประกอบด้วย 1taq, 1aad, 1pii, 1pfc, 1hfh, 451c, kinase, 1aboA และ 1tvxA

ชุดข้อมูล	จำนวนลำดับ	ความยาว(น้อย,มาก,เฉลี่ย)	ร้อยละของเอกลักษณ์
1taq	5	(806,928,865.2)	>35%
1aad	4	(67,79,72.75)	20-40%
1pii	4	(247,259,251.5)	20-40%
1pfc	5	(108,117,112)	20-40%
1hfh	5	(116,132,121.2)	20-40%
451c	5	(70,87,80)	20-40%
kinase	5	(54,70,67.75)	<25%
1aboA	5	(49,80,63.6)	<25%
1tvxA	4	(54,70,67.75)	<25%

ตารางที่ 5.1 รายระเอียด 9 ชุดข้อมูลที่ใช้ในการทดลอง

5.2 การคำนวณความถูกต้องของผลเฉลย

ผลเฉลยที่ได้จากโปรแกรม MOMSA จะเปรียบเทียบกับลำดับอ้างอิงในฐานข้อมูล BAliBASE และคำนวณความถูกต้องกับสมการค่าผลรวมคู่เบส (Sum-of-pairs score, SPS) และสมการค่าแถว (Column score, CS) ซึ่งค่าจากสมการทั้งสองอยู่ระหว่าง 0-1 และค่าที่ได้ยิ่งเข้าใกล้ 1 ยิ่งมีความถูกต้องในการจัดเรียงมาก

5.3 ผลการทดลอง

ผลการทดลองนี้แบ่งออกเป็นสองส่วนหลัก คือการเปรียบเทียบค่าที่ได้จากการใช้ฟังก์ชันวัตถุประสงค์เพียงค่าเดียวกับการใช้ฟังก์ชันวัตถุประสงค์สองค่าในการแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับ และการเปรียบเทียบการใช้คำตอบจากโปรแกรมต่างๆในการสร้างกลุ่มประชากรเริ่มต้นให้กับโปรแกรม MOMSA

5.3.1 ผลการเปรียบเทียบระหว่างการใช้ฟังก์ชันวัตถุประสงค์เดียวกับการใช้สองฟังก์ชันวัตถุประสงค์

เนื่องจากโปรแกรม MOMSA การใช้ฟังก์ชันวัตถุประสงค์สองค่าในการแก้ปัญหาการจัดเรียงลำดับเบสหลายลำดับ เพื่อการเปรียบเทียบคำตอบกับโปรแกรมที่ใช้ฟังก์ชันวัตถุประสงค์เพียงค่าเดียว และมีวิธีใกล้เคียงกันในการหาคำตอบ จึงเสนอโปรแกรมอีกตัวหนึ่งชื่อ SOMSA (Single objective evolutionary algorithm for multiple sequence alignment) ซึ่งปรับปรุงจากโปรแกรม MOMSA โดยปรับเปลี่ยนไปใช้ฟังก์ชันวัตถุประสงค์เพียงตัวเดียว และจากการใช้หน่วยเก็บถาวร (Achive) เป็นการใช้การคัดเลือก $\mu+\lambda$ selection โดยการคัดเลือกชนิดนี้จะเลือกประชากรรุ่นถัดไปจะมาจากผลเฉลยที่ดีที่สุดของประชากรพ่อแม่ และประชากรลูกในรุ่นปัจจุบัน

โปรแกรม SOMSA มีขั้นตอนวิธีดังรูปที่5.1 เมื่อ t คือจำนวนรุ่น และ T คือจำนวนรุ่นที่มากที่สุด

ขั้นที่1: กำหนดค่าเริ่มต้นจากคำตอบของโปรแกรม Clustal W เป็นประชากรเริ่มต้น

$t = 0$

ขั้นที่2: คำนวนหาค่าฟังก์ชันวัตถุประสงค์ให้กลุ่มประชากร

ขั้นที่3: ตรวจเงื่อนไขการหยุดโปรแกรมเมื่อ $t \geq T$

ขั้นที่4: ทำการกลายพันธุ์ หรือการไขว้เปลี่ยนเพื่อปรับปรุงผลเฉลย

ขั้นที่5: คัดเลือกประชากรรุ่นถัดไปด้วยวิธี $\mu+\lambda$ selection

ขั้นที่6: คำนวณฟังก์ชันวัตถุประสงค์ให้กับกลุ่มประชากรรุ่นถัดไป

$t = t + 1$

ทำซ้ำขั้นที่3

รูปที่5.1 รหัสเทียมขั้นตอนวิธีของ SOMSA

การทดลองโดยใช้โปรแกรม SOMSA จะทำการทดลอง 2 แบบโดยใช้ตารางตัวแทนอันเดียวกับโปรแกรม MOMSA แต่การทดลองแบบแรกกำหนดค่าการทำโทษของแก็ปโดยให้ค่า GOP = 10 และค่า GEP = 1 และการทดลองแบบที่สองกำหนดค่าการทำโทษของแก็ปโดยให้ค่า GOP = 8 และค่า GEP = 12

ในการทดลองนี้จะนำคำตอบจากโปรแกรม Clustal W เป็นกลุ่มประชากรเริ่มต้น และทำการประมวลผลทั้งหมด 30 ครั้งแล้วเก็บค่าคำตอบที่ดีที่สุด

Dataset	MOMSA	SOMSA1	SOMSA2
1taq	**0.878/0.819**	0.876/ 0.815	0.875/ 0.816
1aad	**0.833/0.714**	**0.833/ 0.714**	**0.833/ 0.714**
1pii	**0.793/0.631**	0.791/ 0.627	**0.793/ 0.631**
1pfc	0.797/0.610	**0.780/ 0.620**	0.789/ 0.600
1hfh	**0.833/0.670**	0.829/ 0.660	**0.833/ 0.670**
451c	**0.568/0.354**	0.563/ 0.354	**0.568/ 0.354**
kinase	**0.663/0.494**	0.660/ 0.485	0.661/ 0.485
1aboA	**0.696/0.556**	0.707/ 0.533	0.687/ 0.556
1tvxA	**0.227/0.000**	0.223/0.000	0.223/0.000

ตารางที่5.2 เปรียบเทียบค่า SPS/CS ที่ได้จากโปรแกรม MOMSA, SOMSA1, SOMSA2

จากผลการทดลองเปรียบเทียบคำตอบที่ได้ในตารางที่5.2 โดยสังเกตจากค่า CS เป็นหลัก ถ้ามีค่ามากที่สุดแสดงว่ามีคำตอบที่ได้จะมีความถูกต้องมากกว่าโปรแกรมอื่น แต่ถ้าค่า CS เท่ากันจะดูจากค่า SPS โดยถ้า SPS มีค่ามากแสดงว่าคำตอบมีความถูกต้องมากกว่าโปรแกรมอื่น

ผลการทดลองที่ได้ปรากฏว่าโปรแกรม MOMSA ที่ใช้ฟังก์ชันวัตถุประสงค์สองค่ามีการหาคำตอบได้ดีกว่าโปรแกรม SOMSA ที่ใช้ฟังก์ชันวัตถุประสงค์เพียงค่าเดียวในการแก้ปัญหา

5.3.2 ผลการเปรียบเทียบการใช้โปรแกรมต่าง ๆในการสร้างกลุ่มประชากรเริ่มต้นให้กับโปรแกรม MOMSA

ผลการทดลองของโปรแกรม MOMSA ที่ใช้กลุ่มประชากรตั้งต้นจากคำตอบของโปรแกรม Clustal W, Dialign, MAFFT version fft-n-1, MAFFT version fft-n-2 และ T-Coffee โดยโปรแกรมทั้งหมดไม่มีการปรับเปลี่ยนค่าพารามิเตอร์ งานวิจัยนี้ทำการประมวลผลทั้งหมด 30 ครั้ง แล้วทำการเก็บค่าคำตอบที่ดีที่สุด (Best) ค่าเฉลี่ย (Mean) และค่าเบี่ยงเบนมาตรฐาน (Standard deviation, SD)

Dataset	Clustal W		MOMSA					
	SPS	CS	SPS			CS		
	*	*	Best	Mean	SD	Best	Mean	SD
1taq	0.874	0.810	0.878	0.876	0.001	0.819	0.814	0.001
1aad	0.818	0.696	0.833	0.833	0.00	0.714	0.714	0.000
1pii	0.787	0.618	0.793	0.785	0.003	0.631	0.613	0.006
1pfc	0.774	0.600	0.797	0.781	0.007	0.610	0.597	0.010
1hfh	0.820	0.624	0.832	0.816	0.008	0.670	0.641	0.014
451c	0.555	0.338	0.568	0.538	0.011	0.354	0.293	0.026
kinase	0.655	0.485	0.663	0.659	0.003	0.494	0.484	0.004
1aboA	0.693	0.556	0.696	0.682	0.016	0.556	0.529	0.014
1tvxA	0.223	0.000	0.227	0.217	0.007	0.000	0.000	0.000

ตารางที่5.3 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม Clustal W กับโปรแกรม MOMSA

จากตารางที่5.3 เมื่อนำคำตอบจากโปรแกรม Clustal W เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA คำตอบที่ดีที่สุดจากการประมวลผล 30 ครั้งปรากฏว่าค่า SP และ CS ที่ดีที่สุดทุกชุดข้อมูลที่ทำการทดลองมีมากขึ้น ซึ่งดูได้จากค่า SPS และค่า CS ในช่อง Best ที่มากขึ้นเมื่อเทียบกับค่า SPS และ CS ในช่องของโปรแกรม Clustal W ค่าเฉลี่ย SPS ของชุดข้อมูล 1taq, 1aad, 1pfc, kinase มีค่าเพิ่มขึ้น ส่วนชุดข้อมูล 1pii, 1hfh, 451c, 1aboA, 1tvxA มีค่าใกล้เคียงหรือน้อยลง ค่าเฉลี่ย CS ของชุดข้อมูล 1taq, 1aad, 1hfh มีค่าเพิ่มขึ้น ชุดข้อมูล kinase, 1tvxA มี

ค่าเท่าเดิม และชุดข้อมูล 1pii, 1pfc, 451c,1aboA มีค่าลดลง ค่าเบี่ยงเบนมาตรฐานของ SPS และ CS ทุกชุดข้อมูลมีค่าน้อยแสดงถึงความแม่นยำในการประมวลผลครั้งต่อๆไป

Dataset	Dialign		MOMSA					
	SPS	CS	SPS			CS		
	*	*	Best	Mean	SD	Best	Mean	SD
1taq	0.834	0.725	0.843	0.840	0.001	0.742	0.734	0.003
1aad	0.901	0.803	0.893	0.878	0.010	0.803	0.803	0.000
1pii	0.825	0.693	0.842	0.830	0.005	0.729	0.704	0.011
1pfc	0.700	0.520	0.751	0.731	0.008	0.600	0.577	0.011
1hfh	0.325	0.000	0.352	0.340	0.006	0.028	0.012	0.009
451c	0.630	0.400	0.672	0.665	0.004	0.476	0.461	0.006
kinase	0.604	0.445	0.635	0.627	0.006	0.494	0.473	0.008
1aboA	0.257	0.000	0.298	0.272	0.011	0.000	0.000	0.000
1tvxA	0.231	0.000	0.257	0.223	0.013	0.000	0.000	0.000

ตารางที่5.4 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม Dialign กับโปรแกรม MOMSA

จากตารางที่5.4 เมื่อนำคำตอบจากโปรแกรม Dialign เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA คำตอบที่ดีที่สุดจากการประมวลผล 30 ครั้งปรากฏว่าทุกชุดข้อมูลที่ทำการทดลองมีค่า SPS ที่ดีที่สุดมากขึ้นเกือบทุกชุด ยกเว้นชุดข้อมูล 1aad ที่มีค่าลดลงเล็กน้อย ค่า CS ที่ดีที่สุดของทุกชุดข้อมูลมีค่ามากขึ้นหรือเท่าเดิม ค่าเฉลี่ย SPS ของทุกชุดข้อมูลมีค่ามากขึ้น ยกเว้นชุดข้อมูล 1aad, 1tvxA มีค่าน้อยลง ค่าเฉลี่ย CS ทุกชุดข้อมูลมีค่ามากขึ้นหรือเท่าเดิม ค่าเบี่ยงเบนมาตรฐานของ SPS และ CS ทุกชุดข้อมูลมีค่าน้อย

Dataset	Fft-n-1		MOMSA					
	SPS	CS	SPS			CS		
	*	*	Best	Mean	SD	Best	Mean	SD
1taq	0.850	0.765	0.853	0.851	0.001	0.771	0.766	0.002
1aad	0.845	0.786	0.872	0.863	0.005	0.804	0.804	0.000
1pii	0.784	0.613	0.786	0.782	0.002	0.622	0.612	0.005
1pfc	0.783	0.630	0.811	0.795	0.007	0.680	0.640	0.016
1hfh	0.811	0.624	0.828	0.816	0.004	0.670	0.641	0.010
451c	0.584	0.400	0.623	0.598	0.009	0.462	0.398	0.022
kinase	0.596	0.346	0.611	0.591	0.005	0.381	0.347	0.007
1aboA	0.558	0.289	0.571	0.555	0.010	0.311	0.287	0.022
1tvxA	0.178	0.000	0.201	0.181	0.004	0.000	0.000	0.000

ตารางที่5.5 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่างโปรแกรม MAFFT version fft-n-1 กับโปรแกรม MOMSA

จากตารางที่5.5 เมื่อนำคำตอบจากโปรแกรม MAFFT version fft-n-1 เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA คำตอบที่ดีที่สุดจากการประมวลผล 30 ครั้งปรากฏว่าทุกชุดข้อมูลที่ทำการทดลองมีค่า SPS และ CS ที่ดีที่สุดส่วนใหญ่มีค่ามากขึ้น และมีบางส่วนเท่าเดิมค่าเฉลี่ย SPS ของทุกชุดข้อมูลมีค่ามากขึ้น ยกเว้นชุดข้อมูล 1pii, kinase, 1aboA มีค่าน้อยลง ค่าเฉลี่ย CS ชุดข้อมูล 1pii, 451c, 1aboA มีค่าน้อยลง นอกนั้นชุดข้อมูลอื่นมีค่ามากขึ้น ค่าเบี่ยงเบนมาตรฐานของ SPS และ CS ทุกชุดข้อมูลมีค่าน้อย

Dataset	Fft-n-2		MOMSA					
	SPS	CS	SPS			CS		
	*	*	Best	Mean	SD	Best	Mean	SD
1taq	0.885	0.815	0.890	0.887	0.001	0.824	0.819	0.002
1aad	0.848	0.803	0.854	0.854	0.000	0.803	0.803	0.000
1pii	0.764	0.586	0.792	0.787	0.002	0.627	0.621	0.002
1pfc	0.795	0.700	0.810	0.788	0.009	0.720	0.689	0.017
1hfh	0.811	0.624	0.826	0.817	0.005	0.670	0.642	0.011
451c	0.572	0.431	0.583	0.571	0.002	0.431	0.431	0.000
kinase	0.636	0.502	0.648	0.643	0.004	0.519	0.506	0.007
1aboA	0.449	0.244	0.511	0.455	0.024	0.333	0.244	0.049
1tvxA	0.303	0.000	0.307	0.286	0.010	0.000	0.000	0.000

ตารางที่5.6 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่าง MAFFT version fft-n-2 กับโปรแกรม MOMSA

จากตารางที่5.6 เมื่อนำคำตอบจากโปรแกรม MAFFT version fft-n-2 เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA คำตอบที่ดีที่สุดจากการประมวลผล 30 ครั้งปรากฏว่าทุกชุดข้อมูลที่ทำการทดลองมีค่า SPS และ CS ที่ดีที่สุดส่วนใหญ่มีค่ามากขึ้น และมีบางส่วนเท่าเดิมค่าเฉลี่ย SPS ของทุกชุดข้อมูลมีค่ามากขึ้น ยกเว้นชุดข้อมูล 1pfc, 451c, 1tvxA มีค่าน้อยลง ค่าเฉลี่ย CS ชุดข้อมูล 1pfc มีค่าน้อยลง นอกนั้นชุดข้อมูลอื่นมีค่ามากขึ้น ค่าเบี่ยงเบนมาตรฐานของ SPS และ CS ทุกชุดข้อมูลมีค่าน้อย ยกเว้นชุดข้อมูล 1aboA ที่มีค่าค่อนข้างมาก

Dataset	T-Coffee		MOMSA					
	SPS	CS	SPS			CS		
	*	*	Best	Mean	SD	Best	Mean	SD
1taq	0.850	0.773	0.854	0.852	0.001	0.778	0.776	0.002
1aad	0.878	0.768	0.929	0.895	0.024	0.857	0.821	0.024
1pii	0.787	0.618	0.793	0.784	0.004	0.631	0.613	0.008
1pfc	0.817	0.700	0.836	0.827	0.007	0.740	0.707	0.018
1hfh	0.864	0.734	0.878	0.851	0.010	0.761	0.711	0.017
451c	0.571	0.369	0.574	0.563	0.008	0.354	0.350	0.016
kinase	0.696	0.537	0.699	0.697	0.001	0.541	0.541	0.000
1aboA	0.649	0.533	0.688	0.667	0.008	0.556	0.532	0.009
1tvxA	0.322	0.000	0.360	0.311	0.021	0.000	0.000	0.000

ตารางที่5.7 เปรียบเทียบค่าความถูกต้องของคำตอบระหว่าง T-Coffee กับโปรแกรม MOMSA

จากตารางที่5.7 เมื่อนำคำตอบจากโปรแกรม T-Coffee เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA คำตอบที่ดีที่สุดจากการประมวลผล 30 ครั้งปรากฏว่าทุกชุดข้อมูลที่ทำการทดลองมีค่า SPS และ CS ที่ดีที่สุดส่วนใหญ่มีค่ามากขึ้น และมีบางส่วนเท่าเดิม ค่าเฉลี่ย SPS ของชุดข้อมูล 1pii, 1hfh, 451c, 1tvxA มีค่าน้อยลง นอกนั้นทุกชุดข้อมูลมีค่ามากขึ้น ค่าเฉลี่ย CS ชุดข้อมูล 1pii, 1hfh, 451c, 1aboA มีค่าน้อยลง นอกนั้นชุดข้อมูลอื่นมีค่ามากขึ้น ค่าเบี่ยงเบนมาตรฐานของ SPS และ CS ทุกชุดข้อมูลมีค่าน้อย ยกเว้นชุดข้อมูล 1aad ที่มีค่าค่อนข้างมาก

จากผลที่ได้มาทั้งหมดสามารถสรุปได้โดยดูได้จากตาราง 5.8 ในหน้าถัดไป

Dataset	MOMSA with Clustal W	MOMSA with Dialign	MOMSA with Fft-n-1	MOMSA with Fft-n-2	MOMSA with T-Coffee
1taq	+	+	+	+	+
1aad	+	-	+	+	+
1pii	+	+	+	+	+
1pfc	+	+	+	+	+
1hfh	+	+	+	+	+
451c	+	+	+	+	-
kinase	+	+	+	+	+
1aboA	+	+	+	+	+
1tvxA	+	+	+	+	+

ตารางที่5.8 แสดงผลที่ได้จากโปรแกรม MOMSA เมื่อ + คือคำตอบมีค่าดีขึ้น และ - คือคำตอบมีค่าแย่ลง

หลังจากทำการทดลองในการใช้คำตอบจาก Clustal W, Dialign, MAFFT version fft-n-1, MAFFT version fft-n-2 และ T-Coffee มาใช้เป็นกลุ่มประชากรตั้งต้นในโปรแกรม MOMSA คำตอบที่ได้มามีการพัฒนาให้ดีขึ้นโดยดูได้จากตารางที่5.8 ในส่วนคำตอบที่แย่ลงมี 2 ตำแหน่ง คือชุดข้อมูล 1aad โดยเมื่อเปรียบเทียบคำตอบที่ได้จาก Dialign และ MOMSA ที่ใช้คำตอบจาก Dialign ค่า CS จะมีค่าเท่ากัน แต่ค่า SPS จาก Dialign มีค่ามากกว่าเล็กน้อย (มากกว่า 0.008) และชุดข้อมูล 451c โดยเมื่อเปรียบเทียบคำตอบที่ได้จาก T-Coffee และ MOMSA ที่ใช้คำตอบจาก T-Coffee ถึงแม้ค่า SPS จาก T-Coffee จะน้อยกว่า แต่ค่า CS ของ T-Coffee มีค่ามากกว่าเล็กน้อย (มากกว่า 0.004)

Dataset	MOMSA with Clustal W	MOMSA with Dialign	MOMSA with Fft-n-1	MOMSA with Fft-n-2	MOMSA with T-Coffee
1taq	0.878/0.819	0.843/0.742	0.853/0.771	**0.890/0.824**	0.854/0.778
1aad	0.833/0.714	0.893/0.803	0.872/0.804	0.854/0.803	**0.929/0.857**
1pii	0.793/0.631	**0.842/0.729**	0.786/0.622	0.792/0.627	0.793/0.631
1pfc	0.797/0.610	0.751/0.600	0.811/0.680	0.810/0.720	**0.836/0.740**
1hfh	0.832/0.670	0.352/0.028	0.828/0.670	0.826/0.670	**0.878/0.761**
451c	0.568/0.354	**0.672/0.476**	0.623/0.462	0.583/0.431	0.574/0.354
kinase	0.663/0.494	0.635/0.494	0.611/0.381	0.648/0.519	**0.699/0.541**
1aboA	**0.696/0.556**	0.298/0.000	0.571/0.311	0.511/0.333	0.688/0.556
1tvxA	0.227/0.000	0.257/0.000	0.201/0.000	0.307/0.000	**0.360/0.000**

ตารางที่5.9 แสดงค่า SPS/CS ที่ได้จากโปรแกรม MOMSA

จากผลการทดลองกับโปรแกรม Clustal W, Dialign, MAFFT version fft-n-1, MAFFT version fft-n-2 และ T-Coffee คำตอบที่ดีที่สุดส่วนใหญ่ได้จากการนำคำตอบจากโปรแกรม T-Coffee เป็นกลุ่มประชากรตั้งต้น ดังนั้นจึงควรนำคำตอบจากโปรแกรม T-Coffee เป็นกลุ่มประชากรตั้งต้นในชุดข้อมูลอื่นๆในฐานข้อมูล BAliBASE ด้วย ดังที่สังเกตได้ในตารางที่5.9

จากการประมวลผลชุดข้อมูลในฐานข้อมูล BAliBASE ทั้งหมด 142 ชุด โดยทำการใช้คำตอบจากโปรแกรม Clustal W เป็นกลุ่มประชากรตั้งต้นให้กับโปรแกรม MOMSA และทำการประมวลผลเพียงครั้งเดียว ปรากฏว่าผลที่ได้จากโปรแกรม MOMSA มีการพัฒนาขึ้นร้อยละ 70 จากชุดข้อมูลทั้งหมดในฐานข้อมูล BAliBASE