# บทที่ 4

# รายละเอียดงานวิจัย

งานวิจัยนี้พัฒนาโดยใช้ขั้นตอนวิธีเชิงวิวัฒน์แบบหลายวัตถุประสงค์เพื่อแก้ปัญหาการจัดลำดับเบสหลายลำดับ ในบทที่3 แยกเป็น 9 ส่วน โดยส่วนที่1 ทำการอธิบายขั้นตอนโดยย่อของโปรแกรม MOMSA, ส่วนที่2 บรรยายการเข้ารหัสของปัญหาการจัดลำดับเบสหลายลำดับที่ใช้ในโปรแกรม, ส่วนที่3 อธิบายการสร้างประชากรเริ่มต้น, ส่วนที่4 ทำการกำหนดฟังก์ชันวัตถุประสงค์ที่ใช้ โดยสมการที่ใช้ คือสมการผลรวมคู่เบส (Sum-of-pairs), ส่วนที่5 กำหนดลำดับที่, ส่วนที่6 อธิบายการคัดเลือกประชากรของรุ่นถัดไป, ส่วนที่7 อธิบายการคัดเลือกผลเฉลยจากประชากรเพื่อนำไปปรับปรุง, ส่วนที่8 อธิบายการปรับปรุงคำตอบโดยการไขว้เปลี่ยน และการกลายพันธุ์, ส่วนที่9 บรรยายการหาผลเฉลยสุดท้าย และส่วนที่10 แจกแจงค่าพารามิเตอร์ที่ใช้ในงานวิจัย

## 4.1 ขั้นตอนวิธีโปรแกรม MOMSA

โปรแกรม MOMSA (Multiple objective evolutionary algorithm for multiple sequence alignment) จะนำคำตอบของโปรแกรมที่ใช้สำหรับแก้ปัญหาการจัดลำดับเบสหลายลำดับมาเป็นประชากรเริ่มต้น จากนั้นทำการคำนวณลำดับที่ ซึ่งในโปรแกรมนี้ยิ่งลำดับที่ยิ่งน้อยคำตอบจะยิ่งดี การคัดเลือกกลุ่มหน่วยเก็บถาวรรุ่นถัดไป (Archive) จะคัดเลือกจากกลุ่มประชากรพ่อแม่ และจากกลุ่มหน่วยเก็บถาวรรุ่นปัจจุบัน ส่วนประชากรรุ่นถัดไปจะได้จากประชากรที่ได้รับการคัดเลือกและปรับปรุงผลเฉลยจากการไขว้เปลี่ยน หรือการกลายพันธุ์แล้ว ซึ่งการคัดเลือกผลเฉลยนั้นใช้วิธีคัดเลือกโดยการแข่งขัน (Tournament selection) โปรแกรมจะทำการปรับปรุงผลเฉลยซ้ำไปเรื่อยๆจนกระทั่งถึงจำนวนรอบที่ต้องการแล้วจึงจบการทำงาน

ขั้นตอนวิธีของโปรแกรม MOMSA มีดังรูปที่4.1 เมื่อ

เข้ามูลเข้า: $N$ คือจำนวนผลเฉลยในประชากร

$\overline{A}$ คือจำนวนผลเฉลยในหน่วยเก็บถาวร

$T$ คือจำนวนรุ่นที่มากที่สุด

ข้อมูลออก: $A$ คือเซตของหน่วยเก็บถาวร

ขั้นที่1: กำหนดค่าเริ่มต้นโดยใช้เอาต์พุตจากโปรแกรมสำหรับแก้ปัญหาการจัดลำดับเบสหลายลำดับเป็นอินพุตกับประชากรเริ่มต้น

คำนวณฟังก์ชันวัตถุประสงค์ให้กับประชากรเริ่มต้น

$\overline{A} = 0$

$t = 0$

ขั้นที่2: คำนวนหาลำดับที่ให้กลุ่มประชากร และกลุ่มหน่วยเก็บถาวร

ขั้นที่3: คัดเลือกผลเฉลยที่ดีจากกลุ่มประชากร และกลุ่มหน่วยเก็บถาวรรุ่นปัจจุบัน เข้ากลุ่มหน่วยเก็บถาวรรุ่นถัดไป

ขั้นที่4: ตรวจเงื่อนไขการหยุดโปรแกรมเมื่อ $t \geq T$

ขั้นที่5: คัดเลือกประชากรด้วยวิธีคัดเลือกโดยการแข่งขัน

ขั้นที่6: นำประชากรที่เลือกมาทำการกลายพันธุ์ หรือการไขว้เปลี่ยน แล้วนำผลที่ได้เป็นประชากรรุ่นถัดไป

ทำซ้ำขั้นที่5 จนกระทั้งจำนวนประชากรรุ่นถัดไปเท่ากับ $N$

ขั้นที่7: คำนวณฟังก์ชันวัตถุประสงค์ให้กับกลุ่มประชากรรุ่นถัดไป

$t = t + 1$

ทำซ้ำขั้นที่3

รูปที่4.1 รหัสเทียมขั้นตอนวิธีของโปแกรม MOMSA

## 4.2 การเข้ารหัสปัญหา

การเข้ารหัสปัญหาในงานวิจัยนี้จะประกอบไปด้วย 3 ปัจจัยหลักดังนี้

1. ประกอบด้วยตัวอักษร 22 ตัวดังนี้{C,S,T,P,A,G,N,D,E,Q,H,R,K,M,I,L,V,F,Y,W,-,X} ซึ่งตัว C ถึงตัว W แสดงถึงกรดอะมิโน ตัว – แสดงถึงแก๊ปที่เกิดจากการแทรก หรือการลบระหว่างวิวัฒนาการของสิ่งมีชีวิต ตัว X แสดงถึงกรดอะมิโนที่ไม่เคยเห็นมาก่อน

2. ความกว้างที่ใช้จะเท่ากับจำนวนลำดับเบส ซึ่งมีจำนวน $n$ ลำดับ ($n \geq 3$) และแต่ละลำดับอาจมีความยาวเท่า กันหรือไม่เท่ากันก็ได้

3. ความยาวจะเท่ากับความยาวที่มากที่สุดของลำดับเบสทั้งหมดคูณด้วย 1.1 ($w = l_{max} \times 1.1$) การที่คูณ 1.1 นั้นเพื่อเพิ่มพื้นที่เผื่อมีการขยายตัวของผลเฉลยดั้งเดิมขึ้น

### 4.3 การสร้างประชากรเริ่มต้น

ใช้กลุ่มประชากรตั้งต้นที่ได้จากโปรแกรมสำหรับแก้ปัญหาการจัดลำดับเบสหลายลำดับ เพราะผลเฉลยที่ได้นั้นเป็นใกล้เคียงผลเฉลยที่ดี จึงไม่ต้องเสียเวลาในการหาคำตอบเป็นเวลานาน

### 4.4 การกำหนดฟังก์ชันวัตถุประสงค์

ฟังก์ชันวัตถุประสงค์ประกอบด้วยค่าในการให้รางวัล และค่าในการทำโทษ โดยฟังก์ชันวัตถุประสงค์มีสมการดังนี้

$$Objective\ \ function = SPscore - GapPenalty$$

ค่า $SPscore$ เป็นค่าการให้รางวัลที่ได้จากสมการผลรวมคู่เบส (Sum-of-pairs) ซึ่งค่าที่ได้จะขึ้นอยู่กับตารางที่เลือกใช้เช่น ตาราง $PAM$ หรือตาราง $BLOSUM$ ค่าที่ได้จากสมการผลรวมคู่เบสยิ่งมากจะสันนิษฐานได้ว่าการจัดเรียงลำดับเบสหลายลำดับจะมีความถูกต้อง ตารางที่ใช้ในโปรแกรม MOMSA คือตาราง $BLOSUM45$ ค่าการให้รางวัลมีสมการดังนี้

$$SPscore = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} BLOSUM45(l_i, l_j) \tag{4.1}$$

ค่า $GapPenalty$ เป็นค่าการทำโทษ โดยจะทำโทษจากจำนวนของแก็ปที่มี สมการค่าการทำโทษมีดังนี้

$$GapPenalty = GOP + GAPS \times GEP \tag{4.2}$$

เมื่อ $GOP$ คือค่าคงที่เมื่อเจอแก็ปเริ่มต้น $GAPS$ คือจำนวนแก็ปทั้งหมดที่ต่อจากแก็ปเริ่มต้น และ $GEP$ คือค่าคงที่ให้กับแก็ปที่ต่อจากแก็ปเริ่มต้น

### 4.5 การกำหนดลำดับที่

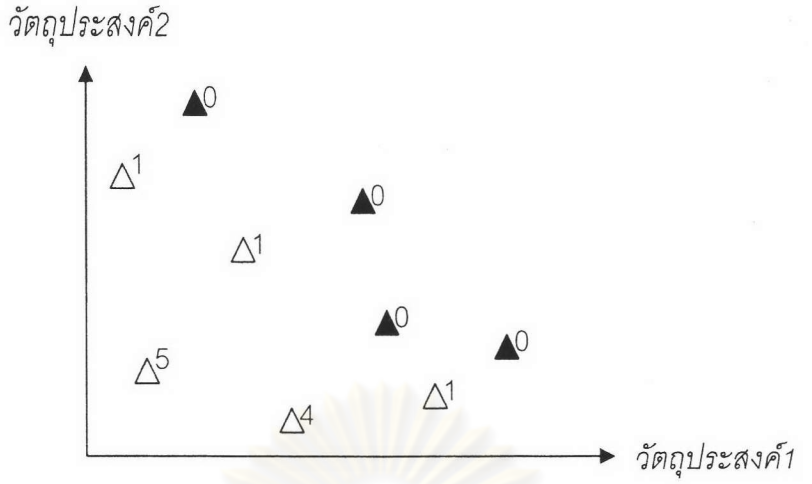
การกำหนดลำดับที่ (Rank) นั้นทำเพื่อบ่งบอกถึงค่าของผลเฉลยที่มีลักษณะดีกว่าตัวอื่น โดยกำหนดให้ตัวที่มีลักษณะดีจะต้องมีฟังก์ชันวัตถุประสงค์ไม่ต่ำกว่าตัวอื่น

มีการกำหนดลำดับที่ดังสมการ

$$Rank(i) = \left| \{ j \mid j \in P_t + A_t \wedge j \succ i \} \right| \tag{4.3}$$

เมื่อ $|\cdot|$ คือจำนวนคำตอบที่อยู่ในเซต $+$ คือการนำเซตมายูเนียน $\succ$ คือเครื่องหมายบอกความเด่น ( Pareto dominance ) และ $i$ คือตำแหน่งของผลเฉลย

ลำดับที่ที่ได้มีค่าน้อยจะเด่นกว่าตัวอื่น ซึ่งค่าที่น้อยที่สุดคือ 0 จะหมายถึงตัวที่เด่นที่สุด




รูปที่4.2 รูปการแก้ปัญหาที่มีฟังก์ชันวัตถุประสงค์สองค่า

โปรแกรม MOMSA ใช้ฟังก์ชันวัตถุประสงค์สองค่าโดยค่าแรกคือ $GOP > GEP$ และค่าที่สองคือ $GOP < GEP$

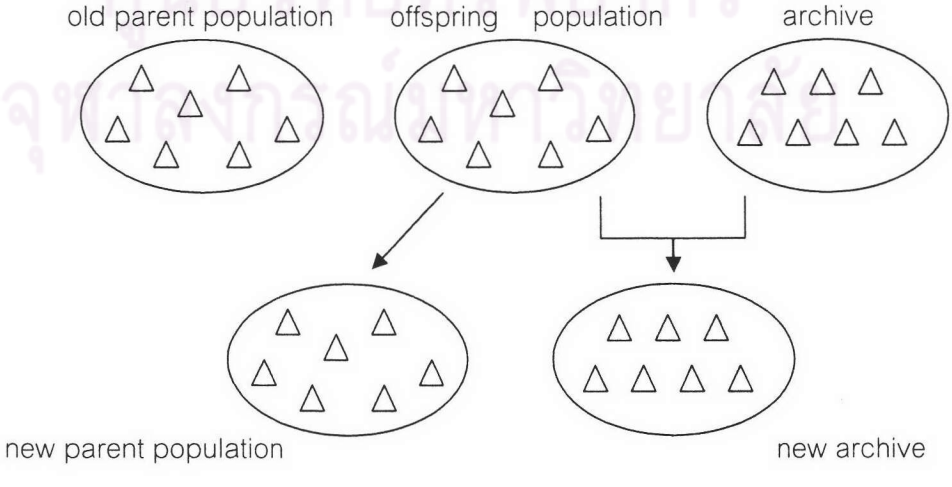
## 4.6 การคัดเลือกประชากรรุ่นถัดไป

ทำการสร้างหน่วยเก็บถาวรรุ่นต่อไปจากประชากรและหน่วยเก็บถาวรในรุ่นปัจจุบันโดยมีเงื่อนไข 2 ข้อคือ

1. ผลเฉลยที่อยู่ในหน่วยเก็บถาวรรุ่นต่อไปจะต้องมีลำดับที่เท่าศูนย์ ดังสมการ

$$A_{t+1} = \{i \mid i \in P_t + A_t \wedge Rank(i) = 0\} \quad (4.4)$$

2. ในหน่วยเก็บถาวรรุ่นต่อไปจะต้องไม่มีผลเฉลยที่ซ้ำกัน




รูปที่4.3 การสร้างประชากร และหน่วยเก็บถาวรรุ่นต่อไป

### 4.7 การคัดเลือกผลเฉลยเพื่อปรับปรุง

การคัดเลือกในโปรแกรม MOMSA ใช้วิธีคัดเลือกโดยการแข่งขัน (Tournament selection) โดยทำการสุ่มเลือกผลเฉลยในประชากร หรือหน่วยเก็บถาวรมาสองตัว จากนั้นเลือกตัวที่มีลำดับที่น้อยกว่าเพื่อนำไปทำการกลายพันธุ์ หรือการไขว้เปลี่ยนต่อไป

ในกรณีที่ที่ผลเฉลยที่สุ่มเลือกมามีลำดับที่เท่ากัน จะทำการเลือกฟังก์ชันวัตถุประสงค์แรก ($GOP > GEP$) ที่มีค่ามากก่อน และเมื่อฟังก์ชันวัตถุประสงค์แรกมีค่าเท่ากันอีกจะทำการเลือกฟังก์ชันวัตถุประสงค์ที่สอง ($GOP < GEP$) ที่มีค่ามากก่อน

### 4.8 การไขว้เปลี่ยน และการกลายพันธุ์

โปรแกรม MOMSA มีการกลายพันธุ์ 3 แบบประกอบด้วย การย้ายเบสแบบสุ่ม การย้ายเบสแบบย้ายฝั่ง และการเลื่อนแถว มีการไขว้เปลี่ยน 1 แบบคือการไขว้เปลี่ยนแบบสองจุด

#### 4.8.1 การย้ายเบสแบบสุ่ม

คัดเลือกผลเฉลยมา 1 ตัวแล้วทำการสุ่มแถว และสุ่มเลือกเบสที่อยู่ในแถวนั้น โดยมีเงื่อนไขให้เบสที่สุ่มเลือกนั้นต้องติดกับแก็ป จากนั้นทำการย้ายเบสแบบสุ่มตำแหน่งในกลุ่มของแก็ป ที่อยู่ติดกัน

```
S1:   AGTTAGTA-T GTTAAA---T
S2:   AGTTAGA--T GTAAAA---T
S3:   ATTTAGTAAT GTAAAAGGTT
```

ก่อนการกลายพันธุ์

```
S1:   AGTTAGTA-T GTTAAA---T
S2:   AGTTAG-A-T GTAAAA---T
S3:   ATTTAGTAAT GTAAAAGGTT
```

หลังการกลายพันธุ์

รูปที่4.4 แสดงการย้ายเบสแบบสุ่ม

จากรูปที่4.4 ขั้นแรกของการย้ายเบสแบบสุ่มจะทำการสุ่มเลือกลำดับโดยลำดับที่เลือกได้คือลำดับ S2 ต่อมาทำการสุ่มเลือกเบสที่อยู่ติดแก็ปในลำดับ S2 ซึ่งตัวที่เลือกได้คือตัว A ในตำแหน่งที่ 7 ทำการสุ่มเลือกตำแหน่งที่ต้องการย้ายที่มีบริเวณอยู่ในกลุ่มของแก็ปที่อยู่ติดกัน ซึ่งสุ่มเลือกได้แก็ปตำแหน่งที่ 8 จากนั้นทำการสลับตำแหน่งระหว่างตัว A และแก็ป

### 4.8.2 การย้ายเบสแบบย้ายฝั่ง

คัดเลือกผลเฉลยมา 1 ตัวแล้วทำการสุ่มแถว และสุ่มเลือกเบสที่อยู่ในแถวนั้น โดยมีเงื่อนไขให้เบสที่สุ่มเลือกนั้นต้องติดกับแก็ป จากนั้นทำการย้ายเบสไปอีกฝั่งของกลุ่มแก็ป

S1: AGTTAGT - AT GTTAAA - - - T
**S2**: AGTTAG**A** - - T GTAAAA - - - T
S3: AT TTAGTAAT GTAAAAGGTT

ก่อนการกลายพันธุ์

S1: AGTTAGT - AT GTTAAA - - - T
**S2**: AGTTAG - - **A**T GTAAAA - - - T
S3: AT TTAGTAAT GTAAAAGGTT

หลังการกลายพันธุ์

รูปที่4.5 แสดงการย้ายเบสแบบย้ายฝั่ง

จากรูปที่4.5 ทำการสุ่มเลือกลำดับ โดยลำดับที่เลือกได้คือ S2 และทำการสุ่มเลือกเบสที่อยู่ติดแก็ปในลำดับ S2 ซึ่งสุ่มเลือกได้ตัว A ในตำแหน่งที่ 7 จากนั้นทำการเลื่อนตัว A ไปอีกด้านหนึ่งของบริเวณแก็ปที่อยู่ติดกัน

### 4.8.3 การเลื่อนแถว

คัดเลือกผลเฉลยมา 1 ตัวแล้วทำการสุ่มเลือกแถว โดยมีเงื่อนไขให้ในแถวนั้นต้องมีแถวใดแถวหนึ่งติดกับแก็ป จากนั้นทำการเลื่อนเบสทั้งหมดในแถวไปยังอีกฝั่ง ถ้าแถวที่เลือกสามารถย้ายไปได้ทั้งด้านซ้ายและด้านขวา จะทำการสุ่มเลือกด้านที่จะทำการย้ายฝั่ง โดยโอกาสในการเลื่อนไปยังด้านขวาหรือด้านซ้ายนั้นมีเท่ากัน

S1: AGTTAGT - AT GTTAA - **T** - TT
S2: AGTTAGA - - T GTAAA - **T** - - T
S3: AT TTAGTAAT GTAAAA**G**GTT

ก่อนการกลายพันธุ์

S1: AGTTAGT - AT GTTAA - - **T**TT
S2: AGTTAGA - - T GTAAA - - - **T**T
S3: AT TTAGTAAT GTAAAA**G**GTT

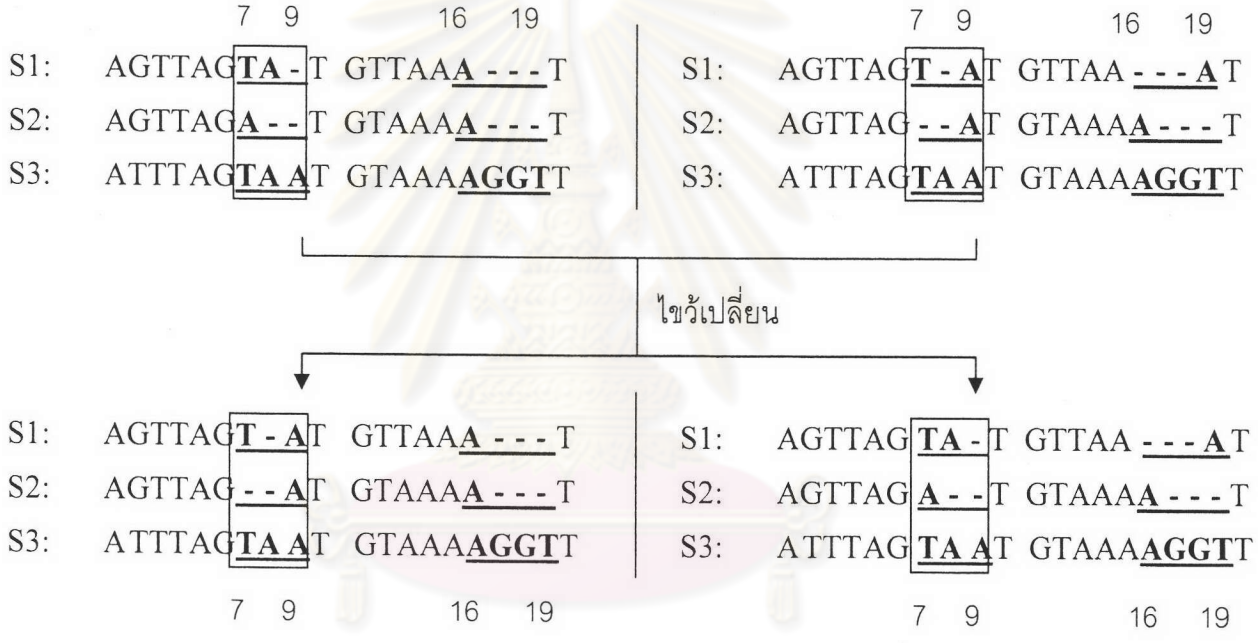
หลังการกลายพันธุ์

รูปที่4.6 แสดงการเลื่อนแถว

จากรูปที่4.6 ทำการสุ่มเลือกแถวที่มีตัวอักษรที่ติดกับแก็ป ซึ่งสุ่มได้แถวที่ 17 ต่อมาสุ่มเลือกด้านที่ต้องการเคลื่อนที่ โดยสุ่มให้เลื่อนไปทางด้านขวา จากนั้นทำการเลื่อนตัวอักษรทั้งหมดไปทางด้านขวาทั้งหมด

### 4.8.4 การไขว้เปลี่ยนแบบสองจุด

คัดเลือกผลเฉลยมา 2 ตัวแล้วทำการเปรียบเทียบหาตำแหน่งเบสที่ไม่เหมือนกันของผลเฉลยทั้งสอง จากนั้นทำการสุ่มเลือกบริเวณที่ตำแหน่งเบสที่ไม่เหมือนกัน และทำการไขว้เปลี่ยนส่วนที่เลือกไว้ ส่วนผลเฉลยที่ได้มาจะคัดเอาตัวที่ฟังก์ชันวัตถุประสงค์มากกว่า




รูปที่4.7 แสดงการไขว้เปลี่ยนแบบสองจุด

จากรูปที่4.7 ทำการเปรียบเทียบตำแหน่งเบสที่ไม่เหมือนกันของผลเฉลยทั้งสองตัว จะได้กลุ่มที่ตำแหน่งเบสไม่เหมือนกันมาสองกลุ่มคือ กลุ่มแรกที่ตำแหน่ง 7-9 และกลุ่มสองที่ตำแหน่ง 16-19 จากนั้นทำการสุ่มเลือกกลุ่มที่ได้มาในตอนแรก โดยกลุ่มที่สุ่มได้คือกลุ่มแรก ทำการไขว้เปลี่ยนชิ้นส่วนของกลุ่มแรกระหว่างผลเฉลยทั้งสองตัว ผลที่ได้จะได้ผลเฉลยใหม่มาสองตัว จากรูป 4.6จะทำการเลือกผลเฉลยตัวแรก เนื่องจากมีฟังก์ชันวัตถุประสงค์มากกว่าผลเฉลยตัวที่สอง

### 4.9 การหาผลเฉลยสุดท้าย

เนื่องจากในพาเรโตฟรอนต์มีคำตอบอยู่เป็นจำนวนมากจึงต้องทำการหาตัวผลเฉลยที่ดีและตัวผลเฉลยที่ได้จะไม่ขึ้นอยู่กับฟังก์ชันวัตถุประสงค์ตัวใดตัวหนึ่ง

ในโปรแกรม MOMSA จะนำฟังก์ชันวัตถุประสงค์ของตัวผลเฉลยในพาเรโตฟรอนต์ทุกตัวมาหาค่าเฉลี่ย ดังสมการ

$$\overline{Objective1} = \sum_{i=0}^{\overline{A}} Objective1(i) \,/\, \overline{A} \tag{4.5}$$

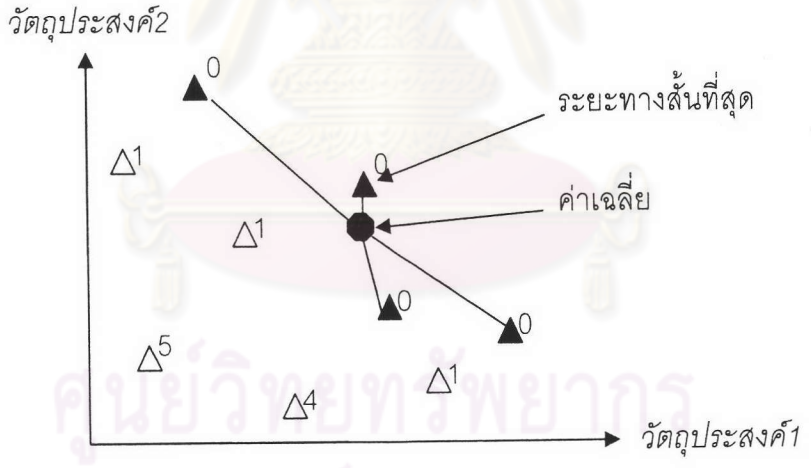
$$\overline{Objective2} = \sum_{i=0}^{\overline{A}} Objective2(i) \,/\, \overline{A} \tag{4.6}$$

เมื่อ $\overline{A}$ คือจำนวนผลเฉลยในหน่วยเก็บถาวร และ $i$ คือตำแหน่งของผลเฉลยในหน่วยเก็บถาวร

จากนั้นทำการคำนวณหาระยะทางที่ผลเฉลยแต่ละตัวห่างจากค่าเฉลี่ย ดังสมการ

$$Distance(i) = \sqrt{\left(Obj1(i) - \overline{Obj1}\right)^2 + \left(Obj2(i) - \overline{Obj2}\right)^2} \tag{4.7}$$

จากสมการ ผลเฉลยแต่ละตัวจะมีค่าระยะทางที่แตกต่างกัน และตัวที่มีค่าระยะทางน้อยที่สุดจะถูกเลือกเป็นคำตอบ




รูปที่4.8 รูปการหาผลเฉลยสุดท้าย

### 4.10 พารามิเตอร์ที่ใช้

งานวิจัยนี้กำหนดขนาดประชากรรุ่นพ่อแม่ และ ประชากรรุ่นลูกเท่ากับ 50 ตัว ใช้จำนวนรุ่นในการปรับปรุงผลเฉลย 200 รุ่น มีค่าความน่าจะเป็นในการไขว้เปลี่ยนร้อยละ 0.25 และมีค่าความน่าจะเป็นในการกลายพันธ์ร้อยละ 0.75 โดยการกลายพันธุ์ทั้ง 3 ชนิดจะมีความน่าจะเป็นในการถูกเลือกใช้เท่ากัน ฟังก์ชันวัตถุประสงค์แรกใช้สมการผลรวมคู่เบสโดยใช้ BLOSUM45 เป็นตารางตัวแทนและกำหนดค่า GOP เท่ากับ 10 และ GEP เท่ากับ 1 ส่วนฟังก์ชันวัตถุประสงค์ที่สอง

ใช้สมการผลรวมคู่เบสโดยใช้ BLOSUM45 เป็นตารางตัวแทนและกำหนดค่า GOP เท่ากับ 8 และ GEP เท่ากับ 12

การกำหนดค่า GOP และ GEP ในฟังก์ชันวัตถุประสงค์ที่หนึ่ง และฟังก์ชันวัตถุประสงค์ที่สองนำมาจากค่าวัถุประสงค์ของโปรแกรมที่ได้รับการยอมรับคือ Clustal W และ SAGA โดยโปรแกรม Clustal W กำหนดค่า GOP เท่ากับ 10 และค่า GEP เท่ากับ 1 และโปรแกรม SAGA กำหนดค่า GOP เท่ากับ 8 และค่า GEP เท่ากับ 12