

บทที่ 4

ทำไมขั้นตอนวิธีพันธุการจึงทำงานได้

บทนี้จะแสดงการพิสูจน์โดยใช้ทฤษฎีสกีมามาเพื่อแสดงให้เห็นว่าขั้นตอนวิธีพันธุการนั้นทำงานได้เป็นอย่างดีเพราะเหตุใด โดยแสดงให้เห็นถึงผลกระทบต่างๆของขบวนการต่างๆในขบวนการวิวัฒนาการคือ การคัดเลือกพันธุ การผสมพันธุ และการกลายพันธุ ที่มีต่ออัตราของการวิวัฒนาการเพื่อให้ได้สายพันธุที่เหมาะสมกับปัญหา โดยเริ่มต้นด้วยการอธิบายทฤษฎีสกีมามา และสรุปท้ายบท

4.1 ทฤษฎีสกีมามา (Schema Theorem)

ในเอกสารของ Michalewicz (1992) ได้แสดงให้เห็นถึงการทำงานของขั้นตอนวิธีพันธุการดังต่อไปนี้

เพื่อศึกษาการทำงานของขั้นตอนวิธีพันธุการ จะพิจารณา ขั้นตอนวิธีพันธุการ ในรูปแบบที่ใช้สายอักขระในการแทนความหมายของวิธีการหาคำตอบจากปัญหาที่กำหนด โดยจะอาศัย ทฤษฎีสกีมามา ซึ่งมีลักษณะเป็นสายอักขระเหมือนกันแต่มีการใช้สัญลักษณ์ "*" (don't care) ในสายอักขระด้วย ตัวสกีมานี้จะใช้ในการเลือกสายอักขระที่เข้ากับสกีมาดังกล่าวโดยเปรียบเทียบในทุกตำแหน่งของสายอักขระยกเว้นสัญลักษณ์ "*" ซึ่งไม่สนใจ(don't care)ว่าจะป็นค่าใด เช่น สกีมามาของสายอักขระที่ยาว 10 ตัวอักษร (* 0 1 0 1 1 1 0 0 1) ซึ่งเทียบได้กับ สายอักขระ[0 0 1 0 1 1 1 0 0 1] และ สายอักขระ[1 0 1 0 1 1 1 0 0 1] เป็นต้น

เป็นที่แน่นอนว่า สกีมามา (1 0 0 0 1 0 1 0 1 0) ซึ่งเทียบได้เพียง สายอักขระ[1 0 0 0 1 0 1 0 1 0] และ สกีมามา (* * * * * * * * *) ซึ่งเทียบได้กับ สายอักขระทุกเส้นที่ยาว 10 ตัวอักษร

ซึ่งมีจำนวน 2^r เส้น โดย r แทนจำนวนของสัญลักษณ์ไม่สนใจ '*' ในสกีมา หรืออีกนัยหนึ่งคือ แต่ละสายอักขระที่ยาว m ตัวอักษร จะสามารถแทนได้ด้วยสกีมาได้ 2^m แบบ

พิจารณาสายอักขระที่มีความยาว m ตัวอักษร จะมีสกีมาที่เป็นไปได้ 3^m แบบ และในประชากรที่มีขนาด n อาจจะถูกแทนด้วยสกีมาระหว่าง 2^m แบบ ถึง $n \cdot 2^m$ แบบ ซึ่งแต่ละแบบจะมีลักษณะเฉพาะตัวที่ต่างกัน จำนวนของสัญลักษณ์ไม่สนใจ '*' ในสกีมาหนึ่งๆจะเป็นตัวบอกจำนวนของสายอักขระที่สามารถเทียบกันได้ด้วยสกีมานั้นๆ

คุณสมบัติที่สำคัญของสกีมามีอยู่ 2 อย่างด้วยกันคือ ส่วนเจาะจง(Order) และ ส่วนกำหนดความยาว(Defining length) ซึ่งใช้เป็นคุณสมบัติพื้นฐานของทฤษฎีสกีมา

ส่วนเจาะจงใช้แทนด้วยสัญลักษณ์ $o(S)$ ใช้แสดงถึงความเฉพาะเจาะจง (specificity) ของสกีมาที่มีต่อสายอักขระที่เทียบได้ เป็นฟังก์ชันที่หาจำนวนของสัญลักษณ์ '0' และสัญลักษณ์ '1' ในสายอักขระ หรืออาจหาได้จาก ค่าของความยาวทั้งหมดของสายอักขระลบด้วยจำนวนของสัญลักษณ์ไม่สนใจ '*' ในสกีมา ตัวอย่างเช่น

$$S_1 = (** * 0 0 1 * 1 1 0), \quad o(S_1) = 6$$

$$S_2 = (** * * 0 0 * * 0 *), \quad o(S_2) = 3$$

$$S_3 = (1 1 1 0 1 * * 0 0 1), \quad o(S_3) = 8$$

สกีมา S_3 เป็นสกีมาที่เฉพาะเจาะจงที่สุด

ส่วนเจาะจงนี้ เป็นคุณสมบัติที่สำคัญในการคำนวณหาความน่าจะเป็นในการมีชีวิตรอดของแต่ละสกีมา ในขบวนการกลายพันธุ์

ส่วนกำหนดความยาว ใช้แทนด้วยสัญลักษณ์ $\delta(S)$ แสดงถึงความกะชับ (compactness) ของข้อมูลที่มีอยู่ในสกีมา เป็นฟังก์ชันที่หาระยะห่างระหว่างตำแหน่งอักขระที่เจาะจงตัวแรก และตำแหน่งอักขระที่เจาะจงตัวสุดท้ายในสกีมา ตัวอย่างเช่น

$$\delta(S_1) = 10 - 4 = 6$$

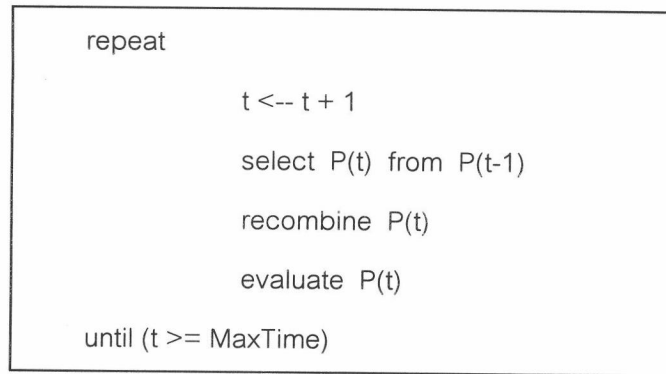
$$\delta(S_2) = 9 - 5 = 4$$

$$\delta(S_3) = 10 - 1 = 9$$

สกีมา S_2 เป็นสกีมาที่กะชับที่สุด

ในสกีมาที่มีอักขระที่เจาะจงเพียงตัวเดียว จะมีค่า $\delta(S) = 0$

ส่วนกำหนดความยาวนี้ เป็นคุณสมบัติที่สำคัญใช้ในการคำนวณหาความน่าจะเป็นในการมีชีวิตรอดของแต่ละสกีมา ในขบวนการผสมพันธุ์



รูปที่ 4.1 แสดงโครงสร้างโดยทั่วไปของขั้นตอนการวิวัฒนาการ

โดยทั่วไปขั้นตอนการวิวัฒนาการ จะมีโครงสร้างง่าย ๆ ดังรูปที่ 4.1 ซึ่ง t แทนเวลาหรือรุ่นของการวิวัฒนาการ, $P(t)$ แทนประชากรในช่วงเวลาปัจจุบันหรือรุ่นปัจจุบัน, $P(t-1)$ แทนประชากรในรุ่นที่แล้ว และ $MaxTime$ แทนเวลาที่มากที่สุดสำหรับการวิวัฒนาการ

ส่วนขั้นตอนแรกคือ $t <-- t + 1$ นั้น เป็นขั้นตอนการเพิ่มตำแหน่งเวลาของการวิวัฒนาการ

ส่วนขั้นตอนท้ายคือ $evaluate P(t)$ นั้น เป็นการประเมินผลค่าของประชากรในช่วงเวลาปัจจุบัน ซึ่งส่วนที่เป็นขั้นตอนการวิวัฒนาการจริงๆ นั้น คือ การคัดเลือกพันธุ์(selection) และการผสมพันธุ์(recombination) ซึ่งจะพิจารณาผลกระทบของแต่ละขั้นตอนต่อไป

4.2 การคัดเลือกพันธุ์

ในส่วนของการคัดเลือกพันธุ์ กำหนดให้ $\xi(S, t)$ แทนจำนวนของสายอักขระในประชากรทั้งหมดที่เทียบได้กับสกีมา S ที่เวลา t

กำหนดให้ค่าความเหมาะสมของสกีมา S ที่เวลา t แทนด้วย $eval(S, t)$ ซึ่งคำนวณได้โดย หาค่าเฉลี่ยของค่าความเหมาะสมทั้งหมดของสายอักขระในประชากรทั้งหมดที่เทียบได้กับสกีมา S ที่เวลา t ซึ่งแทนเป็นสมการได้ดังนี้

$$eval(S, t) = \sum_{i=1}^p eval(v_i) / p$$

v_i : แทนสายอักขระที่ i ซึ่งเทียบได้กับสกีมา S

p : แทนจำนวนของสายอักขระที่เทียบได้กับสกีมา S

เมื่อเวลาผ่านไป ขณะที่ขบวนการคัดเลือกกำลังดำเนินไป แต่ละสายอักขระ v_i จะมีความน่าจะเป็น $p_i = eval(v_i) / F(t)$ โดย $F(t)$ มีค่าเท่ากับค่าความเหมาะสมรวมทั้งหมดของประชากรที่เวลา t นั่นคือ $F(t) = \sum_i^{pop_size} eval(v_i)$ โดยที่ pop_size : แทนจำนวนของประชากรทั้งหมด

หลังจากจบขบวนการคัดเลือกแล้ว สามารถคาดการณ์เพื่อหาค่า $\xi(S, t+1)$ ได้จาก

$$\xi(S, t+1) = \xi(S, t) \cdot pop_size \cdot eval(S, t) / F(t)$$

พิจารณาจากค่าเฉลี่ยของค่าความเหมาะสมของประชากรทั้งหมด

$$\bar{F}(t) = F(t) / pop_size$$

จะได้สมการใหม่คือ

$$\xi(S, t+1) = \xi(S, t) \cdot eval(S, t) / \bar{F}(t) \quad \dots\dots(1)$$

จะเห็นได้ว่า จำนวนของสายอักขระในประชากรทั้งหมดที่เทียบได้กับสก็มา S ในรุ่นถัดไป จะเพิ่มขึ้นตามสัดส่วนของค่าความเหมาะสมของสก็มา ($eval(S, t)$) กับ ค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด ($\bar{F}(t)$) ซึ่งหมายความว่า ถ้าค่าความเหมาะสมของสก็มา มีค่าสูงกว่า ค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด จำนวนของสายอักขระที่ถูกคัดเลือกในรุ่นถัดไปจะมากขึ้น ในทำนองเดียวกัน ถ้าค่าความเหมาะสมของสก็มา มีค่าต่ำกว่าค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด จำนวนของสายอักขระที่ถูกคัดเลือกในรุ่นถัดไปจะน้อยลง และถ้าค่าความเหมาะสมของสก็มา มีค่าเท่ากับ ค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด จำนวนของสายอักขระที่ถูกคัดเลือกในรุ่นถัดไปจะคงจำนวนเดิม

ถ้ากำหนดให้ สก็มา S มีค่าความเหมาะสมสูงกว่าค่าเฉลี่ยอยู่ $\varepsilon\%$ โดยที่

$$eval(S, t) = \bar{F}(t) + \varepsilon \bar{F}(t)$$

จะได้

$$\xi(S, t) = \xi(S, 0)(1 + \varepsilon)^t$$

และ

$$\varepsilon = (eval(S, t) - \bar{F}(t)) / \bar{F}(t)$$

ดังนั้นถ้า $\epsilon > 0$ แสดงว่ามีค่าความเหมาะสมสูงกว่าค่าเฉลี่ย
 $\epsilon < 0$ แสดงว่ามีค่าความเหมาะสมต่ำกว่าค่าเฉลี่ย

จากสมการที่ได้นั้นแสดงให้เห็นว่า เมื่อสกีมา S มีค่าความเหมาะสมสูงกว่าค่าเฉลี่ยแล้ว จะมีจำนวนที่เพิ่มขึ้นของสายอักขระที่ถูกเลือกในรุ่นถัดไป ซึ่งไม่เพียงแต่เพิ่มขึ้นทีละเล็กละน้อยแต่เพิ่มขึ้นเป็นจำนวนมากขึ้นเป็นเอกซ์โพเนนเชียลในรุ่นถัดไป

แต่อย่างไรก็ตาม การใช้ขบวนการคัดเลือกเพียงลำพังไม่ได้ช่วยเสนอแนวทางใหม่ๆในการค้นหาคำตอบที่แทนด้วยสายอักขระใหม่ๆในประชากรเลย ขบวนการคัดเลือกเป็นเพียงแค่การคัดลอกสายอักขระบางชุดที่เทียบได้กับสกีมาในประชากรทั้งหมดเท่านั้น ขั้นตอนถัดไปในขบวนการวิวัฒนาการคือ การผสมพันธุ์ และการกลายพันธุ์

4.3 การผสมพันธุ์

สำหรับการผสมพันธุ์ พิจารณาสายอักขระที่ยาว 30 ตัวอักขระ

[1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0]

ซึ่งสามารถแทนด้วย สกีมา ได้ 3^{30} แบบ พิจารณาสกีมาต่อไปนี้

$S_0 = (****111*****)$ และ

$S_1 = (111*****10)$

สมมุติว่าสายอักขระที่เทียบได้กับสกีมา S_0 ถูกคัดเลือกเพื่อการผสมพันธุ์ ซึ่งถูกสุ่มตำแหน่งที่ใช้สลับสายอักขระ เป็นที่ตำแหน่ง 20 จะสังเกตได้ว่าสายอักขระดังกล่าวจะยังคงมีชีวิตรอดจากการขบวนการผสมพันธุ์ ซึ่งลูกหลานที่ได้มาจะยังคงเทียบได้กับสกีมา S_0 โดยจะยังคงมีอักขระตัวที่ 5,6 และ 7 ในสายอักขระในรุ่นถัดไป ดังตัวอย่างเช่น

$v_0 = [1 1 1 0 \underline{1 1 1} 1 1 0 1 0 0 0 1 0 0 0 1 0 | 0 0 0 1 0 0 0 1 1 0]$

$v_1 = [0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 1 | 0 1 1 0 1 1 1 0 1 1]$

ผสมพันธุ์แล้วจะได้

$v'_0 = [1 1 1 0 \underline{1 1 1} 1 1 0 1 0 0 0 1 0 0 0 1 0 | 0 1 1 0 1 1 1 0 1 1]$

$v'_1 = [0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 1 | 0 0 0 1 0 0 0 1 1 0]$

ในทางกลับกัน สายอักขระที่เทียบได้กับสกีมา S_1 จะไม่มีชีวิตรอดจากการผสมพันธุ์ ซึ่งลูกหลานที่ได้มาจะไม่สามารถเทียบได้กับสกีมา S_1 ทั้งนี้เป็นเพราะอักขระที่ถูกเจาะจง '1 1 1' ซึ่งอยู่ที่ตำแหน่งเริ่มต้นของสกีมา และอักขระที่ถูกเจาะจง '0 1' ซึ่งอยู่ที่ตำแหน่งสุดท้าย

ท้ายของสกีมา ถูกเปลี่ยนตำแหน่งเมื่อเป็นสายอักขระของลูกหลานที่เทียบไม่ได้กับสกีมา S_I ดังตัวอย่างเช่น

$$v_0 = [11101111101000100010|0001000110]$$

$$v_I = [00010100001000100101|0110111011]$$

ผสมพันธุ์แล้วจะได้

$$v'_0 = [11101111101000100010|0110111011]$$

$$v'_I = [00010100001000100101|0001000110]$$

ค่าส่วนกำหนดความยาวของสกีมา S_0 , $\delta(S_0)=2$ และ ค่าส่วนกำหนดความยาวของสกีมา S_I , $\delta(S_I)=29$ ทำให้เข้าใจได้ว่า ค่าส่วนกำหนดความยาวของแต่ละสกีมา จะแสดงถึงความน่าจะเป็นของการมีชีวิตรอดหรือไม่ของสกีมาที่มีต่อสายอักขระที่เทียบได้กับสกีมานั้นๆ

โดยทั่วไปแล้วการผสมพันธุ์ เป็นการเลือกสุ่มจากจำนวน $m-1$ ที่เป็นไปได้ของตำแหน่งในสายอักขระ (m แทนความยาวสายอักขระ) จึงสามารถเขียนค่าความน่าจะเป็นของการทำลาย(destruction)ของสกีมา S ได้ดังนี้

$$p_d(S) = \frac{\delta(S)}{m-1}$$

และในทำนองเดียวกัน ค่าความน่าจะเป็นของการมีชีวิตรอด(survial)ของสกีมา S ได้ดังนี้

$$p_s(S) = 1 - \frac{\delta(S)}{m-1}$$

จากตัวอย่างข้างต้นจะได้

$$p_d(S_0) = 2/32, \quad p_s(S_0) = 30/32, \quad p_d(S_I) = 32/32 = 1, \quad p_s(S_I) = 0$$

ดังนั้นผลลัพธ์ที่ได้จากขบวนการผสมพันธุ์จะสามารถคาดการณ์ได้

สิ่งที่สำคัญคือ สายอักขระที่ถูกเลือกในขบวนการผสมพันธุ์นั้นยังขึ้นอยู่กับความน่าจะเป็นในการคัดเลือกเพื่อผสมพันธุ์ p_c จะได้ค่าความน่าจะเป็นของการมีชีวิตรอดของสกีมา S ใหม่ดังนี้

$$p_s(S) = 1 - p_c \cdot \frac{\delta(S)}{m-1}$$

ในกรณีที่ตำแหน่งของการผสมพันธุ์ เป็นตำแหน่งที่อักขระถูกเจาะจงในสกีมา ซึ่งมีโอกาสที่จะมีชีวิตรอดได้ แต่มีโอกาสน้อยมาก ดังนั้นค่าความน่าจะเป็นของการมีชีวิตรอดของสกีมา S สามารถเขียนใหม่ได้คือ

$$p_s(S) \geq 1 - p_c \cdot \frac{\delta(S)}{m-1}$$

จากสมการ (1) เมื่อรวมผลกระทบ ของทั้งขบวนการคัดเลือก และการผสมพันธุ์ จะได้สมการใหม่คือ

$$\xi(S, t+1) = \xi(S, t) \cdot \text{eval}(S, t) / \bar{F}(t) \left[1 - p_c \cdot \frac{\delta(S)}{m-1} \right] \dots\dots(2)$$

4.4 การกลายพันธุ์

สำหรับการกลายพันธุ์ ซึ่งเป็นการสุ่มเพื่อสลับค่าของอักขระที่ตำแหน่งเดียวด้วยความน่าจะเป็น p_m โดยสลับจาก '1' เป็น '0' หรือ '0' เป็น '1' ซึ่งการที่จะมีชีวิตรอดจากการกลายพันธุ์ได้นั้น อักขระที่ถูกเจาะจงในสกีมาจะต้องไม่ถูกสลับค่า พิจารณาจากตัวอย่างของสายอักขระยาว 30 ตัวอักษร

[1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0]

ซึ่งสามารถแทนด้วย สกีมา ได้ 3^{30} แบบ พิจารณาสกีมาต่อไปนี้

$S_0 = (****111*****)$

สมมติให้สายอักขระที่เทียบได้กับสกีมา S_0 ถูกคัดเลือกเพื่อการกลายพันธุ์ โดยสุ่มตำแหน่งที่จะทำการสลับค่า เป็นตำแหน่งที่ 8 ได้สายอักขระในรุ่นถัดมาคือ

[1 1 1 0 1 1 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0]

ซึ่งผลลัพธ์ที่ได้สายอักขระยังคงเทียบได้กับสกีมา S_0 ถ้าการเลือกตำแหน่งที่จะทำการสลับค่า เป็นตำแหน่งที่ 1 - 4 หรือ 8 - 30 ผลลัพธ์จะได้จะยังคงเทียบได้กับสกีมา S_0 มีเพียงตำแหน่งที่ 5 - 7 เท่านั้นที่จะทำลายสกีมา

ส่วนที่สำคัญคือ การสลับค่าอย่างน้อยเพียงตำแหน่งเดียว ในตำแหน่งของอักขระที่ถูกเจาะจงในสกีมา ก็สามารถทำลายสกีมานั้นได้ ซึ่งมีค่าเท่ากับค่าส่วนเจาะจงของสกีมาคือจำนวนของอักขระที่เจาะจงในสกีมานั้นเอง

เพราะว่าค่าความน่าจะเป็นของการเลือกตำแหน่งในการกลายพันธุ์ คือ p_m ค่าความน่าจะเป็นของการมีชีวิตรอดสำหรับหนึ่งอักขระ คือ $1 - p_m$ ซึ่งแต่ละขบวนการกลายพันธุ์จะไม่ขึ้นต่อกัน ดังนั้น ค่าความน่าจะเป็นของการมีชีวิตรอดของสก็มา S สำหรับการกลายพันธุ์คือ

$$p_s(S) = (1 - p_m)^{o(S)}$$

แต่ค่า p_m มีค่าน้อยมาก สามารถประมาณได้โดย

$$p_s(S) \approx 1 - o(S) \cdot p_m$$

จากสมการ (2) เมื่อรวมผลกระทบของทั้งขบวนการคัดเลือก การผสมพันธุ์ และการกลายพันธุ์ จะได้สมการใหม่คือ

$$\xi(S, t+1) = \xi(S, t) \cdot eval(S, t) / \bar{F}(t) \left[1 - p_c \cdot \frac{\delta(S)}{m-1} - o(S) \cdot p_m \right] \dots\dots(3)$$

4.5 สรุปท้ายบท

สมการ (1) (2) และ (3) สามารถบอกได้ถึงจำนวนที่คาดหวังของสายอักขระที่เทียบได้กับสก็มา S ในรุ่นถัดไป ซึ่งเกี่ยวข้องกับ จำนวนของสายอักขระที่เทียบได้กับสก็มา , ค่าความเหมาะสมของสก็มา , ค่าส่วนกำหนดความยาว และ ค่าส่วนเจาะจง

สมการทั้ง 3 ตั้งอยู่บนสมมุติฐานที่ว่าทุกฟังก์ชันให้ค่าความเหมาะสมที่เป็นบวก

โดยสรุปแล้ว สมการของการเติบโต(The growth equation) แสดงให้เห็นว่าการคัดเลือกทำให้สก็มา S มีค่าความเหมาะสมสูงกว่าค่าเฉลี่ย มีอัตราการเพิ่มขึ้นอย่างมาก อัตราที่เพิ่มขึ้นไม่ทำให้เกิดสก็มาใหม่ๆ แต่เป็นผลมาจากขบวนการผสมพันธุ์ ที่ทำให้เกิดการแลกเปลี่ยนข้อมูลแบบสุ่ม และขบวนการกลายพันธุ์ ที่ทำให้เกิดการเปลี่ยนแปลงครั้งใหญ่ในประชากร เมื่อรวมคุณสมบัติทั้งสามมาใช้ จะทำให้เกิดผลลัพธ์ ถ้าสก็มาที่มีค่าความเหมาะสมสูงกว่าค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด โดยมีค่าส่วนกำหนดความยาวสั้น และ มีค่าส่วนเจาะจงต่ำ ทำให้การวิวัฒนาการสู่เป้าหมายที่สัมพันธ์กับค่าความเหมาะสมที่ได้กำหนดไว้ยังคงเพิ่มขึ้นในอัตราที่สูงเป็นเอกซ์โพเนนเชียล

จากทั้งหมดสามารถสรุปได้ตามทฤษฎีสก็มาได้ว่า

“สก็มาที่มีค่าความเหมาะสมสูงกว่าค่าเฉลี่ยของค่าความเหมาะสมในประชากรทั้งหมด มีค่าส่วนกำหนดความยาวสั้น และ มีค่าส่วนเจาะจงต่ำ จะทำให้การวิวัฒนาการยังคงเพิ่มขึ้นในอัตราที่สูงเป็นเอกซ์โพเนนเชียล ในแต่ละรุ่นของขั้นตอนวิธีพันธุการ”