공학석사학위논문

다중 목적 진화 연산을 위한 Dominance 기반의 개체 선택 방법

Dominance-Based Selection for Multi-Objective Evolutionary Algorithms

2005년 2월

서울대학교 대학원
컴퓨터공학부

신 기 루

다중 목적 진화 연산을 위한 Dominance 기반의 개체 선택 방법

Dominance-Based Selection for Multi-Objective Evolutionary Algorithms

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2004년 10월

서울대학교 대학원

컴퓨터공학부

신 기 루

신기루의 공학석사 학위논문을 인준함

2004년 12월

위원장	유 석 인	印
부위원장	장 병 탁	印
위원	문 병 로	印

초 록

다중 목적 진화 연산(multi-objective evolutionary algorithms)이란 여러 개의 최적화하고자 하는 목적 함수를 가지는 문제를 해결하는 진화 연산을 말한다. 진화 연산은 자연 생태계의 진화의 개념을 채용하여 해집합(population)을 다루어 좋은 성능의 해들로 진화시켜 나가는 탐색 기법이다. 여러 개의 목적 함수를 가지는 경우, 해의 좋고 나쁨을 판별하는 기준과 해집합을 다루는 효율적인 기법들이 많이 알려져 있다. 그러나 더 이상 이러한 다중 목적 진화 연산의 전체 구조에 대한 연구는 주목할만한 발전을 보이지 않고 있다. 이러한 문제를 해결하는 돌파구로서, 진화적 탐색을 일으키는 방향을 결정할 수 있는 개체 선택 방법에 대한 연구가 대안일 수 있다.

본 논문에서는 부모 개체 간의 dominance 관계를 고려한 개체 선택 방법을 제시한다. 성능을 측정하기 위하여 잘 알려져 있는, NP-hard 문제군의 일종인 multiple knapsack 문제를 가지고 실험하여 선택 방법에 따른 결과를 비교한다.

또한 실제 응용 사례로서, 생물학적 문제인 oligonucleotide probe 서열 선택 문제에도 적용시켜 본다. 실례로서, 자궁경부암을 일으키는 19개 HPV(Human Papilloma Virus) 유전자를 목표로 하는 probe 서열을 선택하는 문제에 적용시켜 본다. 전체 유전자 서열 중에서 생물학적으로 HPV의 유일한 특성을 결정짓는 것으로 알려진, 길이 약 150bp 영역의 서열을 가지고 길이 30bp의 probe 서열들을 선택하는 데에 제안한 선택 방법을 적용시켜 얻어진 결과를 제시한다.

위의 실험들을 통해, dominance 기반의 개체 선택 방법을 적용하여 다양성을 유지하며 진화 연산을 수행할 수 있다는 결론을 내릴 수 있었다. 이러한 다양성은 최종 해집합에서 선택할 수 있는 폭을 넓혀줄 뿐만 아니라, 다양한 탐색 공간을 보존함에 따라 수렴하는 해의 품질이 향상될 가능성을 가져온다.

주요어 : 다중 목적, 최적화, 진화 연산, 개체 선택
학 번 : 2003-21613

목 차

1장

서론

1.1 연구 배경

실세계에 존재하는 대다수의 문제는 효율적인 알고리즘이 존재하지 않거나 알려져 있지 않다. 이러한 문제를 해결하는 데에 다양한 근사 알고리즘들이 사용되지만 그다지 좋은 성능을 보여주지 못한다. 하나의 해결책으로 사용되는 것이 진화 연산(evolutionary algorithms)이다. 진화 연산은 특정 알고리즘이 아니라 커다란 탐색 방법의 골격을 제시해주고 있다. 진화 연산은 생물학적 진화의 개념을 모티브로 하고 있다. 기본적인 진화 연산에서 사용되는 해의 좋고 나쁨의 척도로서 일반적으로 하나의 적합도 함수(fitness function)를 사용한다. 단일 목적(single-objective) 함수를 최적화하는 진화 연산에 대한 많은 연구가 이미 이루어져 왔으며, 효율적인 알고리즘들이 알려져 있다.

그러나 실세계에서 부딪히는 대다수의 문제는 여러 개의 목적을 가지며, 이들을 동시에 최적화하고자 하는 다중 목적 최적화(multi-objective optimization) 문제들이다. 최대화 또는 최소화시키고자 하는 목적 함수가 둘 이상인 문제를 말한다. 해의 좋고 나쁨을 판단하는 기준이 여럿이므로, 가장 좋은 해는 일반적으로 하나가 아닌 여러 수의 해집합이 된다. 즉, 여러 목적 함수들 사이의 이

득과 손실의 균형(tradeoff)이 근본적으로 존재하게 되는 것이다. 다중 목적 최적화 문제의 이러한 특성 때문에, 해집합(population)을 가지고 문제를 해결해 나가는 접근 방식인 진화 연산이 자연스럽게 좋은 해결 방안이 될 수 있다. 최종적으로 구해진 여러 해들을 제시하고, 사용자가 필요에 맞는 해를 선택할 수 있도록 한다.

일반적인 문제에 대해 좋은 성능을 보이는 다중 목적 진화 연산의 몇 가지 기본 골격이 알려져 있다. 이러한 잘 알려진 알고리즘들이 제시해주는 것은 해의 좋고 나쁨의 판단 기준을 어떻게 할 것인가, 전체 해집합을 어떻게 운영해 나가면서 진화를 이루어 나가는 가에 대한 다중 목적 진화 연산의 전체 구조에 집중되고 있다. 이러한 전체 골격에 대한 연구는 더 이상 주목할만한 발전이 이루어지지 않고 있으며, 근본적으로 다른 방향의 연구가 절실히 요구되고 있다.

본 논문에서는 이러한 취지 하에 다중 목적 진화 연산의 선택 방법에 대해 연구하였다. 두 부모 개체 간의 dominance 관계를 고려하는 선택 방법을 제안하고, 이를 multiple knapsack 문제에 적용시켜 실험한 결과를 제시한다. 또한, 실세계의 생물학적 문제인 oligonucleotide probe 집합 선택 문제에 적용한 응용 사례를 제시한다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 2장에서는 다중 목적 최적화에 대해 간략히 살펴보고, 본 논문에서 사용하고 있는 대표적인 다중 목적 진화 연산의 구조와 최근의 관련 연구에 대해 알아본다. 3장에서는 본 논문에서 제시한 개체 선택 방법과 그 알고리즘을 설명한다. 4장에서는 제안한 선택 방법을 multiple knapsack 문제에 적용하여 실험한 결과를 제시하고 분석한다. 5장에서는 다중 목적 최적화의 특정 응용 사례로서, 생물학 영역의 문제인 Oligonucleotide Probe 서열 선택 문제를 해결하는 데에 제안한 선택 방법을 적용시켜 본다. 6장은 결론으로서, 본 논문의 연구 내용을 요약하고 앞으로의 연구 과제를 제시하며 끝을 맺는다.

2장

이론적 배경

2.1 다중 목적 진화 연산 (Multi-Objective Evolutionary Algorithms)

진화 연산(evolutionary algorithm)은 일종의 탐색 기법으로서, 생물학적 진화의 개념을 계산학의 영역으로 가져다 문제 해결의 방법 모델로 사용하는 접근 방법이다. 이러한 접근 방법은, 효율적인 문제 해결 방법이 알려지지 않은 문제에 유용하게 적용되고 있다. 계산학의 대표적인 난제인 NP-hard 문제군의 최적화 문제 해결에 사용되는 것이 그 예이다.

다중 목적 진화 연산이란 다중 목적 최적화(multi-objective optimization)를 위한 진화 연산을 뜻한다. 즉, 최적화하고자 하는 목적 함수가 둘 이상인 문제를 해결하는 데에 사용되는 진화 연산을 말한다. 좋은 성능을 나타내는 것으로 알려진 다중 목적 진화 연산으로는 다음과 같은 알고리즘들이 있다. PAES [Knowles et al., 1999], NSGA-II [Deb et al., 2000], SPEA2 [Zitzler et al., 2001], PESA-II [Corne et al., 2001]. 이러한 알고리즘들은 해집합을 다루어 나가는 전략과 해집합에서 품질의 우위를 측정하는 방법 등의 진화 연산의 커다란 기본 골격을 제공한다.

2.2 관련 연구

2.2.1 A Fast Elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II) [Deb et al., 2000]

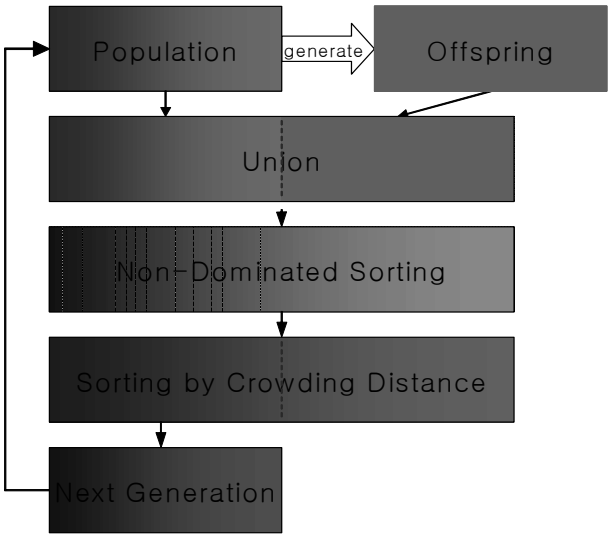

그림 1. NSGA-II 전체 구조도

대표적으로 잘 알려진 다중 목적 진화 연산으로서, 간략히 NSGA-II를 살펴보겠다. 전체적으로 N개의 개체군(population)을 가지고 진화시켜 나간다. 하나의 세대를 거치는 과정을 보면, N개의 개체군에서 tournament 선택 방법을 이용하여 부모를 선택하여 N개의 자식군을 생성한다. 전체 2N개의 개체를 가지고 차례대로 nondominated set을 구해나가는 방식으로 정렬하여 순위(ranking)을 매기고, 같은 nondominated set 안에서의 개체 주위의 밀도를 근사하는 crowding distance를 계산하여 이를 가지고 그 안에서 재정렬한다. 상위 N개의 개체를 다음 세대의 부모 개체군으로 이용하는 방식으로 세대를 거듭한다.

2.2.2 Dominance Based Crossover Operator for Evolutionary Multi-objective Algorithms [Rudenko et al., 2004]

이 논문에서 제안하는 교차 연산의 핵심은 dominance 관계를 가지는 부모에 대해 우선하여 선택함에 있다. 본 논문의 기본 생각과 같은 개념이다. 이들은 잘 알려진 실수 공간의 시험 목적 함수들에 대해 실험한 결과를 제시하고 있다. 두 개의 목적 함수를 가지는 인위적 목적 함수 ZDT1, ZDT2, ZDT3 [Zitzler et al., 2000]에 대해 실험한 결과를 제시하고 있다. 세가지 각기 다른 특성의 함수에 대하여 비교 실험한 결과를 가지고, dominance 관계를 가지는 부모를 선택하여 교차 연산을 수행함에 따라 좀더 빠른 수렴 속도를 얻을 수 있다고 결론내리고 있다.

$$f_1(x) = x_1, \qquad f_2(x) = g(x_2, \ldots, x_m)h(f_1(x), g(x_2, \ldots, x_m))$$

ZDT1: convex Pareto-optimal front

$$g(x_2, \ldots, x_m) = 1 + 9\sum_{i=2}^{m}\frac{x_i}{(m-1)}, \qquad h(f_1, g) = 1 - \sqrt{\frac{f_1}{g}}$$

ZDT2: nonconvex Pareto-optimal front

$$g(x_2, \ldots, x_m) = 1 + 9\sum_{i=2}^{m}\frac{x_i}{(m-1)}, \qquad h(f_1, g) = 1 - (\frac{f_1}{g})^2$$

ZDT3: noncontiguous convex Pareto-optimal front

$$g(x_2, \ldots, x_m) = 1 + 9\sum_{i=2}^{m}\frac{x_i}{(m-1)},$$

$$h(f_1, g) = 1 - \sqrt{\frac{f_1}{g}} - \frac{f_1}{g}sin(10\Pi f_1)$$

3장

제안 알고리즘

3.1 제안한 선택 방법

본 논문에서 제안하는 선택 방법의 기본 개념은 두 부모를 선택하는 과정에서 부모 간의 dominance 관계가 존재하는 쌍에 우선권을 주어, 이러한 부모 쌍이 선택되는 비율을 높이자는 것이다.

dominance 관계란 하나의 해가 다른 해를 dominate 하는 관계를 말한다. 모든 목적 함수에 대해 좋거나 같고, 하나 이상의 목적 함수에 대해 좋은 경우에 'dominate' 한다고 말한다. 일반적인 tournament 선택 방법을 통해 선택되는 부모 개체들을 살펴보면, 하나의 부모가 다른 쪽 부모를 dominate 하는 쌍이 선택되는 비율은 일반적으로 매우 낮다.

본 논문에서 주장하고자 하는 가설은 다음과 같다. 서로 dominate 하는 관계의 부모를 선택하여 자손을 조합해 나가는 경우에 진화해 나가는 공간 탐색의 특성은 그렇지 않은 경우와는 서로 다른 성향을 가질 것이며, 이러한 성향의 부모 비율을 조절함으로써 다중 목적 진화 연산의 좀 더 나은 성능을 얻을 수 있다는 것이다. 구현한 기본 알고리즘을 설명하자면 다음과 같다.

1. 일반적인 tournament selection을 사용하여 하나의 부모를 선택한다.
2. tournament selection을 수행하는 과정 중에 1에서 선택된 부모와 dominance 관계를 가지는 부모가 발견되면, 그러한 부모들 중에서 tournament selection을 수행하여 다른 하나의 부모를 선택한다. 그렇지 않다면 1과 마찬가지로, 일반적인 tournament selection을 사용하는 경우와 동일하다.

dominate 관계를 가지는 부모를 발견할 확률은 tournament size가 커짐에 따라 함께 커지게 된다. 즉, tournament size가 전체 진화 연산의 선택압(selection pressure)에 영향을 주고, 동시에 dominate 관계를 가지는 부모를 선택하는 확률에도 영향을 주게 되는 것이다. 그러므로 이러한 단순한 방식으로 제안한 선택 방법을 적용하는 경우에는 신중하게 tournament size를 결정해야 한다. 4장의 실험에서 살펴보겠지만 너무 부족하거나 과도하게 dominate 관계의 부모를 선택하게 되면, 기대하는 성능 향상을 얻지 못하는 경우도 발생하게 된다.

선택압에 영향을 주지 않으면서 tournament size를 조절하지 않고 이러한 단점을 극복하는 방법으로는 다음과 같은 방법들을 생각해볼 수 있겠다. dominate하는 관계의 부모 쌍이 선택되는 비율이 과도한 경우에는 일정한 비율로 위와 같은 알고리즘을 적용하거나, 또는 부족한 경우에는 일정 회수동안 dominate 관계의 부모 쌍을 찾도록 위의 알고리즘을 반복하는 등의 방법들을 통해 dominate하는 관계의 부모 쌍을 선택하는 비율을 적절하게 조절할 수 있겠다. 이에 대한 고찰은 4장의 실험 결과에서 다시 한 번 살펴보도록 하겠다.

4장

실험 및 분석

4.1 문제 정의

4.1.1 Multiple Knapsack 문제

k 개의 knapsack과 n 개의 item에 대한 multiple knapsack 문제를 설명하면 다음과 같다. n 개의 item중에서 선택한 같은 집합을, k 개의 knapsack에 담아서 얻게 되는 각각의 이득합을 최대화하는 문제이다. 단, 제약이 존재하는데 각 knapsack은 최대로 담을 수 있는 무게가 정해져 있어서 무게의 합이 이를 넘지 않는 범위 내에서 item을 담아야 한다.

각 i 번째 item에 대하여 x_i 는 그것이 선택되었는지를 나타내며, 선택된 경우에는 1의 값을, 선택되지 않았을 경우에는 0의 값을 가진다. 즉, 하나의 vector $\boldsymbol{x} = \{x_1, \ldots, x_k\}$가 하나의 해를 나타내게 된다. 각각의 i 번째 item은 각 j 번째 knapsack에 대하여 선택되었을 경우 얻게 되는 이득(profit)을 나타내는 p_{ij} , knapsack에 더해지는 무게(weight)를 나타내는 w_{ij} 값을 가진다. 즉, 같은 item이라도 서로 다른 knapsack에 대한 이득과 무게는 서로 같지 않고 상관 관계를 가지지 않는 서로 다른 값을 가진다.

각각의 j 번째 knapsack은 허용되는 최대 무게 용량(capacity) c_j 를 가지며, 선택된 item들의 j 번째 knapsack에 대한 무게의 합이 이를 넘어서는 안 된다. 수식으로 표현하자면 다음과 같다.

$$\sum_{i=1}^{n} w_{ij} x_i \leqq c_j \ , \qquad j = 1, \ldots, k \ . \tag{1}$$

j 번째 knapsack에서 얻게 되는 전체 이득 $f_j(\boldsymbol{x})$는

$$f_j(\boldsymbol{x}) = \sum_{i=1}^{n} p_{ij} x_i \ , \qquad j = 1, \ldots, k \ , \tag{2}$$

이 되며, $f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})$ 가 최대화하고자 하는 목적 함수들이 된다.

4.1.2 문제 집합

본 논문에서는 SPEA 그룹에서 제시한 문제 집합[Zitzler et al., 1999a]을 그대로 가져다 사용하였다. 위의 문제 집합들은 2,3,4개의 knapsack에 대하여 100,250,500,750개의 item을 가지고 이득들을 최대가 되도록 선택하는, 총 12개의 multiple knapsack 문제 집합들이다. 하나의 item이 각각의 knapsack에 대해 가지는 profit, weight는 [10,100] 구간의 정수들 중에 균등하게 임의로 선택하여 정해져 있으며, 각각의 knapsack의 capacity는 자신에 대해 정해진, 모든 item들의 weight들을 다 더한 총합의 1/2로 정해져 있다. 이러한 조건 상에서 knapsack의 capacity를 넘지 않고 선택할 수 있는 최적해의 item 개수는 대략 전체 item 개수의 절반 정도가 된다 [Martello et al., 1990].

앞으로 '(knapsack의 개수)-(item의 개수)KN'으로써 위의 문제 집합들을 표시하도록 하겠다. 본 논문의 실험에서 사용한 문제 집합은 2-250KN, 2-500KN 문제들이다. 2-100KN 문제와 함께, 이 두 개의 문제에 대한 실제 Pareto 최적해 집합은 알려져 있다. [Zitzler et al., http://www.tik.ee.ethz.ch/~zitzler/testdata.html/]

4.2 성능 평가 기준

4.2.1 Two Set Coverage (C)

두 개의 nondominated set A, B가 주어진 경우, B의 해 중에서 A에 의해 dominate 되는 비율을 나타내는 평가 기준인 Two Set Coverage (C)이다 [Zitzler, 1999a]. 수식으로 나타내면 다음과 같다.

$$C(A, B) = \frac{\{b \in B | \exists a \in A : a \preceq b\}}{|B|}. \qquad (3)$$

수식 $a \preceq b$ 는 두 개체 사이의 우월 관계(dominance relation)를 뜻하며, 개체 a 가 또다른 개체 b 를 dominate함을 나타낸다. 다음 두 조건을 만족할 때 a 가 b 를 dominate 한다고 말한다. 조건 1, 2는 주어진 최적화 문제가 최대화 문제임을 가정하고 있으며, M은 목적 함수의 개수이다.

1. $f_j(a) \geqq f_j(b)$ for all $j = 1, \ldots, M$.
2. $f_k(a) > f_k(b)$ for at least one $k \in \{1, \ldots, M\}$.

의미를 풀어 설명하자면, 개체 a 의 모든 목적 함수가 b 보다 낫거나 같고, b보다 나은 a의 목적 함수가 적어도 하나 이상이 존재할 때 'a 는 b 를 dominate 한다' 또는 'b 는 a에 의해 dominated 된다'라고 말한다. 다른 모든 해들에 의해서 dominated되지 않는 해집합을 nondominated set이라고 부른다. 가능한 모든 해들에 대해 이를 만족하는 경우에 Pareto-optimal front라고 부르며, 이는 다중 목적 진화 연산이 찾고자 목표로 하는 최적의 해집합을 뜻한다.

4.2.2 Generational Distance (GD)

Generational Distance [Van Veldhuizen, 1999]는 실제 Pareto-optimal front 또는 기준이 될 만한 해집합이 주어진 경우에 사용할 수 있는 평가 기준이다. 수식으로 표현하면 다음과 같다.

$$GD = \frac{(\sum_{i=1}^{n} d_i^p)^{1/p}}{n}, \qquad (4)$$

n 은 측정된 nondominated set 에 속하는 해의 개수이며, d_i 는 각각의 해에서 알고 있는 실제 Pareto Front 까지의 가장 짧은 거리이다. 파라미터 p의 값이 1이면 Hamming distance, 2이면 Euclidean distance 값을 사용하게 된다.

Generational Distance (GD) 값은 구해진 nondominated set에 속하는 모든 해들에 대하여, 실제 최적해까지의 가장 가까운 평균 거리를 뜻하므로, GD 값이 작을수록 좀더 최적해에 가까운 품질 좋은 해를 찾았음을 뜻하고, 이는 수렴성(convergency)을 나타내는 기준이 된다.

그러나, 단 하나의 해를 찾았어도 최적해에 가깝다면 GD 값이 작아지므로 얼마나 다양한 해를 가지는가를 나타내는 적절한 평가 기준과 함께 사용되어야 하겠다.

4.2.3 $D1_R$ Measure

$D1_R$ measure[Czyzak and Jaszkiewicz, 1998]는 GD에 상대적으로 반대되는 개념의 평가 기준이다. 의미를 살펴보면, 알고 있는 실제 Pareto-optimal front 혹은 기준으로 삼고 있는 해집합의 모든 해들로부터 측정된 nondominated set까지의 가장 짧은 거리들의 평균이다.

$$D1_R(A, \Lambda) = \frac{1}{R} \sum_{r \in R} min_{a \in A} \{ d(r, a) \}, \qquad (5)$$

A는 측정된 nondominated set을, R은 기준으로 삼은 해집합을 뜻한다. $d(r, a)$는 GD와 마찬가지로 문제 특성에 맞는 거리 단위를 사용하면 되겠다. 본 논문에서는 모두 Euclidean distance를 사용하여 평가하였다.

물론 측정된 nondominated set이 실제 최적해에 가까울수록 $D1_R$ 값이 작아지겠지만, GD값과 같이 단지 최적해에 가까이 수렴할수록 작은 값을 가지는 것은 아니다. 구해진 nondominated set이 알려진 최적해들 전체를 적절하게 잘 반영하면서 가까워야 $D1_R$값이 작아진다. 즉, $D1_R$ measure는 측정된 해집합의 수렴성뿐만 아니라 다양성(diversity)도 함께 보여주는 평가 기준이 된다.

4.3 실험 결과

4.3.1 실험 방법

본 논문에서 실험에 사용한 진화 연산의 기본 골격은 NSGA-II [Deb et al., 2000]을 채택하였다. 본 논문에서 실험하고자 하는 바는 다중 목적 진화 연산의 전체 구조를 변화시킴으로써 성능을 향상시키는 목적이 아니라, 같은 뼈대 위에서 선택 방법을 달리 하여 얻을 수 있는 성능 향상을 보이고자 함이다. 그러므로 일반적으로 좋은 성능을 나타내는 것으로 알려진 어떠한 알고리즘을 사용해도, 개체 선택 방법의 변화에 따른 성능 향상을 보이는 데에는 무방할 것으로 판단된다. NSGA-II는 대다수의 일반적인 문제에서 좋은 성능을 보이는 것으로 알려져 있다 [Zitzler et al., 2000, 2001].

실제 Pareto 최적해가 알려진 2-250KN, 2-500KN 문제들에 대하여 보통의 tournament 선택 방법과 본 논문에서 제안한 선택 방법을 사용하여 각각 50회씩 실험하여 평가하였다. 성능 평가의 기준으로는 4장 2절에서 언급한 GD, $D1_R$ 값을 사용하였다.

그림 2-1에서 2-8까지 2-250KN 문제에 대해 해집합의 크기를 256,512로, tournament 크기는 4,8로 변화시키면서 실험한 결과를 제시하였다. 일반적인 tournament 선택 방법과 본 논문에서 제안한 dominance 기반의 선택 방법에 대하여 각각 50회 실험한 GD, $D1_R$ 평균값이 세대를 거듭함에 따른 변화하는 과정와 그의 오차 범위-최대값, 최소값, 표준 편차 구간-를 정리하였다.

그림 3-1과 3-2에서는 가장 작은 GD 값을 가지는 경우와 가장 작은 $D1_R$ 값을 가지는 경우에 대하여 두 가지 선택 방법을 비교하여, 최종적으로 얻어진 nondominated set들을 같은 그래프에 그려서 한 눈에 비교할 수 있도록 하였다.

그림 4-1과 4-2에서는 2-500KN 문제에 대한 실험 결과를 보여준다. 2-500KN 문제에 대해서는 해집합의 크기는 512, tournament 크기는 4인 경우에 대해서만 실험한 결과를 제시하였다.

실험에 사용한 환경 변수들은 다음과 같다. 교차 연산, 변이 연산 모두 uniform하게 적용시켜 사용하였다. 교차 연산을 적용하는 확률은 0.8, 변이 연산을 적용하는 확률은 (1/유전자의 길이)을 사용하여 변이가 일어날 유전자 개수의 기대치를 1로 설정하였다.

4.3.2 실험 결과

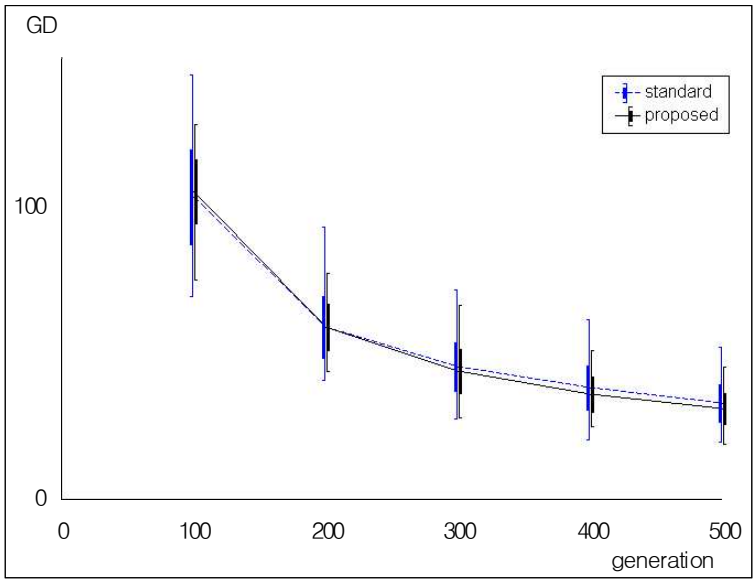

그림 2-1. 2-250KN pop=256, ts=4 GD 비교

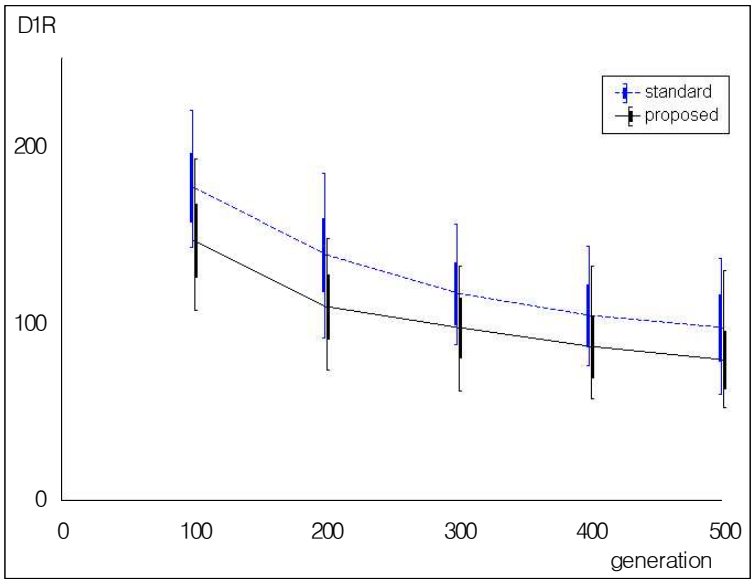

그림 2-2. 2-250KN pop=256, ts=4 $D1_R$ 비교

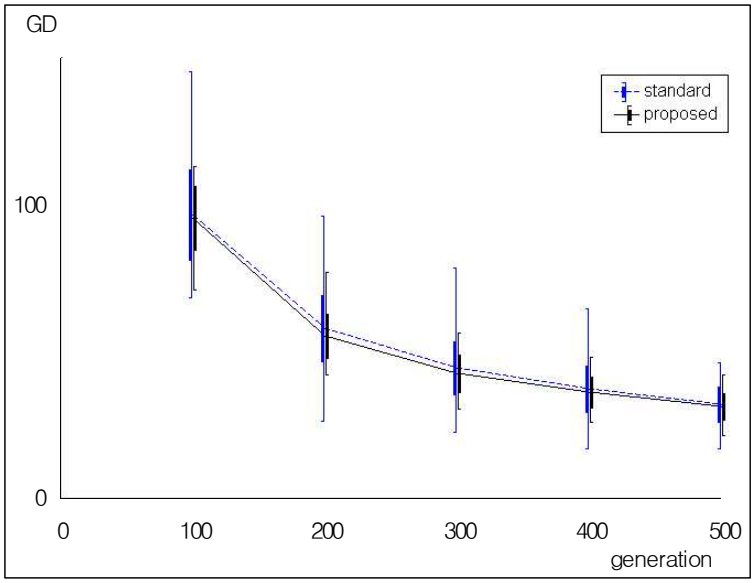

그림 2-3. 2-250KN pop=256, ts=8 GD 비교

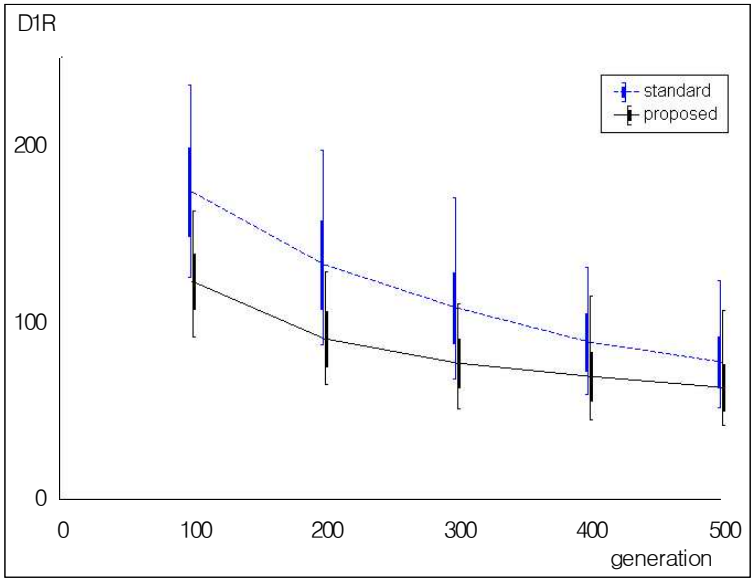

그림 2-4. 2-250KN pop=256, ts=8 $D1_R$ 비교

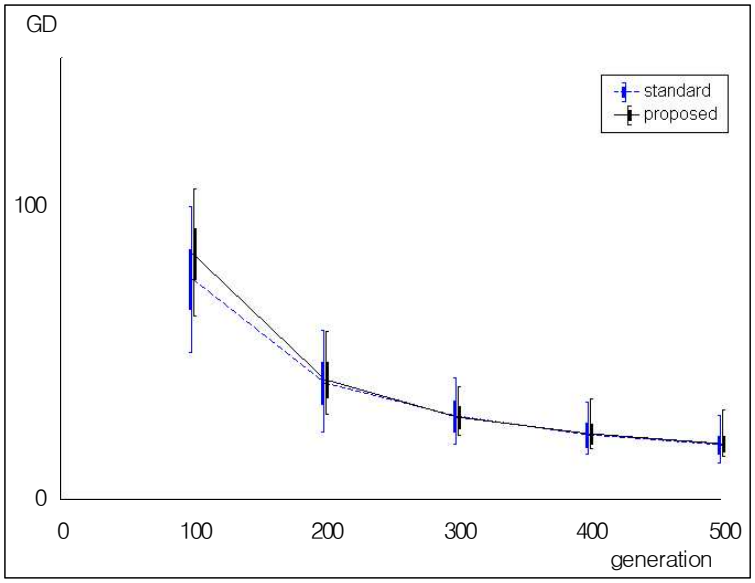

그림 2-5. 2-250KN pop=512, ts=4 GD 비교

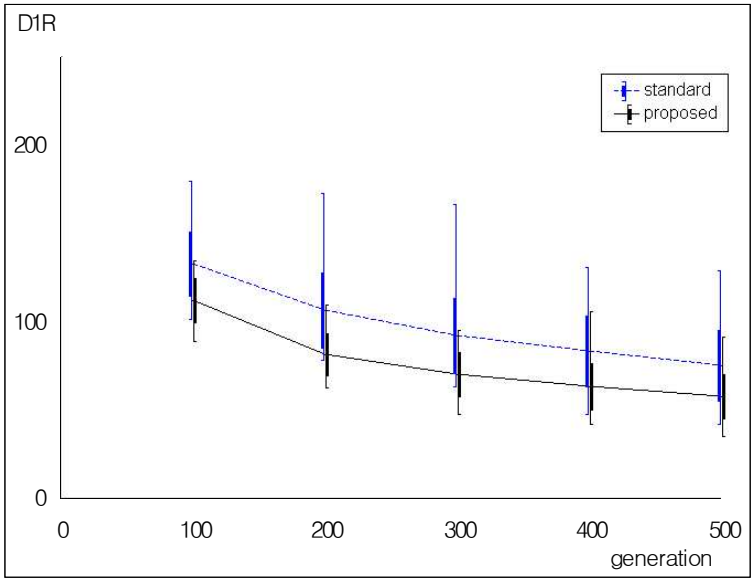

그림 2-6. 2-250KN pop=512, ts=4 $D1_R$ 비교

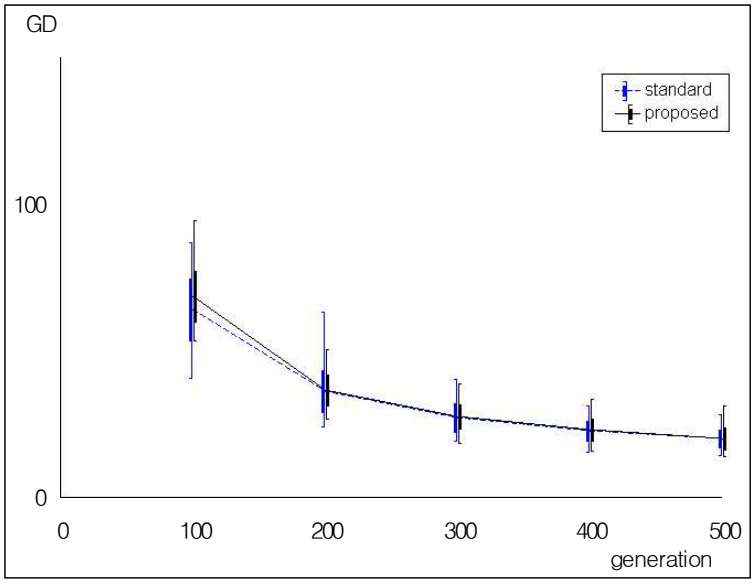

그림 2-7. 2-250KN pop=512, ts=8 GD 비교

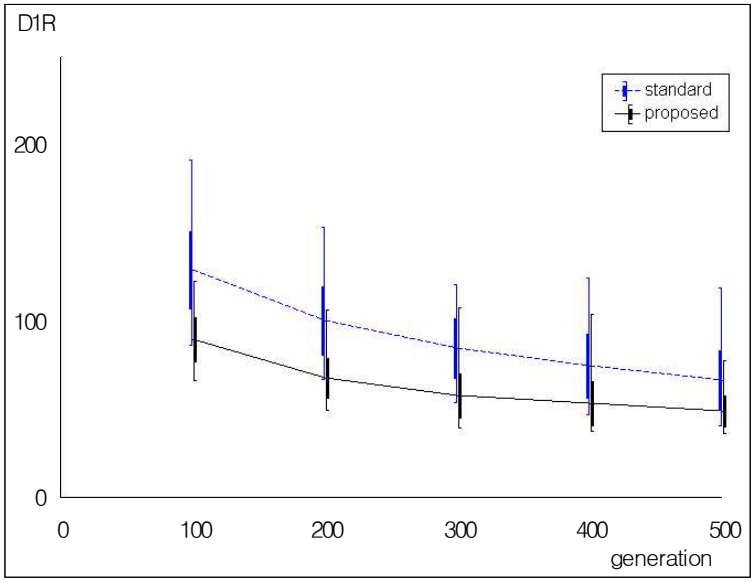

그림 2-8. 2-250KN pop=512, ts=8 $D1_R$ 비교

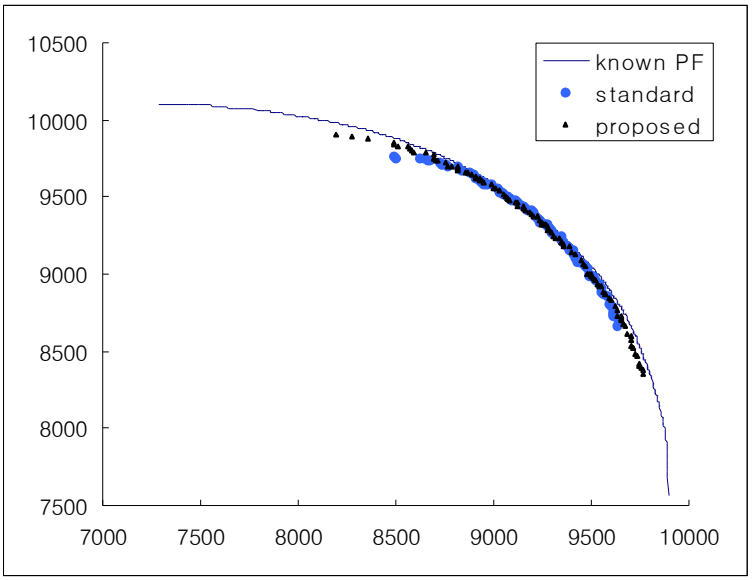

그림 3-1. 2-250KN pop=256, ts=4 best GD 시도의 nondominated set

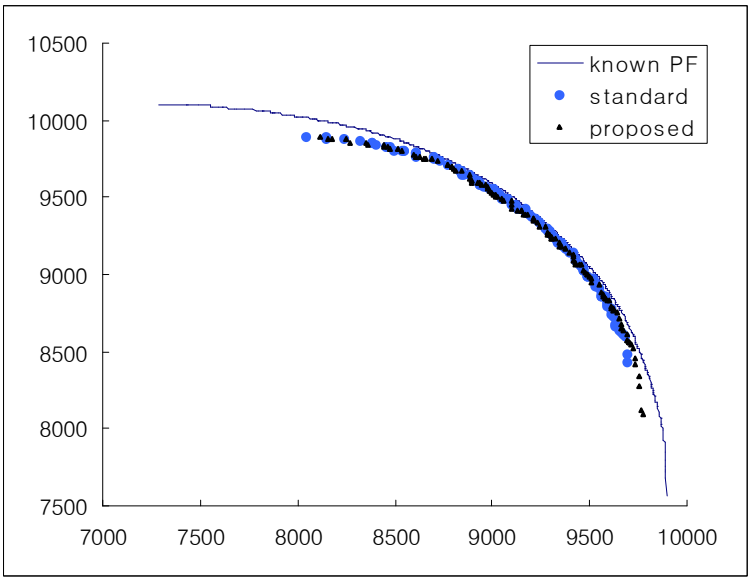

그림 3-2. 2-250KN pop=256, ts=4 best $D1_R$ nondominated set

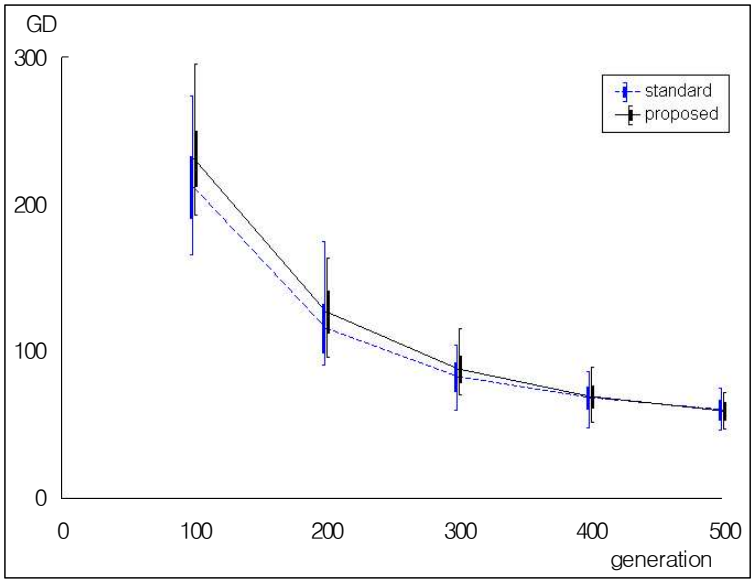

그림 4-1. 2-500KN pop=512, ts=4 GD 비교

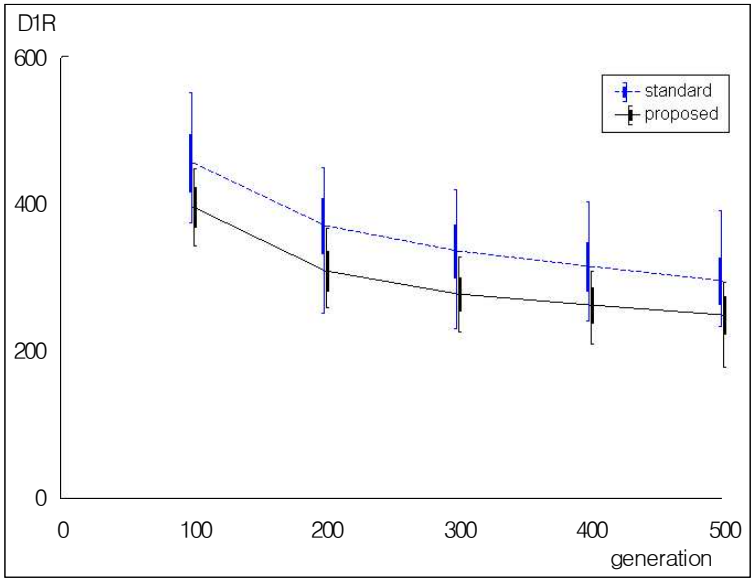

그림 4-2. 2-500KN pop=512, ts=4 $D1_R$ 비교

4.3.3 분석 및 추가 실험

4.3.2의 그래프들을 살펴보며 dominance 기반의 선택 방법을 적용했을 때의 성능을 살펴보겠다. 먼저 GD값들의 변화 추이를 살펴보면, 100에서 200 세대까지의 진화 과정까지는 일반적인 tournament 선택 방법을 사용하였을 때보다 조금 나쁜 성능을 보이고 있다. dominance 기반의 선택 방법의 수행 과정 중에서, tournamnent 크기 내에서 서로 dominate하는 관계의 부모를 발견하지 못하는 경우에는 일반적인 tournament 선택 방법과 같은 성능을 보이게 된다. 하지만, 그러한 dominance 관계를 가지는 두 번째 부모를 발견하게 된 경우에는 tournament 크기보다 훨씬 적은 수의 부모 중에 두 번째 부모를 선택하게 됨으로 선택압이 줄어드는 효과가 발생한다.

제한된 크기의 개체수를 가지고 진화를 이끌어 나감에 따라, 100에서 200 세대 사이 정도가 되면 저장하는 전체 개체들이 nondominated set을 이루게 된다. 이후에는 dominance 관계를 가지는 부모를 발견할 기회는 사라지고, 제안한 선택 방법이 효력을 발휘하지 못할 것으로 생각된다.

그러나 $D1_R$ 값의 추이를 살펴보면 그렇지 않다는 것을 알 수 있다. 물론, 하나의 nondominated set으로 수렴한 이후에는 dominance 기반의 선택 방법이 수행되지 않는다. 그 때까지는 선택압이 낮아지는 효과로 인하여 GD값이 나타내는 수렴성은 조금 떨어지는 성향을 보인다. 그렇지만 $D1_R$ 값이 보여주듯, 제안한 선택 방법으로 인해 얻어진 다양성은 이후의 계속된 진화 과정에도 꾸준한 영향을 준다.

세대를 거듭함에 따라 수렴하는 GD값을 살펴보면, 눈에 띄이는 차이를 찾아볼 수 없다. 오히려, 제안한 선택 방법을 사용한 경우에 더 좋은 수렴성까지 보이며, 극단적으로 여러 세대를 거듭해 더 이

상 해의 성능 향상이 일어나지 않는 시점에서는 초기에 확보된 해의 다양성으로 인해 더 좋은 수렴성을 기대할 수 있다.

특히, 이러한 다양성의 보존 특성은 특정한 목적 함수의 극단적인 값이 중요한 문제에서 강점을 가진다. 다양한 해를 유지하고 있음은, 각 목적 함수 값의 양쪽 경계선이 넓게 유지하고 있음을 뜻한다. 그러므로 최종 얻어진 nondominated set 중에서 하나의 해를 선택하는 tradeoff 위에서, 선택해야 하는 조건이 어떠한 목적 함수의 극단적인 경계 내일 경우에 나머지 목적 함수에 대해 좋은 성능을 보이는 해를 선택할 수 있게 된다.

본 논문에서 추가적으로 살펴보고자 한 것은, dominance 관계를 가지는 부모가 선택되는 비율에 따른 성능 변화이다. 4.3.2에서 보인 결과는 하나의 nondominated set으로 전체 개체가 수렴한 이후의 수렴까지 보이고 있다. 이제는 dominance 관계의 부모가 선택되는 비율을 변화시키며 성능을 평가하기 위해, 하나의 전선(front)으로 수렴하기까지의 변화만 살펴보도록 한다.

2-250KN 문제를 가지고 실험하였으며, 100세대까지만 수행하여 결과를 비교한다. 100세대까지만의 실험에서 수렴성을 높이기 위해 tournament 크기는 8로 고정하였다. 또한 tournament 크기가 적당히 커야만 dominance 관계를 가지는 부모의 비율을 크게 조절하기 용이해진다. 4.3.2에서 tournament 크기 4와 8의 변화에 따른 성능 결과를 보면, tournamnet 크기 8이 주어진 문제의 복잡도에 비해 높은 선택압이 아님을 알 수 있다. 그렇지 않은 경우, 높은 선택압으로 인해 지역적 최적점(local optima)에 빠지는 현상이 발생할 수 있으므로 조심해야 한다.

전체 개체수 256, tournament 크기 8을 고정하고 100세대까지 진화를 수행하였으며, 다른 환경 변수들은 위의 실험들과 동일하다.

표 1. 보통의 tournament 선택 방법, 제안한 선택 방법을 사용한 경우의 dominance 관계의 부모가 선택된 비율과 성능. 아래의 두 가지는 다른 실험 조건은 그대로 두고, dominance 관계의 부모를 선택하는 비율을 인위적으로 조작한 결과. 각각의 실험은 50회 실시하여 평균값과 표준 편차, 그리고 최대값과 최소값에 대한 평가 기준을 제시.

Selection (dominance -related parent %)	GD		$D1_R$		Best GD		Best $D1_R$	
	Avg	Dev	Avg	Dev	GD	$D1_R$	GD	$D1_R$
standard tournament selection (3%)	**93.91**	11.42	173.70	22.69	73.73	188.32	84.08	130.03
proposed selection (30%)	**93.09**	11.16	**119.76**	15.20	69.37	109.70	92.76	91.91
proposed selection (55%)	97.40	11.60	**117.10**	12.16	66.49	139.37	76.49	92.81
proposed selection (80%)	116.56	9.75	121.42	11.84	94.75	100.76	97.92	89.03

표 1을 살펴보면, 극단적으로 선택되는 dominance 관계를 가지는 부모의 비율을 높이는 경우에는 기대하지 않은 나쁜 성능을 보임을 알 수 있다. 세밀한 해상도로 나누어 실험하지는 못하였지만, dominance 관계를 가지는 부모가 뽑히는 비율이 예상보다 넓은 범위에서 적절히 기대하는 성능을 보여줌을 알 수 있었다. 가능한 무조건 dominance 관계를 가지는 부모를 선택하는 경우와 전혀 고려하지 않는, 두 극단적인 경우을 제외하고는 기대하는 성능 향상을 보였다.

이러한 해집합의 다양성을 유지하는 원인으로서, dominance 관계를 가지는 부모의 선택으로 인한 것이 아닌 요인으로는 서로 다른 전선에서 선택되는 것으로 인한 효과를 생각해볼 수 있다. 서로 다른 전선에서 선택됨은 해의 좋고 나쁨의 품질이 차이가 남을 뜻한다. 이러한 두 부모 개체의 성능의 차이를 둠에 따라 얻어지는 해집합의 다양성이 아닌지 알아보는 실험을 하였다. 앞의 실험과 같은 조건에서 NSGA-II에서 수행하는 nondominated sorting 결과의 순위(ranking)을 가지고, 전체 선택되는 부모의 순위가 차이나는 경우의 비율과, 차이가 나는 경우들에서 순위 차이의 평균값을 살펴보았다. 제안한 선택 방법에 의해 나타나는 두 값을 맞춰주기 위해 순위가 다른 부모를 선택하는 비율을 인위적으로 높게 개체 선택이 일어나도록 하여 실험한 값을 표 2에 제시하였다.

표 2의 결과가 보여주듯이, 이러한 두 부모의 성능 차이가 해집합의 다양성을 가져오지는 못하였다. 인위적으로 부모의 순위 차이를 가져온 실험에서 GD값이 약간 높게 측정되었지만, 몇 개의 GD값이 작은 해들이 발견됨에 따라 나타난 값으로 판단되며 조절하기 이전의 일반적인 tournament 선택 방법과 다르지 않은 성능을 보이는 것으로 판단할 수 있다. dominance 관계를 가지는 부모가 서로 nondominated sorting의 순위 차이를 가지는 것은 사실이지만, 그

러한 요인이 해집합의 다양성이라는 성능 향상을 가져온 것이 아님을 확인할 수 있다.

표 2. 선택된 부모가 서로 다른 순위에 존재하는 비율과 그 때의 평균 순위 차이를 분석하고, 제안한 선택 방법을 사용한 경우에 나타난 값과 비슷하도록 일정 비율만큼 서로 다른 전선에서 부모를 선택하여 실험한 결과

Selection (different front %, average difference)	GD		$D1_R$		Best GD		Best $D1_R$	
	Avg	Dev	Avg	Dev	GD	$D1_R$	GD	$D1_R$
standard tournament selection (20%, 1.6)	**93.91**	11.42	173.70	22.69	73.73	188.32	84.08	130.03
proposed selection (35%, 2.7)	**93.09**	11.16	**119.76**	15.20	69.37	109.70	92.76	91.91
control (35%, 2.5)	**91.26**	14.08	171.02	28.03	59.58	131.87	100.26	124.10

5장

특정 응용 사례

5.1 Oligonucleotide Probe 선택 문제

생물학계 또는 의학계에서 실제 존재하는 문제로서 비교적 짧은 길이의 Probe 서열들을 디자인하고자 하는 계산학적 문제가 있다. 서로 다른 여러 개의 목표 유전자 서열이 존재할 때, 알고자 하는 유전자 서열이 어느 유전자인지 알아내기 위한 Probe 서열 집합을 결정하는 문제이다. DNA 간에 일어나는 결합 반응 (hybridization)을 이용하여, 미리 준비된 짧은 DNA 서열과 반응시켜 봄으로써 이를 알아내는 방법이다.

DNA 서열 간의 결합은 Watson-Crick 쌍을 이루고 있다. 염기서열로서, 아데닌(Adenine)과 티민(Thymine)이 서로 이루게 결합을 이루고 구아닌(Guanine)과 시토신(Cytosine)이 서로 결합을 이룬다. 현재 일반화 된 DNA 합성 기술은 500 nt(nucleotide-핵산의 단위) 길이 정도까지 원하는 서열을 만들어 낼 수 있다고 한다. 사람의 경우 유전자의 전체 길이는 3×10^9bp (base pair-결합을 이루고 있는 DNA 서열의 길이를 나타내는 단위) 정도라고 알려져 있다. 염기 서열의 분포가 A,T,G,C 모두 균등하게 이루어져 있다고 가정할 때, 전체 유전자 길이를 구별하는 데에 필요한 Probe 서열의 길이는 16nt가 된다($4^{16}\fallingdotseq4\times10^9$). 일반적으로 20에서 60nt 정도 길이의

범위에서 Probe 서열을 합성하여 사용한다고 한다 [Rouillard et al., 2003]. 이러한 범위 내에서 구별하고자 하는 유전자의 서열 중에서, Probe 서열이 결합을 이루고자 하는 영역을 선택하는 것이 Probe 선택 문제이다.

이 때에 발생하는 문제는 선택한 Probe 서열이 목표하는 유전자 외의 다른 유전자와는 결합을 이루지 않아야 한다는 것이다. 또한 DNA 칩이 아닌 다른 목적으로 Probe 서열이 사용될 때에는, Probe 서열 간에도 결합이 일어나서 목표 유전자와 반응을 일으킬 확률이 줄어드는 일이 발생하지 않도록 해야 하는 문제도 해결해야 한다.

실험한 사례는 여성의 자궁경부암을 일으키는 HPV(Human Papilloma Virus) 유전자 중에서 고위험군과 저위험군에 속하는 19가지 목표 유전자를 정확하게 분석해내기 위한 짧은 길이의 oligonucleotide probe 서열을 디자인하는 문제이다 [Hwang et al., 2003].

19개의 HPV 유전자 전체 서열의 길이는 약 8000bp이지만 전체를 대상으로 probe를 찾지는 않는다. 생물학적으로 알려진 HPV 유전자 각각의 유일한 특성을 결정짓는 L1 region 중에서 GP 5d+/6d+ 영역만을 목표로 삼아 probe 서열을 선택한다. 이 영역의 길이는 대략 150bp 정도이며, 양 끝에는 PCR(Polymerase Chain Reaction)을 통해 서열을 증폭시키기 위해 디자인된 primer 서열이 결합하는 공통의 부분이 존재하며 이 또한 후보 영역에서 제외하여 나머지 약 90에서 100bp 사이에서 probe 서열을 선택해 나간다. 디자인하는 probe 서열의 길이는 30bp이며, 19가지 유전자의 GP 5d+/6d+ 영역의 서열은 다음과 같다.

표 3. 19개 목표 HPV 유전자에 대한 L1 region의 GP 5d+/6d+ 영역의 염기 서열

HPV 6	5 ttt gtt act gtg gta gat acc aca cgc agt acc aac atg aca tta tgt gca tcc gta act aca tct tcc aca tac acc aat tct gat tat aaa gag tac atg cgt cat gtg gaa gag tat gat tta caa ttt att ttt c 3
HPV 11	5 ttt gtt act gtg gta gat acc aca cgc agt aca aat atg aca cta tgt gca tct gtg tct aaa tct gct aca tac act aat tca gat tat aag gaa tac atg cgc cat gtg gag gag ttt gat tta cag ttt att ttt c 3
HPV 16	5 ttt gtt act gtt gtt gat act aca cgc agt aca aat atg tca tta tgt gct gcc ata tct act tca gaa act aca tat aaa aat act aac ttt aag gag tac cta cga cat ggg gag gaa tat gat tta cag ttt att ttt c 3
HPV 18	5 ttt gtt act gtg gta gat acc act ccc agt acc aat tta aca ata tgt gct tct aca cag tct cct gta cct ggg caa tat gat gct acc aaa ttt aag cag tat agc aga cat gtt gag gaa tat gat ttg cag ttt att ttt c 3
HPV 31	5 ttt gtt act gtg gta gat acc aca cgt agt acc aat atg tct gtt tgt gct gca att gca aac agt gat act aca ttt aaa agt agt aat ttt aaa gag tat tta aga cat ggt gag gaa ttt gat tta caa ttt ata ttt c 3
HPV 33	5 ttt gtt act gtg gta gat acc act cgc agt act aat atg act tta tgc aca caa gta act agt gac agt aca tat aaa aat gaa aat ttt aaa gaa tat ata aga cat gtt gaa gaa tat gat cta cag ttt gtt ttt c 3
HPV 34	5 ttt tta act gtt gta gat act act aga agc aca aac ttt tca gtt tgt gta ggt aca caa tcc aca agt aca act gca ccatat gca aac agt aat ttt aag gaa tac ctc aga cat gca gaa gag tat gac ctg cag ttt gtg ttt c 3
HPV 35	5 ttt gtt act gta gtt gat aca acc cgt agt aca aat atg tct gtg tgt tct gct gtg tct tct agt gac agt aca tat aaa aat gac aat ttt aag gaa tat tta agg cat ggt gaa gaa tat gat tta cag ttt att ttt c 3
HPV 39	5 ttt ctt act gtt gtg gac act ac c cgt agt acc aac ttt aca tta tct acc tct ata gag tct tcc ata cct tctaca tat gat cct tct aag ttt aag gaa tat acc agg cac gtg gag gag tat gat tta caa ttt ata ttt c 3
HPV 40	5 ttt gtt aca gtt gta gac acc act cgt agc act aat tta acc tta tgt gct gcc aca cag tcc ccc aca cca acc cca tat aat aac agt aat ttc aag gaa tat ttg cgt cat ggg gag gag ttt gat ttg cag ttt att ttt c 3
HPV 42	5 ttt tta act gtg gtt gat act acc cgt agt act aac atg act ttg tgt gcc act gca aca tct ggt gat aca tat aca gct gct aat ttt aag gaa tat tta aga cat gct gaa gaa tat gat gtg caa ttt ata ttt c 3
HPV 44	5 ttt gtt act gtt gta gat act acc cgt agt aca aac atg aca ata tgt gct gcc act aca cag tcc cct ccg tct aca tatact agt gaa caa tat aag

	caa tac atg cga cat gtt gag gag ttt gac tta caa ttt atg ttt c 3
HPV 45	5 ttt gtt act gta gtg gac act acc cgc agt act aat tta aca tta tgt gcc tct aca caa aat cct gtg cca agt aca tat gac cct act aag ttt aag cag tat agt aga cat gtg gag gaa tat gat tta cag ttt att ttt c 3
HPV 51	5 ttt att acc tgt gtt gat act acc aga agt aca aat tta act att agc act gcc act gct gcg gtt tcc cca acattt act cca agt aac ttt aag caa tat att agg cat ggg gaa gag tat gaa ttg caa ttt att ttt c 3
HPV 52	5 ttt gtc aca gtt gtg gat acc act cgt agc act aac atg act tta tgt gct gag gtt aaa aag gaa agc aca tat aaa aat gaa aat ttt aag gaa tac ctt cgt cat ggc gag gaa ttt gat tta caa ttt att ttt c 3
HPV 56	5 ttt gtt act gta gta gat act act aga agt act aac atg act att agt act gct aca gaa cag tta agt aaa tat gat gca cga aaa att aat cag tac ctt aga cat gtg gag gaa tat gaa tta caa ttt gtt ttt c 3
HPV 58	5 ttt gtt acc gtg gtt gat acc act cgt agc act aat atg aca tta tgc act gaa gta act aag gaa ggt aca tat aaa aat gat aat ttt aag gaa tat gta cgt cat gtt gaa gaa tat gac tta cag ttt gtt ttt c 3
HPV 59	5 ttt tta aca gtt gta gat act act cgc agc acc aat ctt tct gtg tgt gct tct act act tct tct att cct aat gta tac aca cct acc agt ttt aaa gaa tat gcc aga cat gtg gag gaa ttt gat ttg cag ttt ata ttt c 3
HPV 66	5 ttt gtt act gtt gtg gat act acc aga agc acc aac atg act att aat gca gct aaa agc aca tta act aaa tat gat gcc cgt gaa atc aat caa tac ctt cgc cat gtg gag gaa tat gaa cta cag ttt gtg ttt c 3

5.2 실험 결과

5.2.1 목적 함수

Probe 집합의 좋고 나쁨을 나타내는 나타내는 기준으로서 다음의 세 가지 목적 함수를 사용하여 실험하였다.

1. probe와 target 사이의 결합에서 GC pair 비율(GC content)의 편차
2. probe 서열 간의 유사도 (Similarity)
3. melting temperature (Tm) 값의 표준 편차

제약 조건으로는 Probe 서열이 목표 유전자가 아닌 다른 유전자와 결합하지 않도록, edit distance를 기준으로 30% 이상 차이나는 후보 Probe들을 가지고 최적의 Probe 집합을 찾아보았다. 화학적으로 edit distance가 30% 이하인 경우에는 잘못된 결합 반응(cross-hybridization) 이 일어난다고 볼 수 있다 [Rouillard et al., 2003].

평형 상태에서 결합 반응이 50% 일어나는 온도를 뜻하는 melting temperature (Tm)은 nearest-neighbor(NN) method를 사용하여 예측한 값을 사용하였다 [SantaLucia, 1998]. 같은 조건에서 동일하게 실험되어질 Probe 집합의 Tm 값들은 비슷하여야, 동등한 확률을 가지고 결합 반응이 일어난다. 그렇지 않다면 제대로 실험 결과가 얻어지지 않거나, 얻어진 실험 결과의 신뢰도가 떨어지는 문제가 발생한다.

A-T 결합이 2개의 수소 결합을 이루는 데에 반하여, G-C 결합은 3개의 수소 결합을 이루어 좀더 강한 결합력을 가진다. 그렇기 때문에 일반적으로 GC 쌍의 비율이 melting temperature를 결정하는 중요한 요인으로 알려져 있다. nearest-neighbor method를 사용

하여 구하여진 Tm 값 또한 예측값일 뿐이므로, GC content를 통하여 실제 Tm 값의 편차를 다시 한 번 확인해 볼 수 있다.

제약 조건을 통하여 다른 유전자 서열과 결합할 가능성이 적은 후보 Probe 서열만 다루게 되지만 이러한 제약 조건이 절대적으로 잘못된 결합을 방지하는 것은 아니다. 선택된 Probe 집합 내에서 Probe 간에 얼마나 비슷한가 그렇지 않은가를 측정해 봄으로써 잘못된 결합 가능성을 다시 한 번 확인하고 방지한다.

이 밖에도 Probe 서로 간의 잘못된 결합이 일어나는 일이 없도록 조절해야 하지만, 실험해 본 결과 유전자 서열과의 결합 가능성을 배제하는 제약 조건을 통해 충분히 이러한 특성을 얻을 수 있었으므로 목적 함수에서는 제외하였다.

5.2.2 실험 환경 변수

Population size : 1024
Generation : 500
Crossover rate : 0.8
Mutation rate : 1/(gene number=19)
Tournament size : 4

5.2.3 결과 비교

일반적인 tournament 선택 방법과 본 논문에서 제안한 Dominance 기반의 선택 방법을 사용하여 얻어진 Probe 집합을 비교하였다. 구해진 해집합 중에서 Tm 편차가 가장 작은 조건 안에서, Probe 서열 간의 유사도가 작은 기준 (실험값 기준 400,350) 안으로 들어오는 해를 각각 선택하였다. 허용 가능할 것으로 보이는 Tm 표준 편차 범위 내에서, 기준이 될만한 가장 작은 유사도 값이 350으로 판단하였다. 비교를 위해 그 이전에 Tm 표준 편차 위주의 좋은 성능을 평가할 기준으로서 유사도 400 이내의 해 또한 제시하였다. 표 3에서 이를 정리하였으며, 본 논문에서 제안하고 있는 선택 방법을 통해 같은 유사도 조건 내에서 찾아낸 Tm의 표준편차가 작은 해를 찾아내었음을 알 수 있다.

표 3. 각각의 구해진 probe 서열 집합의 목적 함수 값

		GC content variation	Similarity	melting temperature standard deviation
tournament	Probe1	44	392	1.474561487
	Probe2	48	344	1.522910994
dominance-based	Probe3	44	395	**1.466695175**
	Probe4	48	349	**1.49435781**

표 4. 각각의 얻어진 해들의 probe 서열들의 목표 유전자 서열 상에서의 위치

Target Gene	Sequence Length	Candidate Number	Probe1	Probe2	Probe3	Probe4
HPV 6	139	14	50	50	50	50
HPV 11	139	14	32	32	32	32
HPV 16	142	13	40	40	40	40
HPV 18	145	13	50	50	50	50
HPV 31	142	5	48	48	48	48
HPV 33	139	3	39	39	39	39
HPV 34	148	25	24	28	24	28
HPV 35	142	15	34	**37**	34	**35**
HPV 39	145	23	57	57	57	57
HPV 40	145	9	61	61	61	61
HPV 42	139	11	59	59	59	59
HPV 44	145	9	61	60	61	60
HPV 45	145	7	58	58	**50**	58
HPV 51	142	11	60	**59**	60	60
HPV 52	139	16	34	35	34	35
HPV 56	139	8	46	46	46	46
HPV 58	139	4	41	41	41	41
HPV 59	145	10	61	61	61	61
HPV 66	139	6	38	38	38	38

※ (문제 공간 크기; 가능한 모든 조합의 수) = 1.26×10^{19}

표 5. 각각의 얻어진 해들의 probe 서열들의 NN 모델을 이용한 melting temperature 예측 값

	Probe1	Probe2	Probe3	Probe4
1	61.13	61.13	61.13	61.13
2	60.51	60.51	60.51	60.51
3	60.55	60.55	60.55	60.55
4	63.55	63.55	63.55	63.55
5	62.24	62.24	62.24	62.24
6	60.17	60.17	60.17	60.17
7	59.93	59.8	59.93	59.8
8	60.6	**61.77**	60.6	**60.5**
9	58.14	58.14	58.14	58.14
10	62.16	62.16	62.16	62.16
11	60.22	60.22	60.22	60.22
12	61.14	**61.65**	61.14	**61.65**
13	62.01	62.01	**61.84**	62.01
14	62.34	**62.51**	62.34	62.34
15	60.65	60.98	60.65	60.98
16	58.26	58.26	58.26	58.26
17	60.65	60.65	60.65	60.65
18	57.52	57.52	57.52	57.52
19	60.41	60.41	60.41	60.41

6장

결론

6.1 결론 및 향후 과제

본 논문에서는 선택하는 부모 간의 Dominance 관계를 고려한 변형된 tournament 선택 방법을 제안하였다. 4장에서는 multiple knapsack 문제에 적용하여 얻을 수 있는 해집합의 수렴성은 비슷하게 유지하거나 혹은 더 좋은 결과를 가지는 한 편, 해집합의 다양성을 향상시킨 결과를 제시하였고, 5장에서는 본문에서 제시한 선택 방법을 실세계의 문제에 적용하는 예를 보이고 있다. 제시한 응용 사례는 생물학적 문제의 일종인 Oligonucleotide Probe 집합 선택 문제이다. 이러한 실제적 문제를 해결하는 데에도 제안한 선택 방법이 기대한 성능 향상을 가져왔음을 알 수 있었다.

이전의 관련 연구에서 잘 알려진 시험 다중 목표 함수 문제들의 경우에는 쉬운 문제, 즉 다양성이 그다지 중요하지 않고 상대적으로 문제 공간 내에서 골고루 탐색이 일어날만한 문제였다. 이러한 문제에 Dominance 기반의 제시한 선택 방법을 적용한 결과 수렴 속도의 향상을 얻을 수 있었다고 [Rudenko et al., 2004] 논문은 결론맺고 있다. 본 논문에서는 제시한 실험 결과를 토대로, 다양성이 중시될 수 있을만한 조합적 계산학의 문제들에 적용하였을 경우에 제시한 Dominance 기반의 선택 방법의 탐색 성향이 해집합의 다양

성을 향상시킬 수 있다고 결론내릴 수 있겠다. 앞으로 좀더 연구하고자 하는 바를 살펴보면, 다차원의 다중 목적 최적화에서도 마찬가지 성능 향상을 기대할 수 있음을 검증해야 하겠다. 그 외에도 서로 dominance 관계를 가지는 부모를 선택한 유전 과정 중에 기대할 수 있는 탐색 성향이 실제로 어떠한 것인가를 통계적인 기법들을 이용하여 분석해볼 필요가 있다.

참고 문헌

Coello, C. A. C., Van Veldhuizen, D.A. and Lamont, G.B., Evolutionary Algorithms for Solving Multi-Objective Problems, Kluwer Academic Publishers, Boston, 2002

Corne, D. W., Knowles, J.D. and Oates, M. J., The Pareto Envelope-Based Selection Algorithm for Multiobjective Optimization, *Parallel Problem Solving from Nature (PPSN VI)*, pp.839-848, 2000

Corne, D.W., Jerram, N.R., Knowles, J.D. and Oates, M.J., PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization, *Genetic and Evolutionary Computation Conference (GECCO-2001)*, pp.283-290, 2001

Czyzak, P. and Jaszkiewicz, A., Pareto simulated annealing-a metaheuristic technique for multiple-objective combinatorial optimization, *Journal of Multi-Criteria Decision Analysis*, 7, pp.34-47, 1998

Deb, K., Agrawal, S. , Pratap, A., Meyarivan, T., A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimisation: NSGA-II, *Parallel Problem Solving from Nature (PPSN VI)*, p.849-858, 2000

Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms, John Wiley & Sons, Chichester, 2001

Hasen, M. P. and Jaszkiewicz, A., Evaluating the quality of approximations to the non-dominated set, Technical Report IMM-REP-1998-7, Technical University of Denmark, 1998

Hwang, T. S., Jeong, J. K., Park, M., Han, H. S, Choi, H. K. and Park, T. S., Detection and Typing of HPV Genotypes in Various Cervical Lesions by HPV Oligonucleotide Microarray, Gynecol Oncol., 90(1), pp.51-56, 2003

Knowles, J. D. and Corne, D. W., The Pareto Archived Evolution Strategy: A New Baseline Algorithm for Pareto Multiobjective Optimisation, *1999 Congress on Evolutionary Computation (CEC'99)*, 1, pp.98-105, 1999

Knowles, J. D. and Corne, D. W., Approximating the nondominated front using the Pareto Archived Evolution Strategy, *Evolutionary Computation*, 8(2), pp.149-172, 2000

Knowles, J. D. and Corne, D. W., On Metrics for Comparing Non-Dominated Sets, *2002 Congress on Evolutionary Computation Conference (CEC'02)*, pp.711-716, 2002

Martello, S. and Toth, P., Knapsack Problems: Algorithms and Computer Implementations, Chichester, U.K.:Wiley, 1990

Rouillard, J. M., Zuker, M. and Gulari, E., OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach, *Nucleic Acids Research*, 31(12), pp.3057-3062, 2003

Rudenko, O. and Schoenauer, M., Dominance Based Crossover Operator for Evolutionary Multi-objective Algorithms, *Parallel Problem Solving from Nature (PPSN VIII)*, pp.812-821, 2004

SantaLucia, J., Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, *Proc. Natl. Acad. Sci. USA*, 95(4), pp.1460-1465, 1998

Van Veldhuizen, D. A., Multiobjective Evolutionary Algorithms:

Classifications, Analyses, and New Innovations, PhD thesis, Department of Electrical and Computer Engineering, Graduate School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, 1999

Zitzler, E., Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications, PhD thesis, Swiss Federal Institute of Technology (ETH), 1999

Zitzler, E. and Thiele, L., Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach, *IEEE Transactions on Evolutionary Computation*, 3(4), pp.257-271, 1999a

Zitzler, E., Deb. K,and Thiele, L., Comparison of multiobjective evolutionary algorithms: Empirical results, *Evolutionary Computation*, 8(2), pp.173-195, 2000

Zitzler, E., Laumanns, M. and Thiele, L., SPEA2: Improving the Strength Pareto Evolutionary Algorithm, TIK Report 103, Computer Engineering and Networks Laboratory (TIK), Department of Electrical Engineering Swiss federal Institute of Technology (ETH), 2001

Zitzler, E. and Laumanns, M.,
http://www.tik.ee.ethz.ch/~zitzler/testdata.html/

Abstract

Multi-objective evolutionary algorithms (MOEAs) mean evolutionary algorithms for multi-obejctive optimization (MOO) problem that are too complex to be solved by efficient algorithms. Evolutionary algorithms are search strategies, motive of which is from the concept of biological evolution in the natural world. They treat population that consists of individuals. Each individual represents a solution of the given problem. Multi-objective optimization sets the goal at optimizing several objectives at the same time. Recently, several efficient MOEAs have been designed for searching the optimal front. For example, PAES, NSGA-II, SPEA2 and PESA-II show good performance. However, there has been few studies about mating selection method.

This paper introduces Dominance-Based Selection(DBS) for MOEAs, which gives priorities to mating pairs that one dominates the other. To test the performance of DBS, we experiment with multiple knapsack problems which are known as ones of NP-hard class.

As a specific example, we apply MOEAs with DBS to a real-world biological problem; that is the oligonucleotide probe sequence design problem. We make an experiment to design the probe sequences for 19 HPV(Human Papilloma Virus) genes, which cause the cervical cancers.

This paper concludes that it is possible to improve the diversity by increasing the ratio of such dominance-related mating pair. The obtained diversity can support larger choices and cause the possibility of searching for better solution.

Keywords : multi-objective, optimization, evolutionary algorithms, mating selection, dominance relation
Student Number : 2003-21613