다중목적함수 진화 알고리즘을 이용한 마이크로어레이 프로브 디자인

(Microarray Probe Design with Multiobjective Evolutionary Algorithm)

이 인 희 † 신 수 용 †† 조 영 민 ††† 양 경 애 †††† 장 병 탁 †††††
(In-Hee Lee) (Soo-Yong Shin) (Youngmin Cho) (Kyung-Ae Yang) (Byoung-Tak Zhang)

요 약 프로브(probe) 디자인은 성공적인 DNA 마이크로어레이(DNA microarray) 실험을 위해서 필수적인 작업이다. 프로브가 만족시켜야 하는 조건은 마이크로어레이 실험의 목적이나 방법에 따라 다양하게 정의될 수 있는데, 대부분의 기존 연구에서는 각각의 조건에 대하여 각각 독립적으로 정해진 한계치(threshold) 값을 넘지 않는 프로브를 탐색하는 방법을 취하고 있다. 그러나, 본 연구에서는 프로브 디자인을 두가지 목적함수를 지닌 다중목적함수 최적화 문제(multiobjective optimization problem)로 정의하고, ε-다중목적함수 진화 알고리즘(ε multiobjective evolutionary algorithm)을 이용하여 해결하는 방법을 제시한다. 제시된 방법은 19종류의 고위험군 인유두종 바이러스(Human Papillomavirus) 유전자들에 대한 프로브 디자인과 52종류의 애기장대 칼모듈린 유전자군(Arabidopsis Calmodulin multigene family)에 대한 프로브 디자인에 각각 적용되었다. 제안한 방법론을 사용하여 기존의 공개 프로브 디자인 프로그램인 OligoArray 및 OligoWiz에 비해 목표유전자에 더 적합한 프로브를 찾을 수 있었다.

키워드 : 진화 연산, 마이크로어레이 프로브 디자인, 다중목적함수 최적화, ε-다중목적함수 진화 알고리즘

Abstract Probe design is one of the essential tasks in successful DNA microarray experiments. The requirements for probes vary as the purpose or type of microarray experiments. In general, most previous works use the simple filtering approach with the fixed threshold value for each requirement. Here, we formulate the probe design as a multiobjective optimization problem with the two objectives and solve it using ε-multiobjective evolutionary algorithm. The suggested approach was applied in designing probes for 19 types of Human Papillomavirus and 52 genes in Arabidopsis Calmodulin multigene family and successfully produced more target specific probes compared to well known probe design tools such as OligoArray and OligoWiz.

Key words : Evolutionary computation, Microarray probe design, Multiobjective optimization, ε-Multiobjective evolutionary algorithm

· 본 연구는 교육인적자원부 BK21-IT, 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅(MEC) 과제 및 과학기술부 국가지정연구실(NRL) 사업, 그리고 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-214-D00140). 또한 이 연구를 위해 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에도 감사드린다.

† 학생회원 : 서울대학교 컴퓨터공학부
ihlee@bi.snu.ac.kr
†† 비 회 원 : 서울대학교병원 의료정보센터
syshin@snuh.org
††† 비 회 원 : Department of Computer Science and Engineering, University of California, San Diego
yoc002@cs.ucsd.edu
†††† 비 회 원 : 서울대학교 컴퓨터 연구소 연구원
kayang@bi.snu.ac.kr
††††† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@bi.snu.ac.kr

논문접수 : 2008년 2월 22일
심사완료 : 2008년 6월 25일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제8호(2008.8)

1. 서 론

DNA 마이크로어레이(microarray)는 많은 수의 유전자의 발현 정도를 동시에 측정할 수 있는 실험 도구로서 다양한 생물학 실험에 널리 사용되고 있다. 마이크로어레이의 표면은 미세하게 구획이 나누어져 있으며, 각각의 구획 내부에는 발현 정도를 관찰하고자 하는 개별 유전자에 특이적으로 결합할 수 있는 DNA 분자(프로브, probe)가 부착되어 있어서, 세포 샘플과 반응시킨 후, 각 구획별로 프로브와 유전자가 결합한 정도를 측정하여 해당 유전자들의 발현 정도를 알 수 있게 된다. 이 때, 표면에 부착되는 프로브의 생성 방식에 따라 cDNA 마이크로어레이와 올리고뉴클레오타이드(oligonucleotide) 마이크로어레이의 두 가지로 나눠진다. 이 중 cDNA 마이크로어레이는 서열에 대한 제어를 하기 힘들고, 서열 생성 과정상 오류를 완전히 배제하기 어렵기 때문에 정확한 결과를 요구하는 실험에는 잘 쓰이지 않고 있다. 반면에 올리고뉴클레오타이드 마이크로어레이는 DNA 합성 기술의 발달로 인해 프로브 서열을 제어하기 쉽고, 사용자의 용도에 맞게 프로브의 서열을 디자인하여 사용할 수 있다는 장점이 있어서 널리 사용되고 있다. 그러므로 본 연구에서는 올리고뉴클레오타이드 마이크로어레이에 사용될 프로브의 최적화 방법에 한정하여 논의하고자 한다.

마이크로어레이 실험에서 목표 유전자의 발현 정도는 프로브와 목표 유전자와의 결합 정도로부터 추정되기 때문에, 개별 프로브가 각각의 목표 유전자와만 반응하도록 디자인하는 것이 중요하다. 만약, 한 프로브가 둘 이상의 유전자와 결합 가능할 경우, 해당 프로브가 원래 목표했던 유전자의 발현 정도를 정확히 알 수 없게 되어, 부정확한 분석 결과를 유도하게 될 수 있다. 따라서 목표 유전자에 특이적인 프로브를 디자인하는 문제는 마이크로어레이를 사용한 연구에서 가장 중요하고 근본적인 문제 중의 하나라고 할 수 있으며, 이 점은 관련 문제를 다룬 연구 논문들의 방대함에서도 확인할 수 있다[1-8].

대부분의 관련 연구에서 공통적으로 언급되는 좋은 프로브의 조건은 다음의 세 가지이다: 1)목표 유전자와 특이적으로 강하게 상보 결합을 이루되, 2)프로브 자체적인 2차 구조의 형성 가능성이 낮아야 하며, 3)녹는점(Tm, melting temperature) 등의 상호 결합 반응에 영향을 줄 수 있는 화학적인 성질이 균질해야 한다. 이러한 조건을 만족시키는 프로브를 디자인하기 위해 다양한 연구 그룹으로부터 여러 가지 프로브 평가 방법이 제시되어 왔는데, 대부분 DNA의 결합에 필요한 자유에너지(free energy), 목표 유전자가 아닌 서열과 프로브 사이의 BLAST 검색 점수, 프로브의 2차 구조 형성에 필요한 자유에너지와 프로브 자체의 녹는점 등을 각각 검사하여 사용자가 정한 기준치를 모두 만족시키는 프로브를 찾아내는 방식을 취하고 있다[1,2]. 이 외에도 프로브 서열의 빈도수에 따른 방식[3], Shannon entropy에 기반한 정보 이론적인 접근 방식[4], 유전체 정보에 기반한 방식[5] 등이 있다. 그런데 프로브의 평가 방법의 다양함과는 대조적으로 적당한 프로브를 찾기 위한 탐색 방식에 있어서는 대개의 연구에서 공통적인 방식(그림 1의 (a))을 취하고 있다: 먼저 프로브를 디자인하고자 하는 목표 유전자의 서열로부터 슬라이딩 윈도우(sliding window) 방식 등으로 후보 프로브 서열을 생성시킨다. 그 후, 앞서 언급한 여러 평가 방법들을 사용하여 후보 프로브 서열들에 점수를 부여하되, 미리 정의된 한계치(threshold) 값을 만족시키지 못하는 후보 서열은 제거하고 마지막으로 남은 서열들 중에서 각각의 평가 항목에 대한 점수를 통합한 결과가 가장 우수한 후보 서열을 해당 목표 유전자의 프로브로 선정한다. 이 과정을 각각의 목표 유전자마다 수행하여 전체 목표 유전자 집합에 대한 프로브 집합을 구성하게 된다.

그러나 이와 같은 탐색 방식에서는 몇 가지 문제점을 찾을 수 있다. 우선 프로브 평가 점수와 실제 마이크로어레이 실험 결과에서의 형광 강도 값의 연관관계가 아직 불명확하기 때문에 평가 점수에 대한 명확한 한계치를 설정하기가 어렵고, 개별 실험 환경에 따라 여러 조절이 필요하다는 점을 들 수 있다. 그리고 각각의 목표 유전자에 대한 프로브 탐색을 독립적으로 진행하기 때문에 개별 목표 유전자에 대한 프로브들 사이에서 발생할 수 있는 비특이적 결합이 고려되지 않을 수 있다. 또한 최근의 연구 결과에 의하면[6-8], 우수한 프로브는 한두 가지 기준으로 정의될 수 없기 때문에 근본적으로 다중목적함수 최적화(multiobjective optimization)를 필요로 한다.

이와 같은 관찰을 바탕으로 다중목적함수 진화 알고리즘(MOEA, multiobjective evolutionary algorithm)을 올리고뉴클레오타이드 마이크로어레이를 위한 프로브 디자인에 적용하는 방법이 제안된 바 있으며[9], 본 연구에서는 보다 효율적인 탐색과 사용상의 편의를 위하여 추가의 과정을 도입하고, 기존의 프로그램과 비교 분석을 통하여 다중목적함수 진화 알고리즘을 이용한 프로브 디자인의 유용성을 검증하였다. 그리고 온라인 인터페이스(EvoOligo, http://cbit.snu.ac.kr/EvoOligo)를 제공하여 관련 연구자의 편의를 도모하였다.

프로브 최적화를 다중목적함수 최적화 문제(MOP, multiobjective optimization problem)로 보고 이를 다중목

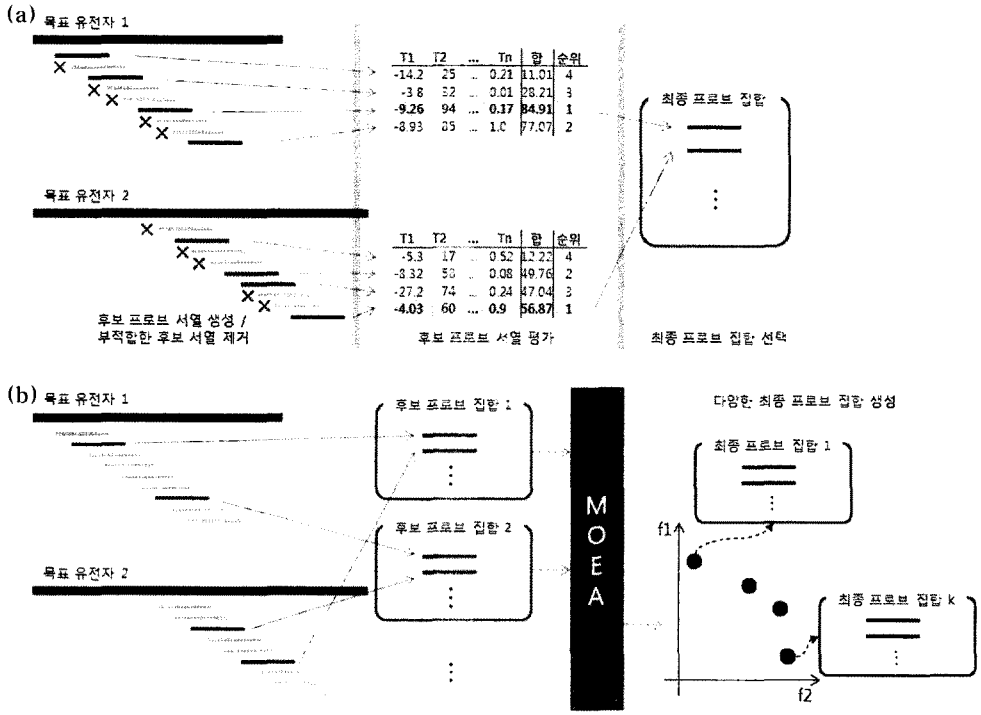

그림 1 프로브 탐색 방법의 비교. (a) 기존의 유전자별 프로브 탐색 방법. (b) 다중목적함수 진화 알고리즘을 이용한 프로브 조합 탐색 방법.

적함수 진화 알고리즘으로 해결할 때의 장점으로는 우선 기존의 방법에서처럼 다양한 평가 항목들을 하나의 수치로 표현하는 데서 발생할 수 있는 왜곡현상을 줄일 수 있고, 적절한 한계점 값이나 각각의 평가 항목들을 하나의 수치로 표현하기 위한 가중치(weight) 조절에 따른 반복적인 작업을 피할 수 있다는 것을 들 수 있다. 그리고 진화 연산의 확률적(stochastic)이고 병렬적인 탐색 특성으로 인하여 여러 평가 항목들 사이의 다양한 타협점(trade-off point)들을 찾아서 한 번에 여러 가지의 프로브 디자인이 가능하다(그림 1의 (b)). 또한, 사용 가능한 사전정보(prior knowledge) 또는 특정 목적에 따른 사용자의 기호 등을 목적함수의 변형 또는 추가를 통하여 손쉽게 반영할 수 있다.

본 논문은 다음과 같이 구성되었다. 2장에서 프로브 디자인을 다중목적함수 최적화 문제의 관점에서 살펴보고, 3장에서 다중목적함수 진화 알고리즘을 이용한 프로브 최적화에 대해 구체적으로 설명한다. 구체적인 적용 사례로서 고위험군 인유두종 바이러스(Human Papillomavirus) 유전자군과 애기장대 칼모듈린 유전자군(Arabidopsis Calmodulin multigene family)에 대한 실험 결과가 4장에서 제시되며, 마지막으로 5장에서 요약 및 결론을 내리고자 한다.

2. 다중목적함수 최적화를 통한 프로브 디자인

2.1 다중목적함수 최적화 문제

다중목적함수 최적화 문제(MOP)란 여러 개의 목적함수(objective)를 한꺼번에 최적화할 수 있는 해를 찾는 문제를 말하는데, 주로 목적함수들 사이에 상충하는 관계가 존재하여 하나의 목적함수에 대한 최적화 방향이 다른 목적함수에 대한 최적화 방향과 일치하지 않아서 동시에 최적화시키기 어렵고, 또한 목적함수들 사이의 우선순위(priority)가 주어지지 않아서 우선순위에 따른 사전적 순서(lexicographic order)대로 최적화하는 것도 불가능한 경우의 문제를 의미한다. MOP의 일반적인 형태는 다음과 같다:

$$\text{optimize } f_i(\boldsymbol{X}), \text{for } i=1,\cdots,M,$$
$$\text{subject to } g_j \geq 0, \text{for } j=1,\cdots,N,$$

where, $\boldsymbol{X}=(x_1,\cdots,x_n)$, $x_k^{(L)} \leq x_k \leq x_k^{(U)}$, $k=1,\cdots,n$.

여기에서 f_i는 개별 목적함수(objective), g_j는 개별 제한조건(constraint), M은 목적함수의 개수, N은 제한조건의 개수를 각각 의미하며, $x^{(L)}$와 $x^{(U)}$는 변수 x가 가질 수 있는 최소값과 최대값을 의미한다.

MOP에서 변수 공간(variable space)에서 하나의 해 $X=(x_1,\cdots,x_n)$에는 이에 상응하는 목적함수 공간

(objective space) 상의 벡터 $f(X)=(f_1(X),\cdots,f_M(X))$가 존재하며, 변수 공간에서 서로 다른 해 사이의 우열은 이에 상응하는 목적함수 공간의 벡터를 비교하여 가리게 된다. 일반적으로 두 해의 비교에는 dominance 관계를 사용하는데, dominace 관계의 정의는 다음과 같다. f_i를 각각 최소화해야 하고(최대화의 경우 부등호의 방향이 반대이다), 제한조건이 모두 동등할 경우, 변수 공간의 서로 다른 해 X와 Y 및 각각에 상응하는 목적함수 공간의 벡터 $f(X)$와 $f(Y)$에 대하여,

$$\forall i\in 1,\cdots,M,\ f_i(X)\le f_i(Y),$$

$$\exists j\in 1,\cdots,M,\ f_j(X)<f_j(Y)$$

의 두 조건을 모두 만족시킬 때, X는 Y보다 더 좋은 해로 평가받고, 이를 'X가 Y를 dominate한다' 또는 $X<Y$로 표현한다. 이것은 여러 개의 목적함수 사이의 우선순위가 존재하지 않기 때문에 하나의 해가 다른 해보다 낫다고 말하기 위해서는 모든 목적함수 값의 면에서 못하지 않고, 또한 하나 이상의 목적함수에서 더 나은 값을 가져야 함을 의미한다. 따라서 MOP에서 최적해는 변수 공간의 어떤 해에도 dominate되지 않는 것을 의미하며 이를 Pareto-최적(Pareto-optimal)해라고 한다. 또한 두 해 중 어느 쪽도 다른 하나를 dominate하지 못하는 경우가 있을 수 있는데, 이 때 두 해는 '서로 non-dominate한다'고 표현하고 서로 동등한 수준의 해로 간주된다. 일반적인 MOP에서 Pareto-최적해는 목적함수 간의 trade-off 관계 때문에 서로 non-dominate인 여러 개의 타협점들로 구성된다.

2.2 다중목적함수 진화 알고리즘

실제 MOP의 응용에서는 최적화 알고리즘을 사용하여 다양한 타협점을 찾아내고, 마지막에 이들 타협점 중에서 해당 응용 문제 및 사용자의 요구에 따라 적당한 것을 선택하게 된다. 이 때 가능한 한 Pareto-최적해에 가까운 타협점들을 찾아내기 위하여 목적함수 공간의 경사면(gradient)을 이용하거나 목적함수들에 임의의 가중치를 부여하여 합산한 값을 최적화하는 등의 여러 가지 방법들이 제시되어 왔다[10]. 그러나 이러한 방법들은 대개 특정 MOP의 형태 또는 최초 시작 위치에 의존적이거나, 지역 최적해(local optimal solution)에서 벗어나지 못한다는 단점을 가지고 있다. 또한, 한 번에 하나의 타협점만을 찾도록 되어 있어서 여러 개의 다양한 타협점을 찾기 위해서는 여러 번의 최적화를 반복해야 한다.

고전적인 다중목적함수 최적화 알고리즘의 이러한 단점을 해결할 수 있는 방안 중의 하나로 주목받고 있는 것이 진화 알고리즘이다[11]. 일반적인 진화 알고리즘은 초기에 일정한 수의 해를 무작위로 생성하여 주어진 문제에 따라 적합도를 평가한 후 이에 따라 좋은 해를 선택하여 변형시키는 과정을 반복하는 방식으로 단계적으로 성능이 좋은 해를 찾아나가게 된다. 이 때 적합도에 따라 해들을 비교하여 좋은 해를 선택하는 대신 dominance 관계에 따라 선택하는 방식으로 MOP를 해결하도록 할 수 있다. 이러한 진화 알고리즘의 한 갈래를 다중목적함수 진화 알고리즘(MOEA)이라 한다[10]. MOEA의 장점으로는, 1) 문제의 형태에 종속되지 않는 일반적인 알고리즘이므로 어떠한 문제에도 적용할 수 있으며, 2) 해 집합을 통하여 여러 개의 타협점들을 한 번에 찾을 수 있고, 3) 진화 알고리즘의 확률적인(stochastic) 탐색 특성으로 인하여 지역 최적해에서 보다 쉽게 빠져나올 수 있다는 점을 들 수 있다.

2.3 다중목적함수 최적화 문제로서의 프로브 디자인

일반적으로 좋은 프로브가 갖추어야 할 조건은 앞서 1장에서 언급한 것과 같이 크게 다음의 세 가지를 들 수 있다.

1) 목표 유전자 이외의 유전자와의 상보 결합 최소화.
2) 프로브 자체의 2차 구조 형성 최소화.
3) 녹는점 등의 화학적인 성질의 균질성 최대화.

이상의 조건 이외에도 기본적으로 프로브의 정의상 프로브와 상보적인 서열이 목표 유전자 이외에 존재해서는 안 된다는 제한조건이 추가된다. 이 조건은 첫 번째 조건에 포함될 수 있으나, 프로브 디자인 문제의 정의를 보다 명확히 하기 위하여 제한조건으로 추가하였다. 그런데 이 중에서 세 번째 조건은 대개 다른 두 조건과 동등한 목적함수라기보다는 별도의 제한조건으로 취급되는 경우가 많다. 따라서 본 연구에서 세 번째 조건은 목적함수가 아니라 MOEA를 통해서 생성한 여러 타협점 중에서 선택할 때의 척도로 사용하였다. 따라서 프로브 디자인은 하나의 제한조건과 두 목적함수를 가진 MOP로 정의될 수 있고, 이를 보다 명확하게 수식화하면 다음과 같다.

목표 유전자의 집합 $T=\{t_1,\cdots,t_n\}$와 이에 대한 프로브의 집합 $P=\{p_1,\cdots,p_n\}$가 있을 때, 각각의 목표 유전자 t_i에 대한 프로브 p_i는 t_i의 부분 서열에 상보적이고 모두 일정한 길이 l인 문자열 $\{A, C, G, T\}^l$로 가정한다. 따라서 프로브의 정의에 따른 제한조건과 두 목적함수는 다음과 같다.

$$g(P)=\sum_i\sum_{j\neq i} subseq(\overline{p_i},t_j),$$

$$f_{Cross-Hyb}(P)=\sum_i \min_{j\neq i} FEnergy(p_i,t_j),$$

$$f_{Self}(P)=\sum_i FEnergy_{self}(p_i).$$

여기서 $subseq(\overline{x},y)$는 두 서열 x와 y에서 x에 상보

적인 서열 $\bar{x}$가 y의 부분 서열로 포함되어 있을 때만 1의 값을 가지고 그 외는 0의 값을 갖는 지시 함수(indicator function)이며, $FEhergy(x,y)$는 두 서열 x와 y가 가장 안정적으로 결합하기 위해 필요한 자유에너지를 나타낸다. 마찬가지로 $FEhergy_{self}(x)$는 한 서열 x가 이루는 2차 구조 중에서 가장 안정적인 구조에 필요한 자유에너지를 나타낸다.

위의 목적함수 값은 OligoArray[2]에 사용된 변형된 Mfold[12] 프로그램을 이용하여 계산하였고 NN 모델(nearest neighbor model)[13]에 기반하였다. 이들 자유에너지의 값이 작을수록 안정적인 결합이라 할 수 있는데, $f_{Cross-Hyb}(P)$와 $f_{Self}(P)$는 목표 유전자가 아닌 유전자와의 결합이나 스스로의 2차 구조 등 프로브의 기능을 저해할 수 있는 결합의 안정성을 표현하므로 이들 함수의 값이 클수록 우수한 프로브이다.

이상에서와 같이 프로브 디자인은 $g(P)=0$이라는 제한조건을 만족시키면서 두 목적함수 $f_{Cross-Hyb}(P)$와 $f_{Self}(P)$를 최대화시켜야 하는 MOP로 간주할 수 있다. 또한 이 두 목적함수는 실제 유전자 서열을 대상으로 테스트한 결과 어느 정도 상충되는 관계에 있음이 관찰되어[9], 프로브 디자인이 MOP의 정의에 잘 맞음을 보여준다.

3. 다중목적함수 진화 알고리즘을 이용한 프로브 디자인

2.3절에서의 정의를 바탕으로 MOEA를 프로브 디자인 문제에 적용하였다. MOEA를 이용한 프로브 디자인 과정은 그림 2에서와 같이 크게 3단계를 거치는데, 첫번째 단계는 적절한 프로브 탐색 영역을 찾는 전처리 과정이고, 두 번째 단계는 찾아진 영역에서 ε-다중목적함수 진화 알고리즘(ε-MOEA, ε-multiobjective evolutionary algorithm)를 이용해서 프로브를 탐색하는 최적화 과정이며, 마지막 단계는 MOEA의 결과로 나온 다양한 프로브 집합들 중에서 우선적으로 사용될 수 있는 프로브를 추천해 주는 후처리 과정이다. 또한 이러한 전체 과정을 web상에서 처리할 수 있는 온라인 인터페이스(EvoOligo, http://cbit.snu.ac.kr/EvoOligo)를 제공한다.

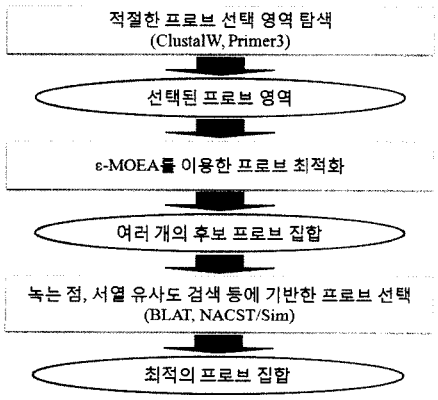

그림 2 프로브 디자인 과정

3.1 프로브 영역 선택을 위한 전처리 과정

목표 유전자 서열의 어느 부분이나 프로브가 될 수 있지만, 효율적인 탐색을 위하여 보다 적합한 프로브 서열을 찾을 수 있을 것으로 예상되는 영역 내로 탐색을 제한하였다. 이를 위하여 목표 유전자들의 서열들을 다중 정렬(multiple alignment)하여 공통적으로 보존되어 있는 영역은 제외하고, 각각의 목표 유전자만의 특색이 있는 영역을 찾는 전처리 과정을 도입하였다. 이것은 목표 유전자들 사이에 연관성이 있어서 서열상에 공통점이 있다는 것을 전제로 하는데, 현재 유전체 전체의 발현 분석이 가능한 상용 마이크로어레이 서비스가 존재하는 것과는 별개로, 특정 유전자군에 대한 추가 분석을 위한 프로브 디자인이 필요한 경우도 많다는 것을 감안한 가정이다.

각각의 목표 유전자 별로 특색이 있을 것으로 예상되는 영역을 찾기 위하여 전처리 과정에서는 중합효소 연쇄 반응(PCR, polymerase chain reaction)에 필요한 프라이머(primer) 디자인 도구인 Primer3[14]를 사용하여 후보 PCR 영역들을 찾은 다음 이를 ClustalW[15]에서 다중 정렬을 통해 생성한 대표 서열과 비교하여 적절한 영역을 선택하여 이를 프로브 탐색 영역으로 사용하였다. 여기서 PCR 영역을 사용하는 것은 PCR 프라이머 선택 과정에서 프라이머 사이의 영역이 본래의 서열이나 프라이머와 결합하는 일을 최소화하도록 선택되므로 프로브 탐색에도 적절하기 때문이다. 그러나 프로브 탐색 영역은 다중 정렬된 서열과 비교하여 가능한 한 여러 유전자에 공통적인 부분을 피하도록 선정하였다(그림 3의 3)). 구체적인 전처리 과정은 그림 3에 자세히 설명되어 있다.

전처리 단계에서 이후의 최적화 단계에서 탐색할 영역이 결정되므로, 필요한 인자들(α, β, γ)은 목표 유전자 서열들 사이의 유사성을 고려하여 결정되어야 한다.

3.2 ε-다중목적함수 진화 알고리즘을 이용한 프로브 최적화

프로브 최적화 단계에서는 다른 MOEA와의 비교에서 탁월한 성능을 보인 ε-MOEA[16]를 적용하였다. ε-MOEA는 steady-state 구조의 진화 알고리즘에 기반한 MOEA로서 ε-dominance를 이용한 일종의 군집화(clustering) 효과와 현재 세대까지 non-dominate인 해들을 별도의 집합(archive)으로 보관하는 방법을 통하여

1) 전체 목표 유전자 서열을 ClustalW를 이용하여 정렬시켜 IUPAC 코드로 표현된 대표 서열을 생성한다.
2) 목표 유전자 각각에 대하여 Primer3로 좌우 프라이머(primer) 쌍의 후보를 모두 생성한 다음, 목표 유전자 별로 다음의 3)~5) 단계를 거쳐 프로브 탐색 영역을 정한다.
3) 각각의 프라이머 쌍에서 좌우 프라이머 각각에 대하여 다음을 계산하여 그 값이 지정된 범위를 만족시키지 못하는 프라이머를 제외한다.
   a) 비-불명료염기 비율(non-degeneracy ratio, β): 프라이머의 길이와 프라이머에 해당하는 대표 서열에서 여러 개의 염기를 가리킬 수 있는 코드로 이루어진 부분의 비율.
   b) 비-공백 비율(non-gap ratio, γ): 프라이머의 길이와 프라이머에 해당하는 대표 서열에서 공백을 나타내는 코드로 이루어진 부분의 비율.
4) 3)의 단계를 통과한 좌우 프라이머를 조합하여 이루어진 쌍에서 프라이머 사이의 거리와 전체 유전자 서열 길이의 비율(α)이 일정 범위를 만족시키지 못하는 프라이머 쌍을 제외한다.
5) 4)의 단계를 통과한 프라이머 쌍에서 좌우 프라이머의 녹는점의 차가 가장 작은 프라이머 사이의 영역을 해당 유전자의 프로브 탐색 영역으로 사용한다.

그림 3 프로브 탐색 영역 선택을 위한 전처리 과정

우수한 성능을 보였다[17]. 여기서 ε-dominance[17]란 ε-MOEA의 가장 큰 특징으로서 하나의 해가 다른 해를 dominate하기 위해서 최소한 ε만큼 더 나은 목적함수값을 가져야 한다는 것을 의미한다. X가 Y를 ε-dominate한다는 것은 $X <_{\epsilon} Y$로 표현하고, 수식으로는 다음과 같다.

$$\forall i \in 1,\cdots,M,\ f_i(X) \geq f_i(Y) + \epsilon,$$
$$\exists j \in 1,\cdots,M,\ f_j(X) > f_j(Y) + \epsilon.$$

따라서 이 ε-dominance 개념을 이용하여 non-dominate하는 해들을 선택하면 각각의 해 사이의 거리가 최소 ε이상이 되어 목적함수 공간에서 지나치게 가까운 해들을 제거해주는 효과를 볼 수 있다. 또한 이렇게 해서 찾아진 해들을 별도의 집합에 보관하여 여러 세대에 걸쳐 진화에 참여하게 함으로써 좋은 형질을 계속해서 유전시킬 수 있도록 하였다.

ε-MOEA를 사용한 프로브 최적화 과정은 그림 4와 같다. 각 세대마다 개체군과 archive에서 각각 선택된 부모 해로부터 생성(그림 4의 3))된 새로운 해를 기존의 해들과 비교하여 더 좋을 경우에만 archive나 개체군에 반영(그림 4의 5)와 6))되도록 되어 있다. 또한 프로브 최적화에는 제한조건이 포함되어 있으므로 5)의 단계에서 archive를 갱신할 때 제한조건을 적게 위반하거나 만족시키는 해를 우선적으로 선택하도록 하였다. 또한 제한조건을 만족시키는 해들 사이에서는 dominate되지 않는 해를 우선시하였다. 따라서 archive의 해들은 제한조건을 만족시키면서 차츰 Pareto-최적해에 가까워지게 되며 ε-dominance의 정의상 목적함수 공간의 특정 구간에 밀집되지도 않게 된다.

3.3 최종 프로브 추천을 위한 후처리 과정

한 번에 다양한 해를 찾아준다는 것이 MOEA의 장점이지만, 비용이나 시간 등의 제약으로 인하여 해 집합의 일부만 필요할 경우를 위하여 실제 실험에 우선적으로 사용해 볼 만한 프로브를 추천해주는 단계도 도입하였다.

EvoOligo에서는 세 가지 기준을 두고 이에 따라 서로 다른 세 프로브 집합을 선택하여 사용자에게 추천하도록 하였다. 첫번째 기준은 BLAT[18]으로 측정한 비-목표 유전자와의 서열 유사도(O_1)의 최소화이고, 두번째는 NACST/Sim[19]으로 측정한 비-목표 유전자와의 결합 안정성(O_2)을 최소화하는 것이다. 결합 안정성의 기준으로는 프로브와 비-목표 유전자가 결합했을 때의 녹는점을 측정하여 37℃보다 높은지의 여부를 사용하였다. 마지막으로 프로브 집합 내의 프로브들의 녹는점의 균일도(O_3)의 최소화를 기준으로 이용하였다. 보다 명확하게 정의하면 다음과 같다.

$$O_1 = \sum_i \sum_{j \neq i} block\_count(p_i^{(k)}, t_j),$$
$$O_2 = \sum_i \sum_{j \neq i} I(Tm(p_i^{(k)}, t_j) > 37),$$
$$O_3 = stdev(Tm(p_i^{(k)}, t_i)).$$

여기서 $p_i^{(k)}$는 ϵ-MOEA의 수행 결과 최종 archive의 k번째 프로브 집합에서 i번째 프로브를 의미하고, $block_count(x,y)$는 BLAT을 이용해서 계산한 두 서열 x와 y의 서열 유사도로서 큰 값일수록 높은 유사도를 의미한다. $I(\cdot)$와 $stdev(\cdot)$는 각각 지시 함수와 표준편차를 나타내며, $Tm(x,y)$는 두 서열 x와 y가 가장 안

1) 개체군(population) P를 무작위로 생성한다.
2) Archive A를 다음과 같이 생성한다. $A=\{P|P\in P, \nexists P'\in P, s.t.\ P' \prec P\}$.
3) P와 A로부터 부모를 선택하여 새로운 개체(프로브 집합) C를 생성한다.
 a) P에서 두 개체를 선택한다.
 b) 만약 두 개체 모두 제한조건을 위반하였다면 더 적게 위반한 개체를, 한 개체만 위반하였다면 제한조건을 만족시키는 개체를, 모두 만족시킨다면 dominate하는 개체를 한쪽 부모로 선택한다. 어느 경우에도 해당되지 않을 경우에는 무작위로 하나를 선택한다.
 c) A에서 하나의 개체를 임의로 선택한다.
 d) b)와 c)에서 선택된 개체에 교차연산(crossover)과 돌연변이연산(mutation)을 적용하여 새로운 개체 C를 생성한다.
4) 새로운 개체 C에 대한 목적함수값들을 2.3절의 정의에 따라 계산한다.
5) Archive A를 갱신한다. 이 때 A의 부분집합 R_1, R_2, R_3은 다음과 같이 정의된다.
 $R_1=\{P|P\in A, g(P)>g(C)\}$, $R_2=\{P|P\in A, C \prec_\varepsilon P\}$, $R_3=\{P|P\in A, P \prec_\varepsilon C\}$.
 a) $g(C)\geq 0$이고 $R_1\neq\varnothing$이면, $A=A\cup\{C\}-R_1$.
 b) $g(C)=0$이고, $R_1=\varnothing$, $R_2\neq\varnothing$이면, $A=A\cup\{C\}-R_2$.
 c) $g(C)=0$이고, $R_1=R_2=R_3=\varnothing$이면, $A=\{P|P\in A\cup\{C\}, \nexists P'\in A\cup\{C\}\ s.t.\ P' \prec P\}$.
 d) 이 외의 경우에는 C는 무시되고 A를 갱신하지 않는다.
6) 현재 개체군 P를 갱신한다. 이 때 P의 부분집합 $S_1=\{P|P\in P, P \prec C\}$이다.
 a) $S_1=\varnothing$이면, $P=P\cup\{C\}-\{X\}$. 여기서 X는 P에서 무작위로 선택한 개체이다.
 b) 이 외의 경우에는 C는 무시되고 P를 갱신하지 않는다.
7) 종료조건이 만족되지 않았으면 3)의 단계로 되돌아간다.

그림 4 ε-MOEA를 사용한 프로브 최적화 과정

정적으로 결합했을 때의 가장 높은 녹는점을 나타낸다.

4. 프로브 디자인 적용 결과

제시된 방법의 검증을 위하여 두 가지 프로브 디자인 문제에 적용해보았다. 하나는 고위험성 인유두종 바이러스(HPV, Human Papillomavirus)에 대한 프로브들을 디자인하는 것인데, (주)바이오메드랩의 협조를 통하여 관련 연구[20]에서 자궁암을 일으킬 확률이 높은 것으로 분류된 19종의 고위험성 인유두종 바이러스를 선별하여 사용하였다. 다른 하나는 52종류의 애기장대 칼모듈린 유전자군(AtCaMs/AtCMLs, Arabidopsis Calmodulin multigene family)에 대한 프로브 디자인이다. 애기장대 칼모듈린 유전자군은 6종류의 애기장대 칼모듈린(*Arabidopsis thaliana* Calmodulins)과 46종류의 애기장대 유사 칼모듈린(*Arabidopsis thaliana* Calmodulin like proteins)으로 구성되어 있는데[21], 기능이나 발현상의 특징에 차이가 있지만 유전자 서열의 유사도는 높은 편이어서 이들의 분석을 위해서는 프로브 디자인에 주의를 기울여야 한다.

프로브의 길이는 HPV의 경우 30 염기, AtCaMs/AtCMLs의 경우에는 24 염기로 지정하였다. ε-MOEA에 사용된 인자로는 교차와 변이연산의 확률을 각각 통상적인 값인 0.9와 0.01로 놓았으며, ε값으로는 1을 사용하였다. 그리고 개체군의 크기와 최대 세대수는 각각 50과 300으로 정하였다. 전처리 과정에 사용된 인자 α, β, γ값은 여러 실험을 거쳐 가장 좋은 성능을 보이는 값을 선택하였는데, HPV의 경우 각각 0.05, 0.5, 0.9, AtCaMs/AtCMLs의 경우 0.25, 0.0, 0.7로 정하였다. 그 외에 Primer3, ClustalW, Mfold, NACST/Sim, BLAT, BLAST에 필요한 인자는 기본값을 사용하였다.

성능 평가를 위하여 널리 알려진 프로브 디자인 프로그램인 OligoArray[2]와 OligoWiz[5]를 사용하여 디자인한 프로브와 EvoOligo에서 최적화한 프로브를 비교분석하고, 전처리 과정의 효과에 대하여 살펴보았다.

4.1 프로브의 비교 평가를 위한 척도

각기 다른 두 방법으로 생성된 두 개의 프로브 집합을 정확하게 비교하여 평가하는 것은 어려운 문제이다. 실제 마이크로어레이 실험을 통하여 평가하는 것이 가장 신뢰성 있는 방법이겠으나, 실험시의 환경 조건 등에 영향을 받기 쉽다는 단점도 있다. 그렇지만 프로브의 길이가 50 또는 70 염기인 경우[6]와 20 염기인 경우[7]에 대한 경험적인 지침을 다룬 문헌에 따르면 목표 유전자와의 결합, 비-목표 유전자와의 결합 및 2차 구조에 필요한 자유에너지와 서열 유사도가 마이크로어레이 상에

서 프로브의 신호 세기와 어느 정도 연관을 갖는 것으로 나타났다.

본 연구에서는 이러한 결과를 참조하여 표 1과 같은 척도를 사용하였다. BLAST match와 ΔG_{25}(non-target)은 모두 프로브의 특이성(specificity)에 관한 척도이다. BLAST match는 비-목표 유전자와 프로브의 서열 유사도($identity(p_i, DB-t_i)$)를 계산하여 특이성을 판단하는데 반하여, ΔG_{25}(non-target)은 프로브를 디자인하고자 하는 유전자 중 프로브의 목표 유전자가 아닌 것과의 결합에 필요한 자유에너지를 계산하여 특이성을 측정하게 된다. BLAST match의 계산에는 BLAST 프로그램[22]을 이용하였는데, HPV의 경우 전역 데이타베이스로 NCBI(http://www.ncbi.nlm.nih.gov/)의 "nr" 데이타를 사용했으며, AtCaMs/AtCMLs의 경우에는 TAIR[23]을 사용했다. ΔG_{25}(self structure)는 프로브 자체의 안정적인 2차 구조 형성 정도를 측정함으로서 프로브의 목표 유전자에 대한 반응 감수성(sensitivity)을 알아본다. 마지막의 Tm(°C)은 프로브의 반응 조건의 유사도에 대한 척도이다. BLAST match는 작은 값일수록, ΔG_{25}(non-target)과 ΔG_{25}(self structure)는 모두 의도하지 않은 결합에 필요한 자유에너지 값의 평균을 의미하므로 값이 클수록 좋은 프로브 집합을 의미한다. Tm(°C)에서 녹는점의 평균은 높고 표준 편차는 작을수록 목표 유전자와 안정적으로 결합하면서 프로브들의 반응 조건이 많이 유사한 좋은 프로브 집합을 의미한다.

4.2 기존 프로그램과의 비교 평가

비교 평가 대상으로는 널리 잘 알려진 프로그램인 OligoArray[2]와 OligoWiz[5]를 사용하였다. OligoArray에서는 서열 유사도와 자유에너지 계산 후 단순한 필터링 방법을 통하여 적당한 프로브를 찾아내는 반면, OligoWiz에서는 dynamic programming 방법을 사용하였다. 디자인할 프로브의 길이와 프로브의 녹는점 범위 등은 EvoOligo와 같도록 하고, 그 외의 설정은 두 프로그램 모두 기본값을 사용하도록 하였다. EvoOligo에서 생성된 최종 프로브 집합 중, NACST/Sim으로 측정한 비-목표 유전자와의 결합 안정성이 가장 낮은(즉, O_2를 최소화하는) 프로브 집합과 비교한 결과만 제시되었으나, 다른 기준으로 선택된 프로브 집합과의 비교 역시 같은 양상을 보였다. 또한 HPV의 경우 (주)바이오메드랩에서 전문가가 디자인하여 상용 바이오칩에 사용된 프로브 집합[9]도 비교 대상에 포함하였다.

표 2의 비교 평가 결과와 같이 HPV나 AtCAMs/AtCMLs의 두 경우 모두에서 EvoOligo를 이용해서 디자인한 프로브가 프로브의 반응 특이성이나 감수성 측면에서 다른 두 프로그램이나 전문가가 디자인한 프로브(Biomedlab)보다 더 좋은 것을 알 수 있다. 유전자 서열의 차이가 보다 큰 편인 HPV의 경우, 비-목표 유전자와의 결합 또는 2차 구조에 필요한 자유에너지뿐만 아니라 BLAST를 통한 서열 비교에서도 EvoOligo와 다른 프로그램이나 전문가가 디자인한 프로브와의 차이가 확연히 남을 알 수 있다. 반면에 유전자 서열의 유사도가 높은 편인 AtCaMs/AtCMLs의 경우, EvoOligo에서 생성한 프로브가 여전히 가장 좋은 것으로 평가되기는 했지만, 다른 프로그램과의 차이는 HPV의 경우보다 적었는데, 이는 AtCaMs/AtCMLs에 대한 프로브 디자인이 쉽지 않은 문제임을 보여준다.

그러나 녹는점의 유사도 면에서는 EvoOligo 쪽의 표준편차가 더 크게 나타나는데, 다른 두 프로그램에서 녹는점을 균일하게 맞추기 위하여 우선 필터링을 한 다음 프로브 최적화를 한다는 것과, EvoOligo의 프로브 최적화 과정에서 녹는점의 균일도가 목적함수에 포함되지 않았다는 점을 감안하면 우수한 성능이라고 할 수 있다.

또한 자유에너지 값이나 녹는점은 어느 정도 서열의 길이에 비례하므로, 프로브의 길이가 30인 HPV의 경우에서 길이가 24인 AtCaMs/AtCMLs의 경우보다 전반적으로 자유에너지 값이 더 낮고 녹는점이 더 높게 나오는 경향이 관찰되었다.

4.3 전처리 단계에 대한 분석

마지막으로 프로브 탐색 영역을 결정하는 전처리 단계의 영향을 알아보기 위하여 전처리 과정을 거쳐서 디자인한 프로브와 그렇지 않은 프로브를 비교하였다(표 3). 또한, HPV의 경우, 전문가들이 경험적으로 선택한 프로브 탐색 영역(L1 영역)이 존재하는데, (주)바이오메

표 1 프로브의 비교 평가에 사용된 척도

이름	정의	설명
BLAST match	$\sum_i I(identity(p_i, DB-\{t_i\}) > 0.85)$	BLAST를 이용하여 계산한 전역 데이타베이스에서의 비-목표 유전자와 프로브의 서열 유사도.
ΔG_{25} (non-target)	$(\sum_i \min_{j \neq i} FEnergy(p_i, t_j))/n$	Mfold를 이용하여 계산한 비-목표 유전자와 프로브의 안정적인 결합 정도.
ΔG_{25} (self structure)	$(\sum_i FEnergy(p_i))/n$	Mfold를 이용하여 계산한 프로브의 안정적인 2차 구조 형성 정도.
Tm (°C)	$(\sum_i Tm(p_i, t_i))/n$과 $stdev(Tm(p_i, t_i))$	프로브의 녹는점들의 평균과 표준편차.

표 2 디자인된 프로브의 비교 결과. Biomedlab으로 표시된 항목은 [9]에서 비교에 사용된 전문가가 디자인한 프로브 집합을 의미한다. 프로브의 녹는점은 나트륨 이온 농도 0.5M를 가정하고 NN 모델을 사용하여 계산한 값이다.

	EvoOligo	OligoArray	OligoWiz	Biomedlab	
HPV	2	3	4	4	BLAST match
	-21.2801	25.9314	27.6639	-21.8649	ΔG_{25} (non-target) (kcal/mol)
	-0.8545	-2.4670	-2.2126	-2.0208	ΔG_{25} (self structure) (kcal/mol)
	67.49±3.22	73.59±2.72	72.93±1.27	73.35±5.00	Tm(°C)
AtCaMs/AtCMLs	4	4	5		BLAST match
	-19.5756	20.5116	21.2913		ΔG_{25} (non-target) (kcal/mol)
	-0.7454	-1.3851	0.8898		ΔG_{25} (self structure) (kcal/mol)
	68.33±4.00	69.73±2.57	70.63±2.10		Tm (°C)

표 3 프로브 탐색 영역의 변화에 따른 프로브 성능 비교. Evo는 EvoOligo의 전처리 과정에서 선정한 영역에서 디자인한 프로브, WS는 전처리 과정을 거치지 않고 전체 유전자 서열에서 프로브 탐색을 수행한 경우, L1은 HPV 서열의 L1 영역에서 디자인한 프로브를 각각 의미한다.

	$EvoOligo_{Evo}$	$EvoOligo_{L1}$	$EvoOligo_{WS}$	$Biomedlab_{L1}$	
HPV	2	7	0	4	BLAST match
	-21.2801	-22.6492	-22.9594	-21.8649	ΔG_{25} (non-target) (kcal/mol)
	0.8545	-2.4497	1.0747	2.0208	ΔG_{25} (self structure) (kcal/mol)
	67.49±3.22	71.30±2.44	67.78±3.85	73.35±5.00	Tm (°C)
AtCaMs/AtCMLs	4		3		BLAST match
	-19.5756		-18.5346		ΔG_{25} (non target) (kcal/mol)
	-0.7454		0.9746		ΔG_{25} (self structure) (kcal/mol)
	68.33±4.00		66.81±4.49		Tm (°C)

드랩의 상용 바이오칩에서 이 영역을 대상으로 하여 디자인한 프로브(표 3의 $Biomedlab_{L1}$)와도 비교하였다.

표 3에서 프로브 탐색 영역이 어떻게 정해지는지에 따라 최종적으로 디자인된 프로브의 성능에 많은 차이가 있음을 알 수 있다. HPV에서 전문가가 선정한 영역(L1 영역)에서는, EvoOligo를 이용하여 디자인한 프로브가 녹는점의 편차는 더 작지만, 반응 특이성이나 감수성 측면에서 약간 뒤떨어지는 것으로 나타났다. 그러나 EvoOligo 자체의 전처리 과정에서 선택한 영역에서 디자인한 프로브가 다른 프로브보다 모든 면에서 더 좋은 값을 나타내는 것을 알 수 있다. 그리고 AtCaMs/AtCMLs의 경우, 전체 유전자 서열을 사용한 경우가 반응 특이성이 더 좋지만, 감수성이나 녹는점의 편차가 더 큰 것으로 나타났다. AtCaMs/AtCMLs의 경우 유전자 서열의 전체적인 다양성이 HPV의 경우보다 커서 전처리 단계에서 좋은 프로브 탐색 영역을 찾기가 어려웠는데, 이런 특성 때문에 좋은 프로브 서열을 찾을 수 있는 영역을 놓쳐서 전처리 단계를 거친 프로브의 반응 특이성이 오히려 낮아진 것으로 생각된다.

또한, 전처리 단계에서 프로브 탐색 영역을 제한함으로 인해서 ε-MOEA의 탐색 공간이 줄어들어 수렴 속도를 빠르게 하는 효과를 관찰할 수 있었다(그림 5). 그

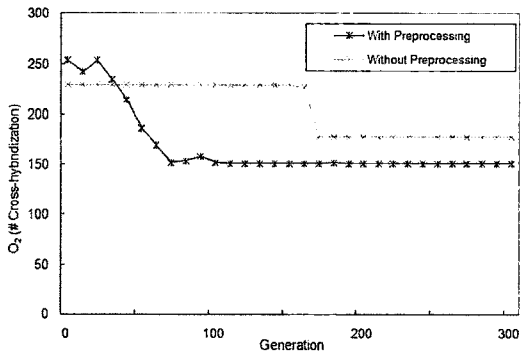

그림 5 전처리 단계에 따른 수렴속도의 변화 비교

림 5는 AtCaMs/AtCMLs의 경우 각각의 세대에서 O_2의 최소값의 변화를 관찰하여 전처리 단계를 사용했을 때와 그렇지 않을 때를 비교한 것인데, 전처리를 사용한 경우가 그렇지 않은 경우보다 더 우수한 프로브를 더 빨리 발견할 수 있음을 알 수 있다. O_1이나 O_3의 경우에도 비슷한 경향성이 나타났으며, HPV에서도 마찬가지의 결과를 보였다.

5. 결 론

본 연구에서는 생물정보학 분야에서 중요한 문제의

하나인 마이크로어레이 실험에 사용될 프로브의 최적화 문제를 다중목적함수 최적화 문제로 정의하고, 이를 다중목적함수 진화 알고리즘인 ε-MOEA를 사용하여 해결하는 방안을 제시하였으며, 19종류의 고위험군 인유두종 바이러스 유전자와 52종류의 애기장대 칼모듈린 유전자군에 대한 프로브 디자인 문제에서 널리 알려진 다른 프로브 최적화 도구와 비교하여 제시된 방법의 성능을 입증하였다. 또한 제시된 방법은 온라인 도구인 EvoOligo(http://cbit.snu.ac.kr/EvoOligo)로 구현되어 관련 연구자가 편리하게 접근하여 사용할 수 있도록 하였다. 진화 알고리즘의 특성상 프로브의 최적화에 약간의 시간이 걸리기 때문에 디자인된 프로브는 사용자에게 e-메일로 전달된다. 또한, EvoOligo는 마이크로어레이의 통합 분석/디자인 플랫폼인 DNAChipBench의 한 부분인 ProbeBench로도 사용되었다(http://cbit.snu.ac.kr/~DNAChipBench/).

현재 EvoOligo에서는 전처리 단계에서 여러 인자를 사용자가 지정해줄 필요가 있는데, 만약 이러한 인자를 입력된 유전자 서열의 분석을 통해서 자동적으로 지정할 수 있다면 더욱 편리할 것이다. 따라서 이후 추가 연구의 방향은 프로브 디자인에 필요한 여러 인자의 자동화 방안을 연구하는 것이 될 것이다.

참 고 문 헌

[1] Gordon, P. M. K. and Sensen, C. W., "Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays," Nucleic Acids Research, Vol.30, No.17, pp.e133, 2004.

[2] 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, Vol.31, No.12, pp.3057-3062, 2003.

[3] Drmanac, S., Stravropoulos, N. A., Labat, I., Vonau, J., Hauser, B., Soares, M. B. and Drmanac, R., "Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes," Genomics, Vol.37, No.1, pp.29-40, 1996.

[4] Herwig, R., Schmitt, A. O., Steinfath, M., O'Brien, J., Seidel, H., Meier-Ewert, S., Lehrach, H. and Radelof, U., "Information theoretical probe selection for hybridisation experiments," Bioinformatics, Vol.16, No.10, pp.890-898, 2000.

[5] Wernersson, R. and Nielsen, H., "OligoWiz 2.0-integrating sequence feature annotation into the design of microarray probes," Nucleic Acids Research, Vol.33, Web Server issue, pp.W611-W615, 2005.

[6] He, Z., Wu, L., Li, X., Fields, M. W. and Zhou, J., "Empirical establishment of oligonucleotide probe design criteria," Applied and Environmental Microbiology, Vol.71, No.7, pp.3753-3760, 2005.

[7] Matveeva, O. V., Shabalina, S. A., Nemtsov, V. A., Tsodikov, A. D., Gesteland, R. F. and Atkins, J. F., "Thermodynamic calculations and statistical correlations for oligo-probes design," Nucleic Acids Research, Vol.31, No.14, pp.4211-4217, 2003.

[8] Wu, C., Carta, R. and Zhang, L., "Sequence dependence of cross-hybridization on short oligo microarrays," Nucleic Acids Research, Vol.33, No.9, pp.e84, 2005.

[9] Shin, S.-Y., Lee, I.-H. and Zhang, B.-T., "Microarray probe design using ε-multi-objective evolutionary algorithms with thermodynamic criteria," Lecture Notes in Computer Science (EvoBio 2006), Vol.3907, pp.184-195, 2006.

[10] Deb, K., Multi-Objective Optimization using Evolutionary Algorithms, John Wiley & Sons, Ltd., 2001.

[11] Bäck, T., Evolutionary Algorithms in Theory and Practice, Oxford University Press, 1996.

[12] Zuker, M., "Mfold web server for nucleic acid folding and hybridization prediction," Nucleic Acids Research, Vol.31, No.13, pp.3406-3415, 2003.

[13] SantaLucia, J. Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," Proceedings of the National Academy of Sciences of the United States of America, Vol.95, No.4, pp.1460-1465, 1998.

[14] Rozen, S. and Skaletsky, H., "Primer3 on the WWW for general users and for biologist programmers," Methods in Molecular Biology, Vol.132, pp.365-386, 2000.

[15] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. and Thompson, J. D., "Multiple sequence alignment with the Clustal series of programs," Nucleic Acids Research, Vol.31, No.13, pp.3497-3500, 2003.

[16] Deb, K., Mohan, M. and Mishra, S., "A fast multi-objective evolutionary algorithm for finding well-spread Pareto-optimal solutions," Kanpur Genetic Algorithm Laboratory, Indian Institute of Technology Kanpur, KanGAL Report 2003002, 2003.

[17] Laumanns, M., Thiele, L., Deb, K. and Zitzler, E., "Combining convergence and diversity in evolutionary multiobjective optimization," Evolutionary Computation, Vol.10, No.3, pp.263-282, 2002.

[18] Kent, W. J., "BLAT-the BLAST-like alignment tool," Genome Research, Vol.12, No.4, pp.656-664, 2002.

[19] Shin, S.-Y., Jang, H.-Y., Tak, M.-H. and Zhang, B.-T., "Simulation of DNA hybridization chain reaction based on thermodynamics and artificial chemistry," Preliminary Proceedings of 9th Inter-

national Meeting on DNA Based Computer, pp.451, 2004.
[20] Walboomers, J. M. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., Snijders, P. J. F., Peto, J., Meijer, C. J. L. M. and Munoz, N., "Human papillomavirus is a necessary cause of invasive cervical cancer worldwide," The Journal of Pathology, Vol.189, No.1, pp.12-19, 1999.
[21] McCormack, E., Tasi, Y.-C. and Braam, J., "Handling calcium signaling: Arabidopsis CaMs and CMLs," Trends in Plant Science, Vol.10, No.8, pp.383-389, 2005.
[22] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, Vol.25, No.17, pp.3389-3402, 1997.
[23] Rhee, S.Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P., "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community," Nucleic Acids Research, Vol.31, No.1, pp.224-228, 2003.

이 인 희

2001년 2월 서울대학교 컴퓨터공학부 학사. 2001년 3월~현재 서울대학교 컴퓨터공학부 석박사 통합과정. 관심분야는 진화연산, 생물정보학, 기계학습, DNA 컴퓨팅

신 수 용

1998년 2월 서울대학교 컴퓨터공학부 학사. 2000년 2월 서울대학교 컴퓨터공학부 석사. 2005년 8월 서울대학교 전기,컴퓨터공학부 박사. 2006년 4월~2008년 3월 NIST 연구원. 2008년 4월~현재 서울대학교병원 의료정보센터 연구교수. 관심분야는 진화연산, data mining, 생물정보학, 의료정보학, 기계학습, DNA 컴퓨팅

조 영 민

2001년 서울대학교 컴퓨터공학부 학사. 2006년~현재 University of California, San Diego Computer Science 박사과정

양 경 애

1998년 제주대학교(학사). 2000년 제주대학교(석사). 2005년 경상대학교(박사). 2006년~현재 서울대학교 바이오정보기술 연구센터(박사후 연수생)

장 병 탁

정보과학회논문지 : 소프트웨어 및 응용
제 35 권 제 6 호 참조