Proceedings of KFIS 2002 Spring Conference, 2002. 5. 25

유전 알고리즘을 이용한 DNA Microarray의 Probe 선택

Probe Selection of DNA Microarrays Using Genetic Algorithms

김선, 장병탁
서울대학교 전기컴퓨터공학부

Sun Kim, Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
E-mail: {skim,btzhang}@bi.snu.ac.kr

요 약

DNA microarray는 분자생물학 및 DNA 컴퓨팅 분야에 널리 사용되고 있는 실험 도구이다. DNA microarray를 이용하는 한 예는 알려진 유전자 집합을 바탕으로 하여 hybridization을 통해 새로운 DNA 서열을 분석하는 것이다. 이를 위한 가장 간단한 방법은 알려진 유전자의 모든 서열을 DNA microarray 상에 올려놓는 것이지만 이는 결과의 정확도 및 칩 제작비용 면에서 비효율적이다. 따라서 일반적으로는 유전자 서열 정보를 파악한 후 일련의 DNA 서열을 선택하는 probe 디자인 과정을 거친다. 그러나 현재 유전자 서열을 바탕으로 최적의 probe 집합을 찾는 결정적인 방법이 존재하고 있지 않다. 이에 본 논문은 oligo DNA microarray을 이용한 DNA 서열 분석 문제에 있어서 가능한 많은 유전자를 인식하면서 최소의 probe 개수를 갖는 집합을 찾는 방법을 제안한다. 제시된 방법은 가능한 probe 집합들로 해집합을 구성한 후, 유전 알고리즘을 이용한 진화 과정을 통해 목적하는 probe 집합을 찾는다. 본 논문에서는 GenBank로부터 얻은 일련의 유전자 집합을 대상으로 실험하였으며 그 결과를 분석하였다.

1. 서론

생물정보학의 최근 연구들은 rRNA(ribosomal RNA) 유전자에 대한 분석을 통해 도움을 받고 있으며, 이를 통해 이전에 설명되지 못했던 많은 유전 조직들을 인식할 수 있게 되었다[1, 2]. rRNA 유전자을 이용하는 문제 중 하나는 DNA microarray를 사용해 알려지지 않은 DNA 서열을 인식하는 것이다. DNA microarray는 유리 표면 위에 일련의 DNA 조각들로 이루어져 있는 점들로 구성되며 hybridization을 통해 유전자 발현 정보 또는 DNA의 유전적인 다형성 등을 알아낼 수 있다. 현재 가장 많이 사용되는 DNA microarray 형식은 cDNA 및 oligonucleotide 방법으로 만들어지는 두 가지 형태의 칩이다. Oligonucleotide 방법의 장점은 사용자가 각 유전자 서열에서 다른 유전자들과 중복되거나 유사한 서열이 있을 경우 이를 피해 일련의 DNA 서열을 선택할 수 있다는 것이다.

본 논문에서는 oligonucleotide microarray를 다루며 probe 집합을 선택하기 위해 cDNA (complementary DNA) 서열을 인식하는 문제를 적용한다[3]. 여기에서 DNA microarray는 일련의 DNA oligonucleotide 집합으로 이루어지며, 분석하고자 하는 rDNA는 rDNA clone library로 구성된 후 hybridization 과정을 통해 분류된다. DNA microarray상의 일련의 DNA oligonucleotide를 probe라고 하며 이 probe 집합을 선택하는 것은 DNA microarray 제작에 있어서 가장 중요한 문제 중 하나가 된다. 한편 DNA microarray상에 존재하는 probe 집합의 개수는 곧 필요로 하는 hybridization 횟수를 의미하고 이는 실험 결과의 정확도 및 microarray의 제작 비용과 밀접한 관계가 있다. 따라서 유전자 서열 분석에 있어서 최적의 probe 집합은 가능한 최소

의 probe 개수로 구성된다고 추측해 볼 수 있고, 이는 곧 hybridization 결과에 대한 정확한 결정 및 최소한의 비용과 연결된다. 그러나 현재 DNA microarray에서 최적의 probe를 디자인하기 위한 결정적인 방법은 존재하지 않는 상황이다.

본 논문에서는 알려진 유전자 집합을 바탕으로 가능한 많은 유전자를 구분하면서 최소한의 개수를 가지는 probe 집합을 찾고자 하며, 이를 위해 유전 알고리즘을 이용한 방법을 제안한다. 제시된 방법은 가능한 일련의 probe들로 해집합을 구성하며 유전적인 진화 연산과정을 통해 최적의 probe 집합을 찾으려 시도한다.

2장에서는 관련 연구를 설명하며, 3장에서는 본 논문에서의 probe 선택 문제를 서술한다. 4장은 유전 알고리즘을 이용한 probe 선택 방법에 대해 설명하고, 5장에서는 GenBank(NCBI)로부터 얻어진 유전자 데이터를 바탕으로 한 실험 및 결과를 분석한다.

2. 관련 연구

DNA microarray에서의 probe 디자인 문제는 microarray 제작 및 실험 결과와 관련해된 가장 중요한 문제 중 하나이다. 따라서 그 중요성에 따라 microarray 제작과 관련된 여러 소프트웨어들이 probe를 디자인하는 기능을 제공한다. 그러나 그 중요성에 비해 학문적인 연구는 상대적으로 많이 수행되어 있지 않은 상태이다.

Drmanac *et al.*은 rDNA clone들의 서열 정보에 기반한 빈도수를 바탕으로 probe를 선택하는 방법을 제시하였다[3]. 그러나, rDNA는 서로 중복된 영역이 많이 존재하며, 따라서 빈도수를 이용한 probe 선택방법에는 한계가 존재하게 된다. Herwig *et al.*은 probe 선택을 최적화 문제로 보고 정보 이론에서의 엔트로피 최대화 문제를 이용한 probe 선택 기법을 제시하였다[4]. 여기에서 probe는 DNA 서열에 대해 클러스터링이 잘되는 방향으로 진행되도록 선택되며, 이를 위해 엔트로피가 최대가 되도록 하는 probe를 선택하도록 하였다. 또한 free energy 및 melting temperature를 기준으로 한 probe 선택 기법도 존재한다[5, 6]. 이 방법은 한 유전자에 대한 최적의 probe 서열은 그 유전자에 대해서는 hybridization free energy가 최소가 되면서 다른 유전자들에 대해서는 최대가 되도록 구성 될 것이라는 것에 착안하고 있다. Bourneman *et al.*은 probe 선택 문제를 두 가지 유형으로 나누어 각 문제를 해결하고자 하였다[7]. 첫째는 probe의 개수를 고정한 상태에서 가능한 많은 clone들을 구분할 수 있도록 하는 probe 집합을 선택하는 문제이며, 이를 위해 simulated annealing을 이용한 방법을 제시하였다. 둘째 문제는 고정된 clone에 대해서 이를 구분할 수 있는 최소한의 개수로 구성되는 probe집합을 찾는 것으로써, Lagrangian relexation을 이용한 동적 프로그램 방법을 제안하였다.

3. Probe 선택 문제

본 논문에서는 DNA microarray의 probe 선택을 위해 oligonucleotide fingerprinting 문제를 정의한다.

DNA microarray에서의 hybridization 과정을 통한 형광 발현 결과(fluorescence response)는 clone 안에서 probe가 발생되는 빈도 수에 비례한다고 말할 수 있다. 즉 clone 안에서 한 probe에 해당하는 시퀀스가 많이 존재할 때와, 적게 존재할 때의 형광 발현 정도는 달라지게 된다.

예를 들어, 다음과 같은 두 개의 clone c_1, c_2가 있다고 하자[7].

c_1 = AAACCTGA, c_2 = AAACATAAA

이 경우 clone 서열에서의 probe 서열 발생 빈도 수를 가지고 형광 발현 정도를 생각해 볼 수 있으며 이는 r단계로 나눌 수 있다. 만약 r = 1, probe p = CCT 라면, probe 서열은 c_1 서열 안에 한번 나타나는데 반해 c_2 에는 나타나지 않으므로 이 때 probe p 에 의해 c_1과 c_2는 구별 가능하다고 말할 수 있다.

위 clone에서 probe p = AAA인 경우를 생각해 보면, r = 1 일 때는 c_1과 c_2를 구별할 수 없지만, 형광 발현 단계를 r = 2로 했을 경우에는 c_1과 c_2는 구별 가능하게 된다는 것을 알 수 있다. 왜냐하면 서열 AAA가 c_1에서는 한번 나타나지만, c_1에는 두 번 나타남으로 인해 hybridization 결과 형광 발현 정도가 달라지기 때문이다.

한편 위와 같은 clone 및 probe의 서열의 일치만을 고려한 hybridization 결과 예측은 실험실 환경에서 얻는 결과와는 일치하지 않는다. 간단하게는 hybridization 과정에서 두 DNA 서열이 서로 일치하는 부분을 가지고 있다는 것만으로 서로 결합하지는 않기 때문이다. 뿐만 아니라 hybridization 자체가 화학적인 반응이기 때문에 실험 조건도 결과에 많은 영향을 끼치게 된다. 그러나 본 논문에서는 DNA 서열 사이의 화학적인 연관성 및 실제 실험 환경을 배제한 이상적인 hybridization 결과만을 가정한다. 그렇지만 본 논문에서 제안한 방법은 주어진 조건에서 유전적인 진화 원리를 이용해 최적화된 답을 얻고자 하는 것이기 때문에 hybridization에 영향을 주는 다른 조건들을 이용한 경우로 확장할 수가 있다.

위에 설명한 hybridization 과정을 Borneman *et al.*에 나와 있는 fingerprint 문제로 정의하면 다음과 같다[7]. clone 집합을 $C=\{c_1,c_2,\cdots,c_m\}$, probe 집합을 $P=\{p_1,p_2,\cdots,p_n\}$ 이라고 하자. 이

때, clone c_i에 대한 hybridization 결과는 probe p_j $(j=1,...,n)$ 에 대한 형광 발현 결과 값으로 구성되는 벡터로 표현할 수 있고, 이를 clone c_i의 fingerprint, 즉, $fingerprint(c_i)$ 라고 한다. 그리고 clone c_1, c_2에 대한 fingerprint 값이 만약 $fingerprint(c_1) \neq fingerprint(c_2)$ 라면, c_1과 c_2는 주어진 probe 집합에 의해 구별 가능하다고 말한다.

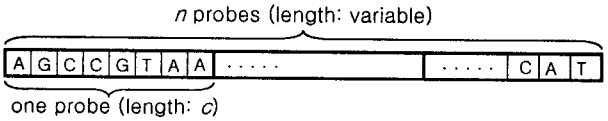
정리하면 probe 선택 문제는 유전자 서열로 구성되어 있는 clone 들을 대상으로 형광 발현 단계 r에 대해 fingerprint 값을 서로 다르게 하는 DNA 서열 조각 집합, 즉, probe 집합을 찾는 문제로 정의할 수 있다.

4. Probe 선택 알고리즘

3장에 설명한 probe 선택문제를 바탕으로 구분할 수 있는 유전자 수를 최대로 하고 집합 개수가 최소인 probe 집합을 찾기 위해 본 논문에서는 유전 알고리즘을 사용한다. 유전 알고리즘은 자연계의 진화 원리를 응용한 것으로써 결정적인 풀이방법이 존재하지 않는 문제들에 대해 좋은 성능을 보여주는 기법이다. 유전 알고리즘을 이용해 문제를 풀기 위해서는 먼저 한 개체에 해당하는 해(solution)를 정의해야 하며, 이를 바탕으로 가능한 해들로 구성된 초기 해집합을 생성한다. 해집합은 유전적인 연산과정(selection, crossover, mutation)을 통해 다음 세대로 진화를 계속해 나가며 최적 해에 접근해 간다. 다음 세대를 생성하기 위한 부모해의 선택은 해집합에 포함되어 있는 개체들의 품질에 비례해서 임의로 선택되며 품질은 적합도 함수(fitness function)에 의해 결정된다. 선택된 부모해는 교차(crossover) 및 변이(mutation)를 통해 자식해를 생성한다. 이렇게 생성된 자식 해들은 해 집단 중 적합도가 낮은 반절의 염색체와 대체된다. 본 논문에서 적용한 개체의 표현방법 및 유전 연산자를 설명하면 다음과 같다.

4.1 개체의 표현

한 개체는 가능한 해, 즉, 가능한 probe 집합을 나타내며 이는 다음 그림과 같이 구성된다. 한 개의 probe는 고정된 길이 c를 가지고 있으며, n개의 probe가 모여 개체가 구성된다(그림 1). 따라서 본 논문에서는 가능한 높은 유전자 분별력을 가지면서 최소한의 n을 갖는 해를 찾는 게 목적이다.



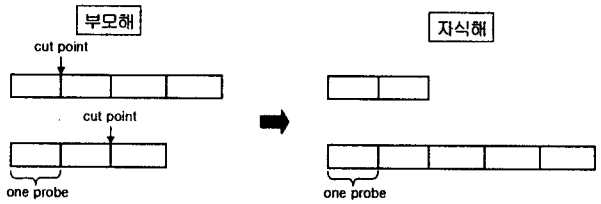
<그림 1> 개체 표현

4.2 적합도 함수

개체의 품질을 나타내는 적합도 값은 fingerprint를 기준으로 삼는다. Hybridization 결과는 clone과 probe의 쌍으로 구성되는 $C \times P$형태의 행렬로 나타내어지며 적합도 값은 probe에 대한 clone의 구별 가능 정도를 비율로 나타낸다. 즉, probe 집합 P가 주어진 모든 유전자 서열을 구별할 수 있으면 그 개체의 적합도 값은 1이 되며, 모든 유전자 서열을 구별할 수 없다면 적합도 값은 0이 된다.

4.3 유전 연산자

부모해의 선택을 위해서 본 논문에서는 적합도 값에 비례해 선택될 확률이 결정되는 roulette wheel 선택 방법을 이용한다. 선택된 부모해는 교차연산을 통해 자식해를 생성하게 되는데 제시된 알고리즘에서의 개체의 특징은 probe의 개수에 따라 다양한 길이를 가지게 된다는 것이다. 따라서 <그림 2>에 나타난 방법과 같이 각 부모해에 별도의 cut point를 적용, 자식해를 생성하는 교차 방법을 이용한다. 이는 각 부모해의 같은 지점을 cut point로 지정했을 때 자식해는 항상 부모와 같은 길이의 해가 나옴으로써 해의 다양성 문제가 제한될 수 있다는 단점을 해결해 준다.



<그림 2> 교차 연산

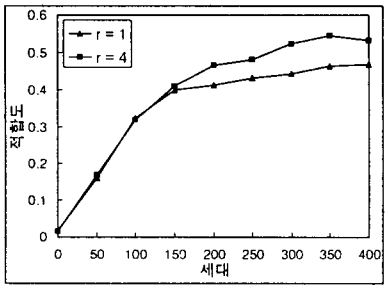
또한 교차 연산에 있어서의 제약조건은 cut point지점은 항상 probe와 probe의 경계선 상이어야 한다는 것과 자식해를 생성했을 때 미리 정의한 최대 probe 개수를 넘지 않도록 해야 한다는 것이다. 만약 최대 probe개수를 지정하지 않으면 자식해의 probe개수가 계속 커진다는 문제점이 발생한다.

교차 연산을 수행한 후에는 해의 다양성을 유지하기 위한 변이 연산을 수행한다. 본 문제에서 수행할 수 있는 변이는 다음 두 가지를 생각해 볼 수 있다. 한가지는 probe단위의 변이를 수행하는 것이며, 다른 한가지는 DNA nucleotide(A, G, C, T) 단위의 변이를 수행하는 것이다. 본 연구에서는 후자의 방법을 통해 변이를 수행한다. probe단위의 변이를 수행하면 해의 다양성을 고려했을 때 장점이 있지만, probe단위의 교체로 인해 주요 정보가 사라져 버릴 수 있는 가능성이 있기 때문이다.

5. 실험 및 결과

실험은 GenBank(NCBI)로부터 뽑은 500여개의 ribosomal gene 데이터를 대상으로 하였고 hybridization의 형광 발현 단계는 binary(r = 1)인 경우와 non-binary(r = 4)인 경우로 나누어 실험하였다. 아래 제시된 실험결과는 해집합의 개체 수 300, 한 probe의 서열 길이 8, 한 개체에서 허용되는 최대 probe 개수를 20으로 제한했을 때 얻은 평균 치수이다.

<그림 3>은 세대에 따른 평균 적합도 변화를 나타낸 것으로써 세대가 계속됨에 따라 적합도 값이 커지는 것을 볼 수 있다. 이는 제시된 알고리즘이 시간이 지날수록 가능한 많은 유전자를 구별하는 probe 집합을 찾아가고 있음을 의미한다.



<그림 3> 적합도 변화

최적해는 일정 세대가 지난 후 최대의 적합도 값을 가지는 개체로 가정하였으며, 선택된 probe 집합의 평균 결과는 <표 1>과 같다. 선택된 probe 집합의 평균 개수는 20개였으며, 그 probe 집합으로 구별할 수 있는 평균 유전자 수는 binary fingerprint의 경우는 331개, non-binary fingerprint인 경우는 366개였다. 실험에서 제한한 최대 probe 개수가 20인데 20미만의 개수를 가지는 probe 집합이 선택되지 못한 것은 해집합의 진화과정에 있어서 적합도의 판단 기준이 구별할 수 있는 유전자의 개수였기 때문이다. 따라서 probe 개수가 최소인 집합을 찾아가는 것을 더 많이 고려하기 위해서는 적합도 함수에 이를 반영해야 할 것이다.

	binary(r=1)	non-binary(r=4)
probe 개수	20	20
구별할수 있는 유전자	331	366

<표 1> 선택된 probe 집합

6. 결론 및 향후과제

본 논문은 DNA microarray의 probe 선택 문제에서 가능한 많은 유전자 서열을 구별하면서 최소한의 크기를 갖는 probe 집합을 찾기 위해 유전 알고리즘을 적용한 방법을 제시하였다. 유전 알고리즘은 문제를 해결하는데 결정적인 방법이 없을 경우 효율적인 기법으로 알려져 있으며 probe 선택 문제에 있어서도 좋은 대안이 될 수 있음을 보여주었다.

한편 본 논문에서는 probe의 길이가 8인 경우에 대한 실험 결과만을 담고 있다. 그러나 실제 microarray 제작에 있어서 probe의 길이는 일반적으로 20이상이다. 따라서 probe의 길이가 늘어남에 따른 알고리즘의 성능 변화를 알아보고 그에 따라 발생되는 문제점 등을 해결해야 한다. 또한 현재 DNA 서열만을 고려한 이상적인 hybridization 결과를 가정하고 있다. 따라서 그 외 hybridization에 영향을 주는 환경을 고려한 probe 선택 문제가 연구되어야 한다.

감사의 글

본 연구는 교육부 BK21 사업과 산업자원부 차세대신기술사업에 의하여 일부 지원되었음.

참고 문헌

[1] Pace, N. R., A Molecular View of Microbial Diversity and the Biosphere, *Science*, 276, pp. 734-740, 1997.

[2] Simon, N., et al., Oligonucleotide Probes for the Identification of Three Algal Groups by Dot Blot and Fluorescent Whole-Cell Hybridization, *The Journal of Eukaryotic Microbiology*, 47(1), pp. 76-84, 2000.

[3] Drmanac, S., et al., Gene Representing cDNA clusters defined by hybridization of 57,419 Clones from Infant Brain Libraries with Short Oligonucleotide Probes, *Genomics*, 37, pp. 29-40, 1996.

[4] Herwig, R., et al., Information Theoretical Probe Selection for Hybridisation Experiments, *Bioinformatics*, 10, pp. 890-898, 2000.

[5] Li, F. and Stormo, G. D., Selecting Optimum DNA Oligos for Microarrays, *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 200-207, 2000.

[6] Li. F. and Stormo, G. D., Selection of Optimal DNA Oligos for Gene Expression Arrays, *Bioinformatics*, 17, pp. 1067-1076, 2001.

[7] Borneman, et al., Probe Selection Algorithms with Applications in the Analysis of Microbial Communities, *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, pp. 39-48, 2001.