염기서열 디자인에 사용되는 적합도 함수 분석

이인희○ 신수용 장병탁
서울대학교 컴퓨터 공학부 바이오 지능 연구실
{ihlee○, syshin, btzhang}@bi.snu.ac.kr

Analysis of Fitness Functions for Sequence Design

In-Hee Lee○ Soo-Yong Shin Byoung-Tak Zhang
Biointelligence Laboratory, School of Computer Science and Engineering
Seoul National University, Seoul 151-742, Korea

요 약

염기서열 디자인은 DNA computing, 생물정보학 등의 분야에서 실험 설계시 고려해야할 중요한 문제 중의 하나이다. 이 문제는 다양한 조건을 만족시키는 최적의 염기서열 집합을 생성하는 조합 최적화 문제로 생각될 수 있으며, 염기서열이 갖추어야 할 조건을 적합도 함수로 사용한 진화 연산 등의 방법이 적용되어 왔다. 본 논문에서는 여러 논문들에서 제시된 적합도 함수의 구체적인 형태를 해 공간 상에서 조사해 보았으며, 각 적합도 함수 간의 관계도 분석해 보았다.

1. 서 론

DNA를 이용한 생물학적 실험 및 DNA microarray 등에 사용되는 염기서열들은 서로 다른 서열을 가지면서도 비슷한 실험 조건을 가져야 하며, 그 외에 실험자가 요구하는 사항을 만족시키도록 디자인되어야 한다. 따라서 염기서열 디자인 문제는 여러 가지 조건을 만족시키는 최적의 염기서열 집합을 생성하는 조합 최적화 문제 중의 하나로 생각될 수 있다. 이 문제를 위해 많은 연구자들이 다양한 기법을 적용해 왔으며[1-2][4-10], 그 중 하나로 진화연산을 들 수 있는데, 이 때 염기서열이 갖추어야 하는 조건을 적합도 함수로 표현할 수 있다.

본 논문에서는 염기서열 디자인 문제에서 주로 사용되는 조건들을 적합도 함수로 표현하고, 그 형태를 해 공간 상에서 조사해 보았다. 염기서열 디자인 문제에서 해 공간은 생성해 낼 수 있는 가능한 모든 염기서열의 집합인데, 이는 지나치게 방대한 공간이므로 적당한 범위로 제한하여 조사해 보았다. 또한, 각각의 조건들 중에는 상호 관계가 있는 조건이 있을 수 있는데 이에 대해서도 분석해 보았다.

2. 염기서열 디자인에 사용되는 적합도 함수

본 장에서는 염기서열 디자인 문제에 주로 사용되는 조건을 적합도 함수로 표현하고, 해 공간에서의 형태를 조사해 보았다. 여기서 해 공간은 생성 가능한 모든 염기서열의 조합이지만, 이는 지나치게 방대하여 길이 5인 염기서열 2개의 집합으로 제한하였다. 여기에서는 염기서열 디자인 문제를 서로 같은 길이의 서로 다른 염기서열 집합을 생성하는 것으로 가정하였다.

2.1 유사도(Similarity)

유사도는 한 염기서열 집합 내에서 각 염기서열 간의 비슷한 정도를 측정한다. 유사도가 높은 경우 서로 다른 염기서열 A, B에 대한 상보 염기서열이 잘못 상보결합할 가능성이 있다. 따라서 염기서열 집합 내의 유사도가 낮을 수록 좋은 염기서열 집합이라고 할 수 있다.

염기서열 집합의 유사도는 집합 내에 포함된 염기서열의 모든 쌍에 대한 유사도의 합으로 정의되며, 한 염기서열 쌍에 대한 유사도는 다음 그림과 같이 shift를 고려하여 상자 안의 두 서열에서 같은 위치에 같은 염기가 위치한 경우의 수로 계산된다. 예를 들어, 아래 그림에서 위의 경우 유사도는 2이고, 아래의 경우 유사도는 1이다.

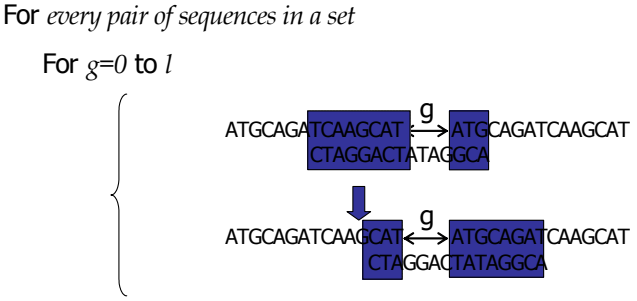

그림 1 염기서열 집합의 유사도 계산

유사도의 해공간 내에서의 형태는 그림 2와 같다. 해공간 내에서의 형태가 매우 불규칙함을 알 수 있다.

2.2 H-Measure [3]

H-Measure 는 한 염기서열 집합 내에서 서로 다른 염기서열간 상보결합 정도를 측정하는 적합도로서, 이전 절의 유사도와 비슷한 방법으로 계산된다. 단, 이 때는 같은 위치에 있는 염기가 같은 경우가 아니라 상보적인 경우의 수로 계산된다. 예를 들어, 위의 그림 1에서 위의 경우 H-Measure는 0이고, 아래의 경우는

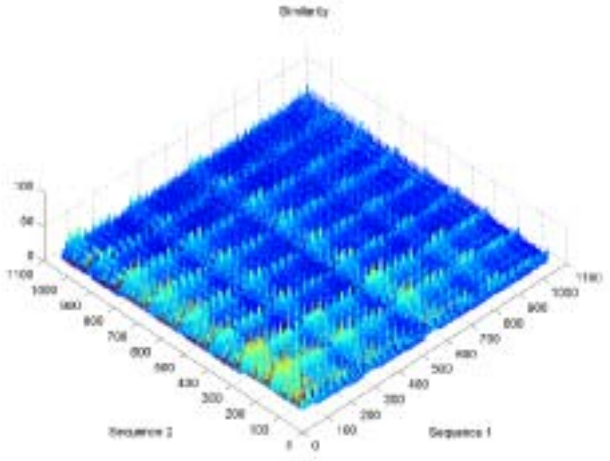

그림 2 길이 5인 염기서열의 모든 쌍에 대한 유사도

5이다.

염기서열 집합의 H-Measure 역시 모든 쌍의 H-Measure의 합으로 정의되므로, H-Measure 값이 작은 염기서열 집합이 서로 다른 서열간 상보결합 가능성이 낮다고 할 수 있다.

H-Measure의 해 공간 내에서의 형태는 다음과 같다.

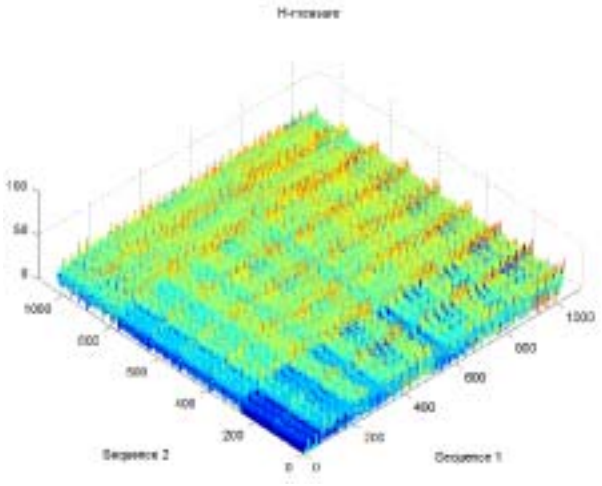

그림 3 길이 5인 염기서열의 모든 쌍에 대한 H-measure

2.3 Hairpin

Hairpin은 하나의 염기서열에 대하여 스스로 휘어져 2차 구조를 형성할 가능성을 측정한다. 염기서열 집합의 Hairpin 값은 집합에 속한 모든 염기서열의 Hairpin 값의 합으로 계산되며, 역시 값이 작을수록 2차 구조를 형성할 가능성이 낮아진다.

하나의 염기서열에 대하여 Hairpin은 다음 그림에서와 같이 계산한다. 즉, 스스로 휘어서 2차 구조를 형성할 수 있는 가장 작은 크기에서부터 시작하여 위치를 이동해 가면서 상자 안 부분의 H-measure를 계산하여 2차 구조 가능성을 계산한다. 이후 주어진 길이에서 형성할수 있는 가장 큰 크기가 될 때까지 휘어진 부분의 크기를 늘려가면서 계산하여 합산한다.

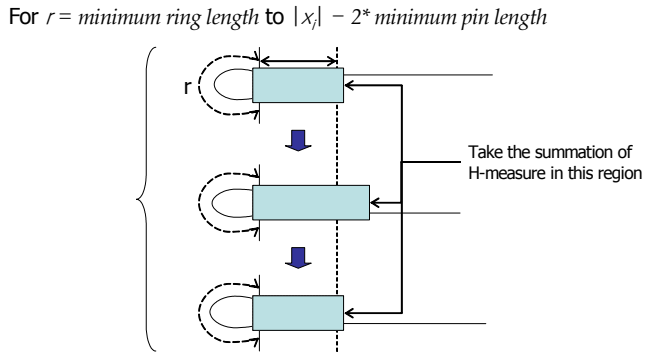

그림 4 염기서열 집합의 2차 구조 가능성 계산

이 적합도 함수의 경우 2차구조를 형성하려면 최소한 휘어진 부분에 5개, 양 끝의 길이 각 6개씩의 염기가 필요하므로 길이 17 이상인 염기서열에서만 계산할 수 있으므로 여기서는 해공간을 구해볼 수 없었다.

2.4 Continuity

Continuity는 하나의 염기서열에서 같은 염기가 연속해서 출현하는 정도를 측정한다. 염기서열에서 같은 염기가 연속해서 출현하면 DNA 가닥이 휘어져서 상보적인 가닥과의 결합을 방해하거나, microarray용 probe의 경우, 주변 cell의 probe의 상보결합을 방해할 가능성도 있다. 따라서 continuity를 줄이는 것이 필요한데, 긴 염기서열을 디자인할 경우 일정 개수 이상 연속해서 나타나는 현상은 거의 피할 수 없다. 따라서 임계값을 정한 후 그 이상 연속해서 나타나는 경우만 continutiy 계산에 포함된다. 임계값은 경험적으로 4정도로 제한된다.

이 경우 해 공간에서의 형태는 매우 단순하므로 생략하였다.

2.5 물리적 조건 (Tm, GC의 비율)

본 논문에서는 비슷한 물리적 조건을 가지는 염기서열의 집합을 생성하는 문제로 가정하였다. 이 때 고려되는 조건으로는 절반 정도의 이중 가닥이 단일 가닥으로 분리되는 온도인 Tm과 염기서열 내에 G나 C가 차지하는 비율이 있다. 두 조건 모두 실험자가 원하는 값에서 일정 이상 벗어나지 않아야 한다.

여기에서 고려한 적합도 함수에서는 집합 내의 각 염기서열의 Tm 값과 GC의 비율이 목표값에서 벗어난 정도를 절대값으로 측정하여 합산한 값을 염기서열 집합의 적합도로 사용하였다. 따라서 이 값이 작을수록 원하는 조건을 잘 갖춘 집합이라고 하겠다.

이 경우 역시 해 공간에서 적합도 함수의 형태가 매우 단순하므로 생략하였다.

3. 적합도 함수 간의 관계

이전 장에서 설명한 각 적합도 함수들은 서로 독립적이 아니라 상호 간의 관계를 가지고 있다.

우선 가장 중요한 적합도 함수에 해당하는 유사도와 H-Measure를 생각해보자. 두 가지 적합도 함수 모두 최소화하는 염기서열 집합이 최적의 집합이다. 그러나 H-Measure를 최소화하는 집합의 경우 유사도를 증가시킬 수 있다. 즉, 모두 같은 염기로 이루어진 염기서열의 집합이 H-Measure를 최소화시킬 수 있다. 그러나 이 경우, 유사도나 continuity가 증가하게 된다. 유사도나 continuity를 증가시키지 않는 범위에서 H-Measure를 줄이려면 위와 같은 경우에서처럼 작은 값을 얻을 수가 없다. 이와 같은 두 적합도 함수 간의 관계를 다음 그림에서 확인해 볼 수 있다.

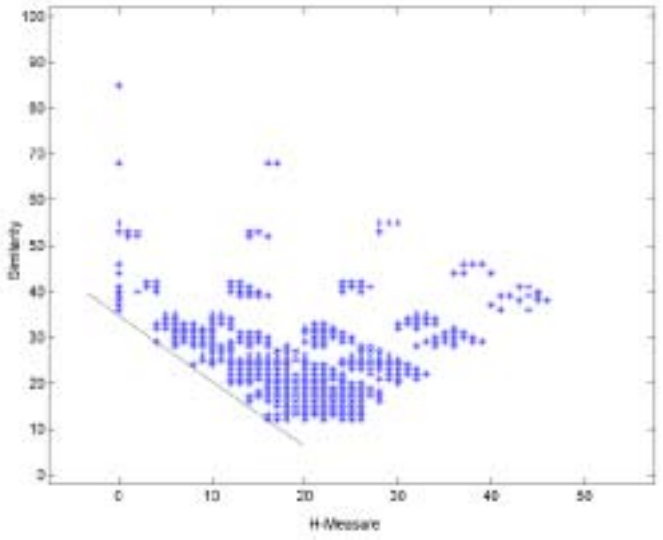

그림 6 H-Measure와 Similarity 사이의 상관도 (직선은 최적값의 쌍을 연결한 것이다).

그림 6은 Tm과 Hairpin을 제외한 네 가지 적합도 함수의 관계도이다. 여기서 GC 비율은 목표값을 50%로 하여 계산한 것이다. 각 그래프의 세로축은 해당 그래프와 같은 행에 있는 적합도 함수의 값이고, 가로축은 같은 열에 있는 적합도 함수의 값이다. 여기서 위에서 설명한 H-Measure와 유사도 간의 경쟁관계 외에도, 유사도와 GC 비율 간에도 약한 경쟁관계가 있음을 볼 수 있다. 즉 그래프에서 왼쪽 아래 부분을 보면 GC 비율이 증가할 때 유사도는 감소하는 양상을 관찰할 수 있다.

4. 결론

본 논문에서는 염기서열 디자인 문제에서 주로 쓰이는 적합도 함수를 정리하고, 해 공간 내에서의 형태를 제시하였다. 또한, 각 적합도 함수간에 음 또는 양의 상관 관계가 있음을 제시하였다.

감사의 글: 본 연구는 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정 연구실 과제에 의해 지원되었음.

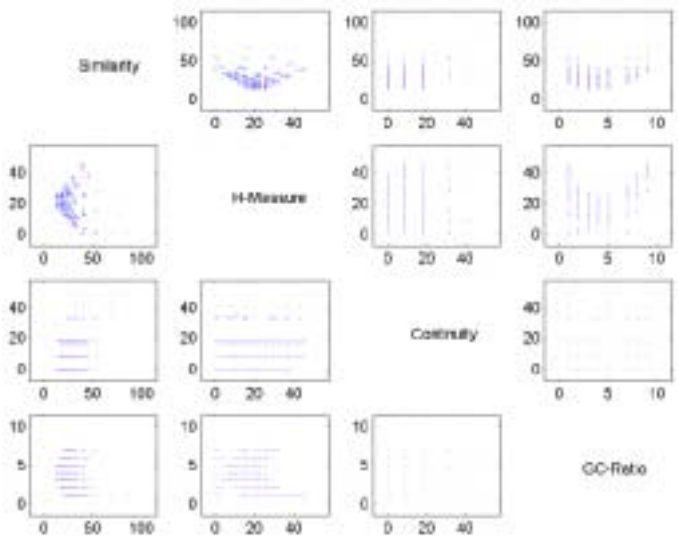

그림 5 네 가지 적합도 함수 (유사도, H-measure, continuity, GC 비율) 사이의 상관도

참고 문헌

[1] B.-T. Zhang and S.-Y. Shin, *"Molecular Algorithms for efficient and reliable DNA computing"*, Proceedings of Genetic Programming 1998, pp.735-742, 1998.
[2] S.-Y. Shin, D.-M. Kim, I.-H. Lee, and B.-T. Zhang, *"Evolutionary sequence generation for reliable DNA computing"*, Proceedings of Congress on Evolutionary Computation, pp. 79-84, 2002.
[3] M. Garzon, P. Neathery, R. Deaton, R. C. Murphy, D. R. Franceschetti, and S. E. Stevens Jr., *"A new metric for DNA computing"*, Proceedings of Genetic Programming 1997, pp.472-478, 1998.
[4] A. Marathe, A. E. Condon, and R. M. Corn, *"On combinatorial DNA word design"*, Proceedings of 5th DIMACS Workshop on DNA Based Computers, pp.75-89, 1999.
[5] A. J. Hartemink, D. K. Gifford, and J. Khodor, *"Automated constraint-based nucleotide sequence selection for DNA computation"*, Proceedings fo 4th DIMACS Workshop on DNA Based Computers, pp.227-235, 1998.
[6] U. Feldkamp, S. Saghafi, W. Banzhalf, and H. Rauhe, *"DNAsequencegenerator - a program for the construction of DNA sequences"*, Proceedings of 7th International Workshop on DNA Based Computer, pp.179-188, 2001.
[7] M. Arita and S. Kobayashi, *"DNA sequence design using templates"*, New Generation Computing, v.20, pp. 263-277, 2002.
[8] R. Deaton, J. Chen, H. Bi, M. Garzon, H. Rubin, and D. H. Wood, *"A PCR-based protocol for in vitro selection of non-crosshybridizing oligonucleotides"*, Preliminary Proceedings of 8th International Meeting on DNA Based Computer, pp.105-115, 2002.
[9] F. Tanaka, M. Nakatsugawa, M. Hagiya, K. Komiya, T. Shiba, and A. Ohuchi, *"Developing surpport system for sequence design in DNA computing"*, Proceedings of 7th International Workshop on DNA Based Computer, pp.340-349, 2001.
[10] D. C. Tuplan, H. Hoose, and A. Condon, *"Stochastic local search algorithms for DNA word design"*, Preliminary Proceedings of 8th International Workshop on DNA Based Computer, pp. 311-323, 2002.