유전자 알고리즘을 이용한 RNA Pseudoknot 예측

이동규[O], 한경숙
인하대학교 컴퓨터 공학부
erlybird@hanmail.net[O], khan@inha.ac.kr

Predicting RNA Pseudoknots Using a Genetic Algorithm

Dongkyu Lee[O] and Kyungsook Han
School of Computer Science & Engineering, Inha University

요 약

RNA 분자의 pseudoknot 구조는 이차 구조의 loop에 있는 염기와 이 loop 외부에 있는 염기와의 결합으로 생성되는 삼차 구조 요소이다. pseudoknot은 삼차 구조 형성에 필수적인 구조 요소일 뿐만 아니라, RNA 분자의 기능에 중요한 영향을 미친다. pseudoknot을 포함한 RNA 구조를 예측하는 문제는 매우 어려우며 많은 계산을 필요로 한다. 현재까지, 병렬 구조를 갖는 수퍼 컴퓨터에서 유전자 알고리즘을 이용한 프로그램의 예측 결과가 가장 우수하다고 알려져 있다. 그러나 이 프로그램은 수퍼 컴퓨터에서만 운용되기 때문에 일반 연구자가 쉽게 사용하기 어려운 단점이 있다. 본 논문은 유전자 알고리즘을 이용한 PC 기반의 pseudoknot 예측 프로그램에 대하여 기술한다. 실험 결과는 PC 기반에서도 유전자 알고리즘을 이용하여 pseudoknot을 포함한 RNA 구조를 효과적으로 예측하고 있음을 보인다.

1. 서 론

RNA의 생물학적 기능을 완전히 결정하는 것은 삼차 구조이다. 그러나 현재까지 대부분 종류의 RNA의 삼차 구조는 아직 밝혀지지 않았다 [1, 2, 3]. RNA의 삼차 구조를 예측하기 위하여 필요한 열역학적 특성 및 시험 데이터가 많이 존재하지 않는 까닭이다. Pseudoknot은 삼차 구조를 형성하는 구조 요소 중 가장 대표적인 것으로서 이에 대한 열역학적 특성은 불완전하게나마 알려진 모델이 존재한다. 따라서 현재까지 알려진 RNA의 pseudoknot 예측 프로그램은 이러한 열역학적 특성을 이용한 컴퓨터 모델링 방법을 사용한다 [4, 5]. 그러나 pseudoknot을 포함한 RNA 구조의 예측은 매우 많은 계산량을 요구하는 어려운 문제이다. 현재 pseudoknot 예측을 위한 프로그램들이 사용하는 알고리즘을 보면 크게 greedy 알고리즘과 dynamic 프로그래밍, 그리고 유전자 알고리즘을 사용하고 있음을 알 수 있다 [6, 7, 8]. 이 중에서 현재 가장 우수한 성능을 나타내고 있는 것은 유전자 알고리즘에 기반한 예측 프로그램이다. 이 프로그램은 병렬 처리를 하는 수퍼 컴퓨터에서 병렬 유전자 알고리즘을 이용한다. 하지만 이 프로그램은 수퍼 컴퓨터에서만 운용되기 때문에 일반 연구자들이나 사용자가 쉽게 사용하기 어렵다는 단점이 있다. 따라서 본 연구에서는 누구나 사용이 편리한 PC 기반의 pseudoknot 예측 프로그램을 개발하고자 한다. 개발한 프로그램은 단일 프로세서를 사용하는 PC에서 운용되며 유전자 알고리즘을 사용한다. 본 연구에서는 이 프로그램을 이용하여 RNA의 pseudoknot 예측과 관련된 유전자 알고리즘의 특성을 알아보고자 하였다.

유전자 알고리즘은 자연계의 진화 현상을 구현하는 알고리즘으로 다윈이 주장한 적자 생존과 자연도태의 원리를 이용하여 개발된 알고리즘이다. 주로 확률적 탐색이나 학습 및 최적화를 위한 기법으로 사용되며 찾고자 하는 해를 표현하기 위한 유전자형과 각 문제에 따라 다르게 정의되는 목적 함수 및 연산자를 필요로 한다. 유전자 알고리즘은 목적 함수 또는 적합도 함수라는 것을 이용하여 새로 형성된 개체의 적합성을 판단하고, 임의로 생성된 모집단으로부터 선택, 교배와 돌연변이라는 과정을 반복해가면서 좀더 나은 적합성을 갖는 전역 해를 찾아가는 과정이라고 할 수 있다. 유전자 알고리즘을 RNA의 구조예측 문제에 연관시켜보면, 적합도 함수는 구조가 갖는 자유 에너지 값을 이용할 수 있고, 각 구조를 나타내는 요소들을 표시할 수 있는 유전자형을 정의해야 한다. 이러한 유전자형과 목적 함수가 정의되면 임의로 생성된 모집단으로부터 유전자 알고리즘의 기본 연산과정인 교배와 돌연변이를 통해 에너지 값을 최소로 하는 구조를 찾아낼 수 있다.

일반적으로 유전자 알고리즘을 이용한 탐색 문제에서 어려운 점은 다음과 같다. 우선 진화를 거듭하면서 해를 찾다가 그 과정을 중단해야 되는데 그러한 종료 조건을 찾아내는 것이 쉽지 않다. 진화를 거듭하면서 구한 해의 수렴도 및 다양성을 이용하여 이러한 조건을 정하고 있으나, 최적의 조건을 부여하는 것은 어려운 문제이다. 그리고 유전자 알고리즘의 진화 오퍼레이터들은 각각 다양한 파라미터와 정책들을 갖고 있다. 선택 오퍼레이터의 경우 여러 가지 선택 전략이 존재하며, 돌연변이 및 교배 오퍼레이터의 경우에도 여러 가지 파라미터와 랜덤 함수에 따라 그 결과 값이 틀려질 수 있는데 이를 최적화하는 것도 쉬운 일이 아니다. 또한 문제에 맞는 유전자형을 결정하는 것도 중요한 고려 사항이 된다.

2. Pseudoknot 예측 알고리즘

유전자 알고리즘은 그 진화 과정 동안 이전 세대의 개체를 교체하고 새로운 개체를 생성하는 방법에 따라서 단순 유전자 알고리즘 (Simple genetic algorithm), 정상적 (Steady-state) 유전자 알고리즘 , 증분적 (Incremental) 유전자 알고리즘, 딤 (Deme) 유전자 알고리즘 등으로 구분할 수 있다. 단순 유전자 알고리즘은 비중복 집단과 선택적 엘리티즘 (Elitism)을 사용하여 전각의 세대에서 완전히 새로운 개체로 이루어진 집단을 형성한다. 엘리티즘이란, 현재 세대의 최적 개체가 다음 세대로 전달되는 것을 의미한다. 정상적 유전자 알고리즘은 중복 집단을 사용한다. 증분적 유전자 알고리즘은 각각의 세대에서 단지 하나 혹은 둘의 자손만 생성하여 새로 생성된 자손을 집단에서 대체하는 방법이다. 딤 유전자 알고리즘은 정상적 유전자 알고리즘을 사용하여 병렬로 다중 집단을 진화시키는 방법을 말한

다 [9]. 이 알고리즘은 병렬 처리를 지원할 수 있는 구조에서만 적용 가능하다. 본 연구에서는 PC 기반의 예측을 수행하기 때문에 단일 프로세서에서 수행 가능한 단순 유전자 알고리즘과 정상적 유전자 알고리즘을 사용하였다.

어떤 유전자 알고리즘을 사용하는가에 따라서 그 초기화 방법 및 유전자 오퍼레이터는 달라질 수 있다. 기존의 연구에서는 병렬로 다중 집단을 진화시키는 등 유전자 알고리즘을 사용하며 이때는 random하게 초기 모집단을 발생시키고, 교배 및 돌연변이 오퍼레이터를 재정의하여 사용하였다 [8]. 본 연구에서는 단일 프로세서에서 유전자 알고리즘을 수행하기 때문에 시스템의 성능을 보완하기 위하여 초기 모집단의 발생과정을 랜덤하게 발생시키지 않고, 가능한 모든 경우를 다 포함 가능하도록 하였다. 그리고 초기 모집단에 모든 가능한 경우가 다 포함되어있기 때문에 진화 과정에서 사용하는 유전자 오퍼레이터의 경우, 별도의 오퍼레이터를 정의하지 않고 일반적인 유전자 알고리즘의 기본 오퍼레이터를 사용하였다.

3. 초기화 및 유전자 연산

본 연구에서는 유전자 알고리즘을 사용하기 위하여 RNA의 구조를 링크드 리스트 (Linked List) 형으로 표현한다. 링크드 리스트를 사용한 것은 구조에 참여할 stem의 수에 제한을 두지 않고 구조를 관리할 수 있고, 구조에 다른 stem의 첨가 및 삭제가 용이하기 때문이다. 본 프로그램에서 사용한 링크드 리스트의 각 노드에 들어가는 정수 값은 해당 염기서열에서 찾은 가능한 stem의 인덱스 값이다. 즉 각 노드는 전처리 과정에서 형성된 stem list의 인덱스 값과 앞, 뒤의 링크를 위한 포인터를 멤버 변수로 갖는 클래스로 구현되었다. 그리고 stem을 리스트에 삽입하기 위한 함수와 리스트로부터의 삭제 기능을 수행하는 함수를 멤버 함수로 갖고 있다.

리스트 형의 유전자는 유전자 알고리즘에서 필요로 하는 기본적인 연산자로서 초기화, 돌연변이, 교배의 세 가지 기본적인 연산자를 갖는다. 이 연산자들을 이용하여 초기화된 집단을 만들고 진화를 수행한다. 다음의 그림 1은 리스트 형태의 유전자를 사용할 때 정의되는 기본적인 연산자들을 나타낸 것이다. 초기화 연산에서는 유전자 알고리즘에서 사용할 초기 모집단을 생성하는 방법을 정의한다. 그리고 교배 방법은 일점 교배, 이점 교배 등의 방법을 사용할 수 있다. 본 연구에서는 일점 교배를 사용한다. 그리고 돌연변이 연산은 노드의 값을 교환하는 교환 돌연변이 연산과 노드를 파괴하는 파괴 돌연변이 연산 등을 사용할 수 있는데 본 연구에서는 파괴 돌연변이 연산을 사용하였다. 본 연구에서 사용하는 리스트 유전자의 경우 교환 돌연변이 연산은 그 의미가 없다. 왜냐하면 본 연구에서 노드에 삽입된 순서는 의미가 없기 때문이다.

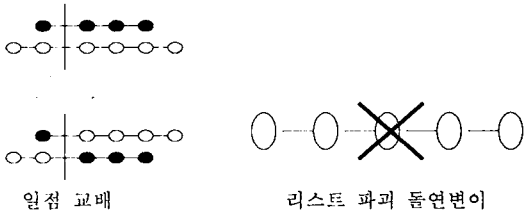

그림 1. 리스트형 유전자의 기본 연산

유전자 알고리즘에서 사용할 초기 집단의 생성을 위해 본 연구에서는 다음과 같은 방법을 사용한다. 우선 구조를 예측할 RNA 염기 서열을 읽어들인 후, 전처리 과정에서 모든 가능한 stem들의 list를 생성한다. 이 때 정해진 개수 이상의 base pair를 갖는 경우만을 stem으로 인정한다. 이렇게 구한 stem list에 대해서 각각의 stacking 에너지를 계산한 후, 에너지 값에 따라 내림차순으로 stem list를 정렬한다. stem list의 정렬 후, 구한 stem list의 크기만큼 초기 모집단을 생성한다. 우선 에너지가 가장 큰 stem을 첫 번째 구조에 참여할 stem으로 선정한다. 그리고 stem list를 탐색하면서 구조에 포함 가능한 모든 stem들을 다 삽입한다. 구조에 포함 가능한 stem은 이미 구조에 참여한 stem과 중복되는 염기가 존재하지 않는 경우를 말한다. 그리고 이 때 pseudoknot의 포함 여부에 따라서 pseudoknot 여부도 구조에 포함될 stem을 결정하는 한 조건이 될 수 있다. 이런 방식으로 첫 번째 구조가 결정되고 나면, 두 번째로 에너지가 큰 stem을 기준으로 하여 다시 stem list를 탐색하여 두 번째 구조에 참여하는 stem들을 결정한다. 이렇게 해서 차례대로 전부 stem list의 크기만큼의 모집단을 생성한다. 이런 방식으로 모집단을 생성하기 때문에 전처리 과정에서 만들어진 모든 stem list를 다 포함하는 초기 모집단을 생성할 수 있다. 이 과정을 그림 2에 나타내었다.

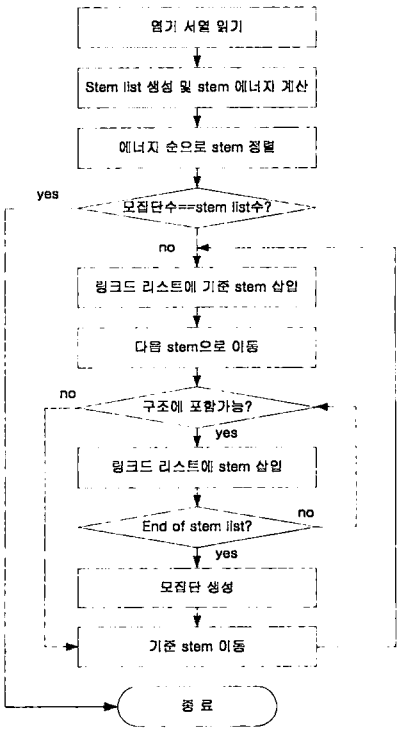

그림 2. 모집단 생성 과정 순서도

모집단이 생성된 후 진화를 시작하여 유전자 알고리즘의 기본 연산인 교배 연산과 돌연변이 연산을 거쳐 새로운 세대를 형성하고 선택하는 과정을 반복한다. 이 때 적용되는 목적함수는 구조의 에너지 값과 구조에 참여한 stem의 수에 일정한 factor를 곱하여 계산하였다. 진화가 종료되는 시점을 결정하기 위해 해의 수렴도나 세대수 등을 이용할 수 있는데 현재는 세대수를 이용하여 종료조건을 결정하였다.

4. 시스템 구현 및 시험 결과

유전자 알고리즘을 이용한 PC 기반의 pseudoknot 예측 프로그램은 192 MByte의 RAM이 장착된 Pentium II PC에서 Inprise 사의 C++ Builder 5.0을 이용하여 개발하였다. 유전자

알고리즘을 지원하기 위한 라이브러리로 MIT의 Mattew Wall에 의해 개발된 C++ 라이브러리인 GALib 2.4.4 [10]를 이용하였다. GALib는 유전자 알고리즘을 지원하기 위하여 약 45개의 클래스와 400여개의 함수로 구성되어 있으며 다양한 형태의 유전자형과 유전자 알고리즘을 지원한다. 예측된 구조를 시각화하기 위하여 본 연구실에서 개발된 시각화 프로그램 [11]을 이용하였다. 개발된 예측 프로그램은 그래픽 사용자 인터페이스를 지원하며 사용자는 마우스를 이용하여 쉬운 조작이 가능하다. 또 별도의 옵션 항목이 있어서 사용자는 pseudoknot의 포함 여부 및 에너지 모델을 선택할 수 있다. 그림 3은 개발한 프로그램을나타낸다

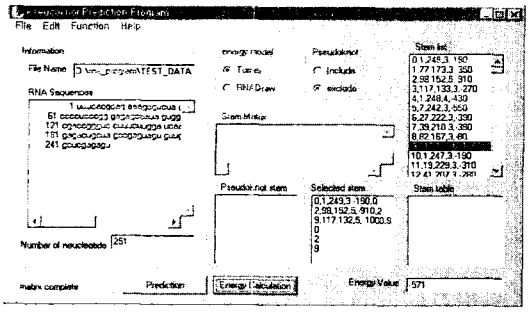

그림 3. 프로그램의 실행 화면

그림 4. MMTV pseudoknot 구조

위의 그림 4는 MMTV의 pseudoknot을 포함한 구조를 나타내고 있다. 총 34개의 염기로 구성되어 있으며 (1, 19, 5), (9, 32, 5) 두 개의 stem으로 구성되어 있다. 이 염기 서열에 대하여 전처리 과정을 거치면 총 13개의 stem list가 형성된다. 그러나 형성된 stem list에는 (9, 32, 5)의 stem이 포함되어 있지 않다. 자세히 살펴보면 이 stem은 원래 (7, 34, 7)의 stem에서 말단의 두 개의 base pair가 탈락된 partially zipped stem이라는 것을 알 수 있다. 위의 염기 서열에 대하여 50회 정도의 시험을 반복했을 때 두 개의 stem이 모두 구조에 포함되는 것을 알 수 있었다. 다른 염기 서열에 대해서도 개발한 프로그램을 이용하여 시험을 수행한 결과, pseudoknot을 포함한 구조의 예측을 만족할 만한 수준으로 예측하는 것을 확인할 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 단일 프로세서에서 사용 가능한 유전자 알고리즘을 이용하여 PC 기반에서 pseudoknot을 포함한 RNA 구조를 예측하는 프로그램을 개발하였다. 개발한 프로그램은 윈도우즈 환경에서 마우스를 사용하는 쉬운 사용자 인터페이스를 지원하며 성능 면에 있어서도 어느 정도의 만족할 만한 성능을 나타낸다. 기존의 방법과 달리 별도의 유전자 오퍼레이터를 사용하지 않고 초기화 방법의 개선을 통하여 만족할 만한 성능을 나타냄을 보였다. 유전자 알고리즘 자체가 요구로 하는 계산량이 많기 때문에 염기 서열의 길이가 매우 커질 경우, 계산시간 및 메모리의 양이 급격히 증가하는 제약이 있지만 염기 서열이 짧은 경우에 대해서는 효과적으로 사용이 가능하다. 본 프로그램을 이용하면 RNA 구조를 예측하기 위한 유전자 알고리즘의 특성 및 입력 파라미터의 최적화를 쉽게 수행할 수 있다.

현재까지 pseudoknot 형성과 관련된 정확한 모델이 알려지지 않았기 때문에 실제 정확한 예측을 위해서는 앞으로의 많은 시험 및 보완이 필요하다. 특히 구조에 참여하는 stem의 unzip 메커니즘은 정확한 예측을 위하여 반드시 필요한 사항이라고 할 수 있다. 향후 관련된 데이터가 밝혀지면 이를 반영하여 좀 더 정확한 예측을 수행하는 알고리즘의 개선 작업이 필요할 것이다.

후기

본 연구는 한국과학재단의 지역대학우수과학자 지원연구 (과제번호 2001-1-30300-018-2)의 지원에 의하여 수행되었음.

참고 문헌

[1] Chen, J.-H., Le, S.-Y., and Maizel, J. V., "A procedure for RNA pseudoknot prediction," CABIOS, 8, 243-248, 1992

[2] Hilbers, C. W., Michiels, P. J. A., Heus, H. A., "New Developments in structure determination of pseudoknots," Biopolymers 48, 137-153, 1998

[3] Ieong, S., Kao, M.-Y., Lam, T.-W., Sung, W.-K., Yiu, S.-M., "Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs," Bioinformatics and Bioengineering Conference, 182-189, 2001

[4] Abrahams, J. P., van den Berg, M., van Batenburg, E. and Pleij, C., "Prediction of RNA secondary structure, including pseudoknotting, by computer simulation," Nucleic Acids Res. 18, 3035-3044, 1990

[5] Freier, S., Kierzek, R., Jaeger, J., Sugimoto, N., Caruthers, M., Neilson, T. and Turner, D., "Improved free-energy parameters for prediction of RNA duplex stability," Proc. Natl Acad. Sci. USA. 83, 9373-9377, 1986

[6] Rivas, E., Eddy S. R. "A dynamic programming algorithm for RNA structure prediction including pseudoknots," Journal of Molecular Biology, 285, 2053-2068, 1999

[7] Akutsu, T., "Dynamic programming algorithm for RNA secondary structure prediction with pseudoknots," Discrete Applied Mathematics, 104, 45-62, 2000

[8] Shapiro, B. A. and Wu, J. C., "Predicting RNA H-Type pseudoknots with the massively parallel genetic algorithm," Computer Applications in the Biosciences, 13, 459-471, 1997

[9] Goldberg, D. E., "Genetic algorithms in search and optimization," Addison-Wesley Pub. Co., 1989

[10] http://lancet.mit.edu/ga/

[11] Han, K. S., Lee, Y. J., Kim, W. T., "Automatic visualization of RNA pseudoknots," ISMB., 2002