RNA의 이차 구조 요소 및 삼차 구조 요소를 추출하기 위한 PDB 구조 데이터 마이닝

임대호° 한경숙
인하대학교 전자계산공학과
kingtiger1°@hanmail.net, khan@inha.ac.kr

Mining the Secondary and Tertiary Structures Elements of RNA from the Structure Data of PDB

Daeho Lim° Kyungsook Han
School of Computer Science and Engineering, Inha University

요 약

이제까지 Protein이나 RNA와 같은 분자의 구조는, 대부분 X-ray crystallography나 Nuclear Magnetic Resonance (NMR) 방법을 통해 분석이 이루어 졌다. 이 방법들은 실제 분자를 직접 원자레벨에서 분석하는 방법으로, 분자를 구성하는 모든 원자의 3차원 좌표 정보를 얻어 낼 수 있다. 원자의 3차원 좌표 정보는 분자의 전체적인 모양과 구조를 이해하는데 유용한 정보이다. 하지만, 분자의 구조를 좀 더 완벽히 이해하기 위해서는 원자 레벨의 좌표 정보 보다는 좀 더 높은 차원에서의 구조 정보가 필요하다. 특히 분자의 구조를 예측하거나, 분자들 사이에 결합 관계를 예측하기 위해서는, 원자 레벨의 정보만으로는 필요한 모든 정보를 얻을 수 없다. 이러한 경우, 분자의 2차원 또는 3차원 구조 요소 (structural elements)가 더욱 좋은 정보를 제공해 줄 수 있다. Protein 분자의 경우, 이미 3차원 좌표 정보를 이용해서, 2차원 구조 요소를 알아내는 자동화된 방법이 알려져 있다. 그러나 RNA의 경우 protein에 비해 알려진 결정 구조가 적기 때문에, 아직까지 2차원 구조 요소나 3차원 구조 요소를 알아내는 자동화된 방법이 알려져 있지 않다. 따라서, 이제까지는 RNA의 구조 요소를 알아내기 위해, 사람이 직접 RNA분자의 3차원 좌표 정보를 분석함으로써 많은 시간과 노력이 필요했다. 이 때문에, 우리는 RNA의 원자들의 3차원 좌표 정보를 이용해서, 2차원 구조 요소와 3차원 구조 요소 정보를 자동화된 방법으로 밝혀내는 알고리즘을 개발하였다. 우리는 분자를 구성하고 있는 원자들의 3차원 좌표 정보를 Protein data bank (PDB)에서 가져왔다. 우리의 알고리즘은 PDB file형태의 데이터라면 protein-RNA 복합체나 RNA 분자 모두에서 RNA의 2차원 구조 요소나 3차원 구조 요소를 얻어낼 수 있다. 우리의 연구는 RNA의 원자레벨의 3차원 좌표 정보를 이용해서 RNA의 구조 요소를 뽑아내는 첫 번째 시도로, 우리의 알고리즘을 통해 얻어진 구조 정보는 RNA의 구조 예측 연구나, protein-RNA complex의 결합 예측 연구에 많은 도움을 줄 수 있으리라 기대된다.

1. 서 론

최근 몇 년간 생물정보학 분야에서는, 생물학에서 이미 밝혀진 데이터를 분석하여, 필요한 정보를 얻어내는 연구가 활발히 진행되어 왔다. 그러나 대부분의 연구가 분자의 서열을 분석하는 연구로 국한되어 왔다. 우리는 이미 알려진 RNA 원자들의 3차원 좌표 정보를 이용해서, 분자의 2차원이나 3차원 구조 요소에 대한 정보를 밝혀내는 알고리즘을 개발하였다. 우리의 알고리즘은 분자의 구조 요소를 알아내기 위해, 분자를 구성하는 원자들 사이의 수소 결합 데이터를 이용한다. 우리는 이 수소 결합 데이터를 분석하여, base-pair를 구성하는 수소 결합들만을 뽑아내고, 이를 base-pair의 종류 [1]에 따라 28가지 종류로 분류 하였다. 그리고 이 정보를 RNA를 구성하는 모든 nucleotide와 비교하여, 분자의 구조 요소 정보를 밝혀냈다. 이제까지 RNA의 2차원 구조 요소나 3차원 구조 요소를 밝히기 위해서는, 분자의 3차원 원자들의 좌표 정보를, 사람이 일일이 분석해서 그 구조 정보를 알아내야만 했다. 따라서 많은 시간과 노력이 필요했다. 우리의 알고리즘을 이용하면, RNA 분자의 구조 정보를 자동으로 짧은 시간에 밝혀냄으로써, 기존의 이러한 문제들을 해결해 줄 수 있다.

2. 배경 지식

PDB: PDB [2]는 Protein Data Bank의 약자로, X-ray crystallography나 NMR 방법을 통해 분석되어진, protein이나 RNA와 같은 분자들의 정보들이 file형태로 저장되어 있는 데이터 베이스이다. 많은 protein-RNA 복합체나 RNA 분자의 3차원 좌표 정보들을 이 데이터 베이스에서 얻을 수 있다.

Base-pair: 두개의 nucleotide의 base가 수소 결합에 의해 안정된 결합을 이룬 형태를 말한다. Base-pair [1]의 분류는 크게 canonical base-pair와 non-canonical base-pair로 나눌 수 있고, 이를 좀 더 세분하여 분류하면, 28가지의 Base-pair로 분류 할 수 있다. 그림 1은 대표적인 base-pair인 G-C Watson-Crick pair이다.

Base-pair 룰: Nucleotide의 base를 구성하는 원자들은 각각 정해진 번호가 있다. 그림 1의 경우 Guanine과 Cytosine의 base 부위의 원자들이, 각각 정해진 번호를 가지고 있는 것을 알 수 있다. Base-pair는 이 정해진 원자들의 번호에 따라 종류별로 구분될 수 있다. 그림 1의 G-C Watson-Crick pair는 Guanine의 6번 산소와 1번 수소, 2번 질소가 Cytosine의 4번, 3번 질소, 2번 산소와 수소 결합을 이룸으로써 base-pair를 형성한다. 이런 식으로, base-pair는 Watson-Crick pair 이외에 Wooble pair, Pyrimidine-Pyrimidine pair, Purine-Purine pair 등 28가지 종류로 분류될 수 있다. 이와 같이 base-pair는 정해진 원자 사이의 수소 결합에 의해 형성되는데, 이를 우리는 base-pair 룰이라고 하였다. 이 룰은 우리의 연구에서 base-pair를 형성하는 수소 결합을 찾아내고, base-pair의 종류에 따라 수소 결합들을 분류 하는데 사용된다.

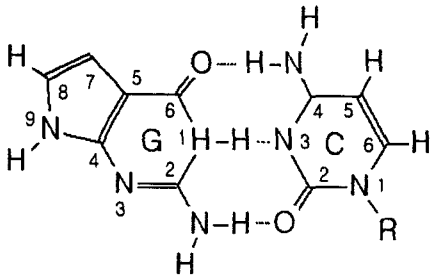

그림 1. G-C Watson-Crick Pair

2. 알고리즘

그림 2는 알고리즘을 단계 별로 나타낸 프로우 차트이다. 알고리즘에서 필요한, RNA 원자들 사이의 수소 결합 정보는, HB-Plus [3]라는 프로그램을 이용하여 뽑아낸다. Step 1은 PDB file을 분석하여, RNA를 구성하는 모든 nucleotide의 정보를 뽑아내는 단계이다. 이 nucleotide 정보는 Step 4에서 사용하기 위해 RNA-SEQ에 저장해 둔다. Step 2에선 HB-Plus를 이용해서 얻어낸 수소 결합 정보를 분석하고, nucleotide의 base 사이에 수소 결합들만을 뽑아낸다. HB-Plus를 이용하여 얻어낸 수소 결합 정보는, 분자를 구성하는 모든 원자들 사이의 수소 결합 정보를 가지고 있기 때문에, base-pair를 구성하는 수소 결합 정보를 얻기 위한 첫 단계로 base사이의 수소 결합만을 뽑아낸다. Step 3에서는, 우선 Step 2에서 뽑아낸 base와 base사이의 수소 결합들 중 base-pair를 이루는데 기여하는 수소 결합들 만을 뽑아내야 한다. 본 논문의 알고리즘은 base사이에 수소 결합이 두개 이상인 경우만을 base-pair로 인정하였다. 따라서 base사이에 수소 결합이라 할지라도, base-pair를 형성하지 못하는 수소 결합은 이 단계에서 제외 시켜야 한다. 이렇게 base-pair를 형성하는 수소 결합들만을 얻어낸 뒤, 이 수소 결합들을 base-pair의 종류에 따라 28가지 종류로 분류 한다. 앞에서 언급한 base-pair 룰은 이 Step 3에서, base-pair를 이루는 수소 결합을 찾아내고, base-pair의 종류에 따라 분류하는 기준으로 사용된다. Step 4에선 최종적으로 base-pair를 이루는 수소 결합들과 Step 1에서 이미 얻어 놓은 RNA-SEQ 정보를 비교하여, 어떠한 nucleotide가 base-pair를 이루는 지를 알아낸다. 그리고 RNA분자를 구성하는 모든 nucleotide들의 순서에 맞게 정렬을 하면, RNA의 2차원 구조 요소나 3차원 구조 요소에 관한 정보를 얻을 수 있다.

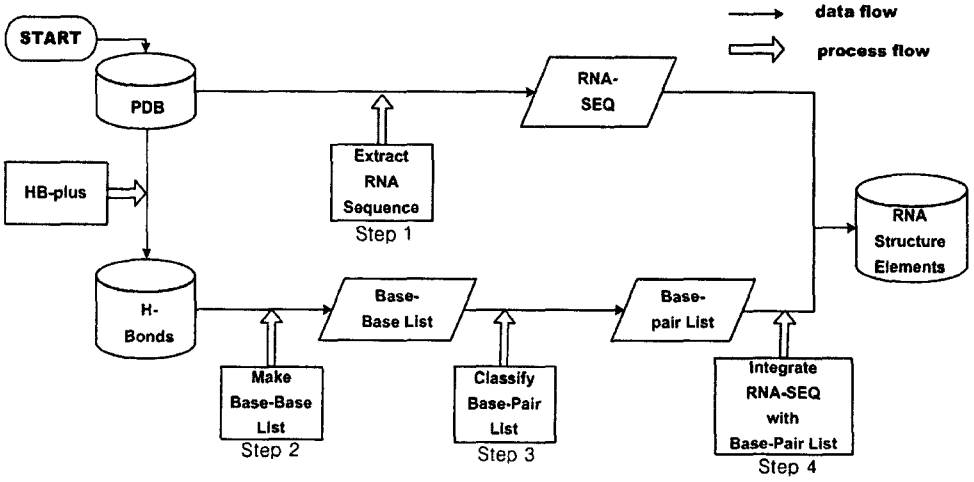

그림 2. PDB file에서 RNA의 2차원 구조나 3차원 구조 요소 정보를 얻어내는 각 단계를 나타낸 플로우 차트.

3. 결과

아래 그림 3은 본 논문의 알고리즘을 이용해서, mouse mammary tumor virus (MMTV)의 pseudoknot 부위 (PDB identifier: 1RNK)를 분석하고, 이 결과를 통해 얻어진 3차원 구조 요소를 입체적으로 나타낸 그림이다. 각 노드들은 nucleotide를 나타내고, 파란 선은 각 nucleotide가 backbone으로 연결되어 있는 모습을 보여 준다. 빨간 점선은 base-pair를 이루는 수소 결합을 나타내 주는데, RNA가 base-pair에 의해 보다 안정된 구조를 취하는 것을 알 수 있다. MMTV virus의 RNA 3차원 구조 요소는 이미 알려져 있다. 우리는 이미 알려진 MMTV virus의 3차원 구조 요소와, 우리의 알고리즘에 의해 얻어진 3차원 구조 요소를 비교해 보았고, 결과가 정확히 일치하는 것을 확인 할 수 있었다.

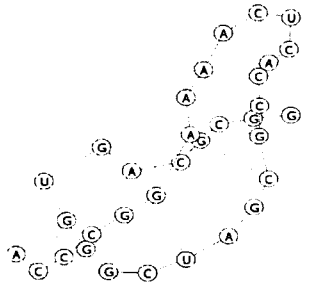

그림 3. Mouse mammary tumor virus (PDB identifier: 1RNK) Pseudoknot 구조

그림 4는 우리의 알고리즘을 이용해서 얻어진, 1DFU (PDB identifier)의 3차원 구조 요소를 나타낸 것이다. 그림에서 보면, RNA가 두개의 chain을 가지고 있고, 이 두개의 chain은 각각 base-pair를 이루는 수소 결합들에 의해 서로 결합되어 있는 것을 알 수 있다. 우리의 알고리즘은 이와 같이 한 개의 chain안에서 이루어지는 base-pair 뿐만 아니라, 서로 다른 chain사이에서 형성되는 base-pair 정보도 뽑아낼 수 있다. 그러므로 두개 이상의 chain을 가진 RNA의 구조 요소도 밝혀 낼 수 있다. 따라서 거의 모든 RNA 분자에 우리의 알고리즘을 적용할 수 있다.

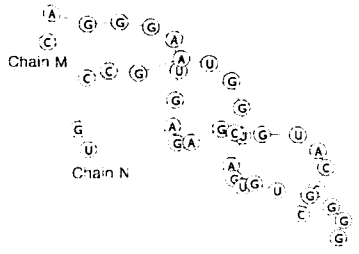

그림 4. Chain을 두개 가지고 있는 RNA (PDB identifier: 1DFU)의 구조.

4. 결론

이제까지는 RNA 분자를 구성하는 원자의 3차원 좌표 정보로부터, 분자의 2차원 구조 요소나 3차원 구조 요소를 얻기 위해서, 사람이 직접 정보를 일일이 분석해야만 했다. 따라서 많은 시간과 노력이 요구되었다. 그러나 RNA 분자의 복잡성과 최근 급격히 불어나는 데이터의 양은 이마저 거의 불가능하게 만들었다. 이러한 문제 때문에, 우리는 RNA 분자의 2차원이나 3차원 구조를 자동으로 쉽게 알아낼 수 있는 알고리즘을 개발하였다. 본 논문에 서술한 알고리즘은 이미 알려진 3차원 원자들의 좌표 데이터에서 필요한 정보를 뽑아내서, 2차원이나 3차원 구조 요소를 얻어내는 데이터 마이닝 알고리즘이다. 이 알고리즘은 몇 단계의 과정을 거쳐 거의 정확하게 RNA의 구조 요소를 알아낸다. 우리의 실험 결과는 우리의 알고리즘이 거의 정확하게 2차나 3차 구조 요소와, base-pair와 base-triple 정보를 뽑아낸 다는 것을 보여 준다. 우리는 우리의 알고리즘이 RNA의 구조 예측 연구와, protein과 RNA의 결합을 예측하는 연구 등에 많은 도움을 줄 수 있을 것이라고 기대한다.

후기

본 연구는 정보통신부 정보통신 선도기반기술개발사업 (과제 번호 01-PJ11-PG9-01BT00B-0012)의 지원에 의하여 이루어졌음.

참고 문헌

1. Tinoco, Jr.: The RNA World (R. F. Gesteland, J. F. Atkins, Eds.), Cold Spring Harbor Laboratory Press, (1993) 603-607
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28 (2000) 235-242
3. McDonald, I.K. Thornton, J.M.: Satisfying Hydrogen Bonding Potential in Proteins, J. Mol.Biol. 238 (1994) 777-793