물리화적인 방법과 생명정보화적인 방법을 활용한 단백질 구조 계산 연구

DOI: 10.3938/PhiT.20.020

이 주 련

Protein Structure Prediction Using Physics and Bioinformatics

Julian LEE

The protein structure prediction can be performed efficiently by combining the methods in physics and bioinformatics. In this article, I review the methods for the computational study of protein structure, based on bioinformatics and physics. Then I discuss how physics and bioinformatics can be combined to perform efficient prediction of protein native structure.

들어가는 글

단백질은 아미노산들이 사슬로 연결되어 만들어지는 분자이다. 아미노산은 20 종류가 있기 때문에, 단백질 서열이라 불리는 아미노산의 연결 순서에 따라 여러 종류의 단백질이 만들어진다. 단백질은 생체 내에서 서열에 따라 고유한 3차원 구조로 접히는데, 이러한 고유 구조는 단백질의 기능을 결정하기 때문에, 계산을 통해 서열로부터 구조를 예측하는 것은 매우 중요한 문제이다. 계산을 통해 단백질의 구조를 예측하기 위한 단백질분자를 비교적 정확히 기술할 수 있는 포텐셜 에너지 함수를 써서 단백질의 물리적인 접힘 과정을 전산 시뮬레이션하는 방법이 있었으나, 단백질 구조 데이터베이스의 정보를 활용하는 생명정보학의 방법을 섞어 씌으로써 더 효율

저자약력

이주련 교수는 미국 브라운대학에서 1994년에 박사학위를 취득하였다. 그 후 이론물리연구센터 연구원, 고등과학원 연구원의 물리학과 계산과학부 연구원 등을 거치고 2003년부터 숭실대학교 의생명시스템학부 및 생명정보학과 대학원 교수로 재직 중이다.
(jul@ssu.ac.kr)

Fig. 1. Relation between biophysics and bioinformatics.

적인 시뮬레이션을 수행할 수 있다. 이 글에서는 생명정보학과 물리학의 단백질 구조 연구 방법에 대해 소개를 하고 이들의 혼합을 통한 단백질 3차원 구조 예측에 대해 설명한다.

생명정보화적인 방법을 활용한 단백질 구조 예측

생명정보학은 분자생물학 실험에서 쏟아져 나오는 방대한 데이터를 분석하여 이로부터 새로운 지식을 얻고자 하는 학문이다. 물리학과 같이 근본적인 법칙에 대한 가정으로부터 출발하여 연역적인 논리에 따라 실험 결과를 예측하는 과학을 가설에 바탕을 둔 과학(hypothesis-driven science)이라 한다면 생명정보학은 데이터를 통계적으로 분석하여 이로부터 귀납적으로 어떤 규칙적인 패턴을 발견해 나간다는 데에서, 데이터에 바탕을 둔 과학(data-driven science)이라 할 수 있다(그림 1). 특히 생명체와 같은 복잡계의 경우 아직까지는 이론적 모형으로부터 계산을 수행하는 생물물리화적인 접근법만으로 현상을 설명하고 예측하는 데에 한계가 있기 때문에, 생명정보화적인 방법은 생물물리학을 보완해주는 역할을 할 수가 있다. 생명정보학의 방법을 단백질 구조 연구에 응용한 예로서, 단백질의 2차 구조와 같은 국소적인 구조의 예측이 있다. 단백질 2차 구조는 단백질의 뼈대 원자들의 수소결합 때문에 생기는 국소적인 구조로서, 크게 α 나선 구조, β 병풍 구조, 코일의 세 가지로 분류할 수 있다.

단백질 2차 구조 예측은 서열은 단백질의 아미노산 서열이

Fig. 2. Secondary structure prediction. For a one-dimensional input string of protein amino acid sequence, the output is the one-dimensional string of secondary structure. The sequence search algorithm PSI-BLAST is used to construct the sequence profile, which contains sequence feature that determines the secondary structure.

라는 일차원적인 정보로부터 2차 구조라는 역시 일차원적인 정보를 예측하는 문제로서(그림 2), 패턴식별 알고리즘을 활용하여 예측을 수행하는데, 이미 답이 알려져 있는 단백질 서열들을 자료로 활용하여 컴퓨터를 학습시킨 후 새로운 서열에 대한 예측을 수행한다. 그런데 단백질의 아미노산 서열 자체에는 단백질의 2차 구조와는 무관한 정보가 많이 들어 있어서 정확한 2차 구조 예측을 방해하는 잡음으로 작용할 수 있기 때문에, 서열로부터 2차 구조를 결정하는 데에 중요한 역할을 하는 성질들을 추출하는 작업을 거쳐야 한다. 예를 들어, 만약 단백질의 2차 구조가 아미노산이 물을 좋아하는가 싫어하는가의 여부로만 결정된다면, 단백질 서열을 20가지 아미노산으로 표현하는 대신 물을 좋아하는 정도에 따라 20보다 적은 몇 가지 글자로 바꿔 쓰는 방법이 있을 것이다. 실제 상황은 이보다 복잡하기 때문에 단백질 서열로부터 중요한 성질을 추출하는 작업 자체도 컴퓨터가 자동적으로 하는 알고리즘이 개발되었는데, 서열 검색 알고리즘인 Position Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST)를 활용하는 방법이다.^[1] PSI-BLAST는 서열 데이터베이스에서 입력 서열과 유사한 서열을 검색해주는 알고리즘으로서 이 역시 일종의 패턴식별이다. PSI-BLAST에서 검색되는 유사 서열은 구조가 밝혀지지 않은 것이 대부분이지만, 입력 서열과 공통의 2차 구조를 갖고 있을 것으로 추정되기 때문에, 검색된 유사 서열들을 나열하여 각 위치에서 아미노산들이 다른 아미노산으로 변이하는 빈도를 표로 나타낸 이른바 서열 옆얼굴(sequence profile)을 계산하면, 서열의 2

Fig. 3. (a) Artificial Neural Network. The pattern space is partitioned according to classes through training with reference data set. The class of the query is determined depending on which region the query falls. (b) k-nearest neighbor methods. The k reference data with the pattern most similar to the query are selected. The class of the query is determined by the majority rule among the k data selected.

차 구조를 결정하는 정보가 들어있다고 생각할 수 있다. 이러한 서열 옆얼굴을 학습 패턴으로 활용하여 패턴식별을 수행하게 되는데(그림 2), 이때 가장 많이 쓰이는 패턴식별 방법이 인공신경망이다. 오늘날 가장 뛰어난 예측 성능을 보이는 Porter^[2]나 PSIPRED^[3]같은 방법이 모두 인공신경망 방법에 바탕을 둔 방법이다. 인공신경망은 패턴의 공간 모든 부분을 소속 클래스로 분할하는데, 이 분할에 쓰이는 곡선을 결정하는 파라미터는 학습 패턴이 제대로 분류되도록 학습을 통해 결정된다. 분할 곡선이 결정되고 나면 2차 구조 예측을 수행할 수 있는데, 새로운 입력 패턴이 어느 영역에 위치하였는가에 따라 2차 구조가 예측된다(그림 3(a)).

패턴식별의 또 다른 방법으로 k-가까운 이웃 방법이 있는데, 입력 패턴과 가장 유사한 학습 패턴 k개를 골라 다수결로 입력 패턴의 클래스를 결정하는 방법이다(그림 3(b)). 예측 시

REFERENCES

- [1] S. F. Altchul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids. Res.* **25**, 3389 (1997).
- [2] G. Pollastri and A. McLysaght, *Bioinformatics.* **21**, 1719 (2005).
- [3] D. T. Jones, *J. Mol. Biol.* **292**, 195 (1999).

에 입력패턴과 모든 학습 패턴과의 거리를 계산해야 하므로 계산 시간이 더 소요되는 단점이 있으나, 인공신경망은 유한 개의 클래스로 분류하는 데에 유용한데 비해 k -가까운 이웃 방법은 입력 패턴과 유사한 학습 패턴을 직접적으로 검색하기 때문에 2차 구조뿐 아니라 양면각들과 같은 연속적인 양들에 대한 예측도 쉽게 수행할 수 있다. 따라서 k -가까운 이웃 방법은 2차 구조와 같은 불연속적이고 부분적인 정보의 예측보다는 단백질의 국소적인 구조 그 자체를 구체적으로 예측하는 데에 더 많이 쓰이고 있으며, 뒤에 상술할 조각 맞추기 방법의 바탕을 이루고 있다.

연구 대상인 단백질 서열이 이미 구조가 밝혀진 단백질 서열과 전체적으로 매우 유사할 경우에는 유사 단백질을 틀로 삼아 단백질의 구조 모형을 만들게 되는데, 이를 비교구조모형(comparative modeling)이라 부른다. 비교구조모형을 실행하는 대표적인 알고리즘으로는 미국 University of California San Francisco(UCSF)의 Andrej Sali가 개발한 modeller^[4]가 있다. 국내에서는 고등과학원 계산과학부의 이주영 교수 연구진이 효율적인 구조 최적화 방법인 구조 공간 볼륨(Conformational Space Annealing: CSA) 방법^[5]에 바탕을 둔 비교구조모형 방법을 개발하여 단백질 구조 예측에서 뛰어난 성과를 내고 있다.^[6]

물리화학적 방법을 활용한 단백질 구조 예측

구조를 예측하고자 하는 단백질 서열이 이미 구조가 실험적으로 밝혀진 단백질 서열과 유사성이 없을 때에는, 물리적인 포텐셜 에너지함수를 활용한 분자동역학이나 몬테칼로와 같은 전산시뮬레이션을 수행하여 고유 구조를 예측할 수 있을 것이다. 그런데 고유 구조의 예측에만 관심이 있고 단백질이 접히는 과정에는 관심이 없는 경우에는 자유에너지를 최소로 하는 구조를 직접적으로 계산함으로써 계산 시간을 단축할 수 있다. 위에서 이미 언급한 CSA 방법은 이러한 최저 에너지 구조들을 효율적으로 찾아주는 강력한 알고리즘으로서, 이미 언급한 비교구조모형 방법은 물론 물리적인 자유 에너지 함수를 활용한 단백질 구조 예측에도 활용되었다. 단백질을 원자 수준까지 자세히 묘사한 에너지 함수로는 전산시뮬레이션의 시간이 너무 많이 소요되기 때문에, 원자를 여러 개씩 묶어서 근사한 대충간(coarse-grained) 에너지 함수가 많이 쓰이는데, 이중 하나인 United Residue(UNRES) 에너지 함수와 CSA 방법을 결합하여 단백질의 구조를 예측하는 연구들이 수행되었다.^[7] 그러나 아직까지 물리화학적 에너지 함수를 쓴 전산 시뮬레이션만으로는 단백질의 구조를 예측하는 데에는 많은 한계가 있다.

Fig. 4. MASH algorithm based on the idea of hierarchical assembly of secondary structure elements.

물리화학적 방법과 생명정보학 방법의 혼합

순수한 물리화학적 방법이나 생명정보학의 방법만으로 단백질의 구조를 예측하기 어려울 때에는 이 두 방법을 섞어서 쓰으로써 효율적인 예측을 수행할 수 있다. 생명정보학적인 방법으로 단백질의 2차 구조를 예측하는 것을 이미 설명하였는데, 사실 연구 대상 서열이 구조가 밝혀진 단백질들과 전체적인 유사성이 없는 경우라도, 부분적인 2차 구조들은 생명정보학적인 패턴식별 방법을 통해 비교적 쉽고 정확하게 예측할 수 있기 때문에 이러한 정보를 활용하면 물리적인 에너지 함수만을 써서 시뮬레이션을 수행하는 것보다 더 좋은 성능을 보여줄 수 있다. 미국 UCSF의 Ken Dill 연구진에서 개발한 Magical Assembly of Strands and Helices (MASH)도 이러한 방법인데, 단백질의 2차 구조가 먼저 형성된 후에 이들의 조합으로 단백질의 전체 구조가 만들어진다는 아이디어에 바탕을 두고 2차 구조의 조각들 간의 조합을 순차적으로 해나가면서 전산 시뮬레이션을 수행한다^[8](그림 4). 이때 2차 구조는 전산 시뮬레이션을 통해 얻는 것보다 생명정보학의 방법을 통

REFERENCES

- [4] A. Sali and T. Blundell, J. Mol. Biol. **234**, 779 (1993).
- [5] J. Lee, H. A. Scheraga and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).
- [6] K. Joo, J. Lee, J. H. Seo, K. Lee, B. G. Kim and J. Lee, Proteins **75**, 1010 (2009).
- [7] S. Ołdziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nianias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson and H. A. Scheraga, PNAS **24**, 7547 (2005).
- [8] G. A. Wu, E. A. Coutsiar and Ken A. Dill, Structure **16**, 1257 (2008).

Fig. 5. The fragment assembly method. The model structure is selected among conformations generated by assembling the fragments.

해 예측하는 것이 훨씬 빠르며, 2차 구조를 처음에는 고정시키지만 이후에 전산 시뮬레이션 과정을 통해 전체적인 에너지 최적화를 수행한다. 미국 U. Washington의 David Baker는 조각맞추기 방법인 Rosetta를 개발하였는데,^[9] 이는 구조데이터베이스의 정보를 MASH보다 더 많이 적극적으로 활용하는 방법이다. 단백질 서열을 일정한 길이로 자른 후, 앞서 2차 구조 예측에서 설명한 k-가까운 이웃 방법을 활용하여 패턴 식별을 수행하면, 단지 2차 구조뿐만 아니라 부분적인 3차원 구조 자체의 후보 개수를 얻을 수 있는데, 이러한 조각들의 조합 중 자유 에너지가 가장 낮은 구조를 예측된 모형 구조로 채택하는 방법이다[그림 5]. 이러한 조각맞추기 방법의 아이디어에 바탕을 두되 에너지 함수와 구조 생성 방법 등을 바꾼 알고리즘들을 필자를 비롯한 여러 연구진들이 개발하였다.^[10,11]

조각맞추기 방법은 단백질 전체 구조의 예측뿐 아니라 이른바 고리(loop) 구조 예측에도 활용할 수 있다. 비교 구조 모형을 써서 단백질의 모형을 만드는 경우에도, 틀로 삼는 단백질과 연구 대상 단백질 간에 진화상의 발산 때문에 유사성이 없는 부분들이 나타나는데, 이러한 부분들을 고리라고 한다. 고리 부분은 단백질이 다른 분자와 상호작용을 하면서 기능을 하는데 중요한 역할을 하기 때문에, 그 구조를 예측하는 것이 매우 중요하다. 고리 구조의 예측 역시 특정 단백질 구조를 틀로 삼아서 모델링할 수 없기 때문에 조각맞추기 방법을 응용할 수 있는 대상이다. 고리의 구조를 예측하는 것은 전체를 예측하는 것에 비해 길이가 상대적으로 짧은 점에서 더 쉬운 면도 있지만, 고리 구조가 비교 구조 모형법으로 결정된 나머지 구조에 들어맞아야 한다는 제약조건은 문제를 더 어렵게 한다. 조각 맞추기를 고리 구조의 예측에 적용한 예는 드물었는데, 일반적으로 데이터베이스의 조각들을 조합하여 구조를 생성하면 고리가 닫혀야 한다는 제약 조건이 만

Fig. 6. Loop closure. The initial loop structure constructed from the fragments are not consistent with the rest of the structure. The torsional angles are adjusted according to the loop closure equation, so that the loop is closed.

Fig. 7. Loop structure prediction server based on fragment assembly and analytic loop closure (<http://falc-loop.seoklab.org/>).

족되지 않기 때문이다. 조각맞추기를 고리 구조의 예측에 적용한 사례가 있기는 하나, 닫히지 않는 상태의 고리 구조들을 생성한 후 고리와 나머지 구조 간의 틈을 점차 최소화시키는 방법을 씀으로써, 최종 구조에도 고리가 완전히 닫히지 않은 채 작은 틈이 남아 있었다.^[12] 반면에 수학적 고리 닫힘 방정식(analytic loop closure equation)을 활용하여, 조각에서 생성된 초기 구조에서 가까운 구조를 유지하면서 고리가 닫

REFERENCES

- [9] R. C. Rohl, C. E. Strauss, K. M. Misura and D. Baker, *Methods Enzymol.* **383**, 66 (2004).
- [10] G. Chikenji, Y. Fujitsuka and S. Takada, *J. Chem. Phys.* **119**, 6895 (2003).
- [11] S.-Y. Kim, W. Lee and J. Lee, *J. Chem. Phys.* **125**, 194908 (2006).
- [12] C. A. Rohl, C. E. Strauss, D. Chivian and D. Baker, *Proteins* **55**, 656 (2004).

Fig. 8. Performance comparison of FALC-Loop with RAPPER. Ave denotes average RMSD of 1000 conformations, best denotes the minimum RMSD among 1000 conformations, and model denotes the RMSD value of the final selected model. All of these are averaged again for each loop lengths.

히도록 양면각들을 조절함으로써[그림 6] 완벽히 닫힌 고리 구조의 공간만을 탐색하도록 한 알고리즘을 필자와 다른 연구자들이 공동 연구를 통해 개발하였으며,^[13] 이를 바탕으로 한 고리 구조 예측 서버를 구축하였다[그림 7]. 새 알고리즘은 기존의 다른 고리 구조 예측 알고리즘보다 우수한 성능을 보이는데, 예를 들어 새 알고리즘의 구조 생성 능력을 시험해 보기 위해 Fiser set이라 불리는, 길이 4–12인 고리 317개로 이루어진 집합에 대해, 고리 이외의 부분을 실험 구조로 고정시킨 상태에서, 이 알고리즘으로 1000개씩 생성한 고리 구조들과, 이들 중에서 DFIRE^[14]라는 경험적 에너지 함수로 최종 선택한 구조를, 기존의 고리 구조 예측 모형인 RAPPER와 비교해 보았다.^[15] 성능은 실제 고유 구조와의 좌표값의 평균제곱근(Root Mean Square Deviation: RMSD)으로 평가할 수 있는데, RMSD 값이 적을수록 성능이 우수한 것이다. 생성된 1000개 중 가장 고유구조에 가까운 구조와 RMSD와 평균

RMSD, 그리고 최종 모형 구조의 RMSD가 모두 RAPPER보다 적다는 것을 알 수 있었다[그림 8].

나오는 글

단백질은 생명활동의 기본을 이루는 물질이므로, 단백질이 서로 어떻게 상호작용을 하여 복잡한 생명 활동을 수행하는지 이해하는 것은 매우 중요하다. 이러한 단백질의 기능을 이해하는 기초가 되는 것이 단백질의 구조의 연구인데, 물리적인 제일원리만으로 단백질의 접힘을 이해하고 고유 구조를 예측하는 것은 아직까지 풀리지 않는 숙제로 남아있다. 대량의 분자생물학 데이터를 분석하여 유용한 정보를 얻어내는 생명정보학은 오늘날 생명현상을 연구하는 중요한 도구인데, 단백질 구조 연구의 경우 이미 실험적으로 밝혀진 단백질 구조 데이터베이스의 분석을 통한 연구가 이에 해당된다. 이 글에서는 순수하게 생명정보학적인 방법 혹은 물리학적인 방법만을 통하여 단백질의 구조를 연구하는 접근법을 살펴보고, 이 두 가지 방법론을 혼합한 연구 방법에 대해 논의하였다. 물리학적인 방법과 생명정보학적인 방법의 혼합을 통한 단백질 접힘 연구는 단백질의 구조 예측은 물론 접힘 과정의 동역학에 대한 이해와 단백질 상호작용 등 많은 분야에서 새로운 결과를 도출할 것으로 기대되고 있다.

REFERENCES

- [13] J. Lee, D. Lee, H. Park, E. A. Coutsias and C. Seok, *Proteins* **75**, 1010 (2010).
- [14] H. Zhou, Y. Zhou and T. L. Blundell, *Protein Science* **11**, 2714 (2002).
- [15] M. A. DePristo, P. I. W. de Bakke, S. C. Lovell and T. L. Blundell, *Proteins* **51**, 41 (2002).