생물물리 난제의 돌파구: 실험과 계산의 결합

DOI: 10.3938/PhiT.23.041

이 주 영

Breakthrough in Bio-physics: A Combination of Experiments and Theory

Jooyoung LEE

Ten years have passed since the Human Genome Project was declared to be complete. Although an immense amount of new information has been released since then, scientific advances arising from the completion of the Human Genome Project have been rather limited. One technical difficulty to overcome is the slow progress of the structural genomics project in uncovering 3D structured information of proteins based on the amino-acid sequence information. We discuss the possibility of combining experimental data with computational methodologies in order to solve difficult biological problems. Nuclear magnetic resonance (NMR) spectroscopy utilizing nuclear overhauser effect (NOE) data is one of the major experimental methods used to determine the 3D protein structure. However, for large proteins with more than 160 amino-acid residues, the appropriate assignment of ambiguous NOE peaks is a bottleneck in determining the 3D protein structure. We demonstrate that this problem can be solved by combining a computational method with a given set of experimental data for the case of Ts763 from the CASP11 protein structure prediction experiment, which was held during the summer of 2014. In the coming decade, computational methodologies that can utilize limited experimental data will play a key role

in tackling challenging problems in biophysics.

물리학의 정의에는 여러 가지가 있겠지만, 그 중에 하나가 problem solving of natural phenomena 즉 “자연현상의 문제해결”이라고 볼 수 있다. 20세기 동안의 눈부신 물리학의 발전으로, 많은 물리학의 문제들에 대한 해답이 제시되었고, 이를 통하여 감히 전통적인 물리학의 문제들에 대해서는 이해하지 못하는 부분보다 이해하는 부분이 더 많다고 말할 수 있다. 한편, 물리학과는 전혀 다른 학문으로 시작된 생물학에 대한 이해는 전통 물리학에 비하면 이제 걸음마 수준이다. 이런 관점에서 20세기가 물리학의 세기였던 것처럼 21세기는 생물학의 세기가 될 것이라고 예상하는 사람들이 많다.

10년 후의 생물물리학은 어떤 모습일까? 이에 대한 답변을 위해서는 10년 혹은 20년 전에 생물학과 생물물리학을 달구었던 최첨단 연구 주제와 그런 연구 결과의 오늘날의 현주소를 살펴볼 필요가 있다. 약 20여 년 전인 1990년, 세계 최대 규모의 국제 생물 연구사업인 인간 게놈 프로젝트가 시작되었다. 그리고 나서 10여 년 동안의 미국, 영국, 프랑스, 호주, 일본, 중국 등의 수많은 과학자들의 연구 결과로, 2000년 미국의 클린턴 대통령과 영국의 블레어 수상이 공동으로 인간 게놈 초안의 완성을 발표하였고, 2003년에 이르러서는 본질적인 인간 게놈의 완성을 공표하였다.

인간 게놈의 완성으로 사람의 유전자 수가 쥐나 다른 포유동물과 크게 다르지 않으며 예상했던 것보다 적은 2만여 개라는 사실이 밝혀졌다. 인간 게놈 지도 완성이 의미하는 바는 여러 가지가 있다. 그 중 가장 실용적인 부분은 개개인의 게놈을 파악함으로써 각자의 여러 질병에 걸릴 성향에 대한 정보를 얻을 수 있다는 것이다. 근본적으로는 10년 전에 이미 게놈으로부터 단백질이 전사되는 과정을 통하여 인간에 존재하는 모든 단백질의 아미노산 서열 순서가 밝혀진 것이나 다름이 없다. 따라서 단백질이 DNA 정보의 마지막 결과물이며 모든 생명현상에 직접 관여하기 때문에, 만약에 어느 한 개인의 특정 단백질의 아미노산 서열이 돌연변이 등으로 인하여 바뀐 것을 알게 된다면, 이로부터 이 사람이 특정 질병에 걸릴 확률이 높

저자약력

이주영 교수는 서울대학교 물리학과를 졸업하고, 1990년에 Brown University에서 상전이 현상의 수치적 접근법이란 주제로 박사학위를 취득했다. Argonne National Lab, Florida State Univ.에서 박사후연구원을, Mt. Sinai School of Medicine, Cornell University에서 연구원으로 지냈고, 2000년부터 고등과학원 계산과학부 교수로 재직 중이며 현재 In Silico 단백질 연구단 단장을 임하고 있다. (jlee@kias.re.kr)

아질 수 있기 때문이다.

한편, 단백질의 3차원 구조는 1차 구조인 아미노산 서열 정보에 비해서 알아내기가 매우 힘들다. 즉 인간 단백질의 경우, 아직, 구조가 밝혀진 단백질보다 그렇지 못한 단백질이 훨씬 더 많은데, 이는 아직 현재의 실험기술로는 밝혀내기 힘든 단백질이 많기 때문이다. 인간 게놈 프로젝트가 한창일 때, CASP(Critical Assessment of protein Structure Prediction) 실험이 제안되었다. 이 실험은 아직 실험적으로 구조가 밝혀지지 않은 단백질의 구조를 컴퓨터 시뮬레이션 등의 계산 방법을 이용해서 예측하고, 나중에 그 단백질의 구조가 실험으로 밝혀지면 이를 정답으로 해서 여러 다른 계산 방법을 공정하게 평가하고, 이를 통하여 이론/계산적인 방법론의 발전을 지향해서 이론/계산방법이 실험방법을 대체하거나 또는 보완하는 수단으로 사용될 가능성을 타진해 보기 위한 것이었다. CASP는 1994년 여름에 처음으로 시작해서 매 2년마다 열리고 있는데 올해는 그 11번째 행사로, 2014년 5월부터 8월 중순까지 CASP11이 진행되었다. 지난 20년 동안의 CASP 실험을 통해서 주형기반(template-based) 모델링의 경우 상당한 단백질 구조 모델링 이론/계산 방법의 발전을 이루었다. 이는 모델링해야 할 단백질과 약 20% 이상 아미노산 서열이 비슷한 단백질의 구조가 이미 실험으로 밝혀진 경우에 해당한다. 하지만 적절한 주형(template)이 없는 경우에는 아직 단백질 구조 모델링을 만족할 만한 수준으로 해낼 수 있는 방법이 밝혀져 있지 않다.

오늘날 정확한 구조 모델링도 불가능하고, 실험으로도 구조를 밝히기 힘든 단백질의 경우 어떻게 그 구조를 구해 낼 수 있을까? 충분한 data를 얻지 못해서 실험만으로는 해결할 수 없고, 또한 계산만으로도 해결할 수 없는 문제를 충분하지 못한 실험 data와 계산방법의 결합을 통하여 해결한 사례를 한 가지 들도록 하겠다.

Nuclear magnetic resonance(NMR) 분광법은 단백질 구조를 결정하는 방법으로 X-ray 결정학 다음으로 자주 쓰이는 실험 방법이다. 여러 가지 NMR 기법이 동시에 쓰이고 있는데, 단백질의 3차원 구조를 결정하는 핵심 기법은 Nuclear Overhauser Effect(NOE) 현상을 이용하는 것이다. 즉, 가까운 거리에 있는 한 쌍의 수소로부터 두 원자핵 사이의 스핀-스핀 relaxation 현상을 측정하여 두 수소 사이의 거리에 대한 정보를 얻을 수 있다. 하나의 단백질에는 보통 수백, 수천 개의 수소 원자가 있는데, 이들 사이의 거리에 대한 정보를 충분히 많이 얻을 수 있다면, 이 정보로부터 단백질의 전체 모양을 알아낼 수 있다. 측정된 각각의 NOE 신호를 그에 상응하는 수소 쌍에 mapping하는 작업을 NOE peak assignment라 부른다. 정확한 NOE 신호 할당을 하나의 아미노산 당 15개 정도 할 수

있다면 (즉 아미노산 100개 크기의 단백질의 경우 1500개 정도), 이로부터 이 모든 거리 정보를 만족시키는 단백질 모델링을 통하여 그 단백질의 구조를 결정하게 된다.

그런데 여기서 문제점은 정확한 NOE 신호 할당이 쉽지 않다는 점이다. 많은 비용과 시간을 투자해서 엄청난 양의 추가 실험을 한 후에만, 정확한 NOE 신호 할당이 가능해진다. 구조를 얻고 싶은 단백질의 크기가 커지면 커질수록 추가해야 할 실험이 많아지는데 현재로는 아미노산 160개 크기의 단백질의 경우가 그 한계점이라고 볼 수 있다. 하지만 적은 양의 추가 실험으로부터 애매모호한(ambiguous) NOE 신호 할당이 가능하다. 즉 각각의 NOE 신호는 평균 수십 개의 수소 쌍 중에 하나일 거라는 정보이다. 예를 들어 1500개의 애매모호한 NOE 신호 할당이 주어진다면, 좀 더 구체적으로는 각각의 NOE 신호가 평균 20개의 주어진 수소 쌍 중에 하나라는 조건이 주어진다면, 이런 불완전한 정보로부터 단백질의 구조를 결정하는 것이 가능할까? 가능한 NOE 신호 할당의 경우의 수는 20^{1500} 이고, 이중에는 정확한 1500개의 NOE 신호 할당도 포함되어 있다. 물론, 각각의 NOE 신호 할당의 경우의 수로부터 위에서 언급한 거리 정보를 사용하여 단백질의 3차원 모델링 작업이 추가로 필요하다. 불가능에 가까운 문제로 보인다. 조합 최적화(combinatorial optimization)의 관점으로 볼 때 이 문제는 두 단계 최적화 문제이다. 첫 단계는 NOE 신호 할당이고, 두 번째 단계는 주어진 NOE 신호 할당으로부터 거리 정보를 최적화하는 단백질 모델링이다.

올 여름에 시행된 CASP11에서는, 처음으로 위에서 언급한 애매모호한 NOE 신호 할당으로부터 단백질의 구조를 예측하는 문제가 제출되었다. 실제로는 x-ray 결정학으로부터 구조가 결정된 단백질에 대해서 NMR의 초기 data를 가정한 애매모호한 NOE data를 전산모사를 통하여 만들어내었는데, 모두 19개의 문제가 제출되었다. 한 문제의 예를 들면 다음 web page에서 Ts763.tar.gz을 내려 받으면 된다. (predictioncenter.org/download_area/CASP11/extra_experiments/). 추가 정보는 다음 web page에서 얻을 수 있다 (predictioncenter.org/casp11/target.cgi?id=79&view=others). 130개의 아미노산에 대한 모델링 문제였는데, 최근에 그 구조가 밝혀졌다. 이 문제의 경우 적절한 주형이 없는 경우였기 때문에, 지금까지 공개된 모델 중에서 가장 정확한 모델도 완전히 틀린 구조를 갖고 있다. 기술적으로는 약 27% 정도 비슷한 구조로 가장 정확한 모델과 실험 구조 둘 사이의 root mean square deviation(RMSD) 값이 14 Å 정도이다. 즉 전체적으로 14 Å 틀린 구조이다(그림 1 참조). 이 경우 위에서 언급한 두 단계 최적화 문제를 풀어서 모델링한 구조의 정확도는 약 87% 정도로 RMSD로 측정한 정확도는 약 1.9 Å이다 그림 2에서 볼 수 있듯이 부정확

Fig. 1. The best server model (green ribbon diagram) for T0763 from CASP11 and the native structure (red ribbon diagram) of T0763 disclosed after all models are submitted for assessment are shown in superposition. The red structure is the most accurate model out of about 200 models generated by over 40 servers tested in CASP11, and it is generated from the nns server led by the author. Although the green structure is the best server submitted model, its RMSD value from the red structure is rather large with about 14 Å, and the accuracy of the model is rather poor.

한 실험 data를 이용해서도 매우 정확한 모델링 결과를 얻을 수 있었음을 보여주는 예이다.

애매모호한 NOE 신호 할당으로부터 얼마나 정확하게 단백질 모델링이 가능한가에 대한 좀 더 정확한 평가는 19개의 문제에 대한 정답 즉 실험 구조가 모두 공개되어야 알 수 있을 것이다. 현재 추가로 3개의 정답이 공개되었는데, 그중 가장 큰 단백질의 경우 274개의 아미노산으로 이루어진 단백질인데 조합 최적화로부터 약 2.3 Å의 RMSD 정확도의 놀랄만한 모델링이 가능하였다. 실제 NMR 실험과 연계할 경우, 이러한 정확한 모델링으로부터 최소한의 추가 실험을 거쳐, 이전에는 불가능했던 NMR로 단백질 구조를 결정하는 연구가 가능해진다.

위에서 보여준 예와 같이, 실험이나 계산/이론만으로는 해결

Fig. 2. A model for Ts763 (blue ribbon diagram) generated from the LEE method led by the author is shown in superposition against the native structure (red ribbon diagram). The blue structure is generated utilizing ambiguous NOE data provided by CASP11. The RMSD value between the two structures is below 2 Å, and the accuracy of the model is quite high. In the real situation, assisted by this kind of accurate model, the actual native structure can be determined with a minimal amount of additional experiments and data analysis.

하기 불가능한 일을 실험 data와 계산연구를 융합하여 해결하는 연구가 앞으로 생물물리뿐만 아니라 여러 분야에서 매우 중요한 연구 방법으로 대두될 것으로 예상된다. 이미 10년 전에 인간 게놈 지도라는 엄청난 정보를 획득하고 나서도, 그 파급효과가 아직 많은 생물학 분야에 미치지 못하고 있다. 이런 문제점의 해결을 위해서는 아직 어렵기 때문에 밝혀지지 못한 중요 단백질의 구조와 기능을 규명하는 구조생물학의 기여가 필요한 시점이다. 50여 년 전, 물리학자 Perutz와 Kendrew가 x-ray 결정학으로 최초로 단백질의 구조를 밝혔듯이, 앞으로 더 많은 물리학자들이 이런 생물물리 문제에 흥미를 얻어서 이 분야에 새로운 패러다임을 얻어낼 수 있는 돌파구를 마련해 주길 기대한다.