단백질 접힘 시뮬레이션: 수치 실험

DOI: 10.3938/PhiT.20.021

이 인 호

Protein Folding Simulations: Numerical Experiments

In-Ho LEE

Understanding how proteins fold has become an attractive research subject for both theoretical and experimental physicists. There has been considerable interest in using computational approaches that generate all-atom reaction coordinates to understand the kinetic properties of protein conformational changes. In this article, I summarize a set of computational methods currently used in protein folding simulation community. The methodologies discussed here are expected to be useful more generally in studies of conformational transitions in complex proteins.

들어가는 글

단백질(흰자질, protein)은 많은 아미노산들로 연결되어진 분자를 말한다. 단백질은 생물의 몸을 구성하고 있을 뿐만 아니라 저마다 고유의 기능을 수행하고 있다. 세포 내의 각종 화학반응의 촉매 물질로서도 단백질의 기능은 중요하다. 각종 화학반응의 촉매로 활동하고 있는 단백질을 특별히 효소라고 부른다. 현재, 약 2200여 종이 알려져 있다. 그밖에 단백질은 면역기능과 에너지 대사에 관여하기도 하고 생체신호 전달의 호르몬을 이루기도 한다. 단백질은 화학반응 이외에 기계적인 기능도 수행하는 경우가 있다. 알이나 종자에 저장된 특수한 단백질도 있다.

DNA 존재와 구조의 확립에 앞서서 1944년 슈뢰딩거 박사는

저자약력

이인호 박사는 한국과학기술원(KAIST, 물리학과) 이학박사(1996)로서 Univ. of Illinois at Urbana-Champaign(Beckman Institute for Advanced Science and Technology, Materials Research Laboratory) 포스트닥 과정 그리고 고등과학원(KIAS, 물리학부) 리서치-펠로우를 거쳐 한국표준과학연구원(KRIS)에서 선임연구원(2001-2005), 책임연구원(2005-)으로 재직 중이다. (ihlee@kris.re.kr)

생명의 핵심 개념으로 네겐트로피(음의 엔트로피)와 코드를 지적했다.^[1] DNA는 단백질을 어떻게 만들 것인가에 대한 정보 저장 형식으로 보이기도 한다. 생명 현상은 물리적으로는 정렬을 이루고 엔트로피를 국소적으로 감소시키는 절차들에 해당한다. 이러한 일들의 단위는 결국 단백질들이 수행한다.

휴먼 게놈 프로젝트로 알려진 DNA 서열 해독은 인류가 이루어낸 생명 현상에 대한 이해의 첫 단추라 부를 수도 있을 것이다. 하지만, 진정으로 생명 현상을 이해하려면 단백질의 기능을 이해해야만 한다. 이것 없이는 생명 현상을 본질적으로 이해하기 힘들다. 대부분의 단백질은 자신만의 특별한 구조를 만들고 고유의 기능을 수행한다. 굉장히 자연스럽게 자신의 모양을 바꾸는 단백질의 운동 특성에 우리는 주목할 필요가 있다.

단백질에 대한 관심이 매우 높기 때문에 많은 연구자들이 단백질의 원자 구조를 실험적으로 밝히고 해당 기능을 찾아내고 있다. 이러한 단백질의 구조를 아미노산 서열로부터 직접 알아낼 수 있다면 획기적인 일이 될 것이다. 하지만, 아직 그 누구도 단백질 구조 예측문제를 물리학적 방법론으로 풀지는 못하고 있다. 때때로 이론가들은 단백질의 구조를 예측하기도 하고 한 발 더 나아가 단백질을 설계하기도 한다. 설계된 단백질 구조가 실험적으로 얻어지는 경우도 있다. 단백질이 어떻게 접힐 수 있는지에 대한 보다 근원적인 이론적 연구도 많이 수행되어 왔다. 상대적으로 간단한 모델계산을 통해서 결론을 이끌어내는 방식이 주를 이룬다.

실험적으로는 X-선 결정학, 핵자기공명(NMR) 분광학 등의 방법이 널리 활용되고 있다. 결정으로 만들어진 단백질은 X-선 회절 무늬를 만들어내는데 이를 분석하여 단백질 속 원자들의 배열을 알아낼 수 있다. NMR 분광학의 경우, 단백질 속의 매우 많은 수의 수소-수소 거리 정보를 실험적으로 얻어서 단백질 구조를 그려낸다. 2011년 3월 현재 약 66720여 개의 단백질 구조가 알려져 있다.

REFERENCES

- [1] Moo Young Choi, Lectures on Physics (Texts in Korean) (2008).

최근 급속히 발전한 컴퓨터 시뮬레이션 방법도 상당한 의미를 가지는 단백질 연구 방법으로 인정받고 있다. 컴퓨터 시뮬레이션은 실험적 연구방법, 이론적 연구 방법과 차별화되는 부분이 있다. 이것은 제 3의 연구방법이다. 실제로 보이지 않는 원자의 이동과 행동을 이해하기 위해서 컴퓨터의 계산 기능을 이용하는 것이다. 다른 응집물질 물리학, 통계 물리학에서 컴퓨터를 이용한 시뮬레이션들과 다를 바 없는 연구 방법이다.

단백질 연구를 위해서 물리학, 화학, 분자 생물학, 컴퓨터 과학 등의 아이디어들이 총 동원되기도 한다. 분자 생물학에 대한 통계학과 컴퓨터 과학의 응용이 바로 생물정보학(bioinformatics)이다. 네트워킹 그리고 코드는 근본적으로 정보로 분류할 수 있는 것들이다. 이러한 관점에서 생물 정보학은 태어날 수밖에 없었던 학문의 존재 이유를 가지고 있다. 생물학적 절차에 대한 이해의 폭을 넓히는 것을 첫 번째 목표로 설정한다고 하더라도 생물정보학의 실제 연구 방식은 물리학의 그것과는 매우 다르다.

특히 이 글에서는 최근에 사용되고 있는 단백질 접힘 시뮬레이션 연구방식들에 대한 소개를 다루고자 한다.

단백질 구조의 유연성

단백질 구조의 놀라운 특징 중의 하나는 그 모양이 쉽게 변형될 수 있다는 것이다. 일반적으로 단백질은 아주 안정한 하나의 구조를 가진다. 특정한 무리의 단백질들은 두 가지 모양을 가지기도 한다. 상황에 따라서 하나의 구조에서 또 다른 구조로 단백질 자신의 모양을 바꾸기도 한다. 특정한 구조를 가지지 않는 단백질도 있다. 때때로 제대로 접히지 않은 상태에서도 고유의 기능을 수행할 수 있음을 의미한다. 단백질은 스스로 자신의 모양을 결정하기도 하지만 다른 분자의 도움을 받아서 자신의 모양을 찾아가기도 한다.

단백질의 접힘 경로를 원자 수준에서 이론적으로 연구하기 위해서 분자동역학(molecular dynamics) 방법이 널리 사용된다. 원자-원자 상호작용을 수식으로 표현하고 이러한 퍼텐셜 함수에서 단백질의 구조 변화를 알아내는 것이다. 단백질은 일반적으로 사슬모양의 확장된 구조가 가능하다. 단백질 내부에서 수소결합이 잘 정렬된 구조는 에너지 관점에서 유리하다. 어떻게 항상 고유의 모양을 가질 수 있는가? 또 어떻게 그 구조를 찾아가는가? 주어진 온도에서 그리고 수용액 안에서 단백질은 자유 에너지가 가장 낮은 구조를 선호하게 될 것이다. 최소 퍼텐셜 에너지 값을 주는 단백질 구조가 열역학적으로 관측되는 구조가 일반적으로 아니다. 이는 자유롭게 변화할 수 있는 구조가 자유 에너지 관점에서 매력적이라는 점을 의미하는 것이다.

하나의 수소결합은 약 5 kcal/mol (0.216 eV) 정도의 퍼텐셜 에너지 이득으로 연결되어진다. 이러한 수소결합이 가능한 단백

Fig. 1. A peptide bond is a covalent bond formed between two adjacent amino acids. Specifically the carboxyl group of one amino acid reacts with the amino group of the other amino acid, thereby releasing a water molecule(H_2O). A peptide bond can be broken by amide hydrolysis.

질 원자구조의 탐색 그리고 접힘 과정에서 어떠한 순서로 수소결합들이 만들어지는지 등에 대한 연구가 분자동역학 시뮬레이션을 통해서 가능하다. 수소결합은 수소원자와 음전하 사이의 결합으로 에너지 관점에서 단백질을 안정한 구조로 유도할 수 있다. 일반적으로 수소결합은 판 데르 발스 결합 에너지보다는 강하다. 공유결합, 이온결합, 금속결합보다는 수소결합이 일반적으로 약한 결합이다. 물을 싫어하는 부분은 접힌 단백질의 안쪽에 위치하고 물을 좋아하는 부분은 접힌 단백질의 바깥쪽에 위치하는 것이 자연스럽다. 수소결합의 세기도 환경에 따라서 바뀔 수 있다. 물을 싫어하는 부분, 즉, 접힌 단백질의 안쪽에 있는 수소결합은 상대적으로 강한 수소결합이다.

단백질 접힘의 방식

아미노산의 일차원 서열로 이루어진 단백질은 자체의 아미노산들 사이의 상호작용으로 에너지가 낮은 구조를 가진다. 그림 1은 일반적인 사슬모양의 단백질 구성에 관한 펩타이드 결합을 설명하고 있다. 아미노산의 종류는 20가지로 한정되어 있다. 아미노산 서열만으로 단백질의 구조가 결정될 수 있다는 믿음이 바로 안핀센(C. B. Anfinsen)의 도그마이다. 자유 에너지 관점에서 단백질은 자유 에너지가 가장 낮은 구조를 선호한다는 것이다. 여기에 추가적인 조건으로 운동학적으로 접근 가능한 경로가 존재한다는 보장이 있어야 할 것이다. 물을 싫어하는 단백질 부분은 안쪽으로 배치되는 것이 자연스럽다. 실제 단백질은 거의 오류 없이 자신의 고유한 모양을 정확히 찾아간다. 몇 가지 질병 중에는 단백질 접힘 과정의 오류와 연관된 것들이 있다.

동서고금을 막론하고 유전은 아주 큰 연구 주제이다. 이러한 주제 또한 단백질 접힘 연구를 통해서 이해할 수 있는 길이 있다. 인종이나 개인별 염기 차이, 즉, DNA 서열의 변이성은 분명 존재한다. 결국 같은 기능을 수행하도록 만들어진 단백질이라고 하더라도 그 단백질의 접힘 방식은 사람에 따라서 다를 수 있음을 의미한다. 사람에 따라서 특정 질병의 발병 비율이 높을 수 있는 것은 잘 알려진 사실이다.

컴퓨터 시뮬레이션을 이용해서 실제 질병과 관련한 기작을 설

명하는 예는 그렇게 많지 않다. 컴퓨터 시뮬레이션이 아직 초보적인 단계에 놓여있는 것도 사실이다. 아주 간단한 단백질의 경우 실험적으로 관측된 단백질의 접히는 방식이 컴퓨터 계산을 통해서 재생산되는 예는 다수 있다. 좀 더 큰 단백질에 대한 심도 있는 컴퓨터 시뮬레이션의 모범적 응용 사례는 증가 일로에 있다. 시뮬레이션에서는 단백질의 모양을 모두 알고 있기 때문에 단백질 내의 특정 위치에서의 수소결합의 생성과 양이온-음이온 거리와 같은 다양한 시그널을 직접 계산할 수 있다.

단백질 접힘 시뮬레이션: 초기 조건 방식

분자동역학 방법은 원자로 구성된 시스템의 변화를 시간에 따라서 계산하는 것을 기본으로 한다. 각 원자는 시간에 따라서 원자-원자 상호작용을 통해서 힘을 받게 된다. 추가적으로 외부와의 상호작용 때문에 숨겨진 자유도들로부터 제공되는 힘도 있다. 이러한 상황에 있는 원자의 운동을 뉴턴-운동방정식을 이용하여 찾아내는 것이다. 현재의 위치, 속도를 이용하고 원자가 받는 힘을 계산하여 미래의 원자 위치를 얻어내는 것이다. 실제 원자-원자 상호작용에 기인한 힘은 양자역학적으로 결정된 힘이어야만 한다. 이는 모든 전자들이 자신들의 바닥상태에 도달했을 때 이온이 받는 힘을 의미한다. 전자들이 바닥상태에 도달했다는 것은 주어진 이온들의 배치에서 전자들은 전자-전자 상호작용 그리고 이온-전자 상호작용 아래에서 슈뢰딩거 방정식을 만족함을 의미한다. 사실, 이렇게 전자와 이온을 분리하여 취급하는 방식이 성립하지 않을 수도 있다. 때때로 수소원자는 양자역학적 터널링 효과를 보일 수도 있다. 예를 들어 수소 원자는 터널링을 통해서 분자 속 두 가지 위치 중 하나를 선택할 수 있다. 대부분의 분자동역학 계산에서는 원자들의 좌표만으로 단백질의 퍼텐셜 에너지가 결정되는 퍼텐셜 함수를 사용한다. 양자역학 계산 결과 그리고 실험 데이터를 동시에 사용하여 만든 퍼텐셜 함수라고 볼 수 있다.

분자동역학 방법은 초기 조건 문제이기 때문에 모든 원자에 대한 초기 위치, 초기 속도를 입력으로 요구한다. 수소-탄소의 진동주기가 대략 11 fs 정도 수준이다. 따라서 수소의 운동을 제대로 기술하기 위해서는 약 1 fs (10^{-15} sec) 정도의 시간 간격으로 수소의 운동을 기술해야만 한다. 그렇지 않을 경우 뉴턴-운동방정식 풀이의 정밀도는 적분 횟수가 증가하면서 급속히 떨어진다. 닫힌계에 대해서 특별히 벨레(Verlet) 알고리즘^[2]을 사용하면 아주 오랜 시간 동안 시스템이 에너지를 얻지도 않고 잃지도 않는다는 것이 알려져 있다. 이것은 특별한 적분 알고리즘(symplectic integrator)으로 알려져 있으며 컴퓨터 게임을 포함한 많은 시뮬레이션에서 널리 사용되는 중요한 알고리즘이다. 또한, 온도 효과를 넣어서 시뮬레이션을 진행할 수 있다. 예를

들어 원자들이 300 K에 해당하는 속도 분포를 가진다고 상상할 수 있다. 실제 시뮬레이션에서는 이러한 분포를 가지도록 해준다. 이것은 실제 실험에서 온도를 고정하고 분자들을 관측하는 상황을 컴퓨터로 연출하는 것이다. 예를 들어 확률론적 미분 방정식인 랑주뱅 방정식(Langevin equation)을 풀어서 브라운 운동을 기술할 수 있다.

단백질의 접힘 과정은 수소-수분 이내에 발생하는 사건이다. 하지만, 이론적으로 접근할 때 이러한 시간은 굉장히 긴 시간이다. 그 이유는 단순히 분자동역학 계산의 시간 단계의 수가 많아야 함을 의미한다. 1 fs를 분자동역학 단위 단계로 가정할 경우, 1초의 시뮬레이션은 10^{15} 단계의 분자동역학 방법을 진행해야만 이루어낼 수 있는 시간이다. 이는 현재의 컴퓨터 계산 능력으로는 얻어내기 힘들다. 현재 최장 단계의 분자동역학 시뮬레이션 시간은 1 msec로 Shaw 박사 그룹에서 이룬 것이다.^[3] 이러한 계산을 위해서 특별한 컴퓨터를 디자인하였다고 한다. 이러한 시뮬레이션 기법은 다른 그룹에서 흉내 내기가 쉽지가 않다. 아주 크기가 작은 단백질의 경우, 1 msec 안에 단백질이 접히는 과정을 관찰할 수 있는 기회가 있을 수 있다.

단일 시뮬레이션이 아닌 다중 시뮬레이션 방식으로 분자동역학을 수행할 수도 있다. 가령, 1 ns (10^{-9} sec) 수준의 분자동역학 계산을 1000번 수행할 수 있다. 이때, 서로 다른 초기 조건을 활용한다. 이러한 계산에서는 단백질이 접히는 과정이 포함될 수 있다. 이것은 단백질 접힘 과정이 시간에 따라서 푸아송 분포(Poisson distribution)를 가진다고 가정할 경우, 단백질의 접힘에 걸리는 시간을 산출할 수 있다. 단백질이 시간상에서 순차적으로 그리고 점진적으로 접히는 것이 아니고 확률론적으로 특정 순간에 접힌다는 것을 여기서 가정한 것이다. 시간이 흘러감에 따라 특정 사건이 일어나는 평균 비율을 생각해 볼 수 있다. 그 중에서 주어진 길이의 시간 안에 사건이 일어날 확률을 생각해 볼 수 있다. 하나의 사건이 일어나고 또 다른 사건이 일어날 때 시간상으로 이전 사건과의 연관성이 없다고 가정하는 분포가 푸아송 분포이다.

독립적으로 수행되는 분자동역학 계산에서 언제 단백질이 접히는지를 모른다. 사실 또 많은 수의 시뮬레이션들 사이에는 아무런 연관성이 없다. 이러한 경우 많은 수의 컴퓨터들에 서로 다른 초기 조건들을 배분한 다음 독립적으로 시뮬레이션을 수행할 수 있다. 이러한 계산을 특별히 분산 컴퓨팅(distributed computing)이라고 한다. 병렬계산 방법 중에서 계산 노드들 사이의 통신이 최소화된 억제된 형태의 계산으로 볼 수 있다. 모든 계산이 수행되고 나면 계산 결과들을 한 곳에 모으고 단백질이 접혀진 경우

REFERENCES

- [2] L. Verlet, Phys. Rev. **159**, 98 (1967).
- [3] D. E. Shaw *et al.*, Science **330**, 341 (2010).

와 단백질이 여전히 풀려져 있는 경우의 수를 각각 계산한다. 이러한 비율과 실제로 시뮬레이션을 수행한 시간을 고려하면 실제 단백질이 어느 정도 시간 안에 접히는지를 추론해 낼 수 있다. 이것은 Pande 교수 그룹에서 개발한 방법으로 널리 알려져 있다.^[4] 특별히 개인 컴퓨터 스크린-세이버 형식으로 단백질 접힘 소프트웨어가 전 세계적으로 보급되어 있다. 단백질 접힘에 관심이 많은 일반 개인용 컴퓨터 사용자들이 자신의 컴퓨터를 연구자들에게 제공하고 있다. 자신의 개인용 컴퓨터가 일을 하지 않을 때 연구자들이 사용할 수 있도록 허용해주고 있는 것이다. 2011년 현재 30만 대 이상의 개인 컴퓨터들이 사용가능한 것으로 보고되었다.

단백질 접힘 시뮬레이션: 경계 조건 방식

단백질 접힘의 과정을 실험 데이터와 직접 비교하기 위해서는 단백질의 자유 에너지를 계산할 수 있어야 한다. 자유 에너지는 통상 반응 좌표(reaction coordinate)를 기준으로 계산하게 된다. 하지만, 반응 좌표를 미리 선택할 수 있는 방법은 흔하지 않다. 이는 단백질 접힘을 하나의 화학 반응으로 보았을 때 이 반응을 잘 기술할 수 있는 반응 좌표를 설정하기가 어렵다는 것이다. 단순한 화학 반응의 경우는 원자간 결합거리, 원자간 결합각, 이면각(dihedral angle) 등을 이용하여 반응의 순차적 과정을 잘 기술할 수 있다. 하지만, 단백질 접힘의 경우는 굉장히 많은 원자들이 집단적으로 움직이기 때문에 반응을 기술할 수 있는 반응 좌표를 설정하기가 매우 어렵다. 단백질의 경우 펩타이드 결합 주변의 질소-탄소-탄소-질소로 이어지는 원자들의 결합을 중심으로 이면각들을 잘 조합하면 단백질의 구조를 효과적으로 탐험할 수 있다. 물론, 이 경우 어떠한 순서로 각각의 이면각을 조절해야 하는지는 여전히 결정되어야만 하는 문제이다. 이면각을 활용하는 방법은 사실 가장 낮은 에너지 비용으로 단백질의 구조를 탐험하는 방법이다.

최근 시뮬레이션 방법으로 각광받고 있는 것이 바로 최소 작용 원리(principle of least action)를 활용하는 것이다. 초기와 말기의 원자 위치를 동시에 이용하는 경계 조건을 활용하는 유사한 기법들도 많이 연구되었다. 이는 고전역학 이론을 컴퓨터 프로그램으로 구현한 것에 지나지 않다. 단백질 안에 있는 각 원자의 이동을 동시에 고려한다. 펼쳐진 단백질과 접힌 단백질 구조를 생각한다. 이들 두 구조들로부터 각 원자는 초기의 위치와 말기의 위치를 가진 것으로 간주할 수 있다. 주어진 시간 안에 단백질 내부의 모든 원자가 자신의 초기 위치에서 말기 위치로 한꺼번에 유한개(P)의 단계를 통해서 이동하는 것을 상상할 수 있다. 이러한 상황에서 우리는 작용(action)을 계산할 수 있다. 이때, 작용은 원자 이동 경로들의 함수가 될 것이다. 사실 최소 작용을 제공하는 원자 이동 경로는 벨레 경로가 됨을 해석적으로 유도할 수도 있다. 결국, 해당 컴퓨터 프로그램은 하나의 함수 최소화 문제

Hamilton's principle

$$\delta \int_0^\tau dt L(\{q\}, \{\dot{q}\}) = 0 \iff$$

Lagrange's equation

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) - \frac{\partial L}{\partial q} = 0$$

$$L(\{q\}, \{\dot{q}\}) = T(\{\dot{q}\}) - V(\{q\})$$

$$S = \Delta \sum_{j=0}^{P-1} \left\{ \frac{M}{2} \left(\frac{q_j - q_{j+1}}{\Delta} \right)^2 - V(\{q_j\}) \right\}$$

$$q_j = q_0 + \frac{j\Delta}{\tau} (q_P - q_0) + \sum_{k=1}^{P-1} a_k \sin \left(\frac{k\pi j\Delta}{\tau} \right)$$

$$j = 0, 1, 2, 3, \dots, P; \quad \Delta = \tau/P$$

Complexity $\rightarrow O(3N_{\text{atom}}(P-1))$

Fig. 2. Protein folding pathway modeling used to be a time-consuming art, where months of CPU and user time were used to make a single model. A significant improvement of folding pathway search over the last few years is due to the use of one type of double ended formulation. A series of conformations, "chain of states," connecting two conformations is a major output. The simulation could not involve selecting a set of reaction coordinate(order parameter). At the first step, we use a straight line between initial position and final position as a trial trajectory. And then, we employ a set of sine functions, which describe an arbitrary deviation of the trajectory from the straight line. Thus, the sine expansion coefficients are the practical parameterization of the trajectory.

를 푸는 것으로 귀결된다. 각 원자가 각각 어떠한 경로를 거쳐서 이동하는지를 프로그램이 직접 결정해 준다. 다시 말해서 단백질의 접힘 과정을 계산해 낼 수 있다. 이를 통상의 분자동역학 방법과 구별하여 작용-유도 분자동역학(action-derived molecular dynamics; ADMD)이라고 부른다.^[5,6]

최적의 경로를 계산하기 위해서 경로의 변화에 대한 작용의 변화도 계산을 해야 한다. 이는 각 원자에 작용하는 힘을 계산함으로써 얻을 수 있다. 확률론적 미분방정식 랑주뱅 방정식을 순차적으로 풀어내는 분자동역학 프로그램과 달리 작용-유도 분자동역학 프로그램은 병렬화가 자연스럽게 이루어진다. 컴퓨터로 작용과 경로 변화에 대한 작용의 변화를 계산할 때 여러 구조들에 대한 에너지와 각 원자에 작용하는 힘들이 동시에 요구된다. 이 부분은 여러 컴퓨터들에 분배하여 전체 계산을 병렬계산 형식으로 처리할 수 있다. 병렬 효율성을 예로 들면 다음과 같다. 1개의 계산 노드를 이용할 때와 비교하여 64개의 계산 노드를 사용하여 약 50배 정도 빨리 전체 계산을 수행할 수 있는 수준에 도달할 수 있다. 다루고 있는 단백질이 N개의 원자로 이루어져 있고 이동 경로가 P 단계로 이루어져 있다고 가정하면 실제로 최적화해야 하는 변수의 숫자는 $3N(P-1)$ 개가 된다. 원자수가 500개이고 2000 단계를 생각하면 최적화해야 하는 변수의 숫자는 약 300

REFERENCES

- [4] C. D. Snow *et al.*, Nature **420**, 102 (2002).
- [5] D. Passerone *et al.*, Phys. Rev. Lett. **87**, 108302 (2001).
- [6] I.-H. Lee *et al.*, J. Comput. Chem. **31**, 57 (2009).

$$\bar{\epsilon}_{I,j} = 2\bar{R}_{I,j} - \bar{R}_{I,j-1} - \bar{R}_{I,j+1} - \frac{\Delta^2 \partial V(\{\bar{R}_{I,j}\})}{M_I \partial \bar{R}_{I,j}}$$

I : 1, 2, 3, ..., N
j : 1, 2, 3, ..., P-1

a.u. = 0.529177 Å

Complexity $\rightarrow 3N(P-1)$,
 N = the number of atoms,
 P = the number of steps

Fig. 3. A typical distribution of the scalar 'error variables' in the text is shown. An example for measuring the path quality of transition paths connecting two stable conformations is shown.

만 개가 된다.

그림 2에서는 질량 M을 가지는 단일 입자의 운동에 대한 작용과 이동 경로에 대한 표현식을 함께 나타내었다. 입자의 수에 따라서 계산 방식이 달라지는 것은 아니다. 편의를 위해서 단일 입자의 작용을 적어본 것이다. 각 입자의 경로는 직선 그리고 sine 함수들의 합으로 표시했다. 총 P 단계를 거쳐서 입자가 말기 위치에 도달한다고 생각하면 찾아야 하는 중간 단계의 수는 P-1 개가 된다. 그림 3에서는 계산된 원자 이동 경로를 분석해 본 것이다. 특별히 벨레 경로로부터 벗어난 정도로 계산된 $3N(P-1)$ 개의 값들을 통계처리해 보았다. 대부분의 경로들은 벨레 경로 조건을 잘 만족하는 것을 알 수 있다. 일부 경로는 벨레 경로로부터 벗어나 있음을 알 수 있다. 이는 계산에 사용한 시간 간격의 값(약 36 fs)이 일반적인 분자동역학에 사용되는 시간 간격(1 fs)보다 크게 잡혀져 있기 때문이다. 결국 36 fs보다 빨리 움직이는 운동들은 작용-유도 분자동역학 계산에서 모두 평균적인 운동으로 처리되었음을 의미한다. 이보다 짧은 시간 안에 일어나는 일들은 본질적으로 관찰할 수 없다.

그림 4에서는 HP-36 단백질의 접힘 경로를 보여 주고 있다. 작용-유도 분자동역학 방법으로 시뮬레이션한 결과이다. 약 36 fs의 단위 시간 간격을 사용했으며 총 2000 단계에 최종적으로 접힌 단백질 구조에 도달하는 과정을 계산한 것이다. 단백질의 유효반경, 고유 구조로부터 떨어진 거리, 고유 접촉의 수, 수소 결합의 수, 그리고 퍼텐셜 에너지 변화를 동시에 보여주고 있다.

Fig. 4. Snapshots of the lowest potential-energy-barrier pathway for HP-36 are shown. ADMD simulations are conducted using AMBER94 all-atom force field with GB/SA solvation model. The variation of colors from blue to red corresponds to the amino-acid sequence of the polypeptide chain from N- to C-terminus. Various reaction coordinates or physical quantities are shown along the folding pathway for HP-36. They are, in the order shown from top to bottom, the radius of gyration, root-mean-square deviation from native structure, native contacts, native hydrogen bonds, the total energy of the system, and the potential energy.

단백질 접힘 시뮬레이션: 복제품-맞바꿈 방식

작용을 최소화시키는 원자 이동 경로로 얻어진 단백질 접힘의 경로는 단백질 접힘 과정을 기술하는 반응 좌표로서 손색이 없다. 다시 말해서 작용-유도 분자동역학 계산은 순차적인 단백질 접힘의 경로를 직접 탐색할 수 있게 해준다. 얻어진 반응 좌표(순차적으로 얻어진 단백질 구조들)를 이용하여 추가적으로 단백질의 접힘 과정에 대한 자유 에너지를 계산할 수도 있다. 이러한 단백질 접힘에 대한 자유 에너지는 복제품-맞바꿈 분자동역학(replica-exchange molecular dynamics; REMD) 방법을 이용하면 효율적으로 얻어낼 수 있다.^[7] 분자동역학 시뮬레이션에서 단백질은 큰 규모의 구조 변화를 경험하기 전에 특정한 퍼텐

REFERENCES

- [7] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

Replica-exchange molecular dynamics

$$p = \min \left\{ 1, \frac{\exp\{-E_j/(k_B T_i) - E_i/(k_B T_j)\}}{\exp\{-E_i/(k_B T_i) - E_j/(k_B T_j)\}} \right\}$$

$$= \min \left\{ 1, \exp\left\{ (1/(k_B T_i) - 1/(k_B T_j))(E_i - E_j) \right\} \right\}$$

Exchanges of conformations are tried periodically.

Fig. 5. REMD method consists of M replicas, covering the temperature range between T_1 and T_M . Periodically, state-exchange tries are attempted, where two neighboring replicas exchange their conformations. The acceptance rule for each state-exchange try between two neighboring states i and j is chosen to be the equation shown in the figure.

설 에너지 덩어리에 걸려 대부분의 시간을 소모하게 된다. 그렇기 때문에 매우 많은 단계의 분자 동역학 시뮬레이션이 필요하다. 하지만, 상대적으로 높은 온도에서 분자 동역학 시뮬레이션을 수행하면 퍼텐셜 에너지 장벽을 상대적으로 쉽게 넘어갈 수 있다. 높은 온도에서는 쉽게 구조 변화를 관찰할 수 있다. 하지만, 높은 온도에서 단백질이 경험하는 구조 변화의 방식은 낮은 온도에서 일어나는 구조 변화 방식과 다르다. 복제품-맞바꿈 분자 동역학은 이러한 상황을 잘 이용한 병렬 계산 방법이다.

관심이 있는 단백질을 여러 개 만들어 낸다. 이들을 우리는 복제품이라고 부른다. 그 다음, 다양한 온도에서 각기 다른 초기 조건으로 분자 동역학 시뮬레이션을 수행한다. 특정한 시뮬레이션 시간이 흐른 후 단백질 복제품들 사이에 맞바꿈의 동작을 수행한다. 이는 온도 분포를 가정한 상태에서 가장 가까운 온도들 사이의 단백질 구조들, 즉, 복제품들의 맞바꿈을 시도한다. 이때, 맞바꿈을 수행할 것인지 말 것인지는 메트로폴리스(Metropolis) 알고리즘을 이용해서 결정한다. 이러한 시뮬레이션은 관심이 있는 모든 온도에서 상대적으로 쉽게 앙상블(ensemble) 평균을 얻어낼 수 있게 해준다. 퍼텐셜 에너지 덩어리에서 벗어난 다양한 단백질 구조를 모든 온도에서 자동으로 관찰할 수 있게 해주는 것이다. 이러한 계산 방법은 단백질의 비열과 같은 열역학적 성질들을 효과적으로 계산할 수 있게 해준다.

복제품-맞바꿈 분자 동역학은 사실 하나의 슈퍼 모의 담금질(super simulated annealing)로 분류할 수 있다. 단 하나의 복제품을 이용하고 높은 온도에서부터 천천히 온도를 낮추어 가면서 퍼텐셜 에너지가 낮은 분자 구조를 찾는 방식인 모의 담금질 방법^[8]을 확장한 것으로도 해석할 수 있다. 복제품-맞바꿈 분자 동역학은 여러 개의 복제품을 동시에 사용하는 특징이 있다. 뿐만 아니라 끊임없이 높은 온도에 적응된 복제품을 만들어내고

Mean first passage times, $\{x_i\}$

$$x_i = \begin{cases} \tau_{lag} + \sum_{j=1}^K x_j P_{ij}, & i \neq K, \\ 0, & i = K. \end{cases}$$

Fig. 6. In a Markov state model, we represent the conformation space by K discrete states, each of which corresponds to some distinct group of protein conformations. Let us define the probability of transition from state i to state j at a lag time as P_{ij} . We can use the transition probability matrix to calculate kinetic properties of the system such as free energy profile of protein folding, mean first passage time to reach the native state.

메트로폴리스 알고리즘을 통해서 낮은 온도 시스템에 복제품을 공급한다. 결국, 특별히 낮은 온도를 하나 추가하게 되면 복제품-맞바꿈 분자 동역학 시뮬레이션은 모의 담금질 시뮬레이션이 될 수 있다. 대부분의 컴퓨터 계산은 주어진 온도에서 독립적으로 랑주뱅 방정식을 푸는 것이기 때문에 복제품-맞바꿈 분자 동역학 방법의 구현은 기본적인 프로그램 병렬화 수준의 것이고 높은 병렬 효율성을 보장한다. 그림 5에서는 복제품-맞바꿈 분자 동역학에서 사용되는 복제품-맞바꿈 확률 조건을 수식으로 표시했다. 결국 복제품-맞바꿈 분자 동역학 계산은 랑주뱅 방정식과 메트로폴리스 알고리즘으로 단백질의 구조 공간을 탐색하는 방법이 된다.

단백질 접힘 시뮬레이션: 마르코프 상태 모델

간단한 단백질의 경우 반응 좌표를 단백질의 유효 반경 또는 고유의 단백질 구조로부터 떨어진 거리 등으로 잡아도 접힘 과정을 잘 기술하는 경우가 있다. 주어진 반응 좌표를 활용하여 각각의 반응 좌표들 사이의 전이 확률을 앞에서 기술한 분자 동역학 방법으로 계산할 수 있다.^[9] 여기서 반응 좌표는 다수의 특징적인 단백질 구조들을 각각 고유한 상태로 취급하는 경우를 말한다. 상대적으로 길지 않은 다수의 분자 동역학 시뮬레이션을 통해서 단백질 상태들 사이의 전이 확률을 계산할 수 있다. 결국에는 선택한 모든 상태들 사이의 전이 확률을 계산할 수 있고 이는 행렬로 표시할 수 있다.

마르코프(Markov) 상태 모델은 한 상태에서 다른 상태로 전

REFERENCES

- [8] S. Kirkpatrick *et al.*, Science **220**, 671 (1983).
- [9] W. C. Swope *et al.*, J. Phys. Chem. B **108**, 6571 (2004).

Fig. 7. Trajectories obtained from molecular dynamics are split or combined to maintain the number of trajectories per bin, while preserving the correct probabilities in each bin.

이하는 것을 총괄적으로 기술하는 수학적 모델이다. 다음의 상태는 오로지 현 상태에만 의존하는 것이 특징인 무작위 절차가 바로 마르코프 절차이다. 마르코프 상태 모델 조건이 만족되고 전이 행렬이 얻어지면 우리는 충분한 시간이 지난 후 각 상태에 단백질이 머물러 있을 확률을 계산할 수 있다. 이는 바로 자유 에너지를 반응 좌표에 대해서 계산하는 것을 의미한다. 또한 특정한 한 상태에서 다른 상태로 전이할 때 걸리는 시간(mean first passage time)을 직접 전이 행렬로부터 계산할 수 있다. 단백질이 완전히 접히는데 걸리는 시간도 직접 계산할 수 있다. 다시 말해서 수학적으로 잘 정의된 마르코프 상태 모델의 특징적인 성질을 그대로 이용하면 단백질 접힘의 과정을 이해할 수 있다. 앞으로 많은 응용 연구가 기대되는 단백질 접힘 연구방식이다. 그림 6에서는 마르코프 상태 모델을 이용하여 단백질 상태들 사이의 이완 시간을 계산하는 공식을 표시했다. 주어진 지연 시간 그리고 전이 행렬을 입력하여 선형대수 계산을 통해서 상태들 사이의 전이 시간을 계산할 수 있다.

단백질 접힘 시뮬레이션: 가중-앙상블

마르코프 상태 모델에서와 유사하게 단백질 접힘의 반응 좌표를 설정한 다음 앞서 이야기한 분자 동역학 시뮬레이션을 수행하여 특정한 시간(지연 시간)이 지난 후 단백질이 어떠한 상태에 있는지를 체크한다. 한 상태에서 서로 다른 초기 조건들을 활용하여 다수의 분자 동역학 계산을 동시에 수행한다. 예를 들어 10개의 서로 다른 계산을 수행했다면 각 분자 동역학 계산에 해당하는 경로는 가중치 1/10을 가지게 되는 것이다. 그리하여 특정한 지연 시간 이후 하나의 분자 동역학 시뮬레이션에서 단백질이 다른 상태로 전이 했다고 가정하자. 이러한 경우 관심이 있는 단백질은 해당 경로가 가지고 있는 가중치만큼 다른 상태로 전이한 것으로 간주한다.^[10] 마찬가지로 전이한 이후의 상태에서 다시 서로 다른 초기 조건들을 가지고 다수의 분자 동역학 시뮬레이션을 수행한다. 이때 선택된 분자 동역학의 가중치는 1/100이 될 것이다. 즉, 하나의 상태에서는 항상 정해진 수, 10개의 분자 동역학 계산을

수행한다. 이러한 방식으로 지연 시간을 가지면서 다수의 분자 동역학 계산을 계속 수행하면 우리는 반응 좌표에 대한 단백질의 전이 과정을 계산된 가중치를 통해서 알 수 있다. 충분한 시간이 지난 후에 접힌 상태의 가중치가 시간에 따라서 어떻게 증가하는지를 관찰하면 단백질 접힘의 비율(1/sec)을 계산할 수 있다. 그림 7에서는 가중-앙상블 시뮬레이션 방법을 도식적으로 나타내었다. 반응 좌표를 통(bin) 형식으로 이산화시켜서 단백질의 상태를 정의했다. 정의된 단백질 상태들 사이의 전이는 분자 동역학계산을 통해서 계산한다. 해당 가중치를 재조정하는 단계를 이용하면 가중-앙상블 계산이 연속적으로 이어질 수 있다.

나오는 글

단백질 분자는 유연하게 자신의 모양을 바꿀 수 있고 단백질 서열에 따라서 특별히 고유한 모양을 정한 다음 고유의 기능을 수행한다. 주어진 아미노산 서열 정보만 가지고 단백질의 모양을 결정할 수 있고 단백질이 구체적으로 어떻게 행동하는지를 알아낼 수 있다면 생명 현상 이해에 한 발 다가서는 꼴이 될 것이다. 하지만, 아직까지 그러한 이론적 기술 방법은 존재하지 않고 있다. 어쩌면 그런 것이 존재하지 않을 수도 있다. 슈뢰딩거 박사의 추측처럼 네겐트로피 그리고 코드가 정보라는 사실을 인정한다면 실험적 데이터베이스를 활용하고 있는 생물정보학이 좋은 대안이 될 수 있을 것이다.

연구자들이 단백질의 행동을 실험적으로 측정하기가 어려운 것도 사실이다. 공간적으로, 시간적으로 충분한 분해능을 얻을 수 없기 때문이다. 이러한 상황에서 분자 동역학 시뮬레이션을 통한 단백질 접힘 과정 연구 또는 단백질 기능 탐구는 생명 현상에 대한 중요한 이해를 제공한다. 다양한 이론 및 컴퓨터 시뮬레이션 연구는 산업적인 응용이 가능한 것으로 평가받고 있다. 보다 더 많은 이론, 컴퓨터 시뮬레이션 연구가 있을 것으로 전망한다. 또한 보다 긴밀한 실험-시뮬레이션 양측 간의 교류가 있어야만 할 것이다.

단백질 연구는 단순한 산업적 응용 분야를 넘어서는 연구 주제로 생명 현상에 대한 사람들의 이해를 증진시킨다. 인위적으로 유전정보를 변화시킬 수 있는 현재의 기술 상태는 우리가 우주의 모든 생명 물체를 어떻게 바라봐야 하는지에 대한 가장 근본적인 관점을 놓치지 않게 해주는 일종의 신호라고 봐야 옳다. 아울러, 인간 사회의 이해와 발전이라는 맥락에서 단백질 연구는 지금보다 더욱더 가중되는 논쟁의 중심에 서게 될 것이 확실하다. 어쩌면 인간은 생물학적으로 잘 살아가기 위해서 단백질을 연구하기 보다는 사회적으로 더 잘 살아가기 위해서 단백질을 연구해야 하는지도 모른다.

REFERENCES

- [10] G. A. Huber and S. Kim, Biophys. J. **70**, 97 (1996).