단백질 구조 결정과 구조 예측의 전산 모사 방법

DOI: 10.3938/PhiT.20.018

주 기 형 · 이 승 중 · 이 주 영

Computational Methods for Protein Structure Determination and Protein Structure Prediction

Keehyoung JOO, Sung Jong LEE and Jooyoung LEE

Proteins are the most essential bio-molecules that carry out vital biological activities in living cells. Protein's function in the cell is usually determined by its unique 3D structure called native structure. In order to understand the phenomena of life at the molecular level, it is essential to know the relationship between native structures of proteins and their functions. In determining the structures of proteins from X-ray or NMR (nuclear magnetic resonance) experiments, it is crucial to perform computational calculations. Also, in recent years, it is getting more popular to combine computational methods of protein structure predictions with experimental structure determination. In this article, we present a brief introduction to computational methods for determining the protein structures from X-ray crystallography and NMR. We also introduce protein structure prediction methods from protein's sequence information.

저자약력

주기형 교수는 성균관대학교 이학박사(입자물리학, 2004)로서 고등과학원 계산과학부 연구원과 조교수를 거쳐 2011년부터 고등과학원 거대수치계산 연구센터의 연구교수로 재직 중이다. (newton@kias.re.kr)

이승중 교수는 미국 시카고 대학 이학박사(동계물리학, 1993)로서 서울대학교 이론물리 연구소 연구원을 거쳐 1995년부터 수원대학교 물리학과 교수로 재직 중이다. (sjree@suwon.ac.kr)

이주영 교수는 서울대학교 물리학과를 졸업하고, 1990년에 Brown University에서 상전이 현상의 수치적 접근법이란 주제로 박사학위를 취득했다. Argonne National Lab, Florida State Univ.에서 박사후 연구원을, Mt. Sinai School of Medicine, Cornell University에서 연구원으로 지냈고, 2000년부터 고등과학원 계산과학부 교수로 재직 중이다. (jlee@kias.re.kr)

“물리학은 앞으로 가장 복잡하고 불가사의한 현상을 규명해야 하는 상황에 직면하게 될 것이다. 그것은 바로 생명이다.”

— Erwin Schrödinger

들어가는 글

복잡한 생명현상을 지배하는 바이오 물질들이 DNA, RNA, 그리고 단백질(protein)임이 밝혀지면서 생명과학은 현재 가장 큰 변화의 시기를 지나가고 있다. 특히 X-ray 결정학과 핵자기 공명법(NMR)을 이용하여 이러한 물질들의 3차원적 구조가 밝혀지게 되어 복잡한 생명현상들을 원자 수준에서 그 원리를 이해하고 설명하는 것이 가능하게 되었다. 1953년 제임스 왓슨(James D. Watson)과 프랜시스 크릭(Francis Crick)에 의한 DNA 이중나선 구조의 발견은 생물학의 모습을 완전히 바꾸어 놓았으며 현대 “분자 생물학(Molecular Biology)”이라는 분야의 시작점이 되었다. 물리학자였던 크릭에 의해 DNA → RNA → Protein으로 연결되는 분자생물학의 중심 원리(Central Dogma of Molecular Biology, Fig. 1)가 제안되었고, 이를 통해 생물학은 인류 역사 이래 가장 큰 변화를 맞이하였으며, 그 중요성으로 현재 21세기의 중심 학문으로 자리를 잡았다.^[1] 이러한 변화의 중심에는 또한 컴퓨터의 계산 능력의 획기적인 발전이 있다. 즉, 컴퓨터를 이용하는 바이오 물질 전산모사 방법들이 개발되고, 광범위하게 사용되기 시작하면서 전산 생물학(Computational Biology)의 분야가 크게 발전되었다. 특별히 생명현상을 이해하는 데에 핵심이 되는 단백질의 구조를 실험적으로 또는 계산으로 알아내고, 어떠한 과정을 거쳐 단백질이 안정된 구조로 접혀 가는지를 밝혀내는 것은 매우 중요한 문제가 되고 있다. 본 기고에서는 위의 바이오 물질 중 특별히 단백질의 3차구조의 결정에서의 전산

REFERENCES

- [1] Francis Crick, *What Mad Pursuit: A Personal View of Scientific Discovery* (New York, Basic Books, 1988).

Fig. 1. A cartoon figure of “The Central Dogma of Molecular Biology” (left) where transfer of information is shown from DNA to RNA to protein via transcription and translation, respectively (from wikipedia). Also shown on the right is the native 3D structure of myoglobin. The protein is spontaneously folded into its stable native structure corresponding to the state of thermodynamic equilibrium.

모사의 방법들과, 실험을 통하지 않고 전산모사의 방법만으로 단백질 3차 구조를 예측하는 방법 그리고 그 응용 가능성에 대하여 소개하고자 한다.

단백질 구조 결정과 전산 모사

단백질은 자연에 존재하는 20가지의 아미노산(amino acid)들이 펩타이드 결합으로 연결되어 짧게는 30여 개에서 많게는 1000개 이상의 아미노산 잔기(residue)의 사슬로 이루어져 있다. DNA 서열에서 gene이라 불리는 영역이 mRNA로 전사되고 생체 내의 단백질 공장인 리보솜에서 단백질로 합성되게 된다. 이 과정에서 단백질은 생체 내에서 각자 고유한 3차 구조로 스스로 접히게 된다. 이를 단백질 접힘(Protein Folding)이라고 한다. 크리스찬 안핀센(Christian Anfinsen)은 이러한 단백질의 안정한 구조가 열역학적 자유에너지를 최소화하는 조건에 의해 결정된다고 하는 이른바 열역학적 가설(Thermodynamic hypothesis)을 제시하였고, 또한 이를 실험적으로 증명하여 1972년 노벨 화학상을 수상하였다.^[2] 단백질은 그 3차원적 구조(Fig. 1)가 매우 중요하다. 그것은 그 3차원적 구조가 생체 내에서의 단백질의 기능과 밀접하게 연관되어 있기 때문이다. 그러므로 생체 내의 수많은 효소반응, 면역반응 등과 같은 여러 생명현상들은 이러한 단백질의 구조를 규명함으로써 그 기능을 원자수준에서 이해할 수 있는 것이다. 현재 단백질의 구조를 밝힐 수 있는 실험적 방법으로 단백질

결정에 대한 X-ray 산란 실험, 핵자기 공명(Nuclear Magnetic Resonance)을 이용한 방법 등이 사용되고 있다. 그러나 이러한 실험 방법을 통해서도 직접 단백질의 구조를 사진 찍듯 바로 알아낼 수 있는 것은 아니며, 실험 데이터에 여러 가지 전산 모사 방법을 이용한 분석 작업을 거친 후에야 그 3차원적 구조를 결정할 수 있다. 1958년 최초로 미오글로빈(myoglobin, Fig. 1)의 구조^[3]를 X-ray 결정학으로 밝혔던 존 켄드류(John Kendrew)에서 시작하여 여러 노벨상이 이 분야에서 나왔으며, 최근까지도 바이오 물질들의 구조 결정과 그 생명현상의 이해에 노벨상이 계속해서 주어지고 있다.

X-ray 결정학과 구조계산

단백질은 기본적으로 그 크기가 수 Å~수십 Å 정도로 매우 작아 전자 현미경으로도 원자 수준의 구조를 규명할 수가 없다. 반면에, X-ray는 그 파장이 0.01~100 Å 정도로 매우 짧아 단백질 결정 내부 구성 원자들에 의한 회절 간섭 패턴을 만들게 되는데, 이를 분석하여 결정을 이루고 있는 원자들의 3차원적 배열을 알아낼 수 있다 (Fig. 2). 아직까지 X선을 휘게 하여 상을 직접 얻게 하는 렌즈는 없어 회절 패턴 데이터로부터 컴퓨터 계산을 통해 그 3차원적 구조를 규명하게 된다. 실험 단계를 간단히 요약하면 다음과 같다. 우선, 실험실에서 얻은 단백질 결정은 그 내부에 단백질이 격자(lattice) 구조 안에 적당한 공간 그룹 대칭(Space Group Symmetry)을 가지고 배열되어 있고 동일한 격자 구조가 반복됨으로써 결정을 이루고 있다. 이 단백질 결정에 X-ray를 통과시키면 X-ray는 규칙적으로 배열되어 있는 원자의 전자구름에서 탄성 산란(Thomson scattering)되어 여러 원자에 의한 산란의 경로차에 따라 서로 간섭하여 회절 현상을 일으키게 된다. 이때 산란 각도에 대한 간섭의 정도에 따라 회절 패턴이 나타나게 된다(Fig. 2). 이 회절 패턴은 결정 내의 원자배열과 직접적인 연관 관계를 갖게 된다. 회절 패턴에서 각 점은 역격자 공간(Reciprocal Space)의 점(h, k, l Miller index)에 대응되어지며 각 점에서의(X-ray 산란의) 강도(intensity)가 측정되어진다. 이러한 실험을 통해 실제 단위격자(unit cell)의 크기와 각도를 나타내는 파라미터들($a, b, c, \alpha, \beta, \gamma$), 그리고 공간 그룹 대칭이 결정되어진다. 그러면 다음의 식에 의해 실제 격자안의 단백질의 전자 밀도 함수(electron density function) $\rho(xyz)$ 를 얻을 수 있다.

REFERENCES

- [2] C. B. Anfinsen, Science **181**, 223 (1973).
- [3] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff and D. C. Phillips, Nature **181**, 662 (1958).

Fig. 2. A schematic representation of an X-ray crystallography experiment is shown. X-ray beams are passed through a protein crystal where atoms in the protein crystal scatter the X-rays producing diffraction patterns on a photographic film. Each spot on the film corresponds to the reflection at a specific point on the reciprocal lattice with indices h, k, l (called miller index) and reflection intensity I . The three axes of the unit cell have lengths a, b, c and three angles α, β, γ (between the axes). Electron density map is calculated by using structure factor $F(hkl)$ with its phase $\alpha(hkl)$ as in the above equation. (taken from the book "Crystallography Made Crystal Clear, Gale Rhodes, Academic Press, 2006").

$$\rho(xyz) = (1/V) \sum_{hkl} |F(hkl)| e^{i\alpha(hkl)} \exp(-2\pi i(hx + ky + lz))$$

$F(hkl)$ 은 구조인자(structure factor)이며 이는 회절 패턴의 Intensity와 $I(hkl) = |F(hkl)|^2$ 의 관계가 있다. 그러나 구조인자의 위상 $\alpha(hkl)$ 는 실험으로부터 주어지지 않는다. 이것을 결정학에서 위상 문제(phase problem)라 부른다. 식에서 보는 것처럼 전자 밀도 함수와 구조인자는 서로 푸리에 변환(Fourier Transform) 관계에 있다. 그래서 X-ray 결정학은 기본적으로 잃어버린 위상을 결정하는 문제로 귀착된다. 위상 문제를 풀기 위한 여러 실험적 방법들이 개발되어 있는데, 예를 들어 IR(isomorphous replacement)이라는 방법에서는, 대상 단백질을 특정한 무거운 원자(heavy atom)들과 함께 결정화한 시료에서 얻는 회절 데이터와 원래의 대상 단백질만으로 이루어진 결정으로부터 얻은 회절 데이터 사이의 차이를 분석함으로써 구조 인자의 위상을 결정할 수 있다. 그러나 여기에는 실험적으로 단백질 결정을 다시 만들어야 하는 등 시간과 비용의 문제가 있게 된다. 여기서는, 이러한 실험적 방법에서의 위상 결정 문제의 어려움을 해결하기 위해서 전산 모사 방법에만 의존하여 위상을 구하고자 하는 이른바 분자 치환(Molecular Replacement, MR) 방법에 대해 소개하고자 한다.^[4]

MR 방법의 기본 아이디어는 비슷한 구조를 갖는 단백질이 같은 격자 안에 있으면 비슷한 회절 패턴을 보인다는 것이다.

그래서 이러한 특성을 이용하여 풀고자 하는 단백질의 단위 격자를 컴퓨터 안에 가상적으로 만들고, 비슷한 구조의 단백질을 그 안에 넣어 적당한 위치를 찾으면 이론적으로 구조인자를 계산할 수 있게 된다. 여기에서 얻은 위상을 실제 실험에서 얻은 산란 세기(구조인자의 제곱값) 결과와 결합하여 푸리에 변환을 통해 전자 밀도 함수를 계산할 수 있게 되는 것이다. 이는 어떤 의미에서는 컴퓨터 안에서 결정화(crystallization) 과정을 모사하는 것이라고 할 수 있다. 이렇게 구조가 주어져 있을 때 구조인자는 다음 식에 의해 이론적으로 계산된다.

$$F(hkl) = \sum_{j=1}^N f_j(hkl) \exp(2\pi i(hx_j + ky_j + lz_j))$$

이는 격자 안의 N 개의 원자들에 대한 합으로 주어지고, f_j 는 j 원자의 종류(단백질의 경우 질소, 탄소, 산소, 황)에 따라 얻어지는 산란 인자(scattering factor)이다. (이 산란인자는 원자의 종류에 대하여 실험적으로 측정되어 있다.) 그러나 격자 안의 단백질의 구성 원자의 위치(x, y, z)에 따라 값이 다르게 계산되므로 전산 모사를 통해 그 위치를 결정하여야 한다. 그 위치를 바꿀 때마다 구조인자의 크기를 계산하여 실제 실험에서 얻은 구조인자의 크기 값과 비교하고 이 과정을 반복하여 일정한 정도의 오차 안에서 실험적 산란 세기로 수학적 수렴이 이루어지는 해(solution)를 찾게 된다. 그러면 그때 계산으로부터 주어지는 구조인자의 위상 정보를 이용하여 전자 밀도 함수를 구하게 된다. 이러한 MR 방법은 단백질 데이터베이스에 비슷한 구조가 있어야만 적용할 수 있는 한계가 있다. 하지만 시간이 지날수록 많은 단백질들의 구조가 밝혀지고 있어 현재는 이러한 MR 방법이 단백질의 구조를 규명하는데 가장 많이 쓰이는 방법이 되었다.¹⁾ 현재 Molrep, Auto_MR, EPMR, Phaser, CNS 등 여러 소프트웨어들이 개발되어 사용되고 있다. 한편 비슷한 단백질이 있더라도 잘 풀리지 않는 경우들이 있는데, 이 경우에는 뒤에 소개할 고해상도 단백질 구조예측 방법을 사용하여 MR 방법에 적용할 수 있으며 구조예측 기술의 정확성이 높아질수록 MR 방법의 성공 가능성이 높아지게 될 것이다. 한편 모델의 3차 구조를 변화시켜가며 MR 해를 찾고자 하는 시도를 할 수 있으며 단백질 서열 정보만 가지고 임의의 구조로부터 직접 전산 모사를 통해 위상정보를 얻고자 하는 시도도 있다(*ab initio* phasing). 또한 X-ray 회절 데이터에는 많은 중복이 있는데, 작은 분자인 경우 이러한 정보를 이용하여 위상정

REFERENCES

- [4] P. Evans and A. McCoy, Acta Cryst. D **64**, 1 (2008).
- [5] Herbert A. Hauptman, Jerome Karle, Direct Methods and Anomalous Dispersion, The Nobel Prize Lecture in Chemistry (1985).

1) 2011년 4월 약 7만2천여 단백질 구조가 알려져 있다. <http://www.pdb.org> (Protein Data Bank Homepage).

보를 직접적으로 얻는 직접법(direct method)이 알려져 있다(Hauptmann-Karle 방법).^[5] 이는 전자밀도가 음수가 될 수 없다는 기본적인 사실과, 서로 다른 밀러 지수에서의 구조인자의 위상 값들 또는 그 절대 값들 사이에 많은 부등식 관계, 즉 특별한 구속 관계(constraint relation)가 성립한다는 수학적 사실을 이용하여 위상을 구하는 방법이다. 아직까지는 너무 많은 계산량을 요구하여 아주 작은 분자나 대칭성이 많은 단백질들에 대해서 적용할 수 있다. 이렇게 위상 문제를 해결하게 되면 얻어진 초기 위상 정보로부터 전자 밀도 함수를 계산하고 여기에 단백질의 서열 정보에 따라 구조를 끼워 맞추고, 밀도 수정(density modification)과 구조 정밀화(structure refinement) 과정을 거쳐 최종 단백질 구조를 완성하게 된다.

NMR 데이터로부터 단백질 구조의 결정

지난 2002년 미국 스크립스 연구소(The Scripps Research Institute)의 Kurt Wüthrich(Kurt Wüthrich) 교수에게 노벨 화학상이 주어졌다.^[6] 그는 지난 1980년대 중반에 핵자기 공명법(Nuclear Magnetic Resonance, NMR)으로 단백질과 같은 바이오 분자의 구조를 규명할 수 있음을 최초로 보였다. 이로서 단백질과 같은 바이오 분자들의 구조를 밝힐 수 있는 방법이 X-ray에 이어 NMR 방법이 추가되었다. 현재 NMR을 이용한 연구에서는 계속해서 새로운 방법들이 개발되고 있으며 그 정밀도도 높아지고 있다. NMR 방법의 장점은 결정이 필요하지 않고 용액 상태에서 데이터를 얻을 수 있다는 점이다. X-ray 결정학과 마찬가지로 NMR 방법 역시 단백질 구조를 직접 얻을 수는 없고, 실험 데이터로부터 전산 모사 방법을 통해 그 3차 구조를 결정하게 된다.

NMR 실험의 원리를 간단히 요약하면 다음과 같다. 먼저 단백질을 구성하고 있는 원자의 핵에서 양성자와 중성자는 $S = 1/2$ 의 스핀 각운동량을 지니고 있어서 각각 고유한 자기모멘트를 가지고 있는데 양성자 또는 중성자의 수가 홀수인 경우라면 핵 전체의 자기 모멘트가 0이 아닌 값을 가지게 되어 외부 자기장에 의해 에너지 준위의 분기(splitting)가 일어나고 이러한 에너지 분기는 외부의 전자기파(주로 라디오파)와 공명(resonance)을 일으킬 수 있다. 이러한 현상을 이용하여 특히 수소(H)의 스펙트럼을 측정하게 되고 이를 분석하여 단백질을 이루고 있는 여러 수소 원자 사이의 거리 정보와 단백질 뼈대(backbone)와 곁사슬(side-chain)의 뒤틀림 각(dihedral angle) 정보를 구속조건으로 얻게 된다. 특히 구조 결정에서 중심 역할을 하는 수소 원자간 거리 정보를 NOE(Nuclear Overhauser Effect) 거리 구속조건(distance restraint)이라 한다. 또한 수용액 상의 단백질들 사이에 바이셀(bicell)과 같은 지질(lipid)을 넣어 주고 자기장을 걸면 자기

Fig. 3. A schematic representation of a typical procedure for structure determination from NMR experiment. After peak assignments from NMR spectra, solution NMR structures are determined by optimizing the NOE distances, dihedral angles and RDC restraints.

장에 따른 바이셀의 정렬을 따라 단백질도 정렬하는 경향이 생기게 된다. 이를 통해 각 결합에 위치한 두 원자의 핵스핀들이 일정하게 정렬하게 되어 쌍극자 커플링이 생기게 되고, 외부 자기장의 주파수에 따라 공명되어 에너지 분기가 관측되게 된다. 이를 잔여 쌍극자 커플링(Residual Dipolar Coupling, RDC)이라 하고 이를 통해 단백질의 도메인(domain) 사이의 방향에 대한 정보를 얻게 된다.

좀 더 진보된 실험들이 계속 개발되고 있지만, 여기에서는 NOE, dihedral angle, RDC 구속조건 정보를 이용하여 단백질의 3차 구조를 결정하는 전산 모사 방법을 소개하고자 한다. 기본적으로 실험에서 얻은 모든 거리 정보와 뒤틀림각 정보, 그리고 RDC 정보를 사용하여 컴퓨터 안에서 이러한 정보를 가장 잘 만족하는 구조를 찾아서 최종적으로 단백질의 3차원 구조를 얻게 된다. 이를 위해 다음과 같이 단백질 구조의 에너지 함수를 정의하여 이를 최적화하는 알고리즘을 사용하게 된다.^[7]

REFERENCES

- [6] Kurt Wüthrich, NMR Studies of Structure and Function of Biological Macromolecules, The Nobel Prize Lecture in Chemistry (2002).

$$E = E_{\text{stereo-chemistry}} + E_{\text{NOE}} + E_{\text{Dihedral}} + E_{\text{RDC}}$$

여기서 $E_{\text{stereo-chemistry}}$ 는 위에서 소개한 것처럼 단백질은 기본적으로 아미노산들이 연속적으로 연결되어 이루어진 사슬 분자여서 원자 간의 결합길이(bond length), 결합각(bond angle) 등이 표준적인 값의 범위 안에 있어야 하는데 이를 단백질이 기본적으로 만족해야 하는 stereo-chemistry 에너지 항이라 부르며 대개 단백질의 구조계산에서 쓰이는 CHARMM, Amber, OPLS 등과 같은 Force Field들이 사용되게 된다. E_{NOE} 항은 NOE 구속조건들에 대한 항으로서 다음과 같다.

$$E_{\text{NOE}} = \begin{cases} k(R - d_{\text{lower}})^2, & \text{for } R < d_{\text{lower}} \\ k(R - d_{\text{upper}})^2, & \text{for } R > d_{\text{upper}} \\ 0, & \text{otherwise} \end{cases}$$

NOE 거리 구속조건이 허용하는 하한값(lower bound)과 상한값(upper bound) 안에서는 0이고, 바깥에서는 조화 포텐셜 형태로 주어진다. 여기에서 k 값은 힘 상수(force constant)이다. 또한, R 은 단백질 구조에서 특정 NOE 신호에 해당하는 두 개의 수소간의 거리를 의미하는데, 경우에 따라서는 여러 개의 수소가 관여해서 가능한 수소-수소 짝이 여러 개가 될 수 있다. 그리고 각각의 NOE 신호는 수소간 거리에 대하여 r^{-6} 에 비례하여 줄어들기 때문에 다음과 같은 합산을 사용하여 계산한다.

$$R = \left(\sum_{ij} |x_i - x_j|^{-6} \right)^{-1/6}$$

여기에서 합산은 가능한 모든 수소 짝(i, j)에 대한 합이다. 그러므로 여러 짝 중에서 가장 가까운 거리가 R 값에 가장 크게 기여하게 된다. 일반적으로 거리의 하한값은 두 원자의 반데르발스 거리의 합을 사용하게 된다. 그리고 이러한 NOE 거리 정보로부터 단백질 뼈대 원자들 간의 수소결합을 유추할 수 있게 되는데, 이는 단백질의 2차 구조(alpha helix, beta sheet) 정보로 활용되어 NOE 거리 구속조건에 추가되기도 한다. NOE 구속조건 중에서 단백질 서열상의 거리가 멀리 떨어져 있는 구속조건들이 단백질의 접힌 구조를 결정하는 중요한 구속 조건들이 된다. 두 번째로 뒤틀림각(Dihedral angle)들에 대한 구속 조건들 역시 하한값과 상한값이 주어지 NOE 에너지 항과 동일한 함수 형태로 계산되어진다. RDC에 대한 에너지는 각 원자 짝에 대한 측정 RDC 값(D_{obs})에 대해 구조로부터 계산된 RDC 값(D_{calc})과 비교하여 역시 비슷한 방법으로 계산되어진다. 주어진 구조로부터 D_{calc} 값을 계산하기 위해서는 단백질 구조의 정렬 텐서(Alignment tensor)를 계산하여야 하는데 이는 측정된 D_{obs} 들과 단백질 구조의 주축 분석(principal axis analysis)을 통해 계산된다.

이처럼 NMR을 이용한 단백질 구조 결정은 주어지는 수많은 NOE 거리들의 모호성과 구속 조건들 간에 서로 불일치할 가능성 등으로 모든 조건을 다 잘 만족하는 구조를 구하기가 쉽지 않을 수 있다. 이를 위해서 전체 에너지 함수에 대한 강력한 최적화 알고리즘이 필요하다. 즉, 가능한 많은 구속조건들을 잘 만족하는 해를 찾는 것이 NMR 구조 결정에서 중요하다. 이를 위해서 Simulated Annealing, Genetic Algorithm 등과 같이 여러 가지 최적화 방법들이 도입되어 있다. NMR 구조 결정 프로그램들도 X-PLOR, CYANA, CNS 등과 같은 프로그램들이 널리 사용되고 있다.

단백질 구조예측

위와 같이 X-ray 회절 실험이나, NMR 방법들을 통해 실험적으로 단백질의 구조를 결정하는 과정은 많은 노력과 비용 그리고 시간이 요구되는 복잡한 과정이다. 그래서 안핀센의 열역학 가설에 따라 서열 정보로부터 단백질의 구조를 전산 모사 방법으로 예측할 수 있다면 이는 이론적으로도 매우 중요하고 또, 실험과는 상호 보완적인 역할을 하게 될 것이다. 이는 현재 생물 물리학(biophysics), 전산 생물학 분야에서 풀어야 할 매우 중요한 문제 중의 하나이다 (Fig. 5 CASP 대회는 구조예측 연구와 방법에 크게 기여하였다). 여기에서는 현재 단백질 구조 예측 방법들에 대해서 소개하고 고해상도 구조예측, 그리고, 그 응용 가능성에 대해서 소개하고자 한다. 단백질의 구조를 전산모사 방법으로 예측하기 위해 사용하는 방법은 크게 두 가지가 있는데, 첫째는 안핀센의 열역학 가설에 기반한 *ab initio* 방법(physics-based modeling) 둘째, 이미 구조가 밝혀진 단백질을 템플릿(template)으로 이용하는 방법(homology modeling, comparative modeling, template-based modeling)이 있다.

먼저 *ab initio* 방법을 시도하기 위해서는 단백질을 이루고 있는 모든 원자들 간의 상호작용 에너지를 기술하고 구조에 대한 에너지를 기술하여 열역학적 자유에너지의 최저점을 주는 구조를 찾으려 한다. 그런데 단백질은 기본적으로 원자 간의 상호작용을 하기 때문에 양자역학의 지배를 받는다. 그리고 수천에서 수만에 이르는 원자로 구성되어 있는 다입자계여서 이를 정확히 풀다는 것은 아직까진 불가능하다. 그래서 많은 경험적 에너지 함수들이 개발되었는데, 사용되는 각 에너지 항만 써보면 대략 다음과 같다.

REFERENCES

- [7] Axel T. Brünger *et al.*, Acta Cryst. D54, 905 (1998).

Fig. 4. A superposition between two protein structures that share structurally similar regions with 15.4% of sequence similarity is shown (protein code 1G90: green, 1N7E: orange). TM-score in structure similarity is 77.68 and RMSD (root mean square deviation) between the two structures is 2.03 Å (generated by pymol program).

$$E = E_b + E_\theta + E_\phi + E_{vdw} + E_{el}$$

여기서 E_b , E_θ , E_ϕ 는 단백질의 모든 원자들이 연결되어 있기 때문에 생기는 화학결합 길이(bond length), 결합 각(bond angle), 뒤틀림 각(torsion angle)에 대한 에너지 함수이며, E_{vdw} 는 원자간 반데르발스 에너지 항이며, E_{el} 는 전하를 띠는 원자에 대한 에너지이다.^[8] 또 여기에 더하여 단백질의 환경은 주로 물속에 있는 것과 비슷하기 때문에 물과의 상호작용(solvation energy)도 고려되어야 한다. 단백질이 가질 수 있는 그 모양의 가짓수를 한번 생각해 보면 아미노산 한 개당 대략 10개의 모양을 가질 수 있다고 가정하면 아미노산이 100개만 되어도 가능한 가짓수가 10^{100} 이 되어, 아무리 빠른 컴퓨터를 사용한다 해도 모든 모양에 대해서 가장 낮은 에너지를 찾으려면 우주의 나이보다 더 긴 시간을 필요로 한다.^[2] 그러므로 주어진 에너지 함수의 자유에너지 최저점을 찾기 위해서는 강력한 광역 최적화(global optimization) 방법이 있어야 한다. 이렇게 해서 경험론적 에너지 함수를 사용하여 분자 동역학(Molecular Dynamics), 몬테카를로 방법을 사용하여 단백질이 접히는 시뮬레이션을 하고자 했던 많은 시도가 있었다. 그러나 아직까지 이렇다 할 만한 좋은 결과를 얻지 못하고 있다. 그래서 등장한 방법이 단백질의 국소 구조(local structure)를 이미 구조가 알려져 있는 단백질의 조각(fragment)에서 가져와 조합(assembly)하는 방법과 이를 기술하는 더 경험적인 에너지 함수들이 도입되었다. 이 방법은 비교적 단백질을 접는

2) 실제의 단백질은 생체 내에서 대개 마이크로 초에서 수 초, 어떤 것은 길어야 수 분 정도 만에 접힌다. 단백질 접힘에서 이런 실제와 이론과의 커다란 장벽을 Levinthal Paradox라고 부른다.^[9]

Fig. 5. Homepage of the protein structure prediction center. CASP (Critical Assessment of Protein Structure Prediction) experiments are organized to provide the means of objective testing of the structure prediction methods through the process of blind prediction. The most recent CASP was CASP9 which was held in December, 2010. There are three major categories of structure prediction in CASP which are called, respectively, high accuracy template based modeling (HA-TBM), template based modeling (TBM), and free modeling (FM) that are classified according to the difficulties of the targets for structure prediction.

데 짧은 시간이면 가능하지만 여전히 그 구조의 정확도에 있어서 만족스러운 결과를 주지 못하고 있다.

그래서 2차적으로 사용하는 방법이 바로 이미 구조가 알려진 단백질을 템플릿으로 이용하는 방법으로 매우 실용적이고 가장 널리 사용되고 있다. 이 방법의 핵심 아이디어는 아미노산의 서열이 비슷한 단백질은 비슷한 모양을 갖는다는 것이다. 실제로 아미노산 서열 유사성(sequence similarity)이 30% 이상이 되면 대체로 비슷한 구조를 갖는다. 또 어떤 경우에는 놀랍게도 10% 미만의 유사성에도 유사한 구조를 갖는 경우도 발견되고 있다 (remote homology).

Fig. 4는 두 개의 비슷한 단백질을 구조적으로 겹쳐본 것(superposition)인데, 이것은 구조를 서로 알고 있을 때 가능하다. 이렇게 구조를 겹쳐 보면 이에 따른 서열 정렬을 얻을 수 있는데, 그 서열의 유사성이 15.4% 정도임을 알 수 있다. (이런 그림은 항상 보는 각도가 조금 다른 그림을 양쪽에 그리는데, 왼쪽 그림은 왼쪽 눈으로 오른쪽은 오른쪽 눈으로 보게 되면 입체감 있게 단백질을 볼 수 있다) 여기서 만약 두 구조 중 하나가 그 3차원 구조가 알려져 있지 않고 아미노산의 서열 정보만이 주어졌다고 가정해보자. 그러면 여기서 그 알려져 있지 않은 구조를 예측하기 위해서는 두 가지 문제에 부

REFERENCES

- [8] Minoru Kanehisa, *Post-Genome Informatics* (Oxford University Press, 2000).
- [9] C. Levinthal, *Journal de Chimie Physique et de Physico-Chimie Biologique* **65**, 44 (1968).

Fig. 6. Multiple sequence alignment of T0290 with 8 homologous proteins. T0290 was a high accuracy modeling target in CASP7 experiment. Many conserved regions between homologous proteins can be found in the alignment. The degree of conservation is indicated by the height of bar curve (below the alignment) where many conserved regions between the proteins as well as remote homologous region can be found. Modeling of T0290 target depends on the structural similarities of the templates. Through multiple sequence alignment, more accurate models can be generated by combining multiple templates. Thus, accurate multiple alignment is crucial to generating high accuracy model.

잊치게 된다. 첫째는 “이미 알려져 있는 구조들 중에서 어떻게 비슷한 구조를 가진 단백질을 찾을 것인가?”이며 두 번째는 찾았다고 했을 때, “어떻게 구조가 없는 상태에서 Fig. 4의 아래에 주어진 정확한 서열 정렬을 찾을 것인가?”이다. 여기서 첫 번째 문제는 구조인식(fold recognition) 문제로 불리고 두 번째는 서열 정렬(sequence alignment) 문제로 일컬어진다. 구조 인식 문제의 경우 서열의 유사성이 높으면 (20~30% 이상) 비교적 쉽게 단백질 데이터베이스(PDB)에서 비슷한 구조들을 찾을 수 있다. 그러나 서열 유사성이 낮은 경우(20% 미만)에는 매우 어려운 문제가 되고 어떤 경우에는 비슷한 구조가 없을 수도 있다. 이것이 템플릿 기반 구조 예측 방법의 약점이다. 유사한 구조를 찾을 수 있는 경우에는 비교적 높은 정확도로 구조예측을 할 수 있지만, 그렇지 못하면 어쩔 수 없이 템플릿이 없이 모델링을 시도하여야 한다. 한편 현재까지 구조가 밝혀진 단백질은 약 7만2천 개 정도이다. 그러나 실제로는 구조의 유사성을 제외하면 5천 개 미만의 구조가 알려져 있다. 그래서 템플릿 기반 구조 예측 방법은 데이터베이스에 의존하게 된다. 두 번째로 서열 정렬 문제도 매우 중요한 문제인데, 현대 생명정보학의 발전으로 이에 대한 많은 연구가 이루어져 왔고, 현재에는 여러 개의 유사한 서열을 동시에 정렬하는 다중 서열 정렬(multiple sequence

Fig. 7. Here is shown a high accuracy model for target T0290 generated by using the templates in Fig. 6 with the multiple sequence alignment, in superposition with its native structure. Backbone RMSD between the model and the native structure is only 0.48 Å for the entire chain of 173 residues. Native backbone and the model backbone structure are indistinguishable from each other with the TM-score 99.08. Native (model) side-chains are shown in green (yellow) lines. Aromatic residues from the core region are well predicted, while some exposed side-chains do not agree with each other. The figure is generated by pymol.

alignment)을 통해 얻게 되는 이른바 프로파일(profile) 정보를 이용하여 점점 정확한 서열 정렬을 얻는 것이 가능해지고 있다. 서열 정렬의 기본 아이디어는 Fig. 4의 서열정렬에서 보는 것처럼 아미노산 서열간의 매치 스코어와 매치되지 않은 부분에 대한 penalty 함수를 도입하여 가장 높은 스코어를 주는 정렬을 동적 프로그래밍(dynamic programming) 방법을 통해 구하는 것이다.^[10]

이렇게 구조인식과 서열 정렬 방법을 통해 템플릿 구조와의 서열정렬을 얻으면 다음 단계로 단백질 각 원자들 간의 거리와 결합 각, 비틀림 각 등의 정보를 이 서열정렬과 템플릿 구조를 통해 얻는다. 이것은 일종의 구속조건으로 보통 re-strain이라고 한다. 그러면 이러한 구속조건들과 단백질 구조의 기본 특성 정보에 대한 에너지 함수를 사용하여 구조를 예측하게 된다. 이러한 방법을 쓰는 대표적인 방법으로 공개 프로그램인 MODELLER가 있다. 또 다른 방법으로 조각 조립방법(fragment assembly)이나 부분 매칭(segment matching) 방법들이 있다. 여기까지 오면 단백질의 뼈대 구조를 얻을 수

REFERENCES

- [10] Sequence Alignment, http://en.wikipedia.org/wiki/Sequence_alignment.

있고 각 아미노산들의 곁가지 구조를 모델링하고 최종 3차원 구조를 얻게 된다.

고해상도 구조예측

고해상도 단백질 구조의 예측은 단백질 구조의 디테일, 예를 들면 뼈대뿐만 아니라 곁가지에 있는 원자들의 위치의 정확성을 더 높이는 것을 의미한다(Fig. 7). 당연히 단백질 전체 구조의 정확성이 높아질수록 그 활용도는 높아지게 되고, 실제 생명현상을 이해하는 데 더 유용하게 될 것이다. 위에서 템플릿 기반 구조 예측 방법에서의 한계는 구조예측의 정확성이 사용한 템플릿보다 높이가 어렵다는 것이다. 이는 실험으로 구조가 얻어진 후에 구조를 비교하여 보면 알 수 있게 되는데, 이는 기본적으로 정답 구조를 알지 못한 채 서열 정렬의 정확성이 완벽할 수 없기 때문에 생기는 문제이다. 그래서 여러 개의 템플릿을 사용하는 다중 서열 정렬을 통해 이를 극복하고자 하는 시도가 계속해서 개발되고 있다. 문제는 다중 서열 정렬의 경우 문제의 복잡성이 이른바 NP(Non-deterministic Polynomial-time) 문제가 되어 정확한 해를 구하는 데 걸리는 시간이 서열의 개수의 증가에 따라 급격히 늘어난다는 점이다.

현재는 다중 서열 중 두 개씩의 서열들에 대한 짝-서열정렬(pairwise alignment)들의 결과에 기반한 휴리스틱 방법들이 많이 사용되고 있으며, 이들 여러 짝-서열 정렬로부터 얻은 정보로부터 서열 매치 라이브러리를 구성하여 가장 많은 라이브러리 일치성(consistency)을 보이는 다중 서열 정렬을 얻고자 하는 시도가 있다.^[11] 실제로 이를 통해 단백질 구조 예측의 결과가 가장 유사한 단백질의 구조보다도 더 좋은 결과를 얻은 예들이 보고되고 있다. 또한 이러한 서열 정렬 후 행해지는 모델 세우기(model building)와 사이드체인 예측 단계에서도 템플릿들과 서열정렬로부터 주어지는 수많은 정보(일종의 구속조건)들로부터 에너지 함수를 정의하여 주어진 정보를 가장 많이 만족하는 구조를 구하는 방법으로 고해상도 구조예측 방법이 개발되었다.^[12] 여기에도 NMR 구조 전산모사에서와 같이 강력한 최적화 방법이 필요하며 이에 대한 계속된 연구가 필요하다.

한편 이러한 전산모사의 방법만으로 단백질 구조예측이 가능해지면서 실제 여러 생명과학 연구에 활용이 가능하게 되었다. 특별히 고해상도 구조예측 방법이 개발되면서 이러한 가능성이 더 열리게 되었다. 예를 들면 X-ray 실험에서 MR 방법에 이러한 예측된 구조를 활용하여 위상문제를 더 잘 풀 수 있는 가능성이 열리게 되었으며 MR 방법의 해의 수렴의 영역을 더 넓힐 수 있게 되었다. 또한 실험으로 얻기 힘든 단백질들에 대해서도 구조예측이 가능하면 그 단백질에 대한 예측 구조로부터 많은 유용한 정보를 얻을 수 있으며, 새로운

접근 방법과 생명현상의 이해에 다가갈 수 있을 것이다. 또한 최근 단백질 고유의 기능을 개선하거나 바꾸는 등의 단백질 디자인에 대한 연구가 진행되고 있는데, 여기에서도 단백질 구조 예측 방법과 전산모사의 방법들은 대상 단백질에 대하여 폭 넓게 응용될 수 있으며 나아가 신약개발과 같은 부분에까지 연결될 수 있다.

맺는말

생명현상에 대한 연구는 매우 광범위하다. 1944년, 유전정보에 대해 거의 아무것도 알려진 것이 없던 그때 “생명이란 무엇인가?(What is Life?)”^[13]의 근원적 질문을 던졌던 슈뢰딩거의 생명에 대한 깊은 통찰력에 대한 답변으로 현대 생명과학은 역사 이래 엄청난 변화와 도전의 시기를 지나가고 있다. 전체의 흐름으로 볼 때, 가장 큰 변화는 생명현상을 물리학에서의 방법론과 같은 물리-화학의 근본적 원리에 따라 이해하고 설명하는 것이 어느 정도 가능해졌다는 것이다. 슈뢰딩거의 언급과 같이, 가장 복잡하고 불가사의해 보이는 바이오 물질 구조와 기능을 규명하고 이해하기 위해서는, 특별히 여러 전산모사의 방법들이 많은 부분에서 중요한 역할을 할 것이다.

본 글에서 단백질 구조결정과 구조예측에서의 전산모사 방법들을 통해 그 3차원적 구조를 규명하는 방법과 그 응용에 대해서 살펴보았다. 아직도 대부분이 베일에 싸여있는 생명과학의 그 넓고 깊은 세계를 이해하기 위한 기본적인 예비 단계로서, 단백질과 같은 중요한 바이오 분자들의 구조와 접합 과정에 대해 전산모사의 방법들이 많은 가능성을 보여 준다고 생각한다. 장차 실험과 이론 및 전산 모사의 복합적 연구를 통해 여러 생명현상의 근본적인 원리와 이해에 더 깊이 다가가게 될 것이 기대된다. 또한 위에서 살펴본 바와 같이 그 응용 가능성으로, 실제 생체 내의 여러 질병을 일으키는 현상의 분석과 이해에 적용되어 신약개발 등 의학의 영역까지 확대가 가능하게 될 것이다. 이를 위해 이러한 분야에 더욱더 많은 관심과 노력이 기울여지기를 기대한다.

We thank Korea Institute for Advanced Study for providing computing resources (KIAS Center for Advanced Computation Linux Cluster System) for this work.

REFERENCES

- [11] Keehyoung Joo, Jinwoo Lee, Ilsoo Kim, Sung Jong Lee and Jooyoung Lee, Biophys. J. **95**, 4813 (2008).
- [12] Keehyoung Joo, Jinwoo Lee, Sunjoong Lee, Joo-Hyun Seo, Sung Jong Lee and Jooyoung Lee, Proteins **69**, Suppl. 8, 83 (2007).
- [13] Erwin Schrödinger, *What is Life?* (Cambridge University Press, 1944).