신약개발을 위한 타겟단백질의 구조계산

이진각 · 한원석 · 윤창노
한국과학기술연구원

만일 단백질 모양을 빠르고 자동적인 방법으로 알아낼 수 있다면, 아직 확인이 안된 많은 생화학적인 메커니즘을 분자 수준에서 더 자세히 살펴볼 수 있게 될 것이며 단백질의 기능을 촉진하거나 방해하는 약물분자의 개발 과정에 응용할 수 있을 것이다. 많은 연구자들이 컴퓨터 계산에 의한 방법으로 특정 타겟 단백질에만 작용하는 약물분자를 개발해낼 수 있는 날이 곧 올 것이라고 예상하고 있다. 인체 내 수많은 단백질 중 지금까지 약의 목표로 사용되는 단백질은 약 4백여개 정도이나 머지 않아 약물의 목표가 되는 타겟 단백질의 수가 1만여개에 이를 것으로 예상하고 있다. 질병을 일으키는 단백질의 정확한 입체 구조를 몰라 이를 치유할 수 있는 약물개발이 늦어지고 있으나 입체 구조가 밝혀지면 새로운 약을 설계하기도 보다 쉬워질 것이다.

단백질은 세포, 기관, 생물체들을 구성하고 있는 생화학적 분자이다. 커다란 건축 구조물을 만들듯이 단백질은 여러 작은 펩타이드 조각들을 모아서 거대한 구조물을 만들어 내며 이러한 과정을 "구조형성(folding)"이라 부른다. 어떻게 이 과정이 이루어지는가를 살펴보려는 시도를 많은 학자들이 하고 있는데 이를 일컬어 "단백질 구조형성 문제(the protein folding problem)"라고 하며 생명과학분야에서 아직 풀리지 않고 있는 중요한 문제중의 하나이다. 1935년에 처음으로 제기된 후 지금까지도 해독되지 않고 있는 암호라고 할 수 있다. 30년전에 크리스챤 안핀센(Christian B. Anfinsen)이란 화학자는 단백질의 구조에 대한 정보가 아미노산 서열에 들어 있다는 것을 증명했다.[1] 보통 단백질은 열이나 화학물질로 인하여 변형될수 있는데 안핀센은 124개의 아미노산으로 이루어진 리보뉴클리아제(ribonuclease)라 불리는 효소 단백질을 변형시킨 다음 화학 분석 실험을 통해 원래의 단백질 모양이 바뀌고 기능이 없어졌다는 것을 확인하였다. 그 후 다시 변형시키는데 사용했던 물질들을 제거함으로써 그 단백질의 기능을 회복시켰다. 그리고 매우 복잡한 과정의 화학 분석을 사용해 안핀센은 그 단백질이 원래의 모양으로 회복되었음을 증명했다. 안핀센은 이 일로 1972년에 노벨 화학상을 타게 되었지만, 그의 실험으로는 어떻게 아미노산 서열정보가 단 한 개의 단백질 모양을 만들어 내도록 구조를 형성시켰는지는 밝혀낼 수 없었다. 최근의 얻어진 연구결과에 의하면 단백질은 엔탈피와 엔트로피[2,3]의 적절한 조화[4]와 van der Waals 인력, 전자기적 인력, 수소결합, 단백질 내부분자간 상호작용, 단백질 외부분자와 용매와의 상호작용 등을 포함하는 약한 상호작용[5]에 의해서 구조가 결정됨이 알려졌다. 따라서 많은 연구자들은 이러한 힘들을 컴퓨터를 사용하여 계산함으로써 단백질 구조형성 문제를 풀기 위해 노력해 왔다.

현재 컴퓨터를 이용해 단백질 구조를 밝히려는 실제적인 시도로서 서열상동성을 이용하거나 최소에너지를 사용하고 있다. 염기서열만 알고 구조를 모르는 단백질의 경우, 단백질 구조 데이터베이스에서 비슷한 염기서열을 가지는 단백질을 찾는다. 염기서열이 비슷하면 구조도 어느 정도 비슷하기 때문이다. 즉 단백질 데이터베이스를 활용해 새로운 단백질의 구조를 어느 정도 짐작하는 것이다. 위와 같이 서열상동성을 이용할 뿐만 아니라 최근에 많은 과학자는 단백질 구조형성 과정을 단백질의 에너지 상태 개념으로부터 풀고 있다. 대부분의 단백질은 전체적으로 보아 둥글게 뭉쳐진 공 모양의 형태로 구조를 만들어간다. 표면에는 물을 좋아하는 아미노산이, 내부로는 물을 싫어하는, 즉 기름을 좋아하는 아미노산이 배치된다. 이때 각각의 아미노산은 비틀리거나 꺾이거나 또는 휘어지며 에너지적으로 가장 안정된 상태를 취한다. 따라서 단백질 구조형성 문제를 해결하기 위해 아미노산으로 이뤄진 긴 사슬의 최소에너지 상태를 찾는 것인데 이런 방법으로 얻어진 모양을 '자연 구조'(native conformation)라 부르고 단백질의 원래의 모양(natural shape)과 똑같을 것이라고 생각한다.

자연계의 모든 단백질은 3차원 입체상태로 존재한다. 입체구조 내에서 어느 특정부분은 매우 단단한 구조를 형성하기도 하는데, 이를 단백질의 2차 구조라 한다. 실제로 가능한 2차 구조는 나선구조(helix), 평면상태(sheet), 꺾인 모양(turn) 등이 있다. 2차 구조는 매우 규칙적이고 반복적인 구조를 가지고 있다. 단백질의 3차 구조는 다양한 2차 구조들이 조합된 결과다. 2차 구조의 조합 결과로 생기는 3차 구조는 거의 무한대에 가까운 수를 생성해낼 수 있다. 많은 연구자들은 무수히 많은 단백질 3차 구조를 빠르고 자동적으로 계산하기 위해 컴퓨터를 사용한다. 컴퓨터는 단백질이 어떤 모양인지 모르지만 화학

의 가장 기본적인 규칙을 적용해 단백질 구조형성 과정을 흉내낼 수 있다(simulation). 이러한 단백질 3차원 구조 시뮬레이션을 위해 컴팩, IBM, 오라클 같은 대형 컴퓨터 회사들은 천문학적 단위의 개발비를 투자하고 있다. 특히 IBM은 단백질 구조연구에 이용될 '블루진'(Blue Gene)이라는 초고속 슈퍼컴퓨터 개발에 1억 달러를 투자하고 있다. 소프트웨어 업체인 오라클도 최근 10만개로 추정되는 인체 내 단백질 목록을 만드는 새 프로젝트의 모든 정보를 저장할 수 있는 데이터베이스를 만들겠다고 발표했다.

컴퓨터를 사용하여서 단백질 구조를 얻는 방법은 크게 *comparative modeling(CM), fold recognition(FR), ab initio* 등으로 나눌 수 있다. 일반적으로 이러한 세 가지 방법에 의해서 예측된 구조의 정확도는 주어진 서열의 상동성의 정도에 의존한다. 서열 상동성이 30~50% 이상인 경우에는 주로 CM 방법을 사용하며, 정확도도 세 방법중 가장 좋다. 서열 상동성이 20~40%인 경우에는 FR 방법과 ab initio 방법이 사용될 수 있으며 FR 방법이 조금 더 좋은 결과를 얻는다. 서열 상동성이 20% 이하인 경우에도 FR 방법과 ab initio 방법은 서로 경쟁적이나, 이 경우에는 ab initio 방법이 조금 더 좋은 결과를 얻으며, 더 낮은 상동성을 가지는 경우에는 ab initio 방법만 가능하다.

Comparative Modeling (CM)

CM 방법은 서열정렬 algorithm을 사용하여서 질의 서열과 관계된 서열과 구조를 찾고, 다중 정렬 algorithm을 사용하여서 family의 구조와 서열을 정렬하여서 구조 template를 얻고 이것을 사용하여서 단백질 구조 결정하는 방법이다. 좋은 구조 template를 얻기 위해서는 서열정렬이 가장 중요하다. 서열정렬은 서열간의 상관관계 즉, 상동성을 나타내기 위해 서열을 정렬하는 것을 말한다. 이러한 서열정렬은 정렬하려는 서열의 수와 상동성의 형태에 따라서 구분한다. 2개의 서열에 대한 정렬은 pairwise alignment, 3개 이상의 서열에 대한 정렬은 multiple alignment라 하고, 상동성 형태에 따라서 global alignment와 local alignment를 구분한다. 서열정렬은 질의 서열과 상동성이 높은 서열을 알아내어서 서열의 기능을 유추하거나, 관련 있는 서열들간의 상관관계 구조 등을 예측하기 위해서 사용된다. 두 서열을 비교할 때 서열을 전체적으로 비교하여 최대의 상동성이 나타나도록 정렬하는 경우, 이러한 정렬을 global alignment라 한다. 반면, 두 서열이 어떤 부분의 서열이 높은 상동성을 가지는지를 알기 위해 정렬하는 것을 local alignment라 한다. Needleman & Wunsch의 dynamic programing 기법에 의한 서열정렬 algorithm[6]이 발표된 이후, Smith & Waterman이 일반적인 길이의 gap에 대해 이 algorithm을 확장[7]함으로써 두 서열간 정렬을 위한 실용 가

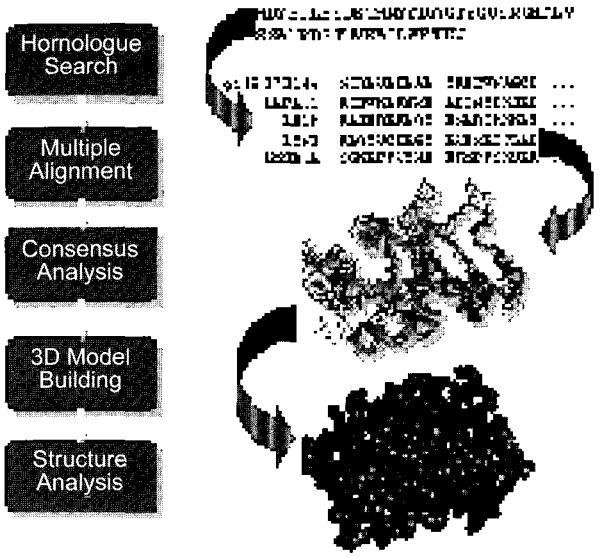

그림 1. 일반적으로 사용되는 CM방법.

능한 프로그램들이 구현되었다. Lipman, Wilbur, Pearson등이 hasing과 window 설정의 방법을 사용하여 고안한 FASTP/FASTN[8]은 local alignment를 포함한 FASTA[9]로 발전하여 데이터베이스 대상의 상동성 검색에 사용되었고, Altschul과 Karlin이 homologous candidate fragment 선정과 extension 방법을 사용하여 개발한 BLAST[10]는 서열정렬을 위해 가장 많이 활용되어 왔으며, gap 처리가 안되던 단점을 해결한 BLAST2[11]가 발표되어 BLAST를 대체하였다. 서열의 다중정렬 algorithm은 gap cost를 dynamic하게 처리하는 방법을 도입한 ClustalW[12]가 가장 많이 사용되고 있다. 그림 1은 일반적으로 사용되는 CM방법을 도표와 그림으로 표현한 것이다.

Fold Recognition (FR)

FR 방법은 질의 서열에 대한 가장 좋은 상동성을 가지는 서열의 상동성 정도가 떨어져서 그 탐색된 구조를 template으로 사용할 만큼의 정확도가 나오지 않으므로, 서열로 얻어진 구조를 사용하여서 구조-구조 정렬을 통하여서 template를 찾는 방법을 사용한다. 구조-구조 정렬 algorithm은 fragment assembly method를 사용하는 Holm & Sander의 Dail, Sipple *et al.*의 ProSup[13]과 단백질 이차구조 벡터를 정렬하는 Gibrat *et al.*의 VAST, 우선 단백질의 일차, 이차, 삼차 구조에 대한 특성 profile 만들고 이것을 정렬한 후 fragment assembly method를 사용하는 Jung & Lee의 SHEBA[14] 등이 있다. FR 방법에 이러한 방법 외에, 알려져 있는 단백질의 domain이나 motif를 cluster하여서 representative domain/motif structure database을 만들어서, 질의 서열을 각 구조에 대응시킨 후 가장 높은 score를 가지는 구조를 결과물로 내는

방법이 있다.

ab initio

ab initio 방법은 CM, FR 방법과 다르게 구조 template을 필요로 하지 않아서, 서열 상동성에 상관없이 사용될 수 있다. ab initio 방법은 실재하는 단백질 구조의 template를 사용하지 않기 때문에 임의의 단백질 구조를 생성하여야 한다. 적은 수의 아미노산으로 이루어진 단백질의 경우에도 단백질이 가질 수 있는 구조의 conformational space의 수는 천문학적인 숫자가 된다. 그래서 ab initio 방법에서는 아미노산을 하나의 원자와 같이 생각하는 unified residue system을 사용하고 이러한 unified-residue 원자로 구조를 표현하는데 lattice model과 off-lattice model을 주로 사용한다. 아래 그림은 lattice model을 사용하여서 단백질 구조를 표현한 한 예이다.

1970년대 후반에 Levitt & Warshel은 아미노산을 2개의 입자로 표현하고 Langevin dynamics와 energy minimization 사용하여 BPTI[15]와 Carp Myogen[16]을 연구하여서 단백질은 이차구조를 먼저 형성함을 밝혔다. lattice model의 장점은 분명하다. 단순화된 모델은 conformational space상에서 효율적인 sampling을 가능하게 한다. 적절하게 설계된 lattice model을 사용하면 분석적으로 계산 가능한 global energy minimum을 얻을 수 있다. 이러한 lattice model은 단백질 구조예측 문제에 적용될 수 있고, Skolnick & Kolinski는 residue-level lattice model[17]을 사용하여서 작은 단백질 구조 형성문제를 성공적으로 계산하였다. lattice model은 목적에 따라서 2가지 형태로 나눌 수 있다. 하나는 Go에 의해서 제기된 단백질 형성과정을 결정하는 기본 물리적 힘을 이해하기 위해 설계[18]된 것이다. Zhou & Karplus는 분산 충돌 모델[19,20]을 사용하여 연구를 수행하여서 단백질 구조형성은 단백질 실제구조로 energetic surface가 체계적으로 편향[21,22]되어 있다고 주장하였다. 이러한 형태의 모델은 실제 단백질을 대상으로 설계되지 않았기 때문에 단백질 구조형성 문제를 연구하는데 제한적이다. 그럼에도 불구하고 이러한 연구를 통해서 단백질 구조형성에 대하여서 알 수 있었다. Miyazawa & Jernigan[23,24] 그리고 Skolnick의 lattice model은 다른 형태에 해당한다. 이러한 형태의 모델은 실제 단백질 구조예측을 추구하므로, 알고 있는 실제 단백질 구조를 template로 사용하여 통계적 기법으로 아미노산간의 상호작용을 계산하였다. 이것은 보통 statistical potential 혹은 knowledge-based potential로 불린다. 이러한 statistical potential은 Crippen[25], Eisenberg [26,27], Sippl[28]에 의해서도 연구되었다. 또한 Scheraga는 reside-based off-lattice model[29]을 개발하였다. 이러한 형태의 lattice model들은 단백질 구조 예측에 큰 힘을 발휘하고 있다.

이렇게 중요하고도 어려운 문제인 단백질 구조예측 문제를 여러 연구자들이 공동으로 대응, 대처하고 현재의 기술수준과 앞으로 나아갈 방향을 확인하기 위하여서 CASP(Critical Assessment of Techniques for Protein Structure Prediction)라는 대회가 있는데 매 2년마다 열리는 이 대회는 아직 구조가 밝혀지지 않은 단백질을 문제로 두고 전세계 과학자들이 구조를 예측하는 전세계적 실험이다. 인간게놈프로젝트를 주도하고 있는 영국의 생거 센터(Sanger Centre)와 미국의 로렌스 리버모어 국립연구소(Lawrence Livermore National Laboratory) 주최로 열리는 이 실험은 인터넷상에서 이뤄지는데, 세계 각국의 단백질 구조예측 연구팀이 모두 참가한다. 가장 최근인 2000년 12월 열린 CASP4(http://predictioncenter.llnl.gov/casp4/)에서의 단백질 구조예측을 위한 방법의 동향은 statistical potential을 사용한 off-lattice model과 구조 motif database를 statistical potential을 사용하여 motif template를 얻고 이것을 조합하여 구조를 예측하는 방법이 주를 이루었다. 아직 서열 상동성이 낮은 경우 아미노산 서열만으로 단백질 구조를 정확히 예측하는 것은 불가능하다. 그러나 예측된 구조로부터 부분적으로 단백질의 기능을 추측할 수 있다. 앞으로

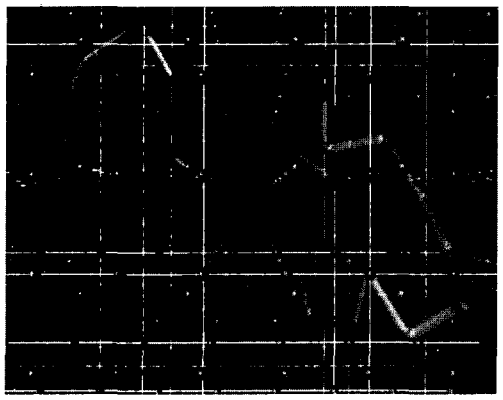

그림 2. 회색 구슬들은 격자를 나타내고 빨간 구슬은 격자를 이용하여 단백질의 구조를 나타낸 것이다. 녹색의 사슬이 자연계에 존재하는 단백질의 실제 구조이다.

그림 3. CASP4 Target T0128, Manganese superoxide dismutase homolog, 222 residues 에 대한 예측구조와 실제구조의 중첩비교(RMSD=1.0A)

단백질 구조 데이터베이스가 더 늘어나고, statistical potential을 개선하면, 서열로부터 계산된 구조가 신약개발을 위한 타겟 단백질 및 그들간의 상호작용에 사용 가능하게 될 것이다.

참고 문헌

1. C. B. Anfinsen, Principles That Govern the Folding of Protein Chains, *Science* 181, No. 96, 223-230 (1973).
2. B. Honig and A -S. Yang, Free Energy Balance in Protein Folding, *Advances in Protein Chemistry* 46, 27-58 (1995).
3. P. L. Privalov, Stability of Proteins: Small Globular Proteins, *Advances in Protein Chemistry* 33, 167-241 (1979).
4. G. I. Makhatadze and P. L. Privalov, Energetics of Protein Structure, *Advances in Protein Chemistry* 47, 307-425 (1995).
5. K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Principles of Protein Folding: A Perspective from Simple Exact Models, *Protein Science* 4, No. 4, 561-602 (1995).
6. S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequences of two proteins, *J. Mol. Biol.* 48, 443-453 (1970)
7. T. F. Smith and M. S. Waterman, *Identification of common molecular subsequences*, *J. Mol. Biol.* 147, 195-197 (1981)
8. D. J. Lipman and W. R. Pearson, Rapid and sensitive protein similarity searches, *Science* 227, 1435-1441 (1985)
9. W. R. Pearson and D. J. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA* 85, 2444-2448 (1988)
10. S. F. Altschul, Amino acid substitution matrices from an information theoretic perspective, *J. Mol. Biol.* 219, 555-565 (1991)
11. S. F. Altschul et. al., Gapped BLAST and PSI-BLAST : a new generation of protein database search programs, *Nucleic Acids Res.* 25, 3389-3402 (1997)
12. J. D. Tompson, D. G. Higgins and T. J. Gibson, Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appilc. Biosci.* 10, 19-29 (1994)
13. M. J. Sipple *et. al.*, ProSup : a refined tool for protein structure alignment *Protein Engineering* 13, 745-752 (2000)
14. J. Jung and B. Lee, Protein structure alignment using environmental profiles *Protein Engineering* 13, 535-543 (2000)
15. M. Levitt and A. Warshel, Computer Simulation of Protein Folding, *Nature* 253, 694-698 (1975).
16. A. Warshel and M. Levitt, Folding and Stability of Helical Proteins : Carp Myogen, *Journal of Molecular Biology* 106, No. 2, 421-437 (1976).
17. J. Skolnick and A. Kolinski, Simulations of the Folding of a Globular Protein, *Science* 250, 1121-1125 (1990).
18. N. Go, Theoretical Studies of Protein Folding, *Annual Review of Biophysics and Bioengineering* 12, 183-210 (1983).
19. D. Bashford, D. L. Weaver, and M. Karplus, Diffusion-Collision Model for the Folding Kinetics of the Lambda-Repressor Operator-Binding Domain, *Journal of Biomolecular Structure and Dynamics* 1, 1243-1255 (1984).
20. M. Karplus and D. L. Weaver, Protein-Folding Dynamics: The Diffusion-Collision Model and Experimental Data, *Protein Sciences* 3, No. 4, 650-668 (1994).
21. Y. Q. Zhou and M. Karplus, Interpreting the Folding Kinetics of Helical Proteins, *Nature* 401, 400-403 (1999).
22. Y. Q. Zhou and M. Karplus, Folding of a Model Three-Helix Bundle Protein: A Thermodynamic and Kinetic Analysis, *Journal of Molecular Biology* 293, No. 4, 917?51 (1999).
23. S. Miyazawa and R. L. Jernigan, Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation, *Macromolecules* 18, 534-552 (1985).
24. S. Miyazawa and R. L. Jernigan, Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term for Simulation and Threading, *Journal of Molecular Biology* 256, No. 3, 623-644 (1996).
25. G. M. Crippen, Easily Searched Protein Folding Potentials, *Journal of Molecular Biology* 260, No. 3, 467?75 (1996).
26. J. U. Bowie, R. Luthy, and D. Eisenberg, A Method to Identify Protein Sequences that Fold into a Known 3-Dimensional Structure, *Science* 253, No. 5016, 164-170 (1991).
27. J. U. Bowie and D. Eisenberg, An Evolutionary Approach to Folding Small Alpha-Helical Proteins That Uses Sequence Information and an Empirical Guiding Fitness Function, *Proceedings of the National Academy of Sciences* (USA) 91, No. 10, 4436-4440 (1994).
28. M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, Identification of Native Protein Folds Amongst a Large Number of Incorrect Models: The Calculation of Low Energy Conformations from Potentials of Mean Force, *Journal of Molecular Biology* 216, No. 1, 167-180 (1990).
29. M. H. Hao and H. A. Scheraga, Designing Potential Energy Functions for Protein Folding, *Current Opinion in Structural Biology* 9, No. 2, 184-188 (1999).