예측된 단백질 구조의 적합성 판별

DOI: 10.3938/PhiT.20.019

이진우

Quality Assessment of Protein Models

Jinwoo LEE

Quality assessment of protein models is to predict the accuracy of a predicted protein model relative to the native structure of a given protein. This technique provides rather critical information useful for desirable decision making during each step of protein structure prediction procedures to the extent that it becomes a synonym for successful protein structure prediction since more accurate assessment would result in more accurate prediction. In this article, various structural properties of protein structures and computational techniques for quality assessment are introduced.

들어가는 글

예측된 단백질 구조의 적합성 판별이란 예측된 단백질 모델이 그 단백질의 실험 구조와 얼마나 비슷할지를 예측하는 문제이다. 이러한 기술은 단백질 구조 예측의 각 단계에서 의사 결정이 필요할 때마다 최적의 선택을 할 수 있는 정보를 제공할 뿐만 아니라, 적합성 판별 기술 자체가 완전해지면 구조 예측의 정밀도 또한 높아지므로 단백질 구조 예측의 또 다른 이름이기도 하다. 이 글에서는 적합성 판별 기술에 사용되는 단백질 구조의 여러 측면과 계산 기술에 대해 살펴볼 것이다.

먼저 적합성 판별에 사용되는 주요한 특징인 2차구조 예측,

저자약력

이진우 교수는 서울대학교 이학박사(2005)로서 고등과학원 계산과학부 연구원을 거쳐 2007년부터 광운대학교 수학과에서 교수로 재직 중이다. (jinwoolee@kw.ac.kr)

PSI-BLAST PROFILE, 통계 에너지 함수, 친수성과 소수성 아미노산, 그리고 Solvent accessible surface area(SASA)라는 개념들에 대하여 살펴보고, 이들을 어떻게 기술적으로 조합하여 구조의 적합성을 판별하는지 차례대로 살펴볼 것이다.

단백질의 2차 구조 예측

단백질의 2차 구조란 수소 결합으로 형성되는 국소적이고

Fig. 1. A typical alpha helix structure is shown. Yellow dots indicate hydrogen bonding between C=O group of i -th residue, and N-H group of $i+4$ -th residue.

Fig. 2. A typical beta sheet structure is shown. White dots indicate hydrogen bonding between C=O and N-H groups.

Fig. 3. Typical alpha helix structures are shown viewed from the top of the helix axis.

규칙적인 구조로서 대표적으로 alpha helix와 beta sheet로 구성되어 있다. Alpha helix는 그림 1에서 보는 바와 같이 i 번째 잔기의 N-H 그룹과 $i+4$ 번째 잔기의 C=O그룹의 연속적인 수소결합을 통해 안정되어 있으며 나선 모양이다. Beta sheet는 그림 2에서 보는 바와 같이 지그재그 모양으로 늘어선 서로 다른 두 부분의 아미노산 사이의 연속적인 수소 결합을 통해 안정되어 있다.

주어진 아미노산 서열만 가지고 이러한 2차 구조를 어떻게 예측할 수 있을까? 일반적으로 아미노산 서열은 물과 상호작용하는 친수성 잔기와 물을 피해 뭉치는 소수성 잔기로 구성되어 있다. Beta sheet 구조를 체계적으로 살펴보면 sheet 구조의 한쪽 면은 친수성 잔기로 구성되어 있으면 반대쪽 면은 보통 소수성 잔기로 구성되어 있는 경우가 많다. 또 beta sheet 구조는 한쪽 잔기가 위로 솟아 있으면 다음 잔기는 반대로 아래쪽으로 나있다.

Alpha helix도 비슷한 특징이 있다. Alpha helix를 그림 3에서 보는 바와 같이 축을 따라 내려다보면 아미노산 잔기가 3 또는 4 방향으로 구분되는데 각 방향은 친수성 잔기로만 또는 소수성 잔기로만 구성되어 있는 경우가 많고, 특정 면을 기준으로 보면 한쪽은 친수성 잔기를 위주로, 한쪽은 소수성 잔기를 위주로 구성되어 있다. 아미노산의 2차 구조 예측은 이러한 특징을 효과적으로 이용하는 아이디어에 기반해 있다.

PSIPRED^[1]라는 프로그램은 대표적인 2차 구조 예측 프로그램으로 그 정확도가 평균 80%에 이른다. 이 프로그램은 위와 같은 관찰을 효과적으로 자동화하기 위해 Artificial Neural Network(ANN)이라는 계산 구조를 사용한다. ANN은 그림 4에서 보는 바와 같이 Input layer, Hidden layer, Output layer로 구성되어 있는데 Input layer에 들어오는 고차원 공간의 벡터를 효과적으로 클러스터링하는 계산 기술이다. 그림에 자세히 표시되어 있지 않지만 실제 PSIPRED는 315개의 노드로 구성된 Input layer, 75개의 노드로 구성된 Hidden layer, 3 노드로 구성된 output layer를 가진 ANN

Fig. 4. A schematical neural network structure is shown.

구조와 이 첫 번째 ANN의 결과를 바탕으로 60개의 노드로 구성된 Input layer, 또다시 60개의 노드로 구성된 Hidden layer, 그리고 3개의 노드로 구성된 output layer를 가진 두 번째 ANN으로 복잡하고 정교하게 구성되어 있다.

단백질 구조 예측 문제에서 2차 구조 예측은 만족스럽게 해결된 것으로 연구자들은 보고 있으며 이 기술은 단백질의 3차 구조 예측에 매우 중요하게 활용된다.

다중 서열 정렬 및 PSI-BLAST PROFILE

다중 서열 정렬(Multiple Sequence Alignment) 및 그것으로부터 추상화된 행렬의 일종인 PSI-blast profile^[2]은 단백질 구조 예측 각 단계의 정확도를 좌우하는 중요하고 유용한 개념이다. 다중 서열 정렬은 공통 조상을 가지며 진화상에서 서로 연관되어 있는 세 개 이상의 아미노산 서열을 서로 비교하여 그림 5와 같은 정렬을 생성하는 절차이다. 다중 서열 정렬은 서열 두 개를 비교하는 단순 서열 정렬보다 서열 간의 관계를 보다 정확하게 추출해 낼 수 있는 것으로 알려져 있다.

PSI-blast profile은 다중 서열 정렬로부터 얻을 수 있는 $20 \times n$ (여기서 n 은 서열의 길이) 행렬로서 그림 6에 표시되어 있다. 이 행렬은 각 위치별 아미노산간의 치환 빈도를 표시하고 있는 행렬이다. 이는 주어진 아미노산 서열의 각종 특

REFERENCES

- [1] D. T. Jones, J. Mol. Biol. **292**, 195 (1999).
- [2] S. F. Altschul, T. L. Madden, A. A. Schaffer, *et al.*, Nucleic Acids Res. **25**, 3389 (1997).

Fig. 5. The first 90 positions of the multiple sequence alignment of the acidic ribosomal protein are shown.

Fig. 6. A representation of psi-blast profile is shown.

성을 예측하기 위해 성질이 비슷한 서열을 찾거나 서로 비교할 때 서열을 대신하여 사용된다. 각 위치별 특성이 반영된 치환빈도를 표시하는 행렬이 국소적인 구조의 특성을 잘 반영하고 있을 것이므로 이를 활용한 예측 및 비교의 정확도는 단순 서열을 사용하였을 때보다 월등히 우수한 것으로 알려져 있다.

앞 절에서 소개된 서열의 2차 구조를 예측하기 위한 프로

그램인 PSIPRED는 artificial neural network (ANN)를 사용하였다. 이 ANN의 input layer에 315개의 노드가 있다고 했는데 이 노드에 들어가는 값들이 다른 아닌 2차 구조를 예측하려는 아미노산 주변 15개의 서열에 대한 Psi-blast profile 이다 ($21 \times 15 = 315$, 이 식에서 20 대신 21이 쓰였는데 이는 부가적인 정보를 표시하는 한 행이 추가되었기 때문이다.) 이 또한 2차 구조 예측에 서열 정보를 직접 사용하는 것보다 서열을 profile 행렬로 표시하고 이를 바탕으로 구분 및 정렬하는 것이 더 효과적이기 때문이다.

통계 기반 에너지 함수

통계 기반 에너지 함수(statistical potential)는 지식 기반 에너지 함수(knowledge-based potential)라고도 불린다. 이 함수는 실험으로 알려진 단백질의 3차원 구조 데이터베이스로부터 얻는데 일반적으로 아미노산을 구성하는 원자의 종류와 거리의 함수이다. 이는 단백질을 구성하는 각 원자들이 서로 어느 정도 거리를 두고 있을 때 가장 안정한지를 이미 알려진 3차원 구조로부터 통계적으로 추출한 것이다 (그림 7 참고).

우수한 통계 기반 에너지 함수는 그것만으로 예측된 구조의 적합성 판별에 사용되기도 하는데, 보다 정확한 예측을 위해 2차 구조 및 Solvent accessible surface area 등과 더불어

Fig. 7. We measure the distances between all pairs of atoms to form a statistical potential.

어 적합성 판별에 필요한 여러 요소 중 한 부분으로서 주로 사용된다.

입자의 친수성, 소수성 및 Solvent Accessible Surface Area

단백질의 3차 구조란 2차 구조가 공간상에서 얹혀 최종적으로 형성하게 되는 구조로 이는 각 원자의 3차원 좌표로 표시되며, 단백질 구조 예측 문제라고 하면 보통 3차 구조 예측을 의미한다. 이러한 3차 구조를 형성하는 힘은 소수성 입자의 뭉침(Hydrophobic collapse)으로 알려져 있는데(그림 8 참고), 이 힘에 대한 정량적 이해는 아직 만족할 만한 수준에 도달해 있지 못하고, 어떤 면에서는 이 힘에 대한 올바른 이해가 단백질 구조 예측 문제에 대한 우리의 이해 수준을 결정한다고 하여도 과언은 아니다.

소수성 입자의 뭉침을 고려하기 위해 구조예측 분야에서 자주 사용되는 개념은 solvent accessible surface area이다. 이는 말 그대로 주어진 분자가 용매와 접촉할 수 있는 표면적이다. 이를 예측하기 위해 주어진 서열 일부의 psi-blast profile을 다시 이용한다. 주어진 서열의 profile과 이미 구조가 알려진 단백질들의 psi-blast profile을 비교하여 가장 가까운 profile을 가진 구조들을 취한 후 accessible surface area를 측정하고 이 측정값을 구조를 모르는, 주어진 서열의 accessible surface area로 추정하는 방식이다.

Solvent accessible surface area의 추정이 3차 구조 형성

Fig. 8. A drop of water on a leaf is shown. The surface of the leaf is hydrophobic, and the water forms into a spherical shape to minimize the contact area.

의 주요한 요소로 알려져 있는 만큼, 실제로 구조예측을 잘하는 그룹은 이 예측에서도 남들보다 앞서는 것으로 알려져 있다.

예측치를 활용한 구조의 적합성 판별

지금까지 psi-blast profile을 활용한 2차 구조의 예측, solvent accessible surface area의 예측, 그리고 통계 기반 에너지 함수들에 대하여 살펴보았다. 그러면 이러한 기술들을 어떻게 조합하여 주어진 단백질 모델의 적합성을 판별하는데 종합적으로 사용할 수 있을까?

가장 간단하지만 그 효과가 뛰어나서 가장 많이 사용되는 방식은 consensus법이라고 불리는 방법이다. 이 절차는 다음과 같다. 먼저 주어진 서열에 대하여 psi-blast profile을 활용하여 위에 상술한 바와 같이 각종 특성을 예측한다. 둘째, 주어진 단백질 모델들은 모든 원자의 3차원 좌표가 주어져 있으므로 위 특성을 직접 측정한다. 즉 어디가 alpha helix이고 어떤 부위의 accessible surface area(그림 9 참조)가 얼마임을 계산할 수 있을 것이다. 셋째, 그 두 값, 즉 이론적으로 예측된 값과 모델의 실측값이 얼마나 일치하는지를 계산한다. 위에 상술한 각 특성의 예측은 구조를 모르는 서열에 대해 우리가 얻을 수 있는 또는 신뢰할 수 있는 유일한 정보이다. 따라서 우리가 할 수 있는 유일한 방법은 예측한 특성을 전적으로 신뢰하고, 추정된 특성치와 가장 가까운 특성치를 갖는 모델을 가장 개연성 있는 모델로 추정하는 일 뿐이다.

위 아이디어는 2차 구조 예측에 쓰인 artificial neural network 같은 도구를 이용하여 구현된다. 즉 input layer의 각 노드는 특별한 특성에 해당하고, 그 값으로는 그 특성의 예측치와 모델의 실측치가 일치하는 정도를 값으로 표현하여 입

Fig. 9. A schematic diagram of solvent accessible surface area is shown. The area is defined as the locus of the center of a solvent molecule (shown in blue) rolled over a protein molecule (shown in red).

력한다. 그리고 정답을 알고 있는 수십만 개의 모델들에 대해 artificial neural network가 올바른 적합성 추정치를 산출하도록 훈련시키는 것이다.

실전에서는 artificial neural network와 비슷하지만 수학적으로 좀 더 세련되고 안정된 support vector machine이라는 계산 기술을 사용하기도 한다. 이 기술은 그림 10에 표시한 바와 같이 평면이나 직선으로 서로 구분하기 어려운 input space의 점들을 보다 고차원의 공간(feature space)으로

Fig. 10. A schematic diagram for support vector machine is shown. SVM maps the points in the input space to a higher dimensional feature space so that the points could be separated by a hyper-plane.

투영한 후 그곳에서 평면이나 직선으로 구분하는 방법을 찾을 수 있도록 해준다.

나오는 글

예측된 구조의 적합성 판별은 위에서 언급한 바와 같이 예측된 성질의 정확도에 전적으로 의지하는 방식으로 되어 있다. 따라서 각 성질을 보다 더 정확하게 예측할 수 있다면 적합성 판별의 정확도도 높아질 것이다.

우리는 현재 행성의 움직임을 예측하고 공기나 물의 흐름 등을 예측할 수 있고 (짧은 미래에 대한 것이지만 이것이 일기예보가 가능한 이유이기도 하다), 원리적으로 조정할 수도 있다. 하지만 단백질 등 생명체에 대한 이해는 이만큼 도달하였다고 말하기 어렵다. 이러한 예측 기술들이 발달하면 할수록, 우리는 그러한 이상에 더 다가갈 수 있을 것이고, 많은 유익한 응용 분야를 열 수 있을 것이다.