

알파폴드 (AlphaFold): 인공지능 기반 단백질 3차구조 예측

서론



주 기 형 연구교수
고등과학원
거대수치계산연구센터
newton@kias.re.kr

생명 현상을 지배하는 중요 물질들이 DNA, RNA, 그리고 단백질(Protein)임이 밝혀지고, 현대의 생명과학은 분자수준에서 복잡한 현상을 규명할 수 있게 되었다. 특별히 여러 생명 현상을 이해하는 데 핵심이 되는 단백질의 생체 내에서의 기능(Function)을 이해하기 위해 그 3차원적 구조(Structure)를 규명하는 것이 현대 분자생물학과 생물물리학에서 가장 기초적인 연구이다[1, 2]. 역사적으로 X-ray 결정학, 핵자기 공명법(NMR) 등의 실험적 결정 방법들이 주로 사용되어 왔고, 이론과 컴퓨터를 이용한 계산방법으로 단백질의 구조를 예측하고자 하는 오랜 시도가 있어 왔다[3, 4]. 일반적으로 단백질의 서열이 유사한 경우($\geq 30\%$), 그 3차구조도 유사하기 때문에 이미 밝혀진 구조가 있으면 그것을 템플릿(Template)으로 사용하여 모델링하는 방법과, 템플릿이 없는 경우 여러 에너지 함수들을 사용하여 모델링하는 방법들이 사용되어 왔다[5, 6]. 템플릿이 존재하지 않는 경우, 예측된 모델의 정확도는 대략 40% 미만으로 오랜 기간 정확도의 향상이 정체되고 어려웠다.

한편 기계학습, 곧 인공지능 연구의 최근의 발전은 그동안 불가능해 보이던 여러 영역들에서 인간의 능력을 뛰어 넘는 결과들을 보여 주고 있다[7, 8]. 자연에 대한 이해와 데이터 분석을 바탕으로 계산과 시뮬레이션으로 주어진 문제를 풀던 과학과 공학 분야에서도, 이러한 인공지능 방법들이 다양하게 적용되고 있으며, 기존 전통적 방법의 한계를 극복할 수 있는 가능성을 보여 주고 있다. 알파고(AlphaGo) 개발을 통해 세계의 주목을 받았던 구글의 DeepMind팀에서 최근 알파폴드(AlphaFold)를 지난 단백질 구조예측 대회 CASP13 (2018년 12월)에서 소개하며 기대를 뛰어넘는 결과와 함께, 과학계에 신선한 충격을 가져왔다[9–11]. 이는 템플릿이 없는 서열에 대한 단백질 구조를 예측하는 데 있어서, 인공지능의 가능성을 보여준 사례라고 볼 수 있다. 본 기고에서는 템플릿을 사용하지 않는 단백질 구조 예측 분야에서 사용된 인공지능 방법들, 특히 알파폴드의 방법을 소개하고, 앞으로의 전망에 대해 소개하고자 한다. 현

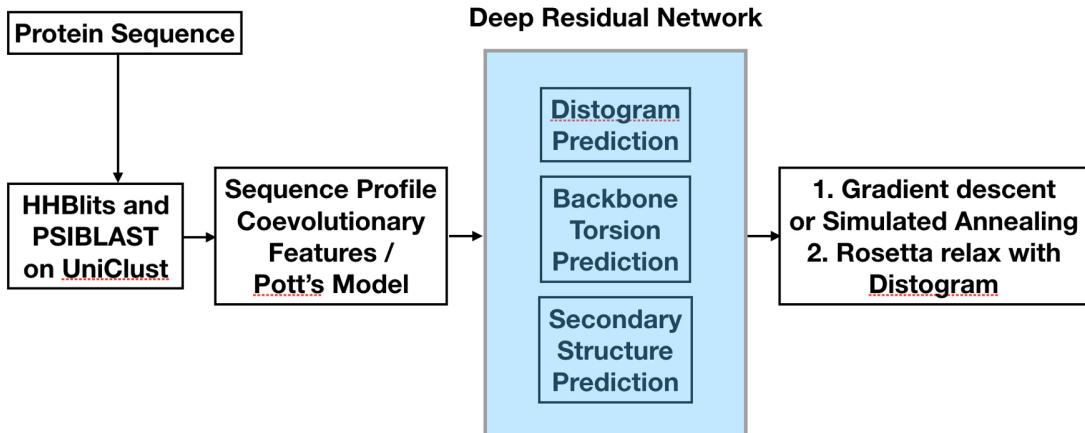


그림 1. 알파폴드 인공지능 기반 단백질 3차구조 예측 다이어그램

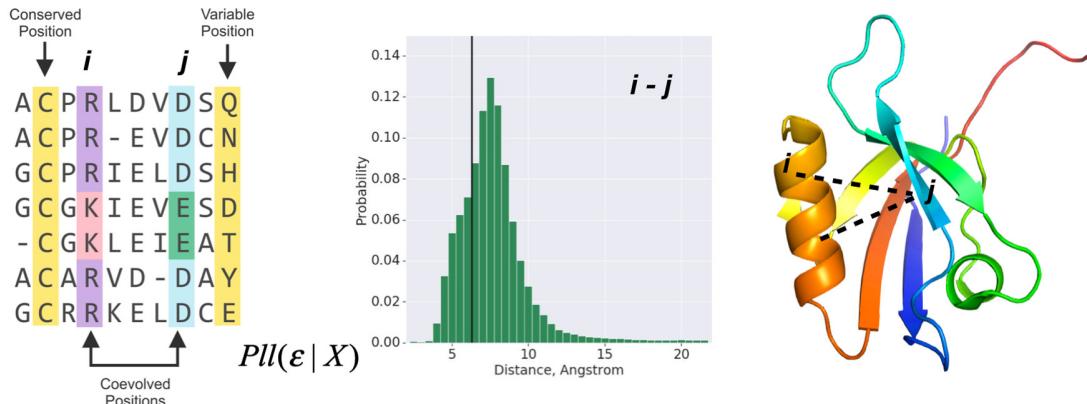
제 알파폴드는 자세한 논문이 나오지 않은 상태이지만, CASP13 학회를 통해 여러 정보들이 공유되었고, 기존의 연구 방법론들의 연장선상에서 설명 되어질 수 있다[12–16].

단백질 구조 데이터베이스와 학습 데이터

일반적인 인공지능 연구 방법에서 핵심적인 부분은 학습(Training)을 위한 양질의 데이터의 확보와 인공지능 모델 파라미터의 최적화이다. 단백질 구조 데이터베이스(Protein Data Bank: PDB)는 1971년 결성된 이래 실험 구조들이 축적되기 시작하여, 현재(2019년 3월) 150,145개의 단백질 구조가 밝혀지게 되었다(<https://www.rcsb.org>) [17]. 이는 그동안 단백질 구조 예측에 대한 여러 이론과 계산적 연구를 가능하게 했으며, 현재의 인공지능 방법을 적용할 수 있는 양질의 데이터를 제공해 주고 있다. 한편 PDB에는 유사한 단백질의 구조가 중복되어 있어 적절한 필터링이 필요하다. 알파폴드에서는 이러한 PDB에 있는 단백질구조를 폴딩의 기본 단위인 도메인으로 나눈 CATH 도메인 데이터베이스 [18]를 사용하였고, 중복을 제거하기 위하여 35퍼센트의 단백질 서열 유사도를 적용하여 필터링 하였다. 이를 통해 29,400개의 단백질 도메인 대표들에 대하여 인공지능 학습 데이터를 얻었다.

거리 히스토그램 (Distance histogram: Distogram)과 백본 비틀림 각 (Torsion angle)의 예측

알파폴드는 전체적으로 그림 1과 같이 여러 단계의 프로세스로 이루어져 있는데, 핵심적인 부분은 Deep Residual Network(DRN) 모델[8]을 사용하여, 단백질을 구성하고 있는 레지듀(Residue) i 와 j 사이의 CB 원자(CA for Glycine) 간의 거리 정보(Distance histogram: Distogram)를 예측하고 각 레지듀들의 백본 비틀림각(Backbone Torsion Angle)과 2차구조 정보(Secondary Structure)를 예측하는 것이다. 이는 그림 2(a)에서 다중서열정렬(Multiple



(a) 다중서열정렬 (MSA)

(b) 거리 히스토그램 (Distogram)

(c) 단백질 구조

그림 2. 다중서열정렬로부터 레지듀 i, j 사이의 거리 히스토그램 (Distogram)을 예측하고, 그것을 통해 단백질의 3차구조를 예측한다.

Sequence Alignment: MSA) 속에 서로 공변(Coevolved)하는 레지듀 i 와 j 는 그림 2(c)와 같이 3차구조적으로도 연관되어 있다는 아이디어에 근거를 두고 있다[12]. 알파폴드의 DRN 모델을 위한 입력 데이터(Input Features)는 CATH 단백질 도메인 대표들에 대한 서열정보로부터 아래와 같은 구성으로 이루어진다.

- 1) 단백질 서열 프로파일(Protein Sequence Profile): 입력서열에 대하여 HHBlits, PSIBLAST를 이용하여 두 가지 서열 프로파일을 생성한다[19,20]. 프로파일 생성을 위한 단백질 서열 데이터베이스는 Uniclust30을 사용했다 [21]. 또한 프로파일 1D 데이터를 행과 열에 대해서 타일링을 하여 2D 데이터로 변환하였다.
- 2) 2D Direct Coupling Analysis(DCA) 정보: 1)의 과정에서 얻어지는 MSA에 대한 파츠 모델(Pott's Model)을 구성하여 레지듀 i, j 사이의 결합(Coupling) 파라미터 ε_{ij} 를 2차원 입력 데이터로 사용한다(그림 2 참조) [22]. MSA X 가 (x_1^n, \dots, x_L^n) (20개의 아미노산 + 1개의 갭 (gap))으로 표현될 때, 이 모델에 대한 Pseudo-likeness는 아래와 같이 정의된다[23, 13].

$$Pll(\varepsilon | X) = \sum_{n=1}^N \sum_{i=1}^L \{ \varepsilon_i(x_i^n) + \sum_{j=1}^L \varepsilon_{i,j}(x_i^n, x_j^n) - \log Z_i^n \}, (j \neq i).$$

L 은 타겟 단백질 서열의 길이이고, N 은 MSA 서열의 개수이다. ε_i 는 단일 레지듀 방출 포텐셜(single-residue emission potential)이며, ε_{ij} 는 결합 파라미터(pair-wise emission potential)이고, Z 는 아래와 같이 주어지는 정규화 상수(normalization constant)이다.

$$Z_i^n = \sum_{c=1}^{20} \exp \left[\varepsilon_i(c) + \sum_{j=1}^L \varepsilon_{i,j}(c, x_j^n) \right], (j \neq i).$$

주어진 MSA에 대해서 $Pll(\varepsilon | X)$ 를 최소화하는 $\varepsilon_i, \varepsilon_{ij}$ 를 구한다. 여기에서 구한 ε_{ij} 와 프로비니우스 놈(Frobenius norm) S_{ij} 를 2차원 입력 데이터로 사용한다. S_{ij} 는 다음과 같이 정의된다.

$$S_{ij} = \sqrt{\sum_{a,b=1}^{20} \varepsilon_{i,j}(a,b)^2}$$

위의 입력레이터에 대해 그림 3과 같은 Deep Residual Network(DRN)모델을 구성한다. 이 모델의 출력 Distogram은 2–22 Å 사이의 거리를 0.5 Å 간격으로 40개의 간격(bins)으로 나누었으며, 목적함수로 소프트맥스(Softmax) 함수를 사용하였다. 알파폴드에서 사용한 DRN은 그림 3과 같은 Residual 블록을 220번 반복하여 쌓아서 구성되었으며, 4개의 블록에 대해서 Dilated Filter의 사이즈를 3×3 , 5×5 , 9×9 그리고 17×17 을 사용/반복 하였다. 최종적으로는 완전 연결 네트워크(Fully Connected Network)를 연결하여 Distogram을 예측하고, 다중작업 학습(Multitask Learning)으로 동시에 백본의 비틀림각의 분포와 8개 상태 2차구조를 예측하였다(그림 3). 이를 위해 알파폴드가 사용한 모델의 총 파라미터의 개수는 대략 2천 1백만이 넘어간다. 이를 학습시키기 위해서 2차원 데이터를 64×64 로 랜덤 크롭(Crop)하여 데이터증대(Data augmentation)하여 사용하였으며, 4개의 다른 모델들을 학습하여 앙상블 평균을 사용하였다. 크롭을 할 때 최소한 32개의 레지듀가 포함되도록 오프셋(offset)을 조정하였고, 데이터 증대효과로 과적합(Overfitting)을 피하는 데 도움을 받을 수 있다. 또한 단백질 서열에 무관한 레퍼런스 모델을 동일한 방법으로 학습시켰으며, 이를 통해 단백질 3차구조 결정을 위한 포텐셜 함수를 구축한다.

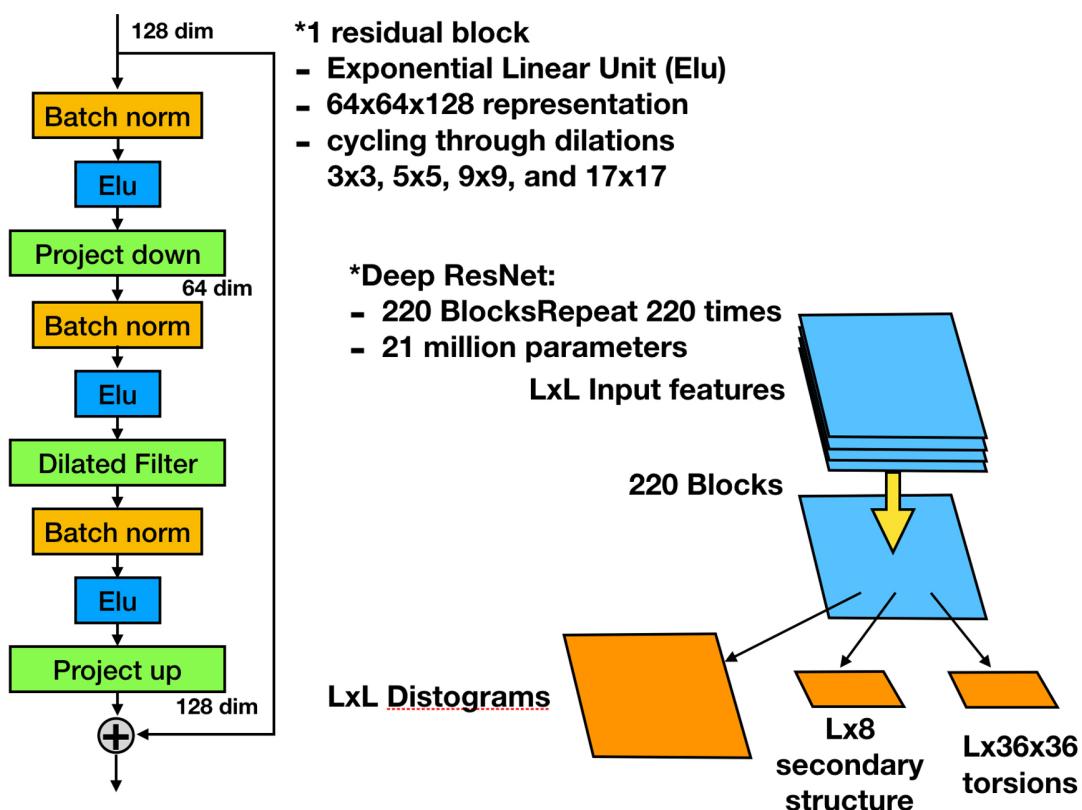


그림 3. Deep Residual Network (DRN) 다이어그램. 마지막 출력 네트워크를 통해 Distogram, 비틀림각 (Torsions), 그리고 단백질 2차구조 (Secondary structure)를 예측한다.

단백질 프레그먼트 생성과 신경망 스코어

알파폴드는 3차구조 구축을 위해 Rosetta 프로그램[24]을 사용하는데, 이를 위해 단백질 프레그먼트(Fragment)를 생성하는 DRAW[25] 기반의 네트워크를 학습시켰으며, 이를 통해 32-레지듀 길이의 프레그먼트를 생성하였다. Rosetta 프레그먼트 조합에 사용하기 위하여 9-레지듀 길이로 잘려진 프레그먼트 데이터베이스를 구축하였다. Rosetta를 사용하면 많은 단백질 3차구조들이 생성되는데, 이로부터 단백질 구조의 정확도를 예측하는 또 다른 Deep ResNet을 학습하였다. 이 네트워크를 위하여서는 위에서 예측한 Distogram, 구조들의 Distogram, 그리고 MSA 정보를 입력 데이터로 사용하여 GDT-TS 스코어(단백질 백분 구조의 정확도)를 출력값으로 사용하였다.

단백질 3차구조 예측

먼저 아래와 같이 예측한 모든 레지듀 짹에 대한 distogram $P(d_{ij} \mid \text{sequence})$ 과 레퍼런스 모델 $P(d_{ij} \mid \text{length})$ distogram의 차이를 사용하여 포텐셜 에너지 함수 V_D 를 구축한다.

$$V_D = \sum_{ij} [\log P(d_{ij} \mid \text{sequence}) - \log P(d_{ij} \mid \text{length})]$$

그리고 Rosetta의 단백질 구조 스코어 함수와 합하여 프레그먼트 삽입으로 풀림 시늉(Simulated Annealing) 방법으로 다양한 단백질 구조를 얻고, 최종적으로 함수의 미분을 사용한 L-BFGS방법을 반복 사용하여 수렴된 단백질 구

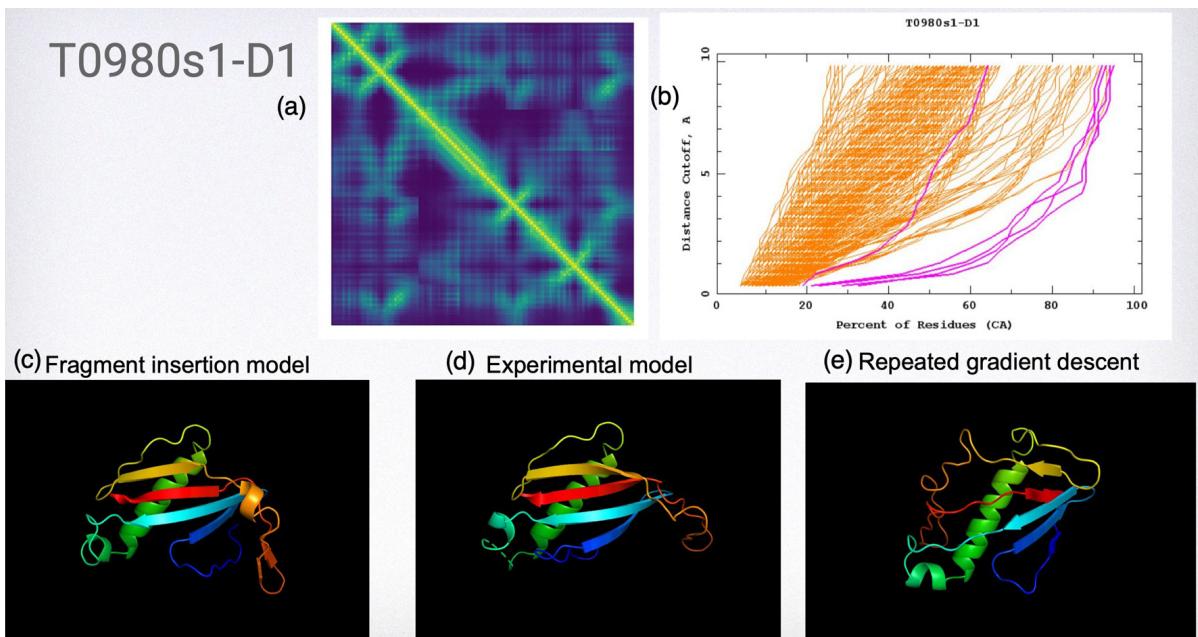


그림 4. CASP13 타겟 T0980s1-D1에 대한 알파폴드 예측 결과 (Photo from CASP13). (a) 예측된 Distogram (b) CASP13 결과 비교 (c) 프레그먼트 삽입 시늉모델 (d) 실험 모델 (e) 미분이용 반복 최적화 모델

조를 예측한다.

그림 4는 알파폴드의 방법을 사용하여 CASP13 타겟 T0980s1-D1의 결과를 보여 주고 있다. (a)는 DRN을 사용하여 예측된 distogram이며 정답구조인 (d)에 대하여 알파폴드의 결과인 (c), (d)의 구조와 (b)의 그림은 전체 제출된 모델들과의 정확도 비교결과(마젠타 색이 알파폴드의 결과)이다.

결론

CASP13 대회에서 알파폴드의 등장은 더 이상 템플릿이 필요 없는 단백질 3차구조 모델링의 가능성을 보여주었다고 할 수 있다. 그림 5는 평균적인 백본 정확도(GDT-TS)에서, 알파폴드(blue line, 1st place)의 결과가 기대 이상이었음을 보여주고 있는데, CASP13이전의 대회에서는 1번째와 2번째 그룹과의 차이가 거의 없었다. 점선 라인이 예측되었던 결과였는데, 알파폴드는 이것과의 차이를 분명하게 만들었다. 이는 알파폴드가 기존의 방법들과 비교하여 의미 있는 성능향상을 이루었으며, 단백질 폴딩 문제의 해결에 한 걸음 더 가까이 갔다는 것을 의미한다. 기술적인 면에서는 단백질의 다중서열정렬(MSA) 속에 있는 정보를 인공지능을 통해 잘 이끌어 낼 수 있다는 것을 의미한다. 하지만, 그래프가 보여주듯이, 템플릿 사용 방법들이 보여주는 ~85% 이상의 정확도 정도까지는 아직도 갈 길이 남아 있다는 것을 볼 수 있다. Distogram 예측의 핵심이 되는 좋은 품질의 DCA 분석정보를 얻기 위해서는 단백질 서열 데이터베이스가 더 확장되어야 하며, 단백질의 실험구조 정보가 더욱 축적 되어야 할 것이다. 현재의 단백질 서열정보 추출 기술과 구조결정 실험들의 지속적 발전을 생각할 때, 이번 알파폴드의 등장과 함께, 단백질 폴딩 문제의 해결이 상당히 희망적이라고 생각되어진다. 이와 함께 이러한 인공지능 방법들이 단백질-리간드 상호작용, 단백질-단백질 상호작용의 연구에도 이미 시도[26-28]되고 있는데, 이는 신약개발 과정에 있어 필수적이며, 의미 있는 기술적 성과를 만들 가능성을 엿보게 한다. 앞으로 여러 단백질 공학과 신약개발 등의 분야에서 인공지능 방법의 활발한 활용과 그 성과를 기대한다.

참고문헌

1. Baker, D. and et al. Protein structure prediction and structural genomics. *Science*, 2001;294 (5540): 93–96.
2. https://en.wikipedia.org/wiki/Protein_structure_prediction
3. Moult, J. and et al. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 1995;23 (3)
4. Protein Structure Prediction Center: <http://predictioncenter.org>
5. Fiser, A. Template-based protein structure modeling. *Methods Mol Biol*. 2010;673:73–94.
6. Eisenhaber F, and et al. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.* 1995;30:1–94.
7. <https://deeplearningblog.com/blog/alphafold/>
8. He, Kaiming and et al. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770–778.
9. http://predictioncenter.org/casp13/doc/CASP13_Abstracts.pdf
10. [https://www.theguardian.com/technology/2015/feb/25/google-develops-computer-program-capable-of-learning-tasks-](https://www.theguardian.com/technology/2015/feb/25/google-develops-computer-program-capable-of-learning-tasks)

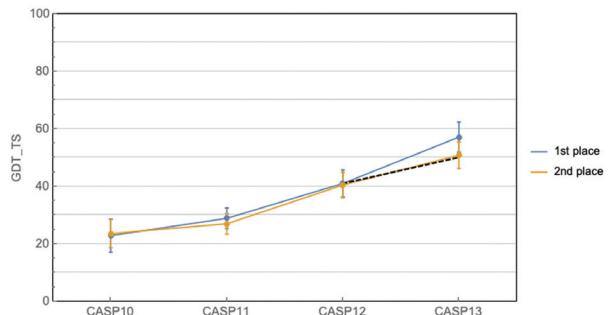


그림 5. 알파폴드 결과 (blue line, 1st place)와 2번째 방법 (orange line, 2nd place)과의 차이 비교 [CASP13].

independently

11. AlphaFold: Using AI for Scientific Discovery: <https://deepmind.com/blog/alphafold/>
12. Altschuh D, and et al. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol.* 1987;193:693–707
13. Seemayer S, and et al. CCMpred: fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics.* 2014;30(21):3128–30.
14. Jones, D.T., et al. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics.* 2012;28:184–190
15. Adhikari B, and et al. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics.* 2017;34(9):1466–1472.
16. Wang S, and et al. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology* 2017;13(1): e1005324.
17. Helen M., and et al. The Protein Data Bank, *Nucleic Acids Research*, 2000;28(1):235–242
18. Dawson, NL and et al. CATH: an expanded resource to predict protein function through structure and sequence". *Nucleic Acids Research* 2017;45: D289 –D295
19. Remmert M and et al. (2011). "HHblits: Lightning-fast iterative protein sequence searching by HMM–HMM alignment". *Nat. Methods.* 9 (2): 173 – 175.
20. Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David (1990). "Basic local alignment search tool". *Journal of Molecular Biology.* 215 (3): 403 – 410.
21. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. UniClust databases of clustered and deeply annotated protein sequences and alignments, *Nucleic Acids Res.* 2016;45(D1):D170–D176.
22. Ekeberg M, and et al. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, *Phys. Rev. E* 2013;87,012707
23. https://en.wikipedia.org/wiki/Direct_coupling_analysis
24. Alford RF, and et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput.* 2017;13(6):3031–3048
25. Gregor, Karol, and et al. Draw: A recurrent neural network for image generation, *arXiv:1502.04623*, 2015.
26. Izhar Wallach, and et al. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery, *arXiv:1510.02855*
27. Joseph Gomes, Atomic Convolutional Networks for Predicting Protein–Ligand Binding Affinity *arXiv:1703.10603*
28. José Jiménez, and et al. KDEEP: Protein –Ligand Absolute Binding Affinity Prediction via 3D–Convolutional Neural Networks. *Journal of Chemical Information and Modeling* 2018;58 (2), 287–296