ZeroFold-Understanding Mutations of SARS-CoV-2 Spike Protein base on Secondary Structure Event Extracting for guiding Vaccine development

[Weibin Chen^{1*}, Zhiyuan Xu^{2*}]

[¹Tsinghua University]

[²University of Southern California]

[cwb96@tsinghua.org.cn, zhiyuanx@usc.edu]

1. Executive Summary

Genes define, Proteins present. This paper is dangerous, please use it carefully, and for everyone. We contributing in four parts: found a problem in protein secondary structure prediction; found some works suggesting maybe there is another latent mechanism inside protein structure; propose a balanced Protein Event extraction method to understand protein secondary structure, test SARS-CoV-2 virus Spike-Protein in this framework. The experiments conduct an over expectation result: event in secondary structure could help developing high quality vaccines to deal with virus mutations.

Table of Contents

1. Executive Summary	2
2. Introduction.....	4
3. Related works.....	5
4. Problem analysis	8
5. Experiments	13
6. Discussion	16
7. References.....	17

2. Introduction

Genes define; proteins present.

Protein is basic unit of live thing(Andrew Senior et al., 2019), depends on its unique 3D structure, protein make it functional, interacting with other protein or ion/molecule. For example, antibody proteins utilized by our immune systems are 'Y-shaped', and form unique hooks. By latching on to viruses and bacteria, these antibody proteins are able to detect and tag disease-causing microorganisms for elimination. Collagen proteins are shaped like cords, which transmit tension between cartilage, ligaments, bones, and skin(Andrew Senior et al., 2019).

Proteins are linear chains of amino acids that adopt unique three- dimensional structures ('native states'), which allow them to carry out intricate biological functions. The protein structure has defined in four levels; all of the information needed to specify a protein's three-dimensional structure is contained within its amino-acid sequence. Given suitable conditions, most small proteins will spontaneously fold to their native states (Anfinsen, 1973).

Over the past five decades, researchers have been able to determine shapes of proteins in labs using experimental techniques like cryo-electron microscopy, nuclear magnetic resonance and X-ray crystallography, but each method depends on a lot of trial and error, which can take years of work, and cost tens or hundreds of thousands of dollars per protein structure. This is why biologists are turning to AI methods as an alternative to this long and laborious process for difficult proteins. The ability to predict a protein's shape computationally from its genetic code alone – rather than determining it through costly experimentation – could help accelerate research.

In this work, we present ZeroFold, a simple but more interpretable baseline for protein secondary structure prediction quest. We design a more balanced paradigm to create both semantically and visualization examples that described the SARS-CoV-2 virus protein structure.

Our main contributions are summarized as follows:

We found a problem in a protein secondary structure prediction quest.

We found some works suggesting maybe there is another latent mechanism inside protein structure.

We propose a balanced Protein Event extraction method to understand protein secondary structure

We test SARS-CoV-2 S-Protein in this framework.

3. Related works

The most related research areas are protein structure prediction, protein structure understanding and event extraction. And the protein structure prediction area can be divided into two small sections: protein secondary structure prediction and protein 3D structure prediction.

Protein secondary structure prediction began in 1951, since then the protein structure prediction problem has been extensively studied during past six decades, a lots of powerful new methods have been brought into this field. Secondary structure prediction techniques have been classified into four generations (Y. Yang et al., 2018).

First generation is more like statistical generation; the secondary structures were predicted according to statistical propensities of amino acid residues towards a specific secondary structure element(Chou & Fasman, 1974).

The second-generation used a sliding window of neighbouring residues and various theoretical algorithms, to make possible as more protein structures became available to estimate pairwise, triplet or longer-segment frequencies(Golmard J. L., 1988).

The third generation of techniques is characterized by using evolutionary information. During this period, new computational algorithms have been implemented. Examples are support vector machines(Cheng et al., 2005), Bayesian(Yao et al., 2008) or hidden semi-Markov network(Aydin et al., 2006) and conditional random fields for combined prediction(Y. Liu et al., 2004).

After the third generation, some recent successes of deep neural networks like convolutional and recurrent neural networks joined this field(Li & Yu, 2016)(Armenteros et al., 2017)(Long & Tian, 2019), then in a recent research, ResNet, Transformer and Pretrain protein secondary structures model also been introduced in this field(Rao et al., 2019).

Besides secondary structures, there are a lot works in protein 3D structure prediction, such as Rosetta, an all-atom energy function(Alford et al., 2017). Based on Rosetta, a deep residual network for predicting interresidue orientations has been studied(J. Yang et al., 2020), and AlphaFold, a protein structure prediction system that was entered in CASP13(Andrew W. Senior et al., 2019)(Andrew W. Senior et al., 2020), then an advanced method combining machine-learning based models from AlphaFold with state-of-the-art physics-based refinement via molecular dynamics simulations further improves predictions to outperform any other prediction method tested during the latest round of CASP(Heo & Feig, 2020).

Most works in protein structure prediction area focus on increasing accuracy or Paired t-test value which is used when researchers are interested in the difference between two variables for the same subject (Y. Yang et al., 2018), but ideal helices and sheets do not exist, and there are no clear boundaries between helix and coil nor sheet and coil states. It was shown that structural homologies differ by about 12% in secondary structure assignment (Burkhard Rost et al., 1994) for those with >30% sequence identity(Zhang et al., 2008). This assignment inconsistency would limit the highest possible accuracy to about 88–90%(B. Rost, 2001).

Structures are not known for the vast majority of protein sequences, but structure is critical for understanding function. Existing approaches for detecting structural similarity between proteins from sequence are unable to recognize and exploit structural patterns when sequences have diverged too far, limiting our ability to transfer knowledge between structurally related proteins(Bepler & Berger, 2019). Therefore, in the protein structure understanding area, a few works about semantics, modularity, unit and amino acid neural network had been discussed(Yan et al., 2014)(Hleap & Blouin, 2016)(Berezovsky, 2017).

Other work, a new end-to-end method has been studied, as end-to-end differentiable models of proteins represent a new paradigm for learning and modeling protein structure, with potential applications in docking, molecular dynamics, and protein design(AIQuraishi, 2018). Most of those research methods focus on the data structure of protein sequences, how to balance biochemical and bioinformatics is more challenging.

Inspired of all those solid research works above, we believe that a novel balanced method to increase interpretability about how the protein sequence change to secondary structures, such as helix and fold, coil, will help bio-community to understand protein better, event extraction of protein structure maybe is a more flexible way to achieve balanced goal.

As an important task of information retrieval in the natural language processing (NLP), event extraction has lots of applications in diverse domains(Xiang & Wang, 2019), in the related domain, Biomedical event extraction is a fundamental and in-demand technology that has attracted substantial interest from many researchers(H. Liu et al., 2010)(L. Zhu & Zheng, 2020), but most works focus on event extraction from Bio text, such as papers and articles.

4. Problem analysis

Although many great research works have been studied about protein structure understanding and prediction, we still found a problem in a protein secondary structure prediction quest. After a period of time researching in related areas, we found some works suggesting maybe there is another latent mechanism inside protein structure. Take it one step further, we propose consider this problem within a novel Nature Protein Processing (NNP) area, which focus on protein understanding, prediction, generation problems, accordingly to the Nature Language Processing (NLP) area. Specifically, we propose a balanced Protein Event extraction method to fulfill protein secondary structure prediction task.

- **4.1 Problem found in a protein secondary structures dataset**

Because of all the great work by brilliant researchers, we can entry this protein predict area easily, now the protein secondary structure can be described in three or eight characters:

C: Loops and irregular elements (corresponding to the blank characters)

E: β -strand

H: α -helix

B: β -bridge

G: 3-helix

I: π -helix

T: Turn

S: Bend

For the purpose of secondary structure prediction, it is common to simplify the aforementioned eight states (Q8) into three (Q3) by merging (E, B) into E, (H, G, I) into H, and (C, S, T) into C(*Protein Secondary Structure*, 2018).

When we practice in a the above dataset about protein secondary structure, the bell is ringing, when we found out: peptide sequence ARTKQTARK can correspond to eight different secondary structures in three types(Q3):

['CCCCCCCC', 'CEECCCCC', 'CEEEEECC', 'CEEEEECC',
'CEECCCCC', 'CEECCCCC', 'CCCECCCC', 'CEEECCCC']

We found there are quarter peptide sequences can correspond to multiple secondary structure when the goal of prediction is accuracy (Found 55135 of a total of 95915 proteins to have multiple secondary structures. The main dataset lists peptide sequences and their corresponding secondary structures. It is a transformation from RSCB PDB into a tabular structure)(*Protein Secondary Structure Contradictions*, 2018).

There are at least two assumptions need be check out about the multiple cases:

1. Maybe a peptide sequences can be according to multiple secondary structure in Q3 types, but no in Q8 types? Turn out the peptide sequence ARTKQTARK (Q8) will according to 16 different secondary structures: ['CTTSCCCCC', 'CEECCCCC', 'EEEEEECC', 'CEEESCSCC', 'CEEEEE ECC', 'CCBSCCCCC', 'CCBSSCCCC', 'CEETTTSCC', 'CEESCCCC', 'CEETTTCCCC', 'CCBTTC', 'CCBSCCCCC', 'CEECCCCC', 'CCCBCCCC', 'CEEECCCC']
2. A protein sequence has many peptide sequences, maybe a peptide sequences only can be according to multiple secondary structure in the same protein? Turn out the sequence ARTKQTARK in different protein sequences are according to 27 different secondary structures.

Those problems surely arise questions, after comparing above task with a NLP translation task, we interestingly found out that in the seq2seq algorithm, the score standard is NLP BLEU instead of accuracy or F1 score (which is combined with recall and accuracy) in the protein secondary structure prediction task.

After we study more deeply in the protein structure area, we found out the accuracy as score standard coming from a Noble prize research in 1973, which built up a basic rule in protein structure - 'One protein sequence has only One structure when the environment condition is equal(Anfinsen, 1973)'. When we check the basic rule discover by Anfinsen which lead to his Nobel prize in 1972, we have to wonder is there any reverse causality in the "thermodynamic hypothesis", that the structure of a native protein in its normal physiological milieu(solvent, pH, ionic strength, presence of other Components such as metal icons or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest(Anfinsen, 1973); In the area, we like to propose a new question that maybe the Gibbs free energy of a protein structure happens to be minimized because the protein in a crystal form when the structure was observed using x-ray crystallography, before new techniques were invented, such as, nuclear magnetic resonance, and cryo-electron microscopy.

We found some strong supporting research about the new hypothesis, there is an interesting thing about protein sequence and structure - homologous protein, which show us how the protein have the same functional in different sequences, because of evolution in the same high-altitude, such as a research in convergent evolution of high-oxygen affinity in hemoglobin from high-altitude dwelling bird species (including some hummingbirds) occurred through different amino acid substitutions(Schuyl, 2016)

Also a study about a surprising simplicity to protein folding showing that the folding kinetics of larger proteins may in some cases be dominated by escape from low-free-energy non-native conformations (Baker, 2000).

All above things suggest that maybe there is a tiny possibility that there can be another different new hypothesis can using to explain all the problems.

4.2 related possible hypothesis researches

Proteomics, was defined in 2000, started a new area about large-scale analysis of proteins, has contributed greatly to understanding into three main areas: (1) protein micro-characterization for large-scale identification of proteins and their post-translational modifications; (2) 'differential display' proteomics for comparison of protein levels with potential application in a wide range of diseases; and (3) studies of protein–protein interactions(Pandey & Mann, 2000), in this area, a new thesis, that protein structures are guided from according gene, is used.

Within progress in the proteomics area, lots of conserved domains were found, which mean those parts of protein sequence does not change during evolution. Protein domains were initially described as stable or autonomously folding units of protein structure, inspired by first results from the experimental characterization of protein three-dimensional (3D) structure. This definition of protein domains tends to coincide remarkably often with what has emerged from systematic analyses of sequence data—the characterization of protein domains as units of molecular evolution, about 76% of protein sequences in NCBI's Entrez query and retrieval system (excluding sequences from metagenomes, which have been obtained via environmental sampling and are not included in the pre-computed annotation set) can be matched to one or more conserved domain models, inferring the function of more than 39 million publicly available proteins. When looking at only the subset of structure-linked protein sequences, which are derived from the processing of 3D structure (currently 213507 entries), the fraction goes up to 94%, and is close to 98% when only considering structure-linked proteins with 50 amino acid residues or more (currently 199602 entries), particularly those that have related sequences, as identified by protein BLAST, with a wide phylogenetic spread(Marchler-Bauer et al., 2013).

Besides Proteomics area, in the deep learning area, the reformulation of complex, human-designed, multi-stage pipelines with differentiable models that can be jointly optimized end-to-end become a key ingredient of the success of deep learning(AIQuraishi, 2018).

The lack of data tends to limit the outcomes of deep learning research, particularly when dealing with end-to-end learning stacks processing raw data. A example is researchers had to invent the phoneme recognition which parts of music to perform good result(Waibel et al., 1989). But when the data was enough, with powerful GPUs, it allows colleagues to unrestrictedly explore two different design paradigms for music auto-tagging(Pons et al., 2018), without the phoneme hypothesis.

A similar study was conducted about prediction of novel protein folds without the use of co-evolutionary information, and the prediction of known protein folds without the use of structural templates. On the first task an end-to-end model achieves state-of-the-art performance, even when compared to methods that rely on co-evolutionary data. On the second task the model is competitive with methods that use experimental protein structures as templates, achieving 3-7Å accuracy despite being template-free(AIQuraishi, 2018).

Of course, there is an interpretability problem about an end-to-end deep learning model, especially about protein structure understanding. We consider to visualize the end-to-end model, using visualization technique that gives insight into the function of intermediate feature layers and the operation of the prediction process(Zeiler & Fergus, 2013). First attempt is using Event Extraction to solve the interpretability problem. This enables us to improve protein understanding on some new diagrams. We can also perform an ablation study to discover the performance contribution from different model layers in the next step research.

4.3 Novel hypothesis about using Event Extraction to solve the interpretability problem about protein secondary structure

When we observe a protein second structure, comparing those visualization structures with event extraction visualization, amazing thing can be seen.

We illustrate biomedical events with Fig. 1 as example. The word “promote” is an event trigger of the event type “Positive Regulation”. This event has a “Theme argument” linked to the word “tumorigenesis”, which is an entity of “Carcinogenesis” type, and a “Cause” argument linked to “over-expression”. Notice that some events can be the argument for other events, i.e., a nested structure, such as “over-expression” serving as a “Gene Expression” event trigger as well as an argument of “Positive Regulation” event. Therefore, the event can be viewed as a directed graph in the text for which the node of graph is the event trigger or entity and the directed edges indicate the arguments(L. Zhu & Zheng, 2020).

Fig.1 A comparison between protein secondary structure and BioNLP sentence with visualized events.

According to Fig.1, we try to consider the helix and beta folds in protein secondary structure as event trigger, and the other parts of sequence as arguments.

Inspired by the conserved domain, the end-to-end deep learning method, and the visualization of end-to-end deep learning model, we propose a novel thesis about the event relationship within protein sequences, and define some experiments to test.

5. Experiments

Task 1: prediction protein secondary structure using an event-extracting algorithm
We first conduct a prediction experiment, using a “cullpdb” filtered dataset³, which have 6133 pairs of protein sequences and secondary structure, using a ProteinNet⁴ to train the model. To evaluate the accuracy of prediction, using a “CB513” dataset(Cuff & Barton, 2009), the results on CB513, Q8 accuracy is 69.7%, Q3 accuracy is 84% for single model. The result is interesting but no impressing.

Task 2: understanding mutations of SARS-CoV-2 spike protein using event trigger and arguments structure

Apparently, there is no baseline about how to understand protein secondary structure using an event-extracting framework, we need to build it up by exploring the data. Currently, the SARS-CoV-2 pandemic is the most important problem all around the world. We try a few methods, then found out simply combining protein secondary structure and related mutations will explain a lot things, the result is very impressing. After carefully observation about SARS-CoV-2 spike protein mutations, we decide to change the Folding items in the protein secondary structure with Events. Therefore, there is Zero Folding item in the protein secondary event structure.

Step1: Using event extraction to understanding the Sequence and secondary structure for SARS-CoV-2 spike protein (6VSB) chain A 1-50, in table1.

position	seq	sec	mutation	numbers	position	seq	sec	mutation	numbers
1	M				28	Y	E	N/H	5
2	F		L	1	29	T	E	I	28
3	V				30	N	E	S	1
4	F				31	S	E		
5	L		F	100	32	F		I	1
6	V		I	1	33	T	S		
7	L				34	R			
8	L		V/W	17	35	G			
9	P		L/S	4	36	V			
10	L		I	6	37	Y	B		
11	V				38	Y			
12	S				39	P	S		
13	S		N/I	3	40	D	S		
14	Q		H	2	41	K	S		
15	C		F	1	42	V		L	1
16	V				43	F	B		
17	N				44	R			
18	L		F	30	45	S			
19	T		S	1	46	S	S	A/L	3
20	T		I	2	47	V	S		

21	R		K/I/T	20	48	L	E		
22	T		I	10	49	H	E	Y	31
23	Q		K/*	2	50	S	E	L	8
24	L		I	3	51	T	E		
25	P		L/S	5	52	Q			
26	P		L/S	10	53	D			
27	A		S/V	11	54	L	E	F	21

Table1 Mutations with Sequence and secondary structure for SARS-CoV-2 spike protein (6VSB) chain A 1-50

Step2: Delete some columns in table2, showing the relationship between Mutations and secondary structure.

sec	numbers	sec	numbers
		E	5
	1	E	28
		E	1
		E	
	100		1
	1	S	
	17		
	4		
	6	B	
		S	
	3	S	
	2	S	
	1		1
		B	
	30		
	1	S	3
	2	S	
	20	E	
	10	E	31
	2	E	8
	3	E	
	5		
	10		
	11	E	21

Table2 Mutations with only secondary structure for SARS-CoV-2 spike protein (6VSB) chain A 1-50,show mutations happen more when there are not secondary structures (with exception structure 'E', which mean β -strand)

In above table, we can understand mutations with event extraction: mutations happen more when there are not secondary structures; with exception structure 'E', which mean β -strand.

6. Discussion

Understanding Protein

Our Novel hypothesis is mainly about using Event Extraction to solve the interpretability problem about protein secondary structure. The results of Task2 in the experiments show some supporting evidences, such as: Blank position in secondary structure is much easier to mutate; Helix in secondary structure is not easy to mutate; β -strand in secondary structure is easy to mutate with some arguments.

We also understand that, the Event Extraction framework maybe one of many ways to solve the interpretability problem about protein secondary structure, there can be a lot of interesting researches happen about this area in the future, we looking for it.

But because of currently pandemic, we have to submit this paper as soon as possible, for everyone.

Guiding vaccine development

Some vaccines are focus on an optimised full-length spike gene based on an early version of SARS-CoV-2 virus (GenBank accession number YP_009724390)(F.-C. Zhu et al., 2020).

But the virus is changing, for example, there are 1013 mutation sites are identified in Spike protein, of which, 539 lead to amino acid changes in chain A of SARS-CoV-2 spike protein (6VSB), which sequence just has 1289 amino acids(Zhao WM, 2019).

Currently, major solution about virus mutation is developing vaccine which focus on several conserved domains, some maybe focus on very conceived-domain. But our understanding about the SARS-CoV-2 virus is limited.

We sincerely hope those event extraction information could help developing high quality vaccines to deal with virus mutations, because the vaccine against seasoning flu is only effective in less than one year.

Conclusion:

In this paper, we present an events extraction framework to understand protein secondary structure. This novel hypothesis integrates protein secondary structure and protein mutation, provide strong evidence to solve the interpretability problem about protein secondary structure. One can use the ZeroFold method help develop high quality vaccine against SARS-CoV-2 virus. In the future, we plan to extend our method on other virus, to develop more high quality vaccines protecting everyone. Also we will transform more NLP techniques into NPP (Natural Protein Process) area, we believe the proposed method has the potential to facilitate bioinformatics in broad, real-world scenarios for researchers.

7. References

- Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6), 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>
- AlQuraishi, M. (2018). *End-to-end differentiable learning of protein structure* [Preprint]. *Bioinformatics*. <https://doi.org/10.1101/265231>
- Andrew W. Senior, Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*, 87(12), 1141–1148. <https://doi.org/10.1002/prot.25834>
- Andrew W. Senior, Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710. <https://doi.org/10.1038/s41586-019-1923-7>
- Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>
- Armenteros, J., Sønderby, C., Sønderby, S., Nielsen, H., & Winther, O. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics (Oxford, England)*, 33. <https://doi.org/10.1093/bioinformatics/btx431>
- Aydin, Z., Altunbasak, Y., & Borodovsky, M. (2006). Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics*, 7(1), 178. <https://doi.org/10.1186/1471-2105-7-178>
- Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, 405(6782), 39–42. <https://doi.org/10.1038/35011000>
- Bepler, T., & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *ArXiv:1902.08661 [Cs, q-Bio, Stat]*. <http://arxiv.org/abs/1902.08661>
- Berezovsky, I. N. (2017). Basic units of protein structure, folding, and function. *Progress in Biophysics and Molecular Biology*, 15.
- Cheng, J., Randall, A., & Baldi, P. (2005). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4), 1125–1132. <https://doi.org/10.1002/prot.20810>
- Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, 13(2), 222–245. <https://doi.org/10.1021/bi00699a002>
- Cuff, J. A., & Barton, G. J. (2009). *Evaluation and improvement of multiple sequence methods for protein secondary structure prediction*. 12.
- Golmard J. L. (1988). A simple method for predicting the secondary structure of globular proteins: Implications and accuracy. *Bioinformatics*, 3, 3. <https://doi.org/10.1093/bioinformatics/4.3.357>

Heo, L., & Feig, M. (2020). High - accuracy protein structures by combining machine - learning with physics - based refinement. *Proteins: Structure, Function, and Bioinformatics*, 88(5), 637–642. <https://doi.org/10.1002/prot.25847>

Hleap, J. S., & Blouin, C. (2016). *The Semantics of the Modular Architecture of Protein Structures*. 10.

Li, Z., & Yu, Y. (2016). *Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks*. 9.

Liu, H., Blouin, C., & Kešelj, V. (2010). *Biological Event Extraction using Subgraph Matching*. 7.

Liu, Y., Carbonell, J., Klein-Seetharaman, J., & Gopalakrishnan, V. (2004). Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics*, 20(17), 3099–3107. <https://doi.org/10.1093/bioinformatics/bth370>

Long, S., & Tian, P. (2019). Protein secondary structure prediction with context convolutional neural network. *RSC Advances*, 9, 38391–38396. <https://doi.org/10.1039/C9RA05218F>

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., & Lanczycki, C. J. (2013). CDD: Conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, 41(D1), D348–D352. <https://doi.org/10.1093/nar/gks1243>

Pandey, A., & Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405(6788), 837–846. <https://doi.org/10.1038/35015709>

Pons, J., Nieto, O., Prockup, M., Schmidt, E., Ehmann, A., & Serra, X. (2018). End-to-end learning for music audio tagging at scale. *ArXiv:1711.02520 [Cs, Eess]*. <http://arxiv.org/abs/1711.02520>

Protein Secondary Structure. (2018). <https://kaggle.com/alfrandom/protein-secondary-structure>

Protein Secondary Structure Contradictions. (2018). <https://www.kaggle.com/alfrandom/protein-secondary-structure/discussion/67072>

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). *Evaluating Protein Transfer Learning with TAPE* [Preprint]. Synthetic Biology. <https://doi.org/10.1101/676825>

Rost, B. (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134(2–3), 204–218. <https://doi.org/10.1006/jsbi.2001.4336>

Rost, Burkhard, Sander, C., & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235(1), 13–26. [https://doi.org/10.1016/S0022-2836\(05\)80007-5](https://doi.org/10.1016/S0022-2836(05)80007-5)

Schuyt, M. (2016). *Predicting the basis of convergent evolution*. 2.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*, 87(12), 1141–1148. <https://doi.org/10.1002/prot.25834>

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3), 328–339. <https://doi.org/10.1109/29.21701>

Xiang, W., & Wang, B. (2019). A Survey of Event Extraction From Text. *IEEE Access*, 7, 173111–173137. <https://doi.org/10.1109/ACCESS.2019.2956831>

Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., & Shen, B. (2014). The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, 46. <https://doi.org/10.1007/s00726-014-1710-6>

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3), 1496–1503. <https://doi.org/10.1073/pnas.1914677117>

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., & Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: The final stretch? *Briefings in Bioinformatics*, 19(3), 482–494. <https://doi.org/10.1093/bib/bbw129>

Yao, X.-Q., Zhu, H., & She, Z.-S. (2008). A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinformatics*, 9(1), 49. <https://doi.org/10.1186/1471-2105-9-49>

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *ArXiv:1311.2901 [Cs]*. <http://arxiv.org/abs/1311.2901>

Zhang, W., Dunker, A. K., & Zhou, Y. (2008). Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins*, 71(1), 61–67. <https://doi.org/10.1002/prot.21654>

Zhao WM, et al. (2019). *Mutation sites in the Spike (S) protein—2019 Novel Coronavirus Resource*. <https://bigd.big.ac.cn/ncov/protein?lang=en>

Zhu, F.-C., Li, Y.-H., Guan, X.-H., Hou, L.-H., Wang, W.-J., Li, J.-X., Wu, S.-P., Wang, B.-S., Wang, Z., Wang, L., Jia, S.-Y., Jiang, H.-D., Wang, L., Jiang, T., Hu, Y., Gou, J.-B., Xu, S.-B., Xu, J.-J., Wang, X.-W., ... Chen, W. (2020). Safety, tolerability, and immunogenicity of a recombinant adenovirus type-5 vectored COVID-19 vaccine: A dose-escalation, open-label, non-randomised, first-in-human trial. *The Lancet*, S0140673620312083. [https://doi.org/10.1016/S0140-6736\(20\)31208-3](https://doi.org/10.1016/S0140-6736(20)31208-3)

Zhu, L., & Zheng, H. (2020). Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks. *BMC Bioinformatics*, 21(1), 47. <https://doi.org/10.1186/s12859-020-3376-2>