Final of Deep Learning

A New Technology for

Fundamental Bioinformatic Researchers:

Visualizing the Pathway of Central Dogma of Molecular Biology

Name: Xing Yicai
Collage of Life Science
Student ID: 2018012452

Name: Lu Dian
Department of Engineering Physics
Student ID: 2016011688

Mentor: Hu Xiaolin
Date: 2020/1/11

A New Technology for Fundamental Bioinformatic Researchers: Visualizing the Pathway of Central Dogma of Molecular Biology

Lu Dian | Xing Yicai | Tsinghua University | 2020 | 1 | 11

Abstract

Organisms on earth follow the central law of copying, passing on, and using their genetic material. Life science and medicine researchers are well aware of the central principle that sequence determines the structure of biomolecules, and that structure determines the function of biomolecules. As the main functional substance of an organism, the sequence of proteins is determined by DNA. Developing a protein with certain function can solve lots of problem. Our team have developed a useful tool for researchers in the field of biology to generate DNA sequence of certain kind of protein.

Background

Central Dogma of Molecular Biology

The central dogma of molecular biology is an explanation of the flow of genetic information within a biological system which often stated as "DNA makes RNA and RNA makes protein,"

There are three main types of biopolymers: DNA and RNA, and proteins. 3*3=9 ways can be imagined the direct information transfer that might occur between these. The theory divides them into three groups of three: three general metastases (thought to occur normally in most cells), three special metastases (known to occur, but only under certain viral or laboratory conditions), and three unknown metastases (thought never to occur).

General transfer describes the normal flow of biological information: DNA can be copied as DNA (DNA replication), DNA information can be copied as mRNA(transcription), and proteins can use the information in mRNA as a template for synthesis (translation).

General	Special	Unknown
DNA → DNA	RNA → DNA	protein → DNA
DNA → RNA	RNA → RNA	protein → RNA
RNA → protein	DNA → protein	protein → protein

Table 1. Table of the three classes of information transfer suggested by the dogma

The progeny of any cell that supplies the genetic material, be it a somatic or a germ cell, must carry out DNA replication, and copying from one DNA to another is a fundamental step in the central dogma.[1]

Transcription is the process by which information contained in a piece of DNA is copied in the form of messenger RNA (mRNA).

Mature mRNA enters the ribosome and is translated. The processes of transcription and translation in prokaryotic cells may be linked without significant separation. The ribosome reads mRNA triplet codons, usually starting with the AUG (adenine-uracil-guanine) or methionine codons downstream of the ribosomal binding site.

Each tRNA carries the appropriate amino acid residues that are added to the polypeptide chain being synthesized. As the amino acid chain enters the growth peptide chain, the peptide chain begins to fold into the correct conformation. The translation ends with a termination codon, which can be a UAA, UGA, or UAG triplet

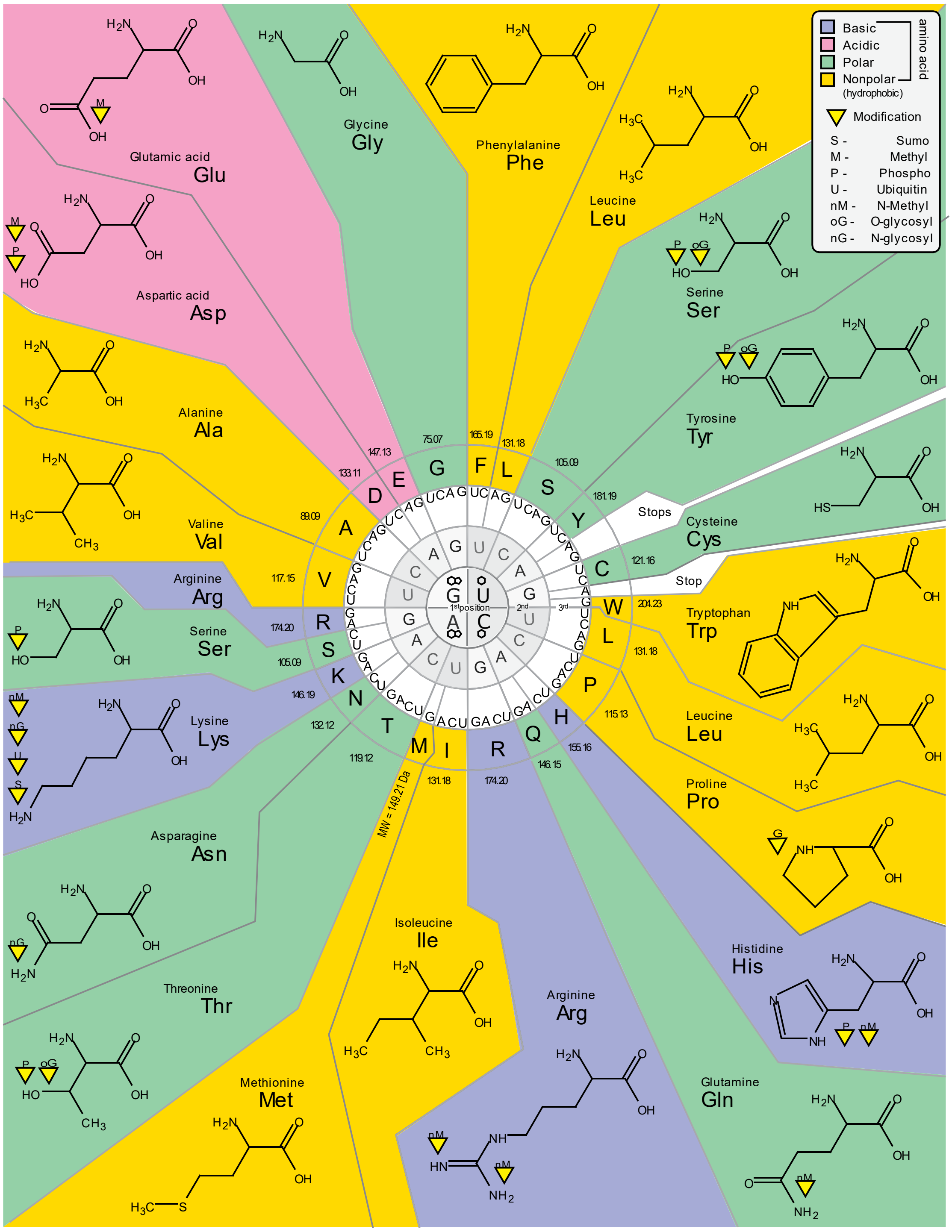


Figure 1. The genetic code is the set of rules used by living cells to translate information encoded within genetic material (DNA or mRNA sequences of nucleotide triplets, or codons) into proteins

After the discovery of the structure of DNA in 1953, efforts began to understand how proteins are encoded. George Gamow hypothesized that three groups of bases must be used to encode the 20 standard amino acids that living cells use to make proteins, allowing a maximum of 4*4*4= 64 amino acids. In 1961, Marshall Nirenberg and Heinrich j. Matthaei first revealed the nature of the codon.

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*)[B]	UGA	Stop (*Opal*)[B]	A
	UUG[A]		UCG		UAG	Stop (*Amber*)[B]	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG[A]		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Table 2. A The codon AUG both codes for methionine and serves as an initiation site: the first AUG in an mRNA's coding region is where translation into protein begins. The other start codons listed by GenBank are rare in eukaryotes and generally codes for Met/fMet

B The historical basis for designating the stop codons as amber, ochre and opal is described in an autobiography by Sydney Brenner and in a historical article by Bob Edgar.[2]

As can be seen from table 2 and figure 1, in fact, there are certain obstacles in the process from protein to DNA. It is not the amino acid sequence of a protein that corresponds to a triplet codon, but more sequence possibilities.

For fewer tryptophan, methionine corresponds to only one possibility, while for leucine, the highest corresponds to as many as six possibilities. This makes it mathematically difficult to convert proteins into DNA precisely.

In addition, we may wonder whether the synthesis of proteins can skip the step of RNA synthesis and directly realize the transformation from DNA to protein. Studies have demonstrated direct translation from DNA to protein using e. coli extracts containing ribosomes rather than whole cells in cell-free systems (that is, in test tubes). These cell fragments can synthesize proteins from the single-stranded DNA templates of other organisms. Neomycin may enhance this effect. However, it is not clear whether this translation mechanism specifically corresponds to the genetic code. [3,4]

A popular approach to molecular biology

At present, 80% of the research is based on the research methods of molecular genetics, including forward genetics .

Forward genetics is a molecular genetic method to determine the genetic basis of a phenotype. This is initially achieved by inducing mutations using naturally occurring mutations or radiation, chemical, or insertional mutations, such as transposable factors. Breeding was then carried out to isolate the mutated individuals and then the gene was mapped. Positive genetics can be thought of as an opposition to reverse genetics, which determines gene function by analyzing the phenotypic effects of altered DNA sequences.[5]

Forward genetics is a molecular genetic method to determine the genetic basis of a phenotype. This is initially achieved by inducing mutations using naturally occurring mutations or radiation, chemical, or insertional mutations, such as transposable factors. Breeding was then carried out to isolate the mutated individuals and then the gene was mapped. Positive genetics can be thought of as an opposition to reverse genetics, which determines gene function by analyzing the phenotypic effects of altered DNA sequences

Synthetic DNA technology

Synthetic gene synthesis or gene synthesis, sometimes called DNA printing [6], is a synthetic biology method used to create artificial genes in the laboratory. Based on solid-phase DNA synthesis, it differs from molecular cloning and polymerase chain reaction (PCR) because it does not have to start with a preexisting DNA sequence. Thus, it is possible to fully synthesize double-stranded DNA molecules, with no apparent restriction on nucleotide sequence or size.

The synthesis of the first complete gene, the yeast tRNA gene, was confirmed in 1972 by Har Gobind Khorana and his colleagues. The first peptide and protein-coding genes were relatively early. Now the technology for synthesizing DNA is maturing.

The most common method of DNA synthesis is oligonucleotide synthesis. Oligonucleotide synthesis refers to the chemical synthesis of relatively short fragments of nucleic acid with a certain chemical structure (sequence). This technique is useful in current laboratory practice because it provides a fast and inexpensive way to obtain customized oligonucleotides of the desired sequence. Enzymes can only make DNA and RNA in the 5' -3 'direction, whereas chemical oligonucleotides have no such restriction, although they are usually made in the opposite 3' -5' direction. Currently, the method is solid-phase synthesis using phosphoamides method and phosphoamides building blocks derived from protected 2 '-deoxynucleosides (dA, dC, dG and T), ribonucleosides (A, C, G and U) or chemically modified nucleosides (such as LNA or BNA).

So far, synthetic DNA has cost $0.50 a base, or less than $0.20 a base if the sequence is short.

Synthetic Protein technology

Although synthetic DNA has been programmed, synthetic proteins remain elusive. 2017 was the first year that a protein anticancer drug, an analogue of the cytokine il-2, was synthesized from scratch. This protein design strategy can be widely used for signaling proteins, which will lead to more and better drug candidates. Neo-2/15 has only

14% of the amino acid sequence of human il-2, but retains most of the functions of il-2, except for the ability to bind CD25, which can cause toxic side effects. In mouse models of colon cancer and melanoma, neo-2/15 strongly inhibited tumor growth and even eliminated tumors entirely in some mice, treating them better than natural il-2. The paper appears in Nature. [7]

From here we can see that if you want to build a protein sequence, you have to start with its DNA.

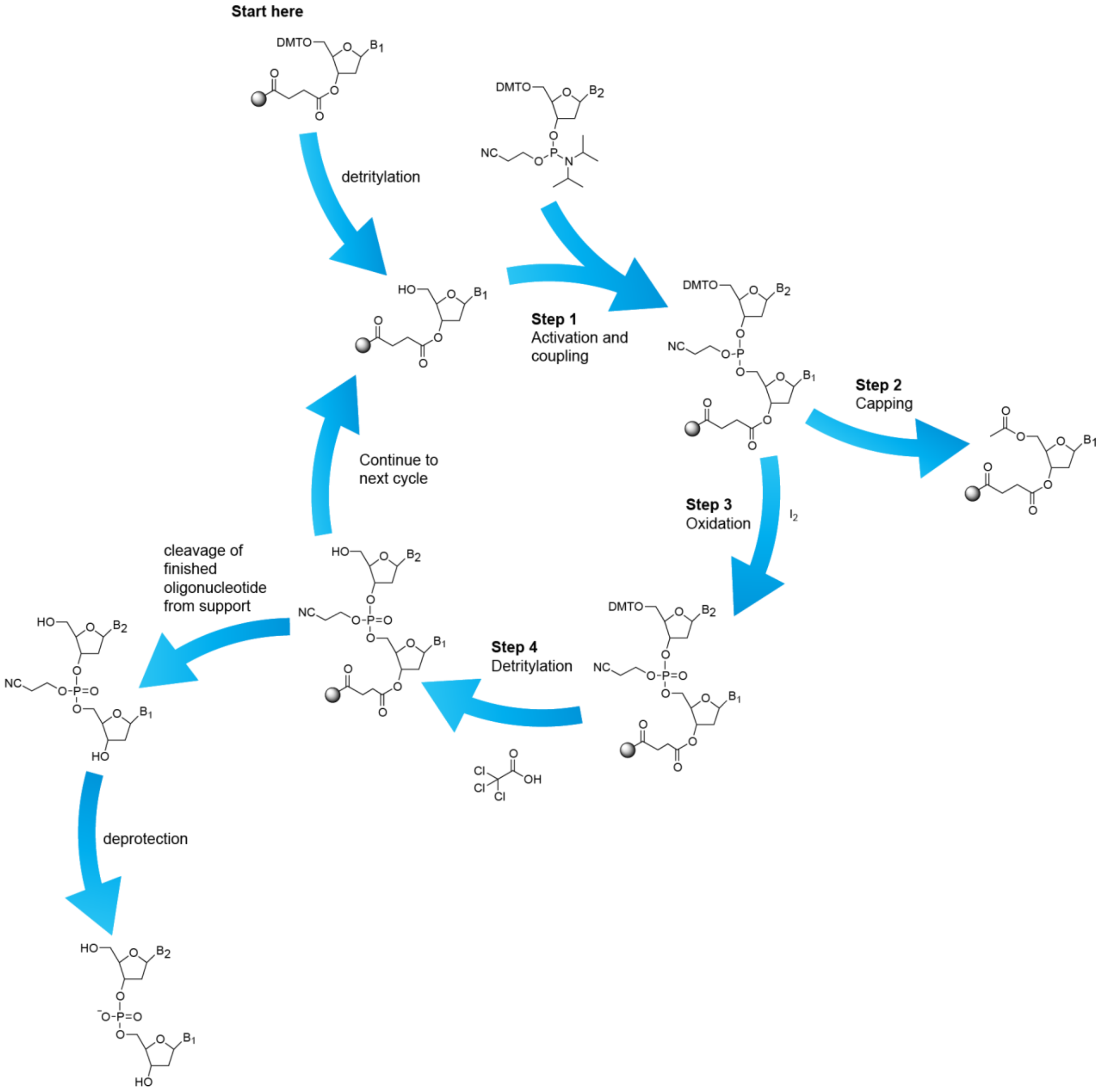


Figure 2. A chemical reaction that occurs during DNA synthesis.

Databank

The database we used this time is UniProt, a freely accessible database of protein sequences and functional information, with many entries from genome sequencing projects. It contains a wealth of information about the biological functions of proteins obtained from the research literature.

Swiss-prot in UniProt is a manually annotated non-redundant protein sequence database in which proteins have biological functions in the organism. The purpose of UniProtKB/ swiss-prot is to provide all known relevant information about a particular protein. Notes are regularly reviewed to keep up with current scientific findings. A manual annotation of an entry includes a detailed analysis of protein sequences and scientific literature.

Sequences from the same genes and species are combined into the same database entries. Identify differences between sequences and record their causes (such as variable splicing, natural variation, incorrect start sites, incorrect exon boundaries, framing, unknown conflicts). A series of sequence analysis tools were used to annotate the UniProtKB/ swiss-prot entries. Computer predictions are evaluated manually and the relevant results are selected as entries. These predictions include post-translational modifications, transmembrane domains and topologies, signal peptides, domain recognition, and protein family classification.[8]

Method

The methods of this study mainly include the following aspects. First, we use the name characteristics of proteins to roughly distinguish proteins into 16 categories, which represent some functional commonality and may also imply some sequence characteristics. And then we try to create a whole new sequence that doesn't exist in the course of biological evolution by using the properties of one of those sequences.

By converting this sequence into DNA, researchers in the life sciences and medicine can use now sophisticated genetic recombination techniques to express similar proteins in living organisms, allowing them to make entirely new proteins.

It is worth considering that the proteins expressed can only be verified by biological experiments, but such biological verification takes an extremely long time. So I want to try to predict the commonality of this protein through some structural biology.

Algorithm

Algorithm 1: Generator & Classifier

In this method, we trained separately a generator and a classifier. The task of a classifier is to determine the label of the given data, that is, to classify the type of the protein out of the sequence of the amino acids. And the task of the generator is to generate a sequence of amino acids given the target type of protein, and that would be sent into the classifier and make the classifier’s output closer to the label offered by the generator. We train the classifier first on the labelled dataset, and then train the generator by fixing the parameters of the classifier and optimize the parameters of the generator applying RMS propagation method. The method is somehow similar to a CGAN method, but we have plenty of labelled real data (the class with least quantity is 166), so we did not use an adversarial model at first. The structure of the method is shown below, (see figure 3) model:

```
generator(
(emb): Embedding(16, 128)
(lstm1): LSTM(128, 23, num_layers=4,
batch_first=True, dropout=1e-06,
bidirectional=True)
)
classifier(
(emb): Embedding(24, 20, padding_idx=0)
(lstm1): LSTM(20, 10, batch_first=True,
dropout=2e-06, bidirectional=True)
(conv1): Conv2d(2, 16, kernel_size=(5, 1),
stride=(1, 1), padding=(2, 0))
(bn1): BatchNorm2d(16, eps=1e-05,
momentum=0.1, affine=True,
track_running_stats=True)
(conv2): Conv2d(16, 32, kernel_size=(3, 1),
stride=(1, 1), padding=(1, 0))
(conv3): Conv2d(32, 1, kernel_size=(3, 1),
stride=(1, 1), padding=(1, 0))
(lstm2): LSTM(10, 10, batch_first=True,
dropout=1e-06, bidirectional=True)
(fc1): Linear(in_features=10,
out_features=16, bias=True)
```

Result of algorithm1:
The best accuracy of the classifier is about 0.08 after 20 Epochs, almost same as a random guess result. We have tried several models in classifier. As the length of sequence of amino acids is usually very long and highly varied (randomly varied from about 200 to 5000), an LSTM net is used. We have tried adding convolutional layers and batchnorm layers, but no difference except that the training time is much longer. As the first step of the method has met great difficulty, the following training of the generator is therefore meaningless. The training curve is shown below. (see figure 4)

The fact is that we underestimated the harshness of the data. Although different types of proteins have their identifiable features in their amino acid sequences, the features are quite elusive and is hard to be captured by the current model.

Algorithm 2: GAN
To simplify the task, our next approach is to focus on a certain type of protein, and to generate amino acid sequence using GAN.

The another reason we turned to GAN aside from that we want to simplify the task is that, the goal of the algorithm above is to generate a sequence that has the certain feature of the type of protein, that is, it would not be recognized as other types of protein, but it does not count for that the generated sequence resembles REAL protein.

The structure of our model is a traditional GAN model, consists of a generator and a discriminator. First we choose a certain type of protein, zinc finger for example. Then train the discriminator to discriminate the real zinc finger sequence and the generated sequence by the generator. At the same time, we train the generator to make the discriminator harder to tell the generated sequence out of the real ones. The optimal result is that the generator could generate sequence so similar to the real ones, that the fully trained discriminator could hardly tell the difference.

In this model, the discriminator has to deal with sequences of different length, so an LSTM model is used. But the length of the generated sequence is fixed, so we have the freedom to use CNN model. To monitor the training process, we defined a fake score and a real score. The output of the discriminator is a value between 0 and 1, 0 means the input sequence is a fake sequence generated by generator, while 1 means it’s a real sequence. The fake score is defined as the mean output when the input of the discriminator is a fake sequence, and the real score is defined as the mean output when the input is a real protein sequence.

```
model1 (RNN generator):
generator(
(emb): Embedding(16, 128)
(lstm1): LSTM(128, 23, num_layers=4,
batch_first=True, dropout=1e-06,
bidirectional=True)
)
discriminator(
(emb): Embedding(24, 20, padding_idx=0)
(lstm1): LSTM(20, 10, num_layers=4,
batch_first=True, dropout=2e-06,
bidirectional=True)
(fc1): Linear(in_features=10, out_features=1,
bias=True)
)
```

```
model2 (CNN generator):
generator(
(lin1): Linear(in_features=1,
out_features=10000, bias=True)
(conv1): Conv2d(1, 23, kernel_size=(3, 3),
stride=(1, 1), padding=(1, 1))
(bn1): BatchNorm2d(23, eps=1e-05,
momentum=0.1, affine=True,
track_running_stats=True)
(conv2): Conv2d(23, 23, kernel_size=(11, 1),
stride=(1, 1), padding=(5, 0))
(avgpool1): AvgPool2d(kernel_size=(1, 20),
stride=(1, 20), padding=0)
)
discriminator(
(emb): Embedding(24, 20, padding_idx=0)
(lstm1): LSTM(20, 10, num_layers=4,
batch_first=True, dropout=2e-06,
bidirectional=True)
(fc1): Linear(in_features=10, out_features=1,
bias=True)
)
```

Results of Algorithm 2:
We have tested both of the RNN model and the CNN model. In both of the cases, the discriminator converges much faster than the generator. But the results from the RNN model tends to be a repeating sequence like "FFFFFFFFF…", while the sequence generated by CNN model looks much more reasonable. The curve of fake score and real score is plotted as followed.(see figure 4)

Classification of protein

Classification	*Number of protein*
receptor	2401
kinase	2272
channel	802
zinc finger	978
transferase	1400
regulator	721
ligase	674
ATPase	302
GTPase	371
Synthase	322
Reductase	281
Dehydrogenase	257
Glycoprotein	263
Nuclease	247
Oxidase	194
Enzyme	166

Table 3. classification of protein to be analysised

In this experiment, the names that often appear in the ID of proteins and their frequency are shown in table 3. I'm going to show a little bit more about these proteins in order to understand the results.

Receptor

Receptors are chemical structures made up of proteins that receive and transmit signals that might be integrated into biological systems. These signals are usually chemical messengers that bind to receptors and cause some form of cellular/tissue response, such as changes in a cell's electrical activity. The role of the receptor can be divided into three main modes: signal translation, amplification, or integration, in which electrical transmission moves the signal forward, amplification increases the role of a single ligand, and integration allows the signal to be integrated into another biochemical pathway. In this sense, a receptor is a protein molecule that recognizes and responds to endogenous chemical signals.[9]

There are many types of receptors, and the sequence characteristics of different types of receptors are actually very different, so it is not rational to categorize these receptors directly.

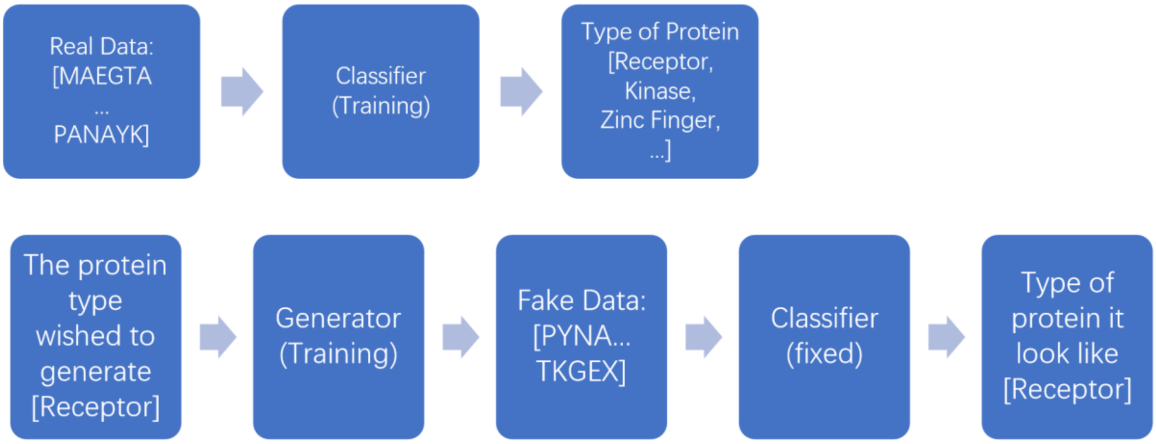
A common class of receptors are ligand gated ion channels, which are similar to proteins such as channels. These receptors are usually targets for rapid neurotransmitters such as acetylcholine and aminobutyric acid. Activation of these receptors results in changes in the movement of ions across the cell membrane. Each subunit consists of an extracellular ligand binding domain and a transmembrane domain consisting of four transmembrane helices. Ligand - binding cavities are located at the interface between subunits.

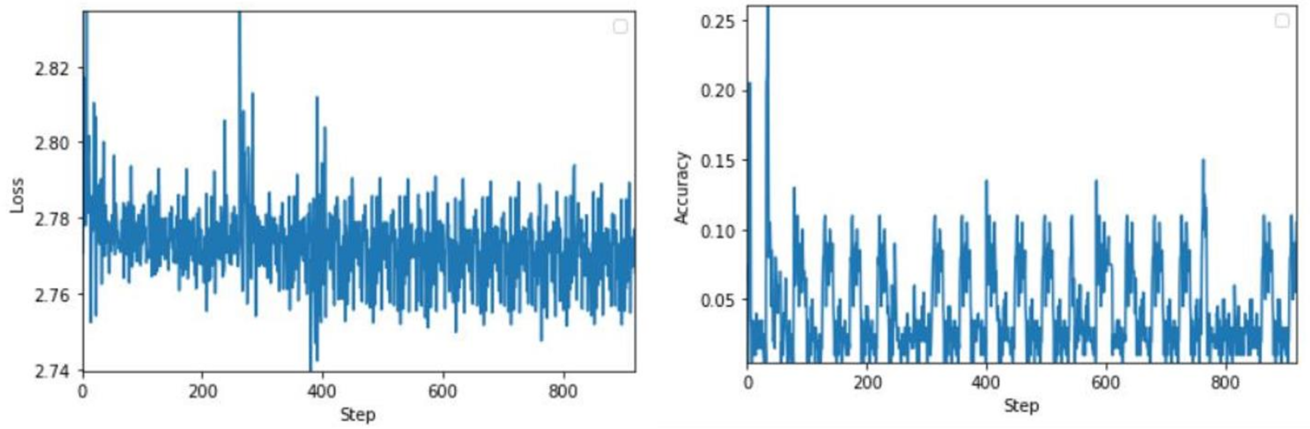
The second type of receptor is the well-known G protein-coupled receptor (denatured receptor). This is the largest family of receptors, including multiple receptors for hormones and slow transmitters such as dopamine and denatured glutamate. They consist of seven transmembrane helices. The rings connecting the helices form extracellular and intracellular domains. The binding sites of the larger peptide ligands are usually located in the extracellular region, while the binding sites of the smaller non-peptide ligands are usually located between seven helices and an extracellular ring. These receptors bind to different intracellular effector systems through G proteins. [10]

Enzyme-linked receptors (see "receptor tyrosine kinases" and "enzyme-linked receptors") are also common receptors that consist of an extracellular domain containing ligand binding sites and an intracellular domain, usually functioning as enzymes, connected by a single transmembrane helix. Insulin receptors are an example.

From this we can see that there are many types of receptors, even if both proteins belong to the receptor, there may not be a common sequence.

Zinc finger is a small protein structure motif characterized by the coordination of one or more zinc ions (Zn2+) to stabilize the fold. Initially, the term zinc finger was used only to describe the DNA binding motif found in the African clawed toad; However, it is now used to refer to any number of structurally related harmonized zinc ions. In general, zinc fingers coordinate zinc ions with combinations of cysteine and histidine residues. Initially, the number and order of these residues were used to distinguish between different types of zinc fingers (e.g. Cys2His2, Cys4, and Cys6). Recently, a more systematic method has been used to classify zinc finger proteins. The zinc finger proteins were divided into "folding groups" according to the overall shape of the protein skeleton in the folding region. The most common zinc finger "folding group" is





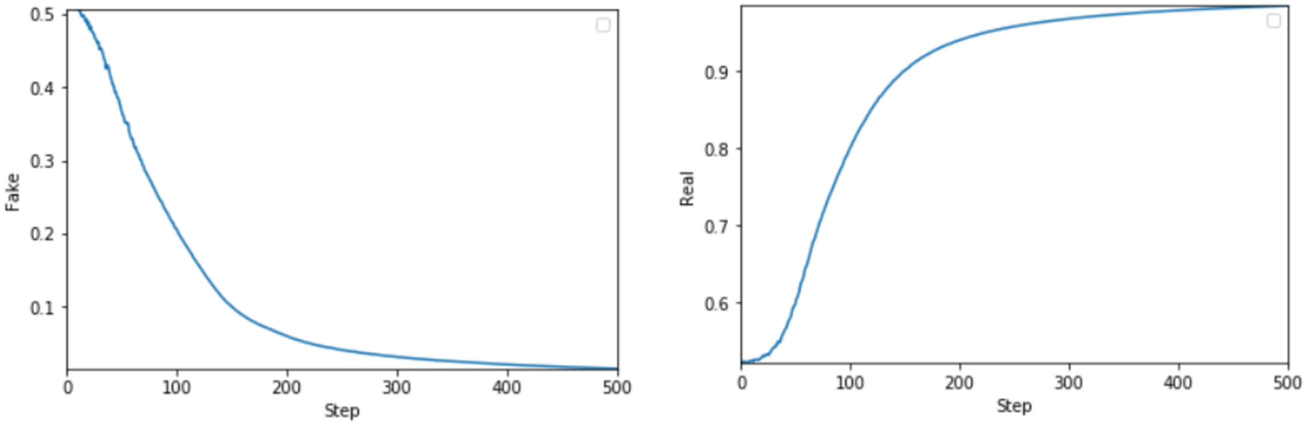


Figure 4: The structure of the G&C method. The method is composed of two separated training procedure. The upper figure shows the training procedure of the classifier. To train the classifier, we put in labelled real protein sequence and expect it will give the correct classification. The lower figure shows the training procedure of the generator. After the training process of the classifier has been done, we fix the parameters of the classifier and train the generator to generate sequences (according to the given label) that could be correctly classified by the classifier.

Figure 5. Training curve of the classifier. Notice the vertical axis. There is a trend of descending of the loss curve, but very slow. Training procedure quickly meets its limitation, indicating great difficulties in the method.

Figure 6. Curves of the fake score and the real score. The curve indicates that the discriminator converges much faster than the generator, it would be more difficult to training the generator.

similar to cys2his2 [11] treble clef, and zinc band. Their sequences are different, but they follow a pattern.

If you choose to use the term receptor for protein sequence generation, a wave of subdivision is needed to get a better result, but the zinc finger protein itself can make a basic division.

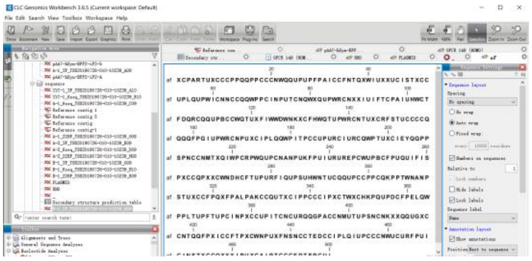
Protein Sequence Generation

By the time we write this report, the generator has not shown clues of generating sequences that is "confusing" enough, and the fake score remains low. One of the result of zinc finger sequence generated by the CNN model is:

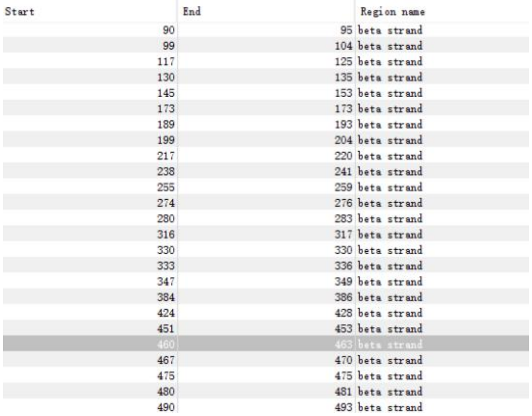
XCPARTUXCCCPPQQPPCCCNWQQUPUP
FPAICCFNTQXWIUXXUCISTXCCUPLQU
PWICNNCCQQWPPCINPUTCNQWXQUP
WRCNXXIUIFTCPAIUHWCTFDQRCQQU
PBCCWQTUXFIWWDWNKXCFHWQTUP
WRCNTUXCRFSTUCCCCQQQQFPGIUP
WRCNPUXCIPLQQWPITPCCUPURCIURC
QWPTUXCIEYQQPPSPNCCNMTXQIWPC
RPWQUPCNANPUKFPUIURUREPCWUPB
CFPUQUIFISPXCCQPXXCWNDHCFTUPU
RFIQUPSUHWNTUCQQUPCCPPCQKPPT
WNANPSTUXCCFPQXFPALPAKCCQUTX
CIPPCCCIPXCTWXCHKPQUPDCFPELQW
PPLTUPFTUPCINPXCCUPITCNCURQQGP
ACCNMUTUPSNCNKXXQQUGXCCNTQ
QFPXICCFTPXCWNPUXFNSNCCTEDCCI
PLQIUPCCCNWUCURFPUICINSTXCCQX
XXIPUXCAIPTCCCFDTEPCUI

However, there are still obvious problems. 'X', 'U', 'B' are rare amino acids, they occurs at most once in thousands, but in the generated sequence they looked like ordinary amino acids. And the first amino acid of a sequence should be 'M', but obviously the generator has not learned of the rule yet.

We simulated the structure of the protein.


CLC Genomics Workbench 3.6.5 (Current workspace: Default)
File Edit Search View Toolbox Workspace Help
Show Bookmark New Save Import Export Graphics Print
Workspace Plug-ins Search
Fit Width 100% Pan Selection Zoom In Zoom Out
Secondary structure prediction table
Alignments and Trees
General Sequence Analyses
Nucleotide Analyses
Sequence layout
Spacing
No wrap
Auto wrap
Fixed wrap
Numbers on sequences
Relative to
Hide labels
Lock labels
Sequence label
Annotation layout
Show annotations

```
af XCPARTUXCCCPPQQPPCCCNWQQUPUPFPAICCFNTQXWIUXXUCISTXCC
af UPLQUPWICNNCCQQWPPCINPUTCNQWXQUPWRCNXXIUIFTCPAIUHWCT
af FDQRCQQUPBCCWQTUXFIWWDWNKXCFHWQTUPWRCNTUXCRFSTUCCCCQ
af QQQFPGIUPWRCNPUXCIPLQQWPITPCCUPURCIURCQWPTUXCIEYQQPP
af SPNCCNMTXQIWPCRPWQUPCNANPUKFPUIURUREPCWUPBCFPUQUIFIS
af PXCCQPXXCWNDHCFTUPURFIQUPSUHWNTUCQQUPCCPPCQKPPTWNANP
af STUXCCFPQXFPALPAKCCQUTXCIPPCCCIPXCTWXCHKPQUPDCFPELQW
af PPLTUPFTUPCINPXCCUPITCNCURQQGPACCNMUTUPSNCNKXXQQUGXC
af CNTQQFPXICCFTPXCWNPUXFNSNCCTEDCCIPLQIUPCCCNWUCURFPUI
```



Start	End	Region name
90	95	beta strand
99	104	beta strand
117	125	beta strand
130	135	beta strand
145	153	beta strand
173	173	beta strand
189	193	beta strand
199	204	beta strand
217	220	beta strand
238	241	beta strand
255	259	beta strand
274	276	beta strand
280	283	beta strand
316	317	beta strand
330	330	beta strand
333	336	beta strand
347	349	beta strand
384	386	beta strand
424	428	beta strand
451	453	beta strand
460	463	beta strand
467	470	beta strand
475	475	beta strand
480	481	beta strand
490	493	beta strand

Figure 7. the result shows that most of the region are beta strand

Unlike many other well-defined superstructures, such as Greek keys or beta hairpins, zinc fingers come in many types, each with a unique three-dimensional structure. The specific zinc finger protein class is determined by this three-dimensional structure, but can also be identified based on the primary structure of the protein or the identity of a ligand that coordinates zinc ions. Interaction modules that bind DNA, RNA, proteins or other useful small molecules, and structural changes are mainly used to change the binding transformation of specific proteins.

Discussion

The classification part was not ideal in this attempt. We have tried many classifier algorithms, but the final result is best at 8% accuracy, which is close to random guessing. Perhaps it's because the classification of proteins is not detailed enough.But we have to note that if we continue to subdivide, in fact, there may be only a hundred proteins under each protein classification entry. This database is currently the largest protein database including all published literature. The main sources are model organisms such as fruit flies, mice, C. elegans, and plants and microorganisms.

With the continuous advancement of sequencing technology, more and more species of proteins will be revealed. Perhaps deep learning algorithms will be more suitable at this time. So far, if you want to do this classification, life information retrieval tools may still Maintain the highest accuracy rate, at least more than half.

If we want to create a protein that has never existed in the world, the best way is for us to understand the proteins of all animals, plants and microorganisms with such functions, so that we can understand the information more comprehensively. This database is still too little for what we want to do.

References

1. Yao NY, O'Donnell M (June 2010). "SnapShot: The replisome". Cell. 141 (6): 1088–1088.
2. Edgar B (2004). "The genome of bacteriophage T4: an archeological dig". Genetics. 168 (2): 575–82. PMC 1448817. PMID 15514035. see pages 580-581
3. McCarthy BJ, Holland JJ (September 1965). "Denatured DNA as a direct template for in vitro protein synthesis". Proceedings of the National Academy of Sciences of the United States of America. 54 (3): 880–6.
4. awa T, Yamagishi A, Oshima T (June 2002). "Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, Thermus thermophilus HB27 and Sulfolobus tokodaii strain 7". Journal of Biochemistry. 131 (6): 849–53.
5. Parsch J. "Forward and Reverse Genetics" Ludwig-maximilians-universitat Munchen. Archived from the original on 13 December 2014. Retrieved 31 October 2014.
6. Kimoto, M.; et al. (2013). "Generation of high-affinity DNA aptamers using an expanded genetic alphabet". Nat. Biotechnol. 31 (5): 453–457.
7. SILVA D-A, YU S, ULGE U Y, et al. De novo design of potent and selective mimics of IL-2 and IL-15[J]. Nature, 2019, 565(7738): 186-191.
8. Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; o'Donovan, C.; Redaschi, N.; Yeh, L. S. (2004). "UniProt: The Universal Protein knowledgebase". Nucleic Acids Research. 32 (90001): 115D–1119.
9. Hall, JE (2016). Guyton and Hall Textbook of Medical Physiology. Philadelphia, PA: Elsevier Saunders. pp. 930–937.
10. Congreve M, Marshall F (March 2010). "The impact of GPCR structures on pharmacology and structure-based drug design". British Journal of Pharmacology. 159 (5): 986–96.
11. Krishna SS, Majumdar I, Grishin NV (January 2003). "Structural classification of zinc fingers: survey and summary". Nucleic Acids Research. 31 (2): 532–50.