МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ПРЕДСКАЗАНИЕ ПРОСТРАНСТВЕННОЙ СТРУКТУРЫ БЕЛКОВ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Гаджиибрагимовой Дианы Адильевны

Научный руководитель
профессор, д.ф.-м.н. _____ В.А. Молчанов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2021

ВВЕДЕНИЕ

Белки - универсальные биополимеры, выполняющие весь спектр биологических функций: от структурной до каталитической. Во всем живом мире, именно белки играют максимум ролей и важность их изучения не ограничивается только фундаментальной наукой: сегодня и медицина и промышленность - потребители знаний о функциях и структуре белка.

Жизненно важно, чтобы белок присутствовал в организме в определенной форме, т.е. его конформация должна быть "правильной". Процесс сворачивания белка называется фолдингом(от англ. folding — сворачивание, укладка) и моделирование этого процесса причисляют к списку крупнейших неразрешенных научных проблем современности. [1]

Информация о том,в какую структуру свернется белок, заложена в самой последовательности аминокислот т.е. для того чтобы принять определенную структуру, белку требуется знать в какой последовательности и какие аминокислотные остатки в нем присутствуют. Проблема же заключается в том, что даже обладая вычислительной мощью и экспериментальными данными, человечество до сих пор не научилось строить модели, описывающие процесс фолдинга.

Взрывной рост геномных проектов привел к тому, что секвенируется все больше геномов, а соответствующие последовательности ДНК и РНК наполняют базы данных по экспоненте. Современные высокопроизводительные технологии секвенирования геномов делают задачу прочтения всей ДНК нового вида лишь вопросом времени. Полученные в ходе секвенирования последовательности аминокислот заносятся в базу UniProt, и на момент написания этой работы в

этой базе насчитывается 180.690.447 последовательностей, тогда как количество известных белковых структур в базе данных PDB(Protein Data Bank) всего 152.003, что составляет менее 0.1 от всех известных последовательностей. Такая огромная разница в значениях связана с относительной сложностью современных методов определения структур. В 2005 году авторитетный журнал Science признал проблему фолдинга белка одной из 125 крупнейших проблем современной науки. [2]

Целью данной работы является создание и обучение нейронной сети для предсказания вторичной структуры белка по перивичной.

Для достижения цели необходимо решить следующие задачи:

1. Построить модель нейронной сети.
2. Подготовить необходимые наборы данных.
3. Реализовать и обучить нейронную сеть.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Теоретическая часть

Белки – непериодические полимеры, мономерами которых являются аминокислоты. Обычно в качестве момномеров белков называют 20 видов аминокислот [3].

Полимеры – органические молекулы, в которых одни и те же звенья(мономеры) повторяются очень много раз. Таким образом, молекула белка представляет собой определенную последовательность аминокислот(рисунок 1).

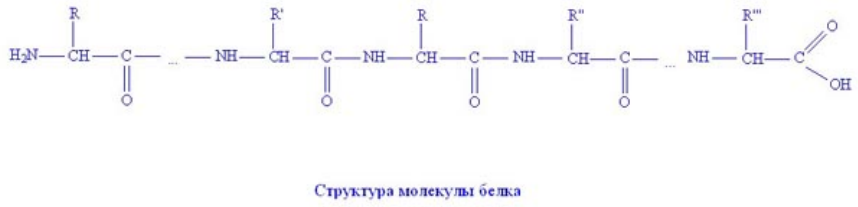

Рисунок 1 – Структура белка

При описании трехмерной структуры белка рассматривают обычно четыре разных уровня организации: первичную, вторичную, третичную и четвертичную структуры. Представление о трехмерном строении белковых молекул необходимо для понимания механизмов химических процессов, протекающих с их участием.

Вторичная структура.Пространственная конфигурация (конформация) полипептидной цепи белка создается благодаря возникновению дополнительных связей — «водородных мостиков», которые образуются как в пределах одной полипептидной цепи, так и между цепями.

2. Формализация задачи

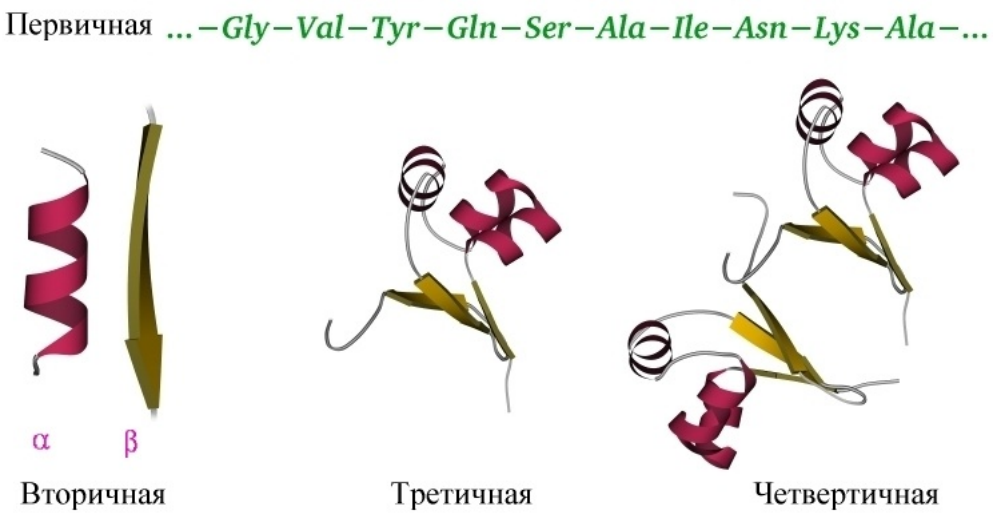

Рисунок 2 – Структура белка

Пусть дано входное множество P,состоящее из 20 аминокислот. $P = \{A, C, D, E...W, Y, X\}$. Для задачи поиска вторичной структуры, выходное множество состоит из девяти элементов: 'H', 'B', 'E', 'G', 'I', 'T', 'S', 'L', 'NoSeq'. (Метки вторичной структуры назначаются в соответствии со словарем меток вторичной структуры белка (DSSP)) [6]:

1. H - $\alpha - helix$
2. B - $\beta - bridge$
3. E - $\beta - strand$
4. G - $3_{10}helix$
5. I - $\pi - helix$
6. T - $Turn$
7. S - $Bend$
8. L - $Coil$
9. NoSeq -

Каждому символу множества P соответствует символ множества

$S = \{s_1, s_2, \ldots, s_n\}$, где $s_i \in \{L, B, E, G, I, H, S, T, NoSeq\}$. Задача ставится следующим образом: найти алгоритм вычисления соответствия $a : P \to S$.

3. Модель нейронной сети для предсказания вторичной структуры

Для предсказания вторичной структуры будем использоваться LSTM-сети (Long short-term memory). LSTM сети (Долгая краткосрочная память) - особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.

При прогнозировании цели x_t однонаправленная LSTM знает только прошлую последовательность, $x_1 ... x_t$. В задачах, где вся последовательность известна заранее, например предсказание вторичной структуры, это нежелательно. Решением этой проблемы являются двунаправленные LSTM сети. Первая модель начинает рекурсии от x_1 и идет вперед, обратная модель начинается с x_n и идет назад. Прогнозы от прямой и обратной сети объединяются и нормализуются.

В нашей модели ячейка LSTM выглядит следующим образом (рисунок 3):

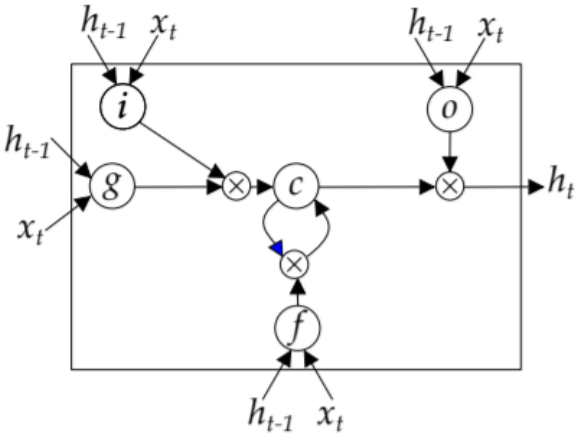

Рисунок 3 – i - входной вентиль, f - вентиль забывания, o - вентиль выхода, g - вентиль входной модуляции, c- ячейка памяти. Обозначения соответствуют уравнениям 1 - 8 таким образом, что W_{xo} соответствует значению x для выходного вентиля, а W_{hf} - это веса для h_{t-1}, чтобы забыть гейты и т.д.

Уравнения, описывающие данную модель:

$$i_t = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \tag{1}$$

$$f_t = \sigma(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \tag{2}$$

$$o_t = \sigma(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \tag{3}$$

$$g_t = \tanh(x_t W_{xg} + h_{t-1} W_{ho} + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

$$h_{t-rec} = h_t + feedforwardnet(h_t) \tag{7}$$

$$\sigma(z) = \frac{1}{1 + exp(-z)} \tag{8}$$

$\odot$ - поэлементное умножение

x_t - ввод из предыдущего слоя h_t^{l-1}

Мы расширяем стандартную сложенную двунаправленную модель LSTM путем введения сети с прямой связью (feedforwardnet, уравнение 7), отвечающей за объединение выходных данных из прямой и обратной сетей в одно предсказание softmax.

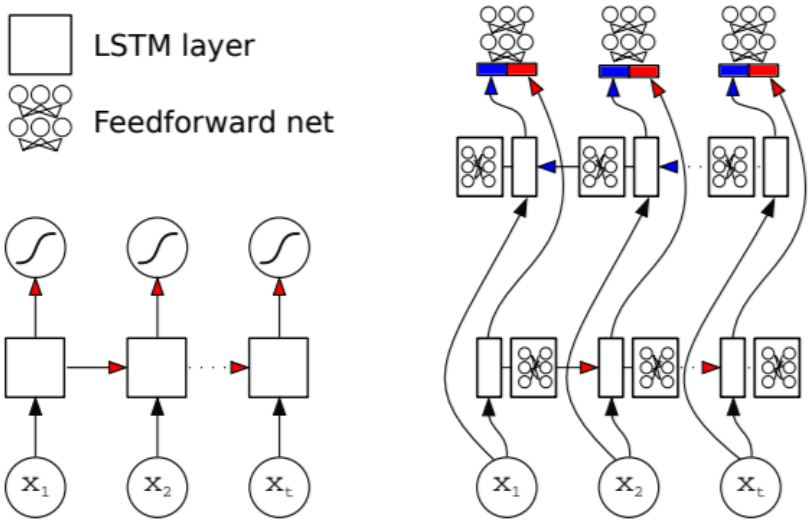

Рисунок 4 – Развернутые рекуррентные нейронные сети. слева: однонаправленный LSTM с одним слоем. справа: Двунаправленный LSTM с одним слоем.

Прямая LSTM (красные стрелки) начинается в момент времени 1, а обратная LSTM (синие стрелки) начинается в момент времени n, затем они идут вперед и назад соответственно. Ошибки из прямой и обратной сетей объединяются с использованием сетей прямой связи,

и результат используется для обратного распространения.На рисунке представлена однослойная модель, но модель легко расширяется.

4. Описание набора данных

Входной поток использует API (Application Programming Interface) набора данных TensorFlow, поэтому файлы данных необходимо преобразовать в TFRecords. Предполагается, что файлы данных представлены в формате .npy.gz и расположены в каталоге datadir.

Набор данных берется с сайта PDB (Protein Data Bank) [8]. Для того, чтобы с ним можно было работать, необходимо перевести данные в корректный (числовой) формат. Таким образом, мы получаем набор данных в формате numpy в виде матрицы (N белков x M аминокислот x k меток).

Для обучающего и тестового датасета:

$N = 595$

$M = 1000$

$k = 9$

Данные представлены следующим образом:

[0,21): аминокислотные остатки в порядке 'A', 'C', 'E', 'D', 'G', 'F', 'I', 'H', 'K', 'M', 'L', 'N', 'Q', 'P', 'S', 'R', 'T', 'W', 'V', 'Y', 'NoSeq'

[22,31): метки вторичной структуры с последовательностью 'L', 'B', 'E', 'G', 'I', 'H', 'S', 'T', 'NoSeq'(скрыты во время тестирования).

5. Оценка производительности

Confusion Matrix - это метод для подведения итогов работы алгоритма классификации.

Из ожидаемых результатов и прогнозов рассчитывают:

1. Количество правильных прогнозов для каждого класса.
2. Количество неверных прогнозов для каждого класса, организованных классом, которое было предсказано.

.

Эти числа затем организуются в таблицу или матрицу следующим образом:

Каждая строка матрицы соответствует прогнозируемому классу.

Каждый столбец матрицы соответствует фактическому классу.

6. Инструменты для построения нейронных сетей

TensorFlow — это ML-framework от Google, который предназначен для проектирования, создания и изучения моделей глубокого обучения. Содержит огромное количество готовых реализаций нейронных сетей и методов работы с ними.

7. Результат

Построенная сеть имеет коэффициент классификации около 76 %. Кроме того, данная модель, с применением двунаправленной LSTM сети работает значительно лучше, чем подобная ей двунаправленная RNN (BRNN) сеть, используемая в SSpro8, и имеющая правильный коэффициент классификации 0,511.

Как видно из рисунка 5, некоторые значения точно вычислить не удалось, а именно: B - $\beta - bridge$, I - $\pi - helix$, NoSeq. Это связано с тем, что чаще всего в белках встречаются H - $\alpha - helix$, B - $\beta - bridge$, L - $Coil$ структуры. Остальные элементы нелегко отследить, это делается при очень больших наборах данных.

Выходные данные записываются в текстовый файл secstruct.txt в формате: аминокислотный остаток -> структура в которую она свернется(рисунок 6).

```
Eval Loss: 110.204010, Eval Accuracy: 0.760241
Confusion Matrix (true label, predicted label):
[[81879     0  3228   170     0   753   417  1473     0]
 [  691     0   294   588     0    75    21    92     0]
 [ 2552     0 94563   554     0  7322    92   433     0]
 [  839     0   231  3591     0   769    26   676     0]
 [    7     0     0     0     0    19     0     4     0]
 [ 1327     0   394   348     0 123046    33  1009     0]
 [ 4225     0  1063   102     0   524 16599  1803     0]
 [ 2203     0  9578   284     0  1467  2255 105226     0]
 [    0     0     0     0     0     0     0     0     0]]
```

Рисунок 5 – Результат настройки и обучения нейронной сети

```
D -> H
K -> H
A -> H
F -> H
Q -> H
K -> C
L -> C
Y -> C
K -> H
L -> H
T -> C
N -> C
F -> C
S -> C
N -> C
L -> C
D -> C
K -> H
A -> H
```

Рисунок 6 – Результат работы нейронной сети

ЗАКЛЮЧЕНИЕ

Таким образом, в ходе данной работы была построена и обучена нейронная сеть, осуществляющая предсказание вторичной структуры белка по первичной.Полученные результаты показывают наличие возможности применения нейронных сетей для решения поставленной задачи, а также открывают простор для дальнейших экспериментов в данном направлении.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Levinthal, C. (1969) How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. J. T. P. DeBrunner and E. Munck eds., University of Illinois Press Pages 22–24.

2 M. Kaleel, M. Torrisi, C. Mooney, G. Pollastri. PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning Amino Acids, 51 (2019), pp. 1289-1296

3 M. Torrisi, G. Pollastri. Protein Structure Annotations. N.A. Shaik, K.R. Hakeem, B. Banaganapalli, R. Elango (Eds.), Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins, Springer International Publishing, Cham (2019), pp. 201-234

4 C.B. Anfinsen. Principles that govern the folding of protein chains.Science, 181 (1973), pp. 223-2307

5 Bastien, Fred eric, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian, Bergeron, Arnaud, Bouchard, Nicolas, Warde- Farley, David, and Bengio, Yoshua. Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590, November 2012

6 Kabsch W, Sander C (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". Biopolymers. 22

7 Keras [Электронный ресурс]:[сайт]. - URL: https://keras.io/about/ - Загл. с экрана.(дата обращения 14.04.2021)

8 - A Structural View of Biology [Электронный ресурс]:[сайт]. - URL: https://www.rcsb.org/ - Загл. с экрана.(дата обращения 14.04.2021)

9 L.H. Holley, M. Karplus. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA, 86 (1989), pp. 152-156

10 M. Kaleel, M. Torrisi, C. Mooney, G. Pollastri. PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning Amino Acids, 51 (2019), pp. 1289-1296

11 M. Torrisi, G. Pollastri. Protein Structure Annotations. N.A. Shaik, K.R. Hakeem, B. Banaganapalli, R. Elango (Eds.), Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins, Springer International Publishing, Cham (2019), pp. 201-234

12 C.B. Anfinsen. Principles that govern the folding of protein chains.Science, 181 (1973), pp. 223-2307

13 C.N. Magnan, P. Baldi. SSpro/ACCpro 5: almost perfect predictionof protein secondary structure and relative solvent accessibility usingprofiles, machine learning and structural similarity. Bioinformatics,30 (2014), pp. 2592-25978

14 D.R. Davies. A correlation between amino acid composition andprotein structure. J Mol Biol, 9 (1964), pp. 605-6099

15 D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 292 (1999), pp. 195-202

16 D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 292 (1999), pp. 195-202

17 Olah C.Understanding LSTM networks.2015[Электронный ресурс]:[сайт].- URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs - Загл. с экрана.(дата обращения 29.04.2021)

18 LONG SHORT-TERM MEMORY, Sepp Hochreiter, Jurgen Schmidhuber[Электронный ресурс]:[сайт].- URL: http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97lstm.pdf - Загл. с экрана.(дата обращения 29.04.2021)

19 Understanding LSTM Networks [Электронный ресурс]:[сайт].- URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ - Загл. с экрана.(дата обращения 13.05.2021)

20 The Unreasonable Effectiveness of Recurent Neural Networks [Электронный ресурс]:[сайт].- URL: http://karpathy.github.io/2015/05/21/rnn-effectiveness/ - Загл. с экрана.(дата обращения 11.05.2021)

21 Ferguson, T. S. A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2):209–230, 1973.[Электронный ресурс]:[сайт].- URL: https://projecteuclid.org/download/pdf1/euclid.aos/1176342360 - Загл. с экрана.(дата обращения 12.05.2021)