Laboratory of functional and comparative genomics

Jakob Denisov, Anton Katsubo,
Daria Pisetskaya, Zoe Chervontseva

>> EXPRENET: A NEURAL NETWORK FOR EXPRESSION ANALYSIS

Анализ признаков, позволяющих предсказывать экспрессию генов у *E. coli*

AIMS

- Выяснить, могут ли нейронные сети глубокого обучения использоваться для анализа данных экспрессии генов *E.coli.*
- Проанализировать, какие признаки влияют на предсказание сети.

- Find out whether neural networks may be used for analysis of expressions
- Identify features considered by the neural network for prediction.

DATA

Из статьи [1] были взяты микрочиповые данные по экспрессии 4189 генов *E. coli* в 1918 условиях.

Microarray expression dataset for 4189 *E. coli* genes in 1918 conditions was taken from [1].

[1] Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. Mol Syst Biol. 2014;10(7):735. Published 2014 Jul 1. doi:10.15252/msb.20145108

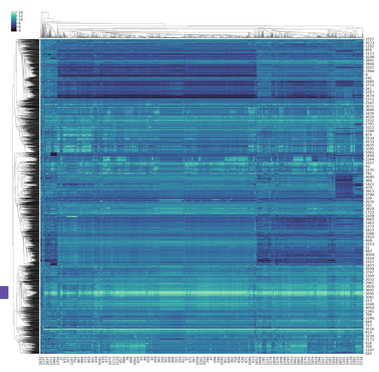
RESULTS

Сеть - denoising автоэнкодер. Сворачивает входной вектор экспрессий в массив с минимальным числом элементов. Анализируется способ свертки.
The network is a denoising autoencoder. Minimizes the input expression vector into an array with a minimal number of elements. The minimization method is analyzed.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	(None, 15552)	0
dense_5 (Dense)	(None, 256)	3981568
dense_6 (Dense)	(None, 32)	8224
dense_7 (Dense)	(None, 256)	8448
dense_8 (Dense)	(None, 15552)	3996864

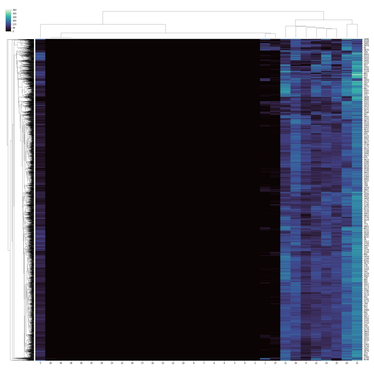
Total params: 7,995,104
Trainable params: 7,995,104
Non-trainable params: 0

Архитектура сети. Использованы полносвязные слои. На входе - несколько раз повторенная таблица экспрессии с добавленным шумом. // The network architecture. Fully connected layers are used. At the input - several repetitions of the expression table with added noise.



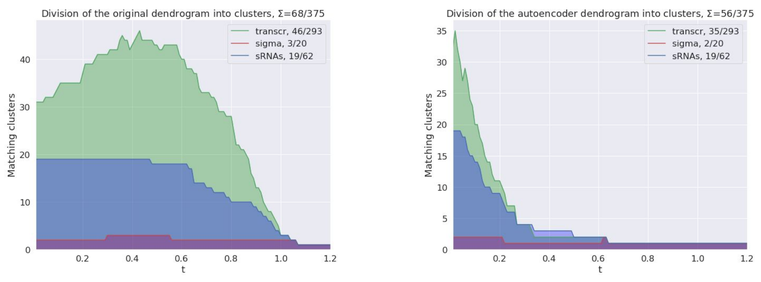
Двойная иерархическая кластеризация исходных данных. Заметны полосы сильно скоррелированных генов, например, рибосомных белков.

Double hierarchical clustering of source data. Bands of highly correlated genes, such as ribosomal proteins, are noticeable.



Кластеризация представления внутреннего слоя автоэнкодера. Многие нейроны просто не используются по неизвестной причине.
Clustering of the autoencoder internal layer representation. Many neurons are simply not used due to unknown reasons.

Кластеризацию генов автоэнкодером можно сравнить с кластеризацией генов исходных данных.
Autoencoder gene clustering can be compared to source data gene clustering.



Дендрограммы для генов нарезались на расстояниях t/2 от конца ветвей для получения кластеров для сравнения их с базой RegulonDB. Кластеры из RegulonDB определялись как множество генов, регулируемых каким-то отдельным сигма-фактором / транскрипционным фактором / малой РНК. Кластеры считались совпадающими, если 9/10 их элементов были одинаковыми.

Dendrograms for genes were cut at a distance of t/2 from the end of the branches to obtain clusters for comparison with the RegulonDB database. RegulonDB clusters were defined as sets of genes regulated by a particular sigma factor / transcription factor / small RNA. Clusters were considered to match if 9/10 of their elements coincided.

transcr - сравнение по транскрипционным факторам
sigma - сравнение по сигма-факторам
sRNAs - сравнение по малым РНК
transcr - comparison by transcription factors
sigma - comparison by sigma factors
sRNAs - comparison by small RNAs

Видно, что автоэнкодеру недостает 9 кластеров по сравнению с алгоритмом кластеризации. Это означает, что результат неудовлетворителен, и полученная нейросеть вряд ли несет много полезной информации о связях между генами.

It can be seen that the autoencoder lacks 9 clusters compared to the clustering algorithm. This means that the result is unsatisfactory, and the resulting neural network is unlikely to provide much useful information about the relationship between the genes.

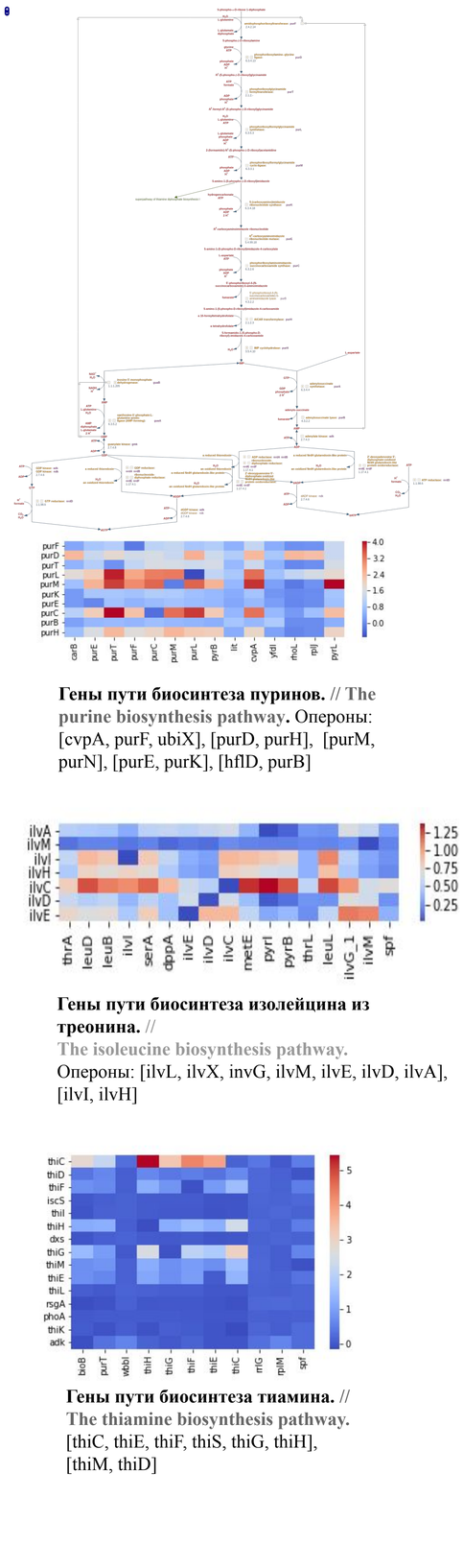
Сеть - предиктор. Предсказывает значение гена, экспрессия которого была скрыта маской. Анализируются гены, на экспрессию которых нейросеть “смотрит” в первую очередь, чтобы предсказать значение скрытой ячейки.
The network is the predictor. Predicts the value of the gene whose expression was hidden by the mask. Genes, the expression of which the neural network "looks at" first to predict the value of the hidden cell, are analyzed.

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 4)	16760
dense_2 (Dense)	(None, 8)	40
dense_3 (Dense)	(None, 8)	72
dense_4 (Dense)	(None, 1)	9
activation_1 (Activation)	(None, 1)	0

Total params: 16,881
Trainable params: 16,881
Non-trainable params: 0

Архитектура сети-предиктора. Использованы полносвязные слои. // The network architecture. Fully connected layers are used.

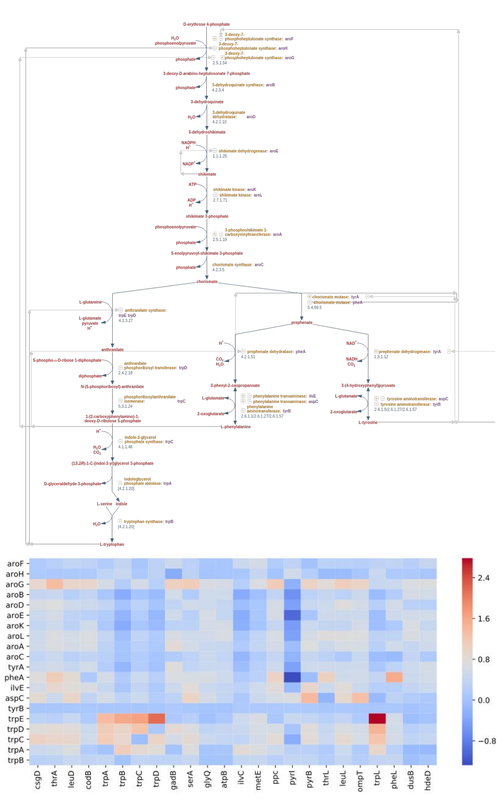
Для каждого гена из метаболического пути (строки) были посчитаны важности всех 4189 генов для предсказания экспрессии. В столбцах оставлены наиболее важные для предсказания гены. Чем больше значение в клетке, тем важнее был ген в столбце для предсказания экспрессии гена в строке. // For each gene in a metabolic pathway (rows), importance of each 4189 genes for the prediction was calculated. Columns represent genes that are the most important for the prediction. The larger is the value in a cell, the more important is the gene in the column is for the prediction of the gene in the row.



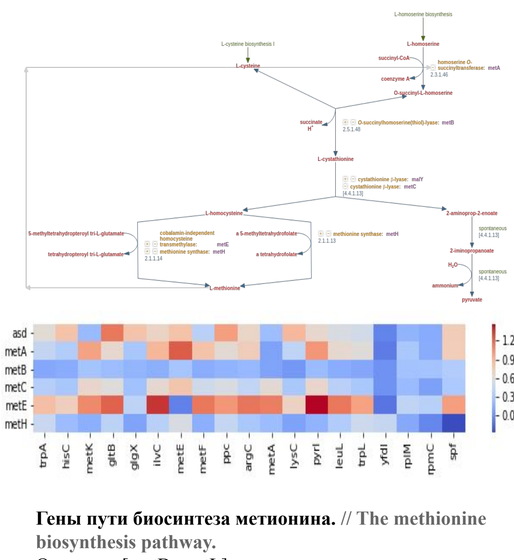
Гены пути биосинтеза пуринов. // **The purine biosynthesis pathway.** Опероны: [cvpA, purF, ubiX], [purD, purH], [purM, purN], [purE, purK], [hflD, purB]

Гены пути биосинтеза изолейцина из треонина. //
The isoleucine biosynthesis pathway.
Опероны: [ilvL, ilvX, invG, ilvM, ilvE, ilvD, ilvA], [ilvI, ilvH]

Гены пути биосинтеза тиамина. // **The thiamine biosynthesis pathway.**
[thiC, thiE, thiF, thiS, thiG, thiH], [thiM, thiD]



Гены пути биосинтеза ароматических аминокислот. // **Biosynthesis of aromatic amino acids.**
Опероны: [aroF, tyrA], [trpA, trpB, trpC, trpD, trpE, trpL], [aroA, serC]



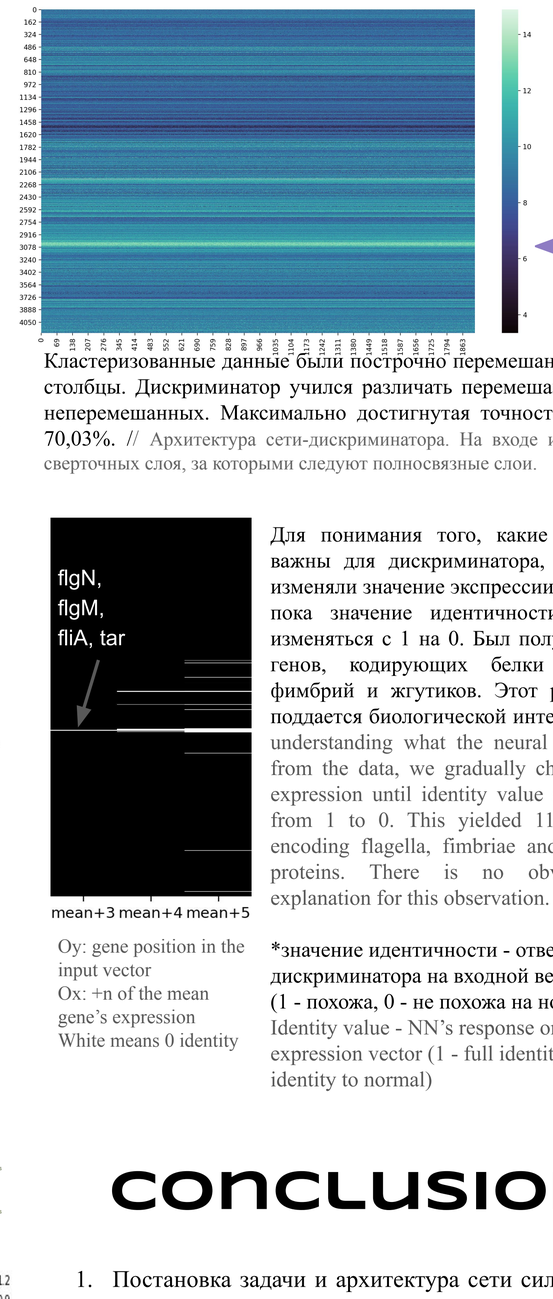
Гены пути биосинтеза метионина. // **The methionine biosynthesis pathway.**
Опероны: [metB, metL].

Сеть-дискриминатор. Определяет, насколько входной вектор экспрессий похож на настоящий.
Network - discriminator. Determines the extent of closeness input vector of data to the real E.coli’s expression vector.

Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 4185, 64)	384
conv1d_4 (Conv1D)	(None, 4181, 16)	5136
conv1d_5 (Conv1D)	(None, 4177, 4)	324
flatten_1 (Flatten)	(None, 16708)	0
dense_5 (Dense)	(None, 100)	1670900
dense_6 (Dense)	(None, 50)	5050
dropout_3 (Dropout)	(None, 50)	0
dense_7 (Dense)	(None, 20)	1020
dropout_4 (Dropout)	(None, 20)	0
dense_8 (Dense)	(None, 4)	84
dropout_5 (Dropout)	(None, 4)	0
dense_9 (Dense)	(None, 1)	5

Total params: 1,682,903
Trainable params: 1,682,903
Non-trainable params: 0

Архитектура сети-дискриминатора. На входе использованы три сверточных слоя, за которыми следуют полносвязные слои. // Discriminator’s architecture. Three convolutional layers were used at the input, followed by dense layers.



Кластеризованные данные были построчно перемешаны и нарезаны на столбцы. Дискриминатор учился различать перемешанные данные от неперемешанных. Максимально достигнутая точность определения - 70,03%. // Архитектура сети-дискриминатора. На входе использованы три сверточных слоя, за которыми следуют полносвязные слои.

Oy: gene position in the input vector
Ox: +n of the mean gene’s expression
White means 0 identity

Для понимания того, какие гены наиболее важны для дискриминатора, мы поочередно изменяли значение экспрессии гена до тех пор, пока значение идентичности* не начнет изменяться с 1 на 0. Был получен список 119 генов, кодирующих белки хеморецепции, фимбрий и жгутиков. Этот результат плохо поддается биологической интерпретации. // To understanding what the neural network extracts from the data, we gradually changed each gene expression until identity value started to change from 1 to 0. This yielded 119 genes, mainly encoding flagella, fimbriae and chemoreception proteins. There is no obvious biological explanation for this observation.

*значение идентичности - ответ дискриминатора на входной вектор экспрессий (1 - похожа, 0 - не похожа на нормальную). // Identity value - NN’s response on input expression vector (1 - full identity, 0 - zero identity to normal)

CONCLUSIONS

1. Постановка задачи и архитектура сети сильно влияют на результат.
2. Обнадеживающие предварительные результаты показал один из трех исследованных подходов.
3. Но и для него требуется больше времени на обучение сети.

1. The problem statement and network architecture strongly influence the result.
2. Only one of three studied approaches yielded promising results.
3. But it requires more computational time to properly train the network.