

Анализ корреляций в протеомах микроорганизмов

Лазарева А.*, Матюшкина Д., Бутенко И., Говорун В.

Институт системной биологии и медицины Роспотребнадзора, Москва, Россия

* a.lazareva@sysbiomed.ru

Ключевые слова: протеомика; системная биология; синтетическая биология

Мотивация и цель: История изучения клетки стартовала еще в XVII веке с рассмотрения Робертом Гуком структуры пробки в микроскоп. С тех пор научные методы претерпели множество изменений, которые позволили изучить все детали: от строения отдельной молекулы до принципиальных связей в каскадах метаболических реакций; от концентраций ферментов и метаболитов в клетке до их точной локализации. Однако современной науке так и не удалось приблизиться к полному пониманию и описанию всех процессов жизнедеятельности клетки.

К настоящему моменту научным сообществом накоплен большой объем «мультиомиксных» данных для различных биологических объектов, из которых исследователи используют только «верхушку айсберга» для проверки конкретных гипотез, при этом гигантский объем информации остается неиспользованным. В свою очередь, тщательный их анализ позволит выявить новые закономерности и механизмы организации и регуляции живой клетки. Эти знания необходимы для возможности создания искусственной упрощенной клетки, которая обеспечит прорыв как в фундаментальной науке о жизни, так и в прикладных исследованиях. И в первую очередь для реализации данного подхода в нашей работе была поставлена следующая цель: выявление минимального белкового «кора» живой клетки и анализ закономерностей его организации и функционирования.

Методы и алгоритмы: Культивирование *Mycoplasma gallisepticum* S6. Лизис клеток. Трипсинолиз в растворе. Количественный протеомный анализ методами MRM и shot-gun. Расчет z-оценки. Расчет коэффициента корреляции методом Пирсона. Анализ методом главных компонент (PCA). Кластеризация методами t-SNE, K-средних и DBSCAN.

Результаты: В системном подходе к изучению микроорганизмов есть несколько ключевых этапов, в том числе секвенирование генома и сопутствующего генетического материала, аннотация генома, установление механизмов регуляции элементов клетки, аннотация метаболических путей и различных биохимических процессов и т. д. Существуют как общепризнанные модельные микроорганизмы, для которых получен большой набор мультиомиксных данных [1–3], так и достаточно хорошо изученные минимальные бактериальные клетки [4, 5]. В ходе работы были отобраны протеомные данные для *Mycoplasma gallisepticum* S6, *Escherichia coli*, *Bacillus subtilis* и *Saccharomyces cerevisiae* в различных экспериментальных условиях, оценено необходимое значение количества образцов, после которого коэффициент корреляции переставал изменяться [1–4]. В зависимости от экспериментальных выборок (точнее степени их гомогенности) требовалось от 20 до 50 образцов для достижения ошибки определения коэффициента корреляции < 0.5 %. Далее были построены взвешенные графы, где в вершинах располагались

белки, а ребрами отмечены корреляционные связи, при этом вес ребра была обратно пропорционально связана с модулем коэффициента корреляции. Для удобства анализа скоррелированных и антискоррелированных пар ребра были покрашены в различные цвета (красный – для положительного коэффициента корреляции, синий – для отрицательного). Еще на этом этапе было обнаружено двудольное распределение микоплазменного протеома, где две группы плотно скоррелированных белков были антискоррелированы между собой. Повышение модуля значений коэффициентов корреляции у отображаемых ребер с 0.75 до 0.85 привело к исчезновению связей между кластерами, что позволяет сделать вывод о более плотном взаимодействии внутри обнаруженных групп.

При этом интересным результатом оказалось то, что первый кластер сформировали в основном белки, обладающие различной ферментативной активностью, а во втором кластере обнаружилось много белков, взаимодействующих с нуклеиновыми кислотами. При этом наложение метаболических путей или данных о белковых комплексах не позволило объяснить подобную кластеризацию. Поэтому полученные результаты дают возможность предположить, что белки в кластерах организуются по определенным правилам, а не случайным образом. Возможно, такая протеомная кластеризация необходима живой клетке для поддержания ее метастабильного состояния.

Для определения кластеризации протеома микоплазмы другим способом, с помощью метода главных компонент, были определены основные белки, участвующие в формировании реакции. Для микоплазмы оказалось достаточно всего 15 белков, чтобы обеспечить 70 % ответа на внешний стимул. Снизив с помощью метода t-SNE размерность белкового пространства до двумерного, все белки были отображены на координатной плоскости, а затем были размечены по результатам кластеризации с помощью методов k-means и DBSCAN. С помощью первого обнаружился оптимальный диапазон числа кластеров: от 3 до 5. Метод DBSCAN разделил протеом на 4 группы. Однако эти результаты не удалось соотнести с результатами корреляционного анализа, оставляя вопрос группировки белков открытым. Аналогичное исследование протеомов *Escherichia coli* и *Bacillus subtilis* также показало двудольную корреляционную организацию, образованную плотными связями внутри групп и более слабыми между. Однако усложнение организации и регуляции относительно *Mycoplasma gallisepticum* привело к увеличению модулей коэффициентов корреляций между белками. В случае же эукариотического протеома *Saccharomyces cerevisiae* подобная двудольная организация была утрачена.

Выводы: Были отобраны и подготовлены протеомные данные в различных условиях для *Mycoplasma gallisepticum* S6 (данные, полученные нами), *Escherichia coli*, *Bacillus subtilis* и *Saccharomyces cerevisiae* (результаты из баз данных).

Были применены методы кластеризации протеомных данных с помощью алгоритмов корреляционного анализа, а также с помощью PCA, k-means, t-SNE, DBSCAN, а также проанализированы результаты корреляционного анализа и выявлены общие закономерности. Для полученных кластеров характерна двудольная организация в протеомах прокариот (*Mycoplasma gallisepticum*, *Escherichia coli*, *Bacillus subtilis*) и отсутствие подобного разделения для эукариот (*Saccharomyces cerevisiae*). А методы кластеризации k-means и DBSCAN на данных для *Mycoplasma gallisepticum* показали существование групп белков со схожим поведением при изменении условий существования клетки.

Для *Mycoplasma gallisepticum* S6 были дополнительно подготовлены протеомные данные для культур в нормальных и осмотических условиях. В каждом из наборов данных проведен корреляционный анализ, показавший динамику изменения соотношений в протеомах при стрессовых воздействиях.

Финансирование: Исследование поддержано Управлением Роспотребнадзора ГЗ № 1022040800170-3-1.6.23 «Создание искусственных клеточных систем».

Analysis of correlations between microbial proteomes

Lazareva A.*, Matyushkina D., Butenko I., Govorun V.

Scientific Research Institute for Systems Biology and Medicine, Moscow, Russia

* a.lazareva@sysbiomed.ru

Key words: proteomics; systems biology; synthetic biology

Motivation and Aim: The history of cell research started back in the XVII century with Robert Hooke's examination of the structure of a cell in a microscope. Since then, scientific methods have undergone many changes that have made it possible to study all the details: from the structure of a single molecule to the fundamental connections in the cascades of metabolic reactions; from the concentrations of enzymes and metabolites in the cell to their exact localization. However, modern science has not been able to come close to a complete understanding and description of all the processes of cell activity.

Up to the present time, the scientific community has accumulated a significant amount of “multiomics” data on various biological objects, of which researchers have only utilized a small fraction to test specific hypotheses, while a huge amount of remains untapped. A thorough analysis of this data could reveal novel patterns and mechanisms of organization and regulation of living cells. This knowledge would be essential for the development of an artificial simplified cell that could lead to a breakthrough in both fundamental life sciences and applied research. And first, in order to implement this approach, we set the following objective in our work: to identify the minimum protein “core” of a living cell and to analyze the patterns of their organization and function.

Methods and Algorithms: Cultivation of *Mycoplasma gallisepticum* S6 cells. Cell lysis. Trypsinolysis in solution. Quantitative proteomic analysis using MRM and shotgun methods. Calculation of the Z-score. Pearson correlation analysis. PCA (Principal Component Analysis). TSN (T-SNE) (t-distributed stochastic neighbor embedding). K-means clustering. DBSCAN (Density-based spatial clustering of applications with noise).

Results: As previously mentioned, there are several crucial stages in the examination of an organism. These include sequencing the genome and associated genetic material, establishing a connection between a gene and protein, identifying the functions and methods of regulating cellular components, annotating metabolic pathways and diverse biochemical processes, among others. There are both widely recognized model microorganisms, for which a significant volume of multiomics data has been collected [1–3], and well-researched minimal bacterial cells [4, 5]. During the course of this work, proteomic data was collected for *Mycoplasma gallisepticum* S6, *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae* under various experimental conditions. The required number of samples was then estimated, following which the correlation coefficient stopped changing [1–4]. Depending on the experimental samples and the degree of their

homogeneity, it took between 20 and 50 samples to achieve an error of less than 0.5 % in determining the correlation coefficient. Next, weighted graphs were created, with proteins as vertices and correlation connections represented by edges. The edge length was inversely proportional to the absolute value of the correlation coefficient. To facilitate analysis of correlated and anticorrelated pairs, edges were color-coded, with red indicating a positive correlation and blue indicating a negative correlation. Even at this stage, a bipartite distribution of the *Mycoplasma* proteome was observed, where two groups of highly correlated proteins were anticorrelated with each other. An increase in the magnitude of the correlation coefficient for the observed edges from 0.75 to 0.85 resulted in the disappearance of connections between clusters, indicating a denser interaction within the identified groups.

Interestingly, the first cluster consisted primarily of proteins with diverse enzymatic activities, while the second cluster contained proteins interacting with nucleic acids. The overlap of metabolic pathways or information on protein complexes was insufficient to explain this clustering. Consequently, the findings suggest that proteins within clusters are organized in a manner consistent with specific rules, rather than randomly. It is possible that such proteomic clustering plays a role in maintaining the metastable state of a living cell.

To identify the clustering of mycoplasma proteomes in another manner, using the principal components analysis, the key proteins involved in reaction formation were identified. For mycoplasmas, only 15 proteins were sufficient to account for 70 % of the response to external stimuli. By reducing the protein space dimension to two dimensions through the t-SNE method, all proteins were plotted on a two-dimensional coordinate plane and then categorized according to the clustering results obtained from the k-means and DBSCAN methods. Using the first one, the optimal number of clusters range was found to be between 3 and 5. However, the DBSCAN approach divided the proteome into four groups, which could not be correlated with correlation analysis results, leaving the issue of protein grouping unresolved.

A similar study of the proteomic profiles of *Escherichia coli* and *Bacillus subtilis* also revealed a bipartite organization, characterized by strong intra-group interactions and weaker inter-group connections. However, due to the complexity of the regulatory mechanisms and organization in *Mycoplasma gallisepticum*, there has been an increase in the number of modules of correlation coefficients between proteins compared to other organisms. In contrast, the eukaryotic proteome of *Saccharomyces cerevisiae* exhibits a more complex organization and lacks the bipartite structure observed in other organisms. **Conclusion:** Proteomic data were collected and processed under various conditions for *Mycoplasma gallisepticum* S6 (data generated by our team), *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae* (data retrieved from publicly available databases).

Methods of clustering proteomic data were applied using correlation analysis algorithms as well as PCA, k-means, t-SNE and DBSCAN. The results of correlation analysis were used for identifying general patterns. The clusters obtained were characterized by a bipartite organization in prokaryotic proteomes, such as *Mycoplasma gallisepticum*, *Escherichia coli* and *Bacillus subtilis*, and the absence of such separation in eukaryotic proteomes like *Saccharomyces cerevisiae*. Additionally, clustering methods such as k-means, t-SNE and DBSCAN on data for *Mycoplasma gallisepticum* showed the existence of groups of proteins that exhibit similar behavior under varying conditions of cellular existence.

Proteomic data for cultures under normal and osmotic conditions was additionally prepared for *Mycoplasma gallisepticum* S6. For each data set, a correlation analysis was conducted, which revealed the dynamics of changes in proteome ratios under stress conditions.

Funding: The study is supported by Rospotrebnadzor service state task No. 1022040800170-3-1.6.23.

Список литературы/References

1. Mori M. et al. From coarse to fine: the absolute *Escherichia coli* proteome under diverse growth conditions. *Mol Syst Biol.* 2021;17(5):e9536. doi 10.15252/msb.20209536
2. Shi L. et al. Tyrosine 601 of Bacillus subtilis DnaK Undergoes Phosphorylation and Is Crucial for Chaperone Activity and Heat Shock Survival. *Front Microbiol.* 2016;7;533. doi 10.3389/fmicb.2016.00533
3. Ho B. et al. Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome. *Cell Syst.* 2018;6(2):192-205. doi 10.1016/j.cels.2017.12.004
4. Butenko I. et al. Data-independent proteome profile of *Mycoplasma gallisepticum* under normal conditions and heat stress. *Data Brief.* 2017;16:700-704. doi 10.1016/j.dib.2017.11.093
5. Hutchison C. et al. Design and synthesis of a minimal bacterial genome. *Science.* 2016;351(6280):aad6253. doi 10.1126/science.aad6253