

Анализ молекулярных последовательностей гена CREBBP методом главных компонент (PCA-seq)

Гун Жуйхань^{1*}, Ефимов В.^{1, 2}

¹ Новосибирский государственный университет, Новосибирск, Россия

² Институт цитологии и генетики СО РАН, Новосибирск, Россия

* z.gun@g.nsu.ru

Ключевые слова: crebbp; pca-seq

Ген CREBBP расположен в хромосоме 16p13.3, имеет 33 экзона и кодирует белок СВР из 2442 аминокислот [1]. Это коактиватор, который играет ключевую роль в эмбриональном развитии, контроле роста и стабильности внутренней среды, взаимодействуя с различными факторами транскрипции. Он может влиять на различные физиологические процессы и широко экспрессируется в тканях человека, включая семенники, костный мозг, яичники и эндометрий. Спектр мутаций CREBBP состоит из точечных вариантов (30–50 %) и делеций (~10 %) [2]. СВР экспрессируется повсеместно и, как было показано, обладает внутренней активностью гистон-ацетилтрансферазы (НАТ), которая влияет на активность хроматина посредством модуляции нуклеосомных гистонов [3]. Кроме того, согласно исследованиям, точечные мутации в основном затрагивают домен НАТ СВР, а потеря активности НАТ достаточна, чтобы вызвать RSTS – наследственное заболевание, которое характеризуется комплексом врожденных аномалий. Хромосомные транслокации с участием этого гена также связаны с острым миелолейкозом.

Мы проанализировали взаимосвязь между тремя положениями каждого кодона гена и нуклеотидом (ACTG), и данные показали, что в целом С (цитозин) показал положительную корреляцию, а Т (тимин) – отрицательную корреляцию, когда А (аденин) и G (гуанин) расположены в первой позиции кодона, корреляция мало чем отличается, но, когда они расположены во второй и третьей позициях соответственно, корреляция прямо противоположная.

Затем мы дополнительно проанализировали корреляцию между различными кодонами каждой аминокислоты, и данные показали, что корреляция глутаминовой кислоты, кодируемой CAG, выше, чем у других кодонов, кодирующих глутаминовую кислоту, поскольку число заболеваний, связанных с геном CREBBP, достигает 600. многие, включая НТТ (болезнь Хантингтона). Основным патогенным кодоном НТТ является cag. Аналогично, другие кодоны с более высокой корреляцией могут быть связаны с другими заболеваниями.

Поэтому мы выбрали ген CREBBP, используя новый статистический метод анализа PCA-Seq. Для достижения поставленной цели были определены следующие задачи: сначала с помощью метода главных компонент проанализировать нуклеотидные и аминокислотные последовательности гена CREBBP, а также изучить взаимосвязь между главными компонентами этих последовательностей и физическими и химическими свойствами аминокислот. Во-вторых, изучить связь между главными компонентами этих последовательностей и вторичной структурой белка.

PCA (Principal Component Analysis) – метод уменьшения размерности матриц данных. Суть этого метода заключается в построении ортогональной системы координатных осей (главных компонент), которая наиболее экономично описывает взаимное расположение объектов в евклидовом пространстве. Главные компоненты не коррелируют друг с другом и расположены в порядке убывания дисперсии. Изобретенный Карлом Пирсоном в 1901 году, он сейчас используется в различных областях для минимизации размеров данных при сохранении максимального количества информации. Для последовательностей PCA-Seq включает разрезание последовательности на фрагменты с использованием скользящего окна, расчет матрицы евклидовых расстояний для всех фрагментов и применение метода РС для расчета главных компонент [4]. AAindex – это база данных числовых индексов, отражающая физические, химические и биологические свойства аминокислот и их пар. База данных AAindex широко используется в биоинформатических исследованиях и постоянно обновляется. Примеры включают предсказание субклеточной локализации белков, сайтов модификации и выравнивания последовательностей нескольких белков. База данных AAindex широко используется в исследованиях в области биоинформатики и постоянно обновляется [5].

Analysis of molecular sequences of the CREBBP gene by principal component analysis (PCA-seq)

Gong Ruihan^{1*}, Efimov V.^{1, 2}

¹ Novosibirsk State University, Novosibirsk, Russia

² Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

* z.gun@g.nsu.ru

Key words: crebbp; pca-seq

The CREBBP gene is located on chromosome 16p13.3, has 33 exons and encodes the CBP protein of 2442 amino acids [1]. It is a coactivator that plays a key role in embryonic development, growth control and stability of the internal environment by interacting with various transcription factors. It can influence various physiological processes and is widely expressed in human tissues, including testes, bone marrow, ovaries and endometrium. The spectrum of CREBBP mutations consists of point variants (30–50 %) and deletions (~10 %) [2]. CBP is expressed ubiquitously and has been shown to have intrinsic histone acetyltransferase (HAT) activity, which influences chromatin activity through modulation of nucleosomal histones [3]. Additionally, research suggests that point mutations primarily affect the HAT domain of CBP, and loss of HAT activity is sufficient to cause RSTS, an inherited disorder characterized by a complex of congenital abnormalities. Chromosomal translocations involving this gene are also associated with acute myeloid leukemia.

We analyzed the relationship between the three positions of each gene codon and the nucleotide (ACTG), and the data showed that overall, C (cytosine) showed a positive correlation and T (thymine) showed a negative correlation when A (adenine) and G (guanine) are located at the first codon position, the correlation is not much different, but when they are located at the second and third positions respectively, the correlation is exactly the opposite.

We then further analyzed the correlation between different codons of each amino acid, and the data showed that the correlation of CAG-encoded glutamic acid is higher than other glutamic acid-encoding codons as the number of diseases associated with the CREBBP gene reaches 600, including HTT. (Huntington's disease). The main pathogenic codon of HTT is cag. Likewise, other codons with higher correlations may be associated with other diseases.

Therefore, we selected the CREBBP gene using a new statistical analysis method, PCA-Seq. To achieve this goal, the following tasks were identified: first, using the principal component method, analyze the nucleotide and amino acid sequences of the CREBBP gene, and also study the relationship between the main components of these sequences and the physical and chemical properties of amino acids. Second, study the relationship between the main components of these sequences and the secondary structure of the protein.

PCA (Principal Component Analysis) is a method for reducing the dimension of data matrices. The essence of this method is to construct an orthogonal system of coordinate axes (principal components), which most economically describes the relative position of objects in Euclidean space. The principal components do not correlate with each other and are arranged in decreasing order of variance. Invented by Karl Pearson in 1901, it is now used in various fields to minimize the size of data while maintaining the maximum amount of information. For sequences, PCA-Seq involves cutting the sequence into fragments using a sliding window, calculating the Euclidean distance matrix for all fragments, and using the PCo method to calculate principal components [4]. AAindex is a database of numerical indexes reflecting the physical, chemical and biological properties of amino acids and their pairs. The AAindex database is widely used in bioinformatics research and is constantly updated. Examples include prediction of protein subcellular localization, modification sites, and sequence alignment of multiple proteins. The AAindex database is widely used in bioinformatics research and is constantly updated [5].

Список литературы/References

1. Pérez-Grijalba V. et al. New insights into genetic variant spectrum and genotype-phenotype correlations of Rubinstein-Taybi syndrome in 39 CREBBP-positive patients. *Mol Genet Genomic Med.* 2019;7(11):e972. doi 10.1002/mgg3.972
2. Rusconi D. et al. Characterization of 14 novel deletions underlying Rubinstein-Taybi syndrome: an update of the CREBBP deletion repertoire. *Hum Genet.* 2015;134(6):613-626. doi 10.1007/s00439-015-1542-9
3. Chan H.M., La Thangue N.B. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *J Cell Sci.* 2001;114(Pt. 13):2363-2373. doi 10.1242/jcs.114.13.2363
4. Efimov V.M., Efimov K.V., Kovaleva V.Y. Principal component analysis and its generalizations for any type of sequence (PCA-Seq). *Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov Journal of Genetics and Breeding.* 2019;23(8):1032-1036. doi 10.18699/VJ19.584
5. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 2008;36(Database issue):D202-D205. doi 10.1093/nar/gkm998