БИОФИЗИКА, 2023, том 68, № 5, с. 920–925

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

ПОИСК АНОМАЛИЙ СИГНАЛА ПОКРЫТИЯ СЕКВЕНИРОВАНИЯ, АССОЦИИРОВАННЫХ СО СТРУКТУРНЫМИ ВАРИАЦИЯМИ ГЕНОМА

© 2023 г. И.В. Бездворных*, Н.А. Черкасов*, А.А. Канапин*, А.А. Самсонова*,#

**Санкт-Петербургскии государственный университет, Университетская набережная, 7–9, Санкт-Петербург, 199034, Россия*

#E-mail: a.samsonova@spbu.ru

Поступила в редакцию 04.10.2022 г.
После доработки 04.10.2022 г.
Принята к публикации 25.10.2022 г.

Структурные вариации генома являются одним из основных источников генетического разнообразия. Как мутагены структурные варианты могут оказывать значительное влияние на здоровье человека, являясь причинами наследственных и онкологических заболеваний. Существующие методы поиска структурных вариантов основываются на анализе данных высокопроизводительного секвенирования, и, несмотря на значительный прогресс в их развитии, не позволяют определять структурные вариации с точностью, достаточной для применения в диагностике. Новые возможности для разработки методов поиска структурных вариаций представляет анализ сигнала покрытия секвенирования (т.е. количества фрагментов секвенирования, выравненных в каждой точке генома), который может рассматриваться как временной ряд. В работе представлен метод для поиска повторяющихся паттернов в сигнале покрытия, разработанный с использованием алгоритмов, применяемых для анализа временных рядов, а именно: KNN- (K-nearest neighbour) и SAX-преобразования (Symbolic Aggregation Approximation) сигнала. С использованием данных проекта Human Genome Diversity Project, включающих полногеномное секвенирование 911 человек разного этнического происхождения, нами были построены обобщающие паттерны сигнала покрытия в окрестностях точек разрыва, соответствующих структурным вариациям. В дополнение был разработан программный пакет для быстрого поиска аномалий в сигнале покрытия с применением полученных паттернов.

Ключевые слова: геномика, секвенирование, структурные вариации, анализ сигналов, SAX-трансформация.

DOI: 10.31857/S0006302923050113, **EDN:** PHCXCT

Структурные вариации включают в себя такие масштабные перестройки генома, как протяженные делеции, инсерции, транслокации, инверсии, а также изменения уровня копийности (copy number variation, CNV) [1]. В отличие от однонуклеотидных замен и коротких (менее 30 п.о.) инсерций и делеций, структурные вариации, в силу своего размера, могут оказывать значительное влияние на фенотип организма хозяина и, в некоторых случаях, провоцировать возникновение заболеваний и наследственных патологий [2, 3]. Современные алгоритмы поиска структурных вариаций базируются на использовании данных высокопроизводительного секвенирования (полногеномного или полноэкзомного) и используют две основных методологии: анализ уровня сигнала покрытия (depth of coverage) и анализ расщепленных фрагментов секвенирования (split reads) [4].

Несмотря на значительный прогресс, достигнутый в области разработки способов обнаружения структурных вариантов, задача точной их локализации до сих пор не имеет удовлетворительного решения, что выражается в значительном уровне разногласий в результатах поиска [5, 6]. Иначе говоря, цель, заключающаяся в нахождении точки разрыва с точностью в плюс/минус одно основание, необходимая для применения результатов в клинической практике не достигнута. Одним из способов решения проблемы неточности обнаружения координат структурных вариаций в геноме, является разработка алгоритмов, анализирующих и интегрирующих результаты предсказания различных программ (так называемых meta-callers), таких, как svclassify и подобных [7].

Сокращение: SAX – аппроксимация с помощью символьной агрегации (Symbolic Aggregation Approximation).

В то же время анализ уровня сигнала покрытия с точки зрения поиска аномалий и аберраций изучен относительно мало, и лишь немногие программные пакеты используют этот источник информации о форме, протяженности, амплитуде, тренде и пр. сигнала для поиска структурных вариантов [8, 9]. Под сигналом покрытия генома понимается вектор целочисленных значений, значение которого в каждой точке соответствует числу выровненных на референсный геном фрагментов секвенирования, покрывающих данную позицию. Такой сигнал может рассматриваться как временной ряд и, следовательно, к нему могут быть применены разнообразные алгоритмы и методы, изначально разработанные для анализа временных рядов и, например, поиска разладок и характеристических паттернов (так называемых мотивов) в них.

В данной работе мы представляем новый метод поиска устойчивых паттернов, ассоциированных с точками разрыва, соответствующих различным типам структурных вариаций, разработанный с использованием алгоритма KNN_Search (K-nearest neighbours) [10] и аппроксимации с помощью символьной агрегации (Symbolic Aggregation Approximation, SAX) [11]. На основе анализа данных, содержащихся в ранее разработанной нами базе данных SWaveform [12], были построены устойчивые повторяющиеся паттерны (мотивы) для имеющихся типов структурных вариаций. Также нами был разработан инструмент быстрого измерения SAX-расстояния между профилем покрытия генома и найденными нашим алгоритмом (или любыми другими) мотивами, позволяющий обнаруживать аберрантные регионы в сигнале покрытия, потенциально связанные со структурными вариациями.

МЕТОДЫ

Источники данных. База данных SWaveform [12] содержит профили сигнала покрытия секвенирования, полученные на основании анализа данных полногеномного секвенирования 911 образцов консорциума Human Genome Diversity Project [13, 14] со средним значением уровня покрытия 30x, общее число профилей составляет 13106216. Для всех имеющихся в базе данных типов структурных вариаций (делеции, дупликации, инверсии, инсерции и варианты копийности) нами были отобраны группы профилей сигнала покрытия в окне ±256 п.о. от точки разрыва (breakpoint). Для анализа были использованы только варианты с длиной, большей размера окна, т.е. превышающих 512 п.о. В случае протяженных вариаций, а именно делеций, дупликаций, инверсий и вариантов копийности, использовались профили, соответствующие их левой и правой границам. Также профили были сгруппированы в соответствии с генотипом (гомо- или гетерозигота) структурных вариантов, что позволило в конечном итоге получить 22 группы для поиска устойчивых паттернов.

SAX-преобразование и модифицированный алгоритм KNN. Аппроксимация с помощью символьной агрегации позволяет осуществить преобразование временного ряда, существенно снижающее размерность входных данных. Классификаторы, построенные на преобразованных данных, обладают обобщающей способностью без ущерба для качества классификации [10]. Для данного преобразования требуется определить два параметра — ширину окна W и размер алфавита A. Преобразование включает в себя следующие этапы:

1. Нормализация исходного сигнала на среднее и дисперсию.

2. Для заданной длины алфавита A генерируется интервальная сетка, где значения распределены по нормальному закону. Каждое i-е значение сетки является i/A-квантилем функции плотности нормального распределения.

3. Сигнал разбивается на W непересекающихся одинаковых сегментов и значения в каждом сегменте усредняются.

4. Каждое усредненное значение сегмента округляется до ближайшего значения из сетки из шага 2. Этому сегменту присваивается соответствующий номер сетки (от 0 до A–1).

Метод поиска K-ближайших соседей (K-Nearest Neighbour, KNN_Search) для заданного вектора v, заданного максимального расстояния T и известной функции расстояния ρ, возвращает список векторов, которые находятся на расстоянии от v не далее, чем на T. При анализе сигнала покрытия, последний рассматривается как одномерный временной ряд и разбивается скользящим окном на векторы, каждый из которых преобразуется в свое SAX-представление $\{v_0, ..., v_V\}$. Между этими векторами рассчитываются попарные расстояния. Из них 1.5% самых малых расстояний между векторами считаются максимальным значением T для алгоритма KNN_Search.

Все векторы, которые оказались ближе друг к другу, чем T, объединяются в группы, которые сортируются по убыванию числа элементов в них. Таким образом, группа, содержащая максимальное число объединенных векторов, соответствует преобладающему мотиву.

Поскольку входящий временной ряд может оказаться достаточно длинным, возникает достаточно ресурсоемкая задача измерения попарных SAX-расстояний между всеми векторами $\{v_0, ..., v_V\}$. Чтобы существенно снизить число операций по измерению расстояния, нами была произведена следующая модификация исходного алгорит-

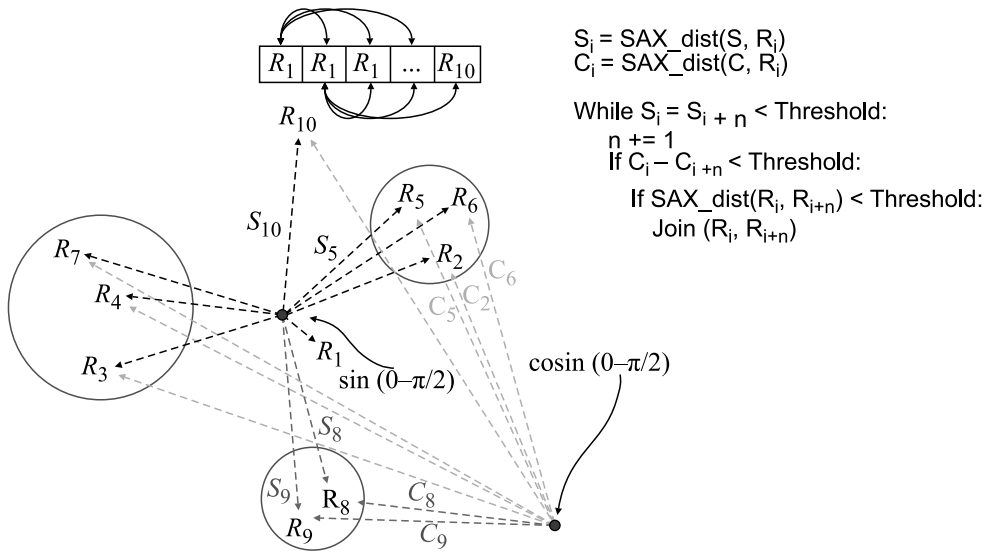

Рис. 1. Схематическое представление модифицированного алгоритма KNN. Расстояния от базисных векторов *S* и *C* до всех остальных векторов – S_i и C_i. Расстояние между любыми двумя векторами *R* может быть меньше порогового тогда и только тогда, когда разница их расстояний до каждого из референсных векторов меньше порогового.

ма KNN_Search. На первом этапе производится предварительный подсчет расстояний от произвольно выбранных нетождественных референсных векторов $\{b_0, ..., b_B\}$ до всех векторов списка $\{v_0, ..., v_V\}$. Затем, для каждой пары векторов v_i и v_j проверяются расстояния до референсных векторов: если существует хотя бы один референсный вектор b_k, такой, что $|\rho(v_i,b_k) - \rho(v_j,b_k)| > T$, то, согласно положениям, представленным в работе [10], $\rho(v_i,v_j)| > T$, т.е. расстояние между данной парой векторов v_i и v_j также больше порогового. Схематически данный метод показан на рис. 1. Такая модификация исходного алгоритма позволяет существенно сократить количество операций вычисления попарных расстояний. Вопрос выбора референсных векторов $\{b_0, ..., b_B\}$ остается открытым, однако с математической точки зрения их выбор может быть произвольным, поэтому в нашей модификации алгоритма используются два референсных вектора – SAX-трансформации функций sin и cos в диапазоне от 0 до π/2.

Построение повторяющихся паттернов (мотивов) на основе профилей сигнала покрытия. Поиск паттернов для каждой из 22 групп профилей сигнала покрытия, описанных выше, проводили следующим образом. Исходные профили группировались с применением алгоритма кластеризации K-means с числом кластеров равным 2. Число *k*, определяющее количество кластеров, оптимальное применительно к данной задаче, было определено в ходе серии вычислительных экспериментов с выборкой профилей, соответствующих структурным вариантам одного класса и генотипа. В ходе кластеризации с различным числом искомых групп рассчитывали силуэтный индекс, который послужил критерием выбора итогового разбиения. Для оценки расстояния между профилями сигнала был использован метод расчета, основанный на принципе динамической трансформации временной шкалы (Dynamic Time Warping, DTW) [15–17]. Существенной его особенностью является устойчивость к сдвигу расстояния между сравниваемыми сигналами, что позволяет сравнивать профили, в которых общий паттерн сигнала сохраняется, но существуют локальные искажения формы (т.е. растяжения или сжатия). Применение этого метода абсолютно оправдано в случае оценки расстояния между интервалами, содержащими точки разрыва, порожденные структурными вариантами, так как дают возможность компенсировать эффекты, вызванные неточностями в работе алгоритмов предсказания структурных вариаций. В зависимости от класса структурного варианта алгоритм выделяет, как правило, либо один преобладающий кластер (т.е. содержащий более 2/3 профилей сигнала) или же два примерно равных по размерам кластера. В случае наличия преобладающего кластера наименьший исключается из дальнейшего анализа.

На следующем этапе все профили подвергаются SAX-преобразованию, где размер алфавита составляет 24 символа, а число сегментов выбрано равным 64. На завершающем этапе внутри полученного кластера или кластеров производится поиск мотива, преобладающего в полученных SAX-трансформированных профилях с использованием описанного выше модифицированного алгоритма KNN. Итоговый мотив сохраняется в виде вектора усредненных значений для каждой

позиции. Также для каждого мотива сохраняются параметры SAX-преобразования.

Использование SAX-мотивов для поиска аномалий в сигнале покрытия. Для поиска участков генома, в которых сигнал покрытия обладает максимальным сходством с полученными SAX-мотивами нами был разработан программный комплекс saxmf (https://github.com/latur/saxmf), реализованный на языках C и Python и способный работать в многопоточном режиме. Входными данными для работы программы является сигнал покрытия в специальном формате BCOV [12], получаемый путем преобразования стандартных файлов выравнивания фрагментов секвенирования в форматах SAM/BAM/CRAM. Преобразование осуществляется с помощью программ mosdepth [18] и bed2cov (см. репозиторий (https://github.com/latur/saxmf)). Поисковая программа saxmf включает в себя SAX-мотивы, рассчитанные из данных Human Genome Diversity Project, однако имеется возможность поиска любых других мотивов, закодированных в формате JSON.

В ходе работы программы сигнал покрытия подвергается SAX-трансформации в скользящем окне, после чего вычисляется SAX-расстояние между ним и искомым мотивом. Полученные в ходе первичного поиска геномные интервалы, потенциально содержащие мотив, соответствующий точке разрыва, порожденной каким–нибудь структурным вариантом, на следующем этапе, объединяются в кластеры для идентификации региона, имеющего максимальную близость с искомым мотивом. Положение центральной точки данного региона ищется как взвешенное среднее по SAX-расстоянию по формуле:

$$BND = \frac{\sum_i Pos_i * SAX_i}{\sum_i (S - SAX_i)},$$

где S – максимальное SAX-расстояние внутри кластера, Pos_i — центр i-го региона, SAX_i – SAX-расстояние i-го региона до мотива. Результаты поиска выводятся в формате BED-файла, включающего координаты региона и SAX-расстояние до искомого мотива.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Проверка предложенного метода поиска паттернов, ассоциированных со структурными вариантами, была проведна на двух независимых наборах данных. В первом случае были использованы результаты полногеномного секвенирования проекта «Геномы России» [19]. Для 360 образцов были определены структурные варианты (делеции и дупликации) с использованием программы DELLY2 [20]. Затем нами был проведен поиск мотивов, ассоциированных с вышеуказанными типами структурных вариаций, предсказанных с использованием данных ресурса SWaveform. Мы исходили из того, что мотив может встречаться в геноме не только в позициях, идентифицированных DELLY2, но и в других, поэтому для нас было критически важным увеличить чувствительность, но не так существенно снизить специфичность нахождения точек разрыва. Регионы, обнаруженные в ходе анализа с помощью описанного пакета saxmf, были отфильтрованы по пороговому значению SAX-расстояния, изменявшегося от 0 до 10 с шагом 0.2. Для каждого случая рассчитывали расстояние Жаккарда между регионами, найденными saxmf и DELLY2, а также процент локусов, найденных только с помощью DELLY2. Результаты анализа демонстрируют высокий уровень совпадения результатов поиска saxmf с данными, полученными одной из самых распространенных программ для поиска структурных вариаций (см. рис. 2).

Таким образом, можно предварительно оценить пороговое значение SAX-расстояния для фильтрации результатов последующего анализа данных с помощью описанного метода.

В дополнение был проведен поиск регионов, содержащих паттерн, ассоциированный с левой границей делеций в эталонном образце генома HG002 [21, 22] (проект Genome In a Bottle [23] (GIAB), NIST). Под эгидой консорциума GIAB в этом геноме был проведен всесторонний анализ структурных вариаций с использованием различных технологий секвенирования, включающих использование длинных и коротких фрагментов, а также программ для предсказания структурных вариаций. Полученные результаты были проанализированы кураторами, что позволило охарактеризовать наиболее достоверные вариации [24]. В частности, для выбранного нами мотива (левая граница делеций) в данном геноме описано 5372 региона, из которых нами был успешно обнаружен 4651 регион (~86%). Этот результат показывает, что разработанная методика поиска паттернов сигнала покрытия может быть успешно применена для первичного поиска структурных вариантов в данных полногеномного секвенирования.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского научного фонда (грант No 20-14-00072).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

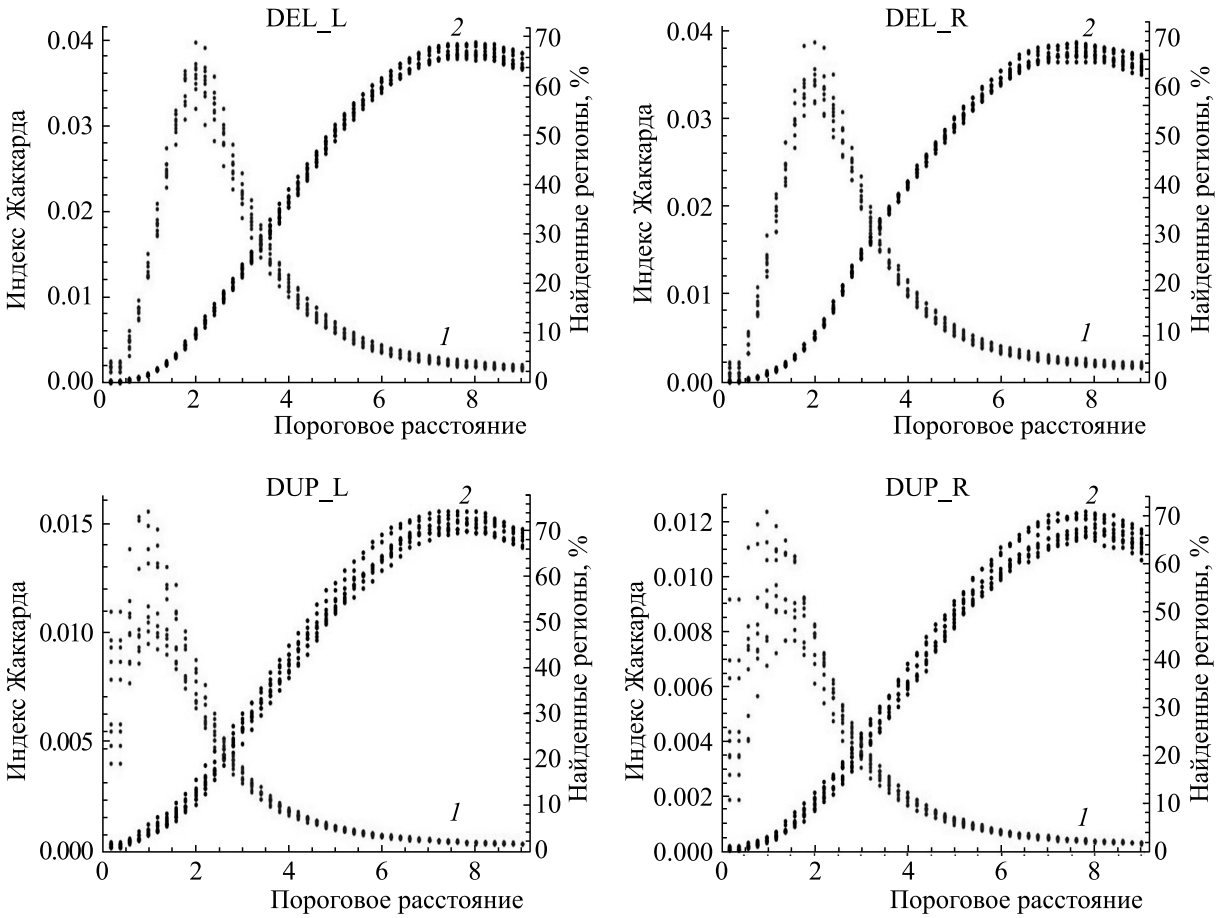

Рис. 2. Графики зависимости индекса Жаккарда (*1*) и процента найденных регионов (*2*) от порогового расстояния (ось абсцисс).

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. R. L. Collins, et al., Nature, **581** (7809), 444 (2020).
2. Y. R. Li, et al., Nature Commun., **11** (1), 255 (2020).
3. S. Girirajan, et al., Am. J. Human Genetics, **92** (2), 221 (2013).
4. M. Mahmoud, et al., Genome Biol., **20** (1), 1 (2019).
5. S. Kosugi, et al., Genome Biol., **20** (1), 117 (2019).
6. Z. Liu, et al., Genome Biol., **23** (1), 68 (2022).
7. H. Parikh, et al., BMC Genomics, **17** (1), 64 (2016).
8. A. Abyzov, et al., Genome Res., **21** (6), 974 (2011).
9. M. Rapti, et al., Brief Bioinform., **23** (2), bbac049 (2022).
10. Z. A. Aghbari, Data Knowl. Eng., **52** (3), 333 (2005).
11. S. Malinowski, et al., Lect. Notes Comput. Sci., 273 (2013).
12. *762 BGRS/SB-2022 Swaveform: a genome-wide survey of structural variation profiles*, Thirteen Int. Multiconference (2022).
13. A. Bergström, et al., Science, **367** (6484), eaay5012 (2020).
14. M. A. Almarri, et al., Cell, **182** (1), 189 (2020).
15. H. Sakoe and S. Chiba, IEEE Trans. Acoust. Speech Signal Process., **26** (1), 43 (1978).
16. F. Petitjean, A. Ketterlin, and P. Gançarski, Pattern Recogn., **44** (3), 678 (2011).
17. R. Tavenard, et al., J. Mach. Learn. Res., **21** (118), 1 (2020).
18. B. S. Pedersen and A. R. Quinlan, Bioinformatics, **34** (5), 867 (2018).
19. D. V. Zhernakova, et al., Genomics, 1 (2019).
20. T. Rausch, et al., Bioinformatics, **28** (18), i333 (2012).
21. J. M. Zook, et al., Nat. Biotechnol., 1 (2020).
22. A. Shumate, et al., Genome Biol., 1 (2020).
23. J. M. Zook, et al., Sci. Data, **3**, 160025 (2016).
24. L. M. Chapman, et al., PLoS Comput. Biol. **16** (6), e1007933-20 (2020).

Searching for Sequencing Signal Anomalies Associated with Genome Structural Variations

I.V. Bezdvornykh*, N.A. Cherkasov*, A.A. Kanapin*, and A.A. Samsonova*

**St. Petersburg State University, Universitetskaya nab. 7–9, St. Petersburg, 199034 Russia*

Genomic structural variations (SVs) are one of the main sources of genetic diversity. Structural variants as mutagens may have a significant impact on human health and lead to hereditary diseases and cancers. Existing methods of finding structural variants are based on analysis of high-throughput sequencing data and despite significant progress in the development of the detection methods, there is still a need for improving the identification of structural variations with accuracy appropriate for use in a diagnostic procedure. Analysis of the signal of sequencing coverage (i.e., the number of sequencing fragments that aligned to every point of a genome) holds new potential for the design of approaches for structural variations discovery, and can be used as time-series analysis. Here, we present an approach for identification of patterns in the coverage signal. The method has been developed based on algorithms used for analysis of time series data, namely KNN (K-nearest neighbour) search algorithm and the SAX (Symbolic Aggregation Approximation) method. Using the rich dataset encompassing full genomes of 911 individuals with different ethnic backgrounds generated by the Human Genome Diversity Project initiative, we constructed generalized patterns of signal coverage in the vicinity of breakpoints corresponding to various structural variant types. Also, with the benefit of the SAX models of the motifs we developed a software package for fast detection of anomalies in the coverage signal.

Keywords: genomics, sequencing, structural variations, signal analysis, SAX transformation