social networks relevant to infection dissemination. It allows to infer transmission direction without additional case-specific epidemiological data, identify transmission clusters and reconstruct transmission history. QUENTIN was validated on data from 33 epidemiologically curated HCV outbreaks, yielding an accuracy of 87% for inference of transmission directions, 90% and 99.6% for detection of outbreak sources and transmission clusters, 78% and 98% for reconstruction of transmission links and ancestries. It was applied to investigate HCV transmissions within communities of hosts with high-risk behavior using the data collected during the investigation of several HIV/HCV outbreaks associated with drug abuse and commercial sex work. QUENTIN allowed to extract information on structures of transmission clusters, roles of transmission mode, co-infection and host's gender in the infection spread.

In conclusion, study of intra-host viral populations, evolutionary modelling and complex network analysis allow for accurate inference of disease transmissions. QUENTIN is most useful for investigation of extensively sampled outbreaks caused by RNA viruses. Its superior performance over consensus-based approaches indicates importance of quasispecies analysis for molecular surveillance and outbreak investigation.

References

1. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data / P. Skums et al. // Bioinformatics. – 2017. – V. 34, N 1. – P. 163–170.

FAST ESTIMATION OF GENETIC RELATEDNESS BETWEEN MEMBERS OF HETEROGENEOUS POPULATIONS OF CLOSELY RELATED GENOMIC VARIANTS

Tsyvina V. A.

Georgia State University, Atlanta, e-mail: vyacheslav.tsivina@gmail.com

Many biological analysis tasks require extraction of families of genetically similar sequences from large datasets produced by Next-generation Sequencing (NGS). Such tasks include detection of viral transmissions by analysis of all genetically close pairs of sequences from viral datasets sampled from infected individuals or studying of evolution of viruses or immune repertoires by analysis of network of intra-host viral variants or antibody clonotypes formed by genetically close sequences. The most obvious naive algorithms to extract such sequence families are impractical in light of the massive size of modern NGS datasets. In this paper, we present fast and scalable k-mer-based framework to perform such sequence similarity queries efficiently, which specifically targets data produced by deep sequencing of heterogeneous populations such as viruses. The tool is freely available for download at **https://github.com/vyacheslav-tsivina/signature-sj**

Further we will use the following notation: $S=\{s_1,s_2,\dots,s_L\}$ – a sequence over the alphabet $\{A,C,G,T\}$; k -mer – any subsequence of length k ; k -segment – k -mer that starts at a position $1+ik$, $i=0,1,2,\dots$; $K(S)$ – the set of all k -mers of the sequence S ; $R(S)$ – the family of all k -segments of the sequence S (possibly with repetitions); $h(S,Q)$ – Hamming distance between two sequences S and Q ; $l(S,Q)$ – edit distance (Levenshtein distance) between two sequences S and Q . We say that two sequences S and Q are related if $l(S,Q)\le t$ or , where t is a given threshold.

Proposition. If $l(S,Q)\le t$, then $|K(Q)\cap R(S)|>m-t$, where $m=\left\lfloor \frac{L}{k} \right\rfloor$.

For Hamming distance, this proposition can be simplified to only k-segments that allow us to use k-segments of unequal size based on positions entropy. Using this proposition and specialized data structures in conjunction with several optimizations to reduce calculations we can obtain a set of related pairs of sequences in inter-sample as well as an intra-sample case in a short time period even for large samples.

We present an efficient signature-based tool to solve problems of edit or Hamming distance sequence retrieval for NGS data obtained from heterogeneous viral populations. It outperforms analog tools for edit distance [1] and for hamming distance [2] in several times.

References

1. Efficient detection of viral transmissions with next-generation sequencing data. / I. Rytsareva [et al.]; BMC genomics. – 2017. – V. 18, N 4. – P. 372.

2. Reconstructing antibody repertoirs from error-prone immunosequencing datasets. / A. Shlemov [et al.]; in Research in Computational Molecular Biology, Springer. – 2017. – P. 396.

ОБНАРУЖЕНИЕ НЕОДНОРОДНОСТЕЙ БОЛЬШИХ ДАННЫХ НА ОСНОВЕ ВЕЙВЛЕТ-АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

Абрамович М. С., Мицкевич М. Н.

Научно-исследовательский институт прикладных проблем математики и информатики, Минск, Беларусь, abramovichms@bsu.by

Пусть данные большого объема описываются моделью временного ряда $x_t, t=1,..,L$ с неоднородностями (разладками) в моменты времени $t_1\le t_2\le ...\le t_K$, K , K - число моментов разладки. Предложим следующую процедуру для обнаружения разладок временного ряда.

Пусть есть $N+1$ вычислительных устройств каждое с размером памяти 2^M. Одно выделенное устройство необходимо для сведения результатов обработки данных со всех остальных N вычислительных устройств. Считаем также, что $N<<2^M$.

Для построения вейвлет-разложения требуется объем памяти, сравнимый с длиной временного ряда, поэтому на каждом устройстве может быть обработан фрагмент временного ряда длиной 2^{M-1}. Тогда общая длина обрабатываемого временного ряда может достигать $N\times 2^{M-1}$.

Предлагаемый алгоритм состоит из стадий масштабирования и обобщения.

Стадия масштабирования включает следующие основные шаги:

– на N вычислительных устройств с размером памяти 2^M записываются фрагменты временного ряда длиной 2^{M-1};

– на каждом устройстве выполняется вейвлет-разложение до максимального уровня J , такого что $N\times 2^{M-J-1}<2^{M-1}$;