Optimizing pooling strategies for the massive next-generation sequencing of viral samples

Pavel Skums * Olga Glebova † Alex Zelikovsky † Ion Mandoiu ‡ Yury Khudyakov *
* Centers of Disease Control and Prevention
Atlanta, GA, USA
email: {kki8,yek0}@cdc.gov
† Georgia State University
Atlanta, Georgia 30303
email: alexz@cs.gsu.edu
‡ University of Connecticut
Storrs, Connecticut 06269
email: ion@engr.uconn.edu

***Abstract*—Next-generation sequencing (NGS) allows for analyzing a large number of viral sequences from infected patients, presenting novel prospects for studying the structure of viral populations and understanding virus evolution and epidemiology. It potentially provides an opportunity to implement large-scale molecular surveillance of viral diseases, which offers more precise estimations of epidemiological parameters, detection of transmissions and studying the structure of transmission networks, prediction of the epidemics progress and development of more effective vaccination strategies.**

A large-scale molecular surveillance requires sequencing of unprecedentedly large sets of viral samples. Although NGS has recently become less expensive and is expected to further decrease its cost in the future, massive NGS of tens of thousands of samples is still highly cost- and labor-intensive. Therefore it is highly important to develop a framework for identification of viral sequences from large number of samples using the smallest possible number of NGS runs.

Here we present a mathematical and algorithmic foundation of this framework.

I. Problem formulation and algorithms

The number of NGS runs for sequencing of n samples could be reduced by generation of m pools (i.e. mixtures of samples) with $m << n$ in such a way that every sample is uniquely identified by the pools to which it belongs. In the most basic case optimal pooling design problem could be formulated as follows:

Problem 1 (optimal pooling design problem).

Given: the set of samples $S = \{S_1, ..., S_n\}$ and a natural number $m << n$.

Output: the set of pools $P = \{P_1, ..., P_m\}$, $P_i \subseteq S$ represented by an $n \times m$- incidence matrix $M(P)$ with ith column equal to the characteristic vector of a pool P_i, such that the rows of $M(P)$ are pairwise different.

We show that Problem 1 has a solution with $m = \log(n) + 1$. However, under more realistic conditions the set of pools P should satisfy additional restrictions: $|P_i|$ is bounded above for every $i = 1, ..., m$ (since numbers of reads which could be obtained by every NGS technology is bounded and if large number of samples are mixed in one pool, some of these samples may be lost due to a PCR bias) and the number of pools containing each sample should be bounded below (to ensure sufficient coverage for each sample). Moreover, some samples could intersect (if they belong to the same transmission cluster). Obviously, putting potentially intersecting samples into same pools should be avoided. The relation of potential intersection between samples could be represented by a graph $G(S)$ with $V(G(S)) = S$ and $S_iS_j \in E(G(S))$ if and only if there is a confidence that samples S_i and S_j do not intersect.

Taking into account these constraints, the optimal pooling design problem could be reformulated as the following Minimum Clique Test Set Problem:

Given: an n-vertex graph $G = G(S)$, natural numbers m, l, k.

Output: the set of cliques $P = \{P_1, ..., P_m\}$ of G represented by an $n \times m$- incidence matrix $M(P)$ with ith column equal to the characteristic vector of a clique P_i, such that $|P_i| \leq k$ for every $i = 1, ..., m$, every vertex of G belongs to at least l cliques from P and the rows of $M(P)$ are pairwise different.

Minimum Clique Test Set Problem is a generalization of the known Minimum Test Set Problem, and, therefore, is NP-complete. We present ILP formulation and a heuristic algorithm for Minimum Clique Test Set Problem.

Acknowledgments

AZ and IM have been partially supported by two Collaborative Research Grant from Life Technologies, awards IIS-0916401 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture.

978-1-4673-1321-6/12/$31.00