Prediction of hosts from metagenomic viral sequences using alignment-free algorithms

Pathogenic bacteria are becoming more and more resistant to antibiotics, which often leads to ineffective treatment of diseases. Currently, there is a number of so-called superbacteria that show complete insensitivity to all types of antibiotics, even those of the last chance. This is particularly true for bacteria that cause hospital-associated infections, e.g. in infections caused by colibacilli or staphylococci. In the fight against superbugs, viruses called bacteriophages or phages can help. They are able to feed on bacteria leading to the total destruction of their host, while being harmless for humans. Phages have very specific hosts - usually one species of virus can multiply only in one bacterial species or only within a specific strain. For this reason, phages can be a precise tool designed to eliminate specific strains of superbacteria, without causing damage to other bacteria, for example to those responsible for the proper digestion of foods or the production of vitamins.

Although phages are the most abundant organic entities on Earth, knowledge of what specific bacteria they attack is very limited. Experimental techniques to detect the potential host of the virus are low-throughput as they require robust growth of the target host strain in laboratory. Computer programs have provided significant progress in predicting the bacterial host based on the DNA sequences of the virus. However, these programs are still not very effective - in the last two years, DNA sequences of more than half a million viruses have been discovered, of which only 8% have managed to find a probable bacterial host.

The main goal of the proposed project is to create an extensive computer program (metaprogram), which for the DNA sequence of a given virus will indicate the species of bacteria attacked by this virus. In contrast to existing programs that use single computational method for sequence comparisons, the proposed metaprogram will use the full arsenal of bioinformatic algorithms, and based on their calculations will make predictions of the most likely bacterial host. This type of tool can provide significant support in phage therapy during the design of mixtures of phages suitable for the treatment of a particular bacterial infection.

Given the increasing availability of molecular data about viruses, the next task of my project will involve application of the metaprogram on large scale - to identify bacterial hosts for almost a million viruses from a wide variety of environments: oceans, soil, freshwater, human gut and skin. Based on the obtained annotations of interaction pairs between phages and bacteria, I will reconstruct the phage-bacteria infection networks specific for a given environment (e.g., human gut). These interactive maps of viral infections will allow to define the ability of a given virus to infect different species of bacteria (host range) and thus allow the identification of cross-infections. Using mathematical models of graph theory, I will not only be able to isolate groups of bacteria subject to the most common viral infections, but also to predict how the functioning of the whole system will change when one or more bacterial species are eliminated from the network. Based on the computer simulation of the viral and host sequences, I will indicate sequence fragments that have the greatest impact on the phage-bacterial interaction.

As the final product of this project I will create a publicly available web application containing all information on virus-host interactions obtained in this project. The comprehensive information can be a reference for molecular biologists studying phage infections in various strains, species and taxonomic groups of bacteria. In addition, the tools included in this application will enable, among others, selection of the bacteriophages or groups of bacteriophages that the metaprogram predicted to be the most effective in attacking a specific bacterial groups of user interest. Such functionality of the web application can be a valuable source of information useful in developing therapeutic phage cocktails aimed at inactivation of specific pathogenic bacteria.