# Approach to mutation analysis of genetic sequences and building mutational database in context of HIV/AIDS and Tuberculosis treatment

Sergeev, R.S., Tuzikov, A.V.

*United Institute of Informatics Problems, National Academy of Sciences, Minsk, Belarus,*
*roma.sergeev@gmail.com, tuzikov@newman.bas-net.by*

Emergence of drug resistance has been recognized as a global threat since the era of chemotherapy began. This problem is extensively discussed in the context of HIV/AIDS as well as Tuberculosis treatment. Alterations in pathogen genomes are among the main mechanisms by which microorganisms exhibit drug resistance. Analysis of the reported cases and discovering new resistance-associated mutations may contribute greatly to the development of new drugs and effective therapy management.

Here we introduce a framework for analyzing data on drug-resistance obtained in own researches and supplemented by information accessible from other databanks. Our approach allows one to keep the knowledgebase up to date by taking into account that information is constantly refreshing since more and more results become available due to different research projects in this field. Data analysis procedure is arranged into a chain of steps or levels so that the output of each level is the input of the next.

On the lower level new original pathogen's genome sequences are analyzed. This step is aimed to find site covariations and identify signals of recent positive selection in target genes under certain conditions (e.g. specific drug or treatment regimen). Methods for genome-wide studies relying on haplotype likelihood ratio test [1] or finding coevolving sites using evolutionary models [2, 3] can be applied to determine correlated residues.

The next level of analysis is supposed to reveal associations of genome variations with results of phenotype resistance tests to known drugs. The underlying methods imply approaches starting from modifications of Fisher's exact test to advanced statistical techniques like efficient mixed-model association test [4] which can adjust for confounding effects from phylogeny and site covariations.

The higher-level algorithms are purposed to construct a probabilistic dependency network in order to structure associations discovered at the previous levels. As soon as associations are based on probabilities, they are represented as weighted arcs between variables. In this context variables correspond to presence or absence of amino acids in codons, received drugs and treatment outcomes. Information on drug resistance from other studies and public databases can be added at this level as supplementary associations in the network which gives an advantage of taking into account all available data. Inference algorithms designed for Bayesian and Markov networks are used to retrieve information trough queries to the networks composed from sets of observed and requested variables. The dependencies inside the network can be updated as soon as new data appear.

Elements of this approach are used in current project performed in collaboration with NIAID of NIH through a CRDF BOB-31120-MK-13 project to establish the Belarus tuberculosis database (http://tuberculosis.by) and conduct comprehensive study of obtained MDR and XDR TB strains.

Acknowledgements

1. P.C.Sabeti et al. (2002) Detecting Recent Positive Selection in the Human Genome from Haplotype Structure, *Nature*, **419:**832–837.

2. D.D.Pollock et al. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure, *J. Mol. Biol.*, **287:**187–198.

3. R.S.Sergeev et al. (2011) Algorithms for mutation analysis of HIV-1 subtype A primary protein sequences, *Informatics*, **3(31):**88-97.

4. X. Zhou, M. Stephens (2012) Genome-wide efficient mixed-model analysis for association studies, *Nature Genetics*, **44(7):**821-824.