Generating hit molecules against pathogenic DNA motifs in silico

Genome (all the information that is encoded in DNA or RNA and is capable of being passed on to an offspring) is the blueprint of an organism. It consists of a sequence of letters A, G, C, T/U, namely the nucleic acid bases. With recent advancements in sequencing technologies, genomes of pathogens (a bacterium, virus, or other microorganism that can cause disease), non-pathogens (organisms incapable of causing disease), humans, etc. are becoming available (www.ncbi.nlm.nih.gov). One can use this genome information to identify DNA/RNA sequences which are unique to pathogen but not for non-pathogens and humans and target these sequences or their product (RNA/protein) for drug discovery to eliminate disease. This is the scope of this research work presented here (www.scfbio-iitd.res.in/PSDDF/).

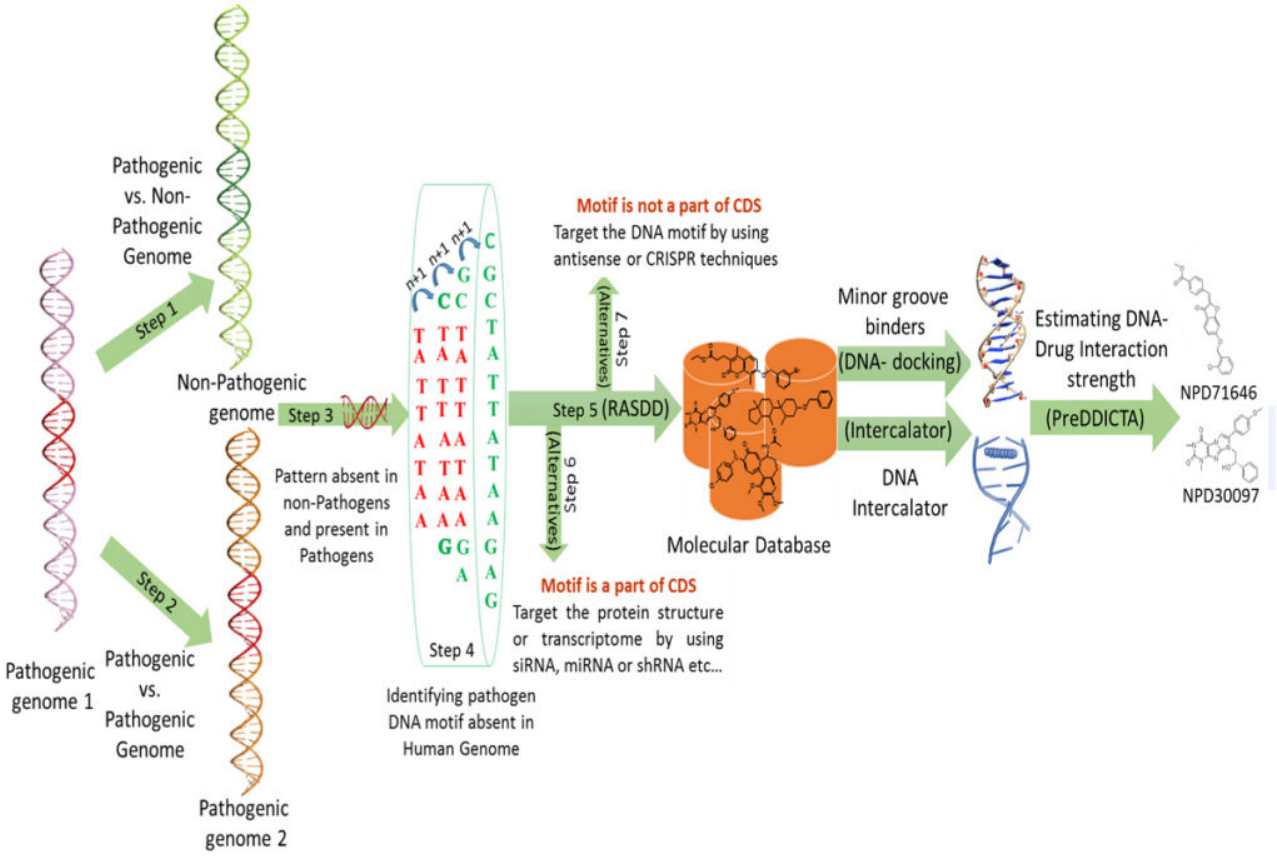

Fig. 1. A computational protocol for identifying unique DNA targets and their potential binder(s).

The first step of the protocol is the identification of unique DNA sequence(s) by comparing the pathogenic genome of interest with the non-pathogenic genome. In the second step i.e. “Pathogen

vs. Pathogen", DNA sequences unique to pathogens of interest are compared with other pathogenic genomes as sub-targets. If the unique DNA sequence (motif) is present in a majority of the selected pathogenic species/strains of interest and absent in all the non-pathogenic species/strain, then that motif should have a crucial role in pathogenicity. In the third step, these unique sequences are also compared with the human genome in "Pathogen vs. Human" step in order to obtain occurrence/non-occurrence of the motif in the human genome. If the sequences thus identified occur in the human genome, in step four, the sequence of the pathogen is extended until its occurrence in the human genome becomes zero.

In order to use the unique sequence as a drug target, we integrated several in-house developed computational tools into a drug discovery pipeline which automates the journey from the DNA motif to potential lead molecules as illustrated in Figure 1. If one desires to target the identified DNA sequence unique to the pathogen, step 5 is performed, which comprises scanning identified motif against a million compound library using RASDD methodology (www.scfbio-iitd.res.in/software/drugdesign/rasdd.jsp) followed by docking (www.scfbio-iitd.res.in/dock/dnadock.jsp) or intercalation (www.scfbio-iitd.res.in/intercalate) to generate potential lead/candidate drug molecules. Alternately, If the identified DNA motif is a part of a sequence coding for protein (CDS), then one can target its protein (www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp) or the mRNA by using siRNA and shRNA (step-6), while if the sequence is not a part of CDS then, PNAs, CRISPR or other molecular biology techniques can be used for specifically targeting the DNA motif (step-7).

As a case study, we analyzed the genome of a pathogenic strain of *Mycobacterium tuberculosis* H37Rv. We compared this genome with genomes of ten other non-pathogenic and pathogenic strains to identify those motifs which are exclusively present in H37Rv. We found one of the motifs, namely TATTATAA, was absent in all the ten non-pathogenic strains and present in eight out of ten pathogenic genomes of Mtb which makes it a unique drug target. The motif identified is a part of PPE family gene PPE54. This protein probably plays an important role in host phagosome maturation arrest and hence identified as a high confidence drug target (Brodin, P. *et al.*, *PLoS Pathog.* 2010, Sassetti, C. M. *et al. Mol. Microbiol.* 2003).

By using the DNA-targeted drug discovery pipeline described in Figure 1, a few potential lead molecules were identified against the selected DNA motif. In this, DNA sequence is converted into its structure by using *DNA Sequence to structure* tool (www.scfbio-iitd.res.in/software/drugdesign/bdna.jsp). Rapid scanning was performed against the natural product database by using *RASDD* tool (www.scfbio-iitd.res.in/software/drugdesign/rasdd.jsp). Top most candidate molecules were docked in the minor groove of the DNA by using *DNA-Dock* tool (http://www.scfbio-iitd.res.in/dock/dnadock.jsp). DNA-drug complex interaction strength was calculated by using *PreDDICTA* tool (www.scfbio-iitd.res.in/software/drugdesign/preddictanew.jsp). MD simulation was performed using *AMBER* software package (Case, D. A. *et al.*). On the basis of the above study, we found that molecules NPD71646 and NPD30097 with -8.5 kcal/mol and -9.6 kcal/mol binding free energy values respectively are good potential lead molecules that can

specifically bind to the identified pathogenic DNA motif.

Overall, DNA targeted drug discovery protocol reported here can identify sequences unique to pathogens and generate new drug candidates against them by utilizing genome level information.

Akhilesh Mishra[1,2], ***Pradeep Pant***[1,3], ***B. Jayaram***[1,2,3]
[1]*Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, India*
[2]*Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, India*
[3]*Department of Chemistry, Indian Institute of Technology Delhi, India*

Publication

A computational protocol for the discovery of lead molecules targeting DNA unique to pathogens.
Mishra A, Pant P, Mrinal N, Jayaram B
Methods. 2017 Dec 1