Supplementary Materials

Targeting Plague Virulence Factors: A Combined Machine Learning Method and Multiple Conformational Virtual Screening for the Discovery of *Yersinia* Protein Kinase A Inhibitors

Xin Hu[#], Gerd Prehna[#], C. Erec Stebbins*

[#]These authors Contributed Equally to this work

Laboratory of Structural Microbiology,The Rockefeller University, New York, NY 10021
E-mail: stebbins@rockefeller.edu

A. Supplementary Figures

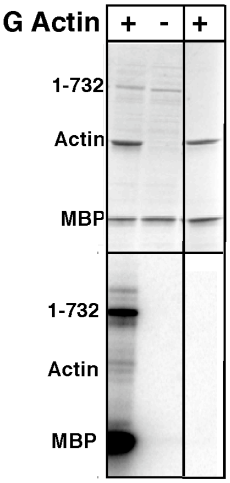

Figure S1: Activity assay of purified YpkA 1-732. YpkA was incubated in G reaction buffer with MBP and G actin. Full length (1-732) YpkA was assayed for autophosphorylation and phosphorylation of MBP with radio-labeled γ-^{32}P ATP (experimental methods). The upper gel shows the coomassie stain of the experiment whereas the bottom gel shows the autoradiograph of the assay. Positions of the various proteins are labeled on the left of the gel images. The presence or absence of G actin form is indicated with a + or – respectively.

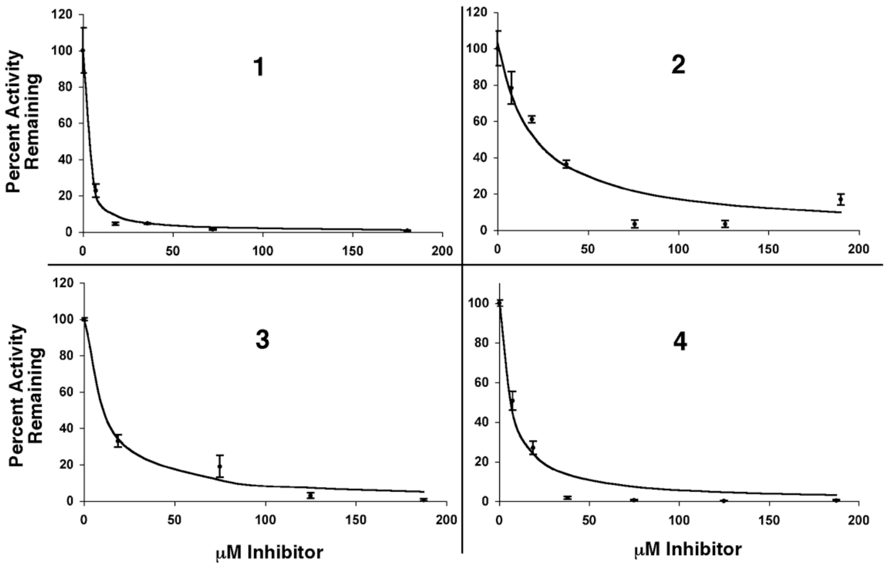

Figure S2. Inhibitor data of YpkA and the four top inhibitors as listed in figure 4. Activity was assayed as percent total phosphorylation remaining versus inhibitor concentration. The data was generated and the curve fit was performed as described in experimental methods. Each curve is labeled with its respective compound from figure 4.

B. Methods

Sequence Alignment and Homology Modeling of the YpkA Kinase Domain. The amino acid sequence of *Yersinia pestis* YpkA was obtained from SWISS-PROT database (accession number Q9RI12). The minimum kinase domain of YpkA (sequence (115-428)) was used. Multiple sequence alignment was performed applying several structure-based approaches (FUGUE, 3D-PSSM, and SUPERFAMILY) in order to get a reliable alignment and suitable templates. The mitogen-actived protein kinase (MAPK) showed the best homology to YpkA with 20% of sequence similarity, as illustrated in Figure 3. The sequence alignment schema was also optimized during model construction. Two structural models were constructed based on different templates of MAPK. Model A used the apo structures of p38 (PDB id 1p38 and 1erk), and model B used the templates of ligand-bound complexes with induced fit at the ATP binding site (PDB id 1a9u and 3erk). The structural models were built using program MODELLER version 7.0. The 3D fold of the generated models was verified by the program PROSA II, and problematic regions were spotted from the analysis of PROSA II energy profile. The resulting models were further optimized with an iterative approach until no significant improvements were obtained. The stereochemical quality and protein structure of the final models were validated by the programs PROCHECK and PROSA II.

Support Vector Machine and Database Filtering. The machine learning SVM model was derived using the program ADMET/Predictor (SimulationsPlus, Lancaster, CA). Molecular descriptors were calculated consisting of 276 descriptors from the 3D structure. The non-linear SVM model was derived from the molecular descriptors of the training set of 4584 compounds in distinguishing the active (364) and inactive (4220) compounds. The testing data set comprising 844 compounds with 175 active inhibitors. Default parameters of ADMET/Predictor were used for the SVM model training and testing. For database filtering with the SVM model, an in-house database collections consisting of more than 2 million compounds was screened, and a kinase-focused library of ~200,000 compounds was obtained.

MD Simulations. MD simulations were performed using the SANDER module of the AMBER 8.0 package. The modeled systems of YpkA were carried out in vacuo and a minimization of 1000 steps (250 steps of steepest descent followed by conjugate gradient) was conducted prior to MD simulations. The simulated system was first subjected to a gradual temperature increase from 0 K to 300 K over 100 ps, and then equilibrated for 500 ps at 300 K, followed by the production run of 2 ns length in total. Constant temperature was maintained using Langevin dynamics method in AMBER 8.0. The resulting trajectories were analyzed using the PTRAJ module. For the conformational sampling, 500 conformers were extracted from the trajectories at 20 ps interval, and clustered using the hierarchical clustering method of MMTSB Tool Set.

Database Virtual Screening. Multiple conformational virtual screening was performed using the program FlexE of Sybyl 7.0 running on a 36-processor LINUX cluster in parallel. A total of eight conformers of YpkA sampled from MD simulations were used. The kinase focused library was docked to the ensemble of protein structures and ranked according to the FlexX score. The top 5% compounds were extracted and re-ranked using X-Score. The top 1000 compounds from both original FlexX score and X-Score were visually inspected in terms of overall fit, key interactions in the binding site, as well as the structural complexity and diversity of compounds. A total of 45 compounds were finally selected for biological evaluation.

Protein Purification. The *Y. pseudotuburculosis* YpkA gene was cloned from the *Yersinia* virulence plasmid (generously provided by J. Bliska, State University of New York at Stony Brook) as an N-terminal fusion with glutathione-S-transferase (GST). GST-YpkA was further purified by affinity chromatography using glutathione sepharose beads. To remove YpkA from GST and the affinity resin, YpkA was cleaved using an N-terminally tagged GST site specific protease construct, liberating YpkA into the supernanent buffer (20mM Tris pH7.5 200mM NaCl 1mM DTT 1mM EDTA). The resulting material was decanted from the affinity resin, concentrated, and flash frozen for storage.

Radiological Assays. Purified YpkA constructs were tested for autophosphorylation and phosphorylation of MBP (Mylein basic protein) in the presence of G actin (Cytoskeleton, Inc.). Reactions were prepared with an excess of MBP relative to YpkA and preincubated in G reaction buffer (20mM Hepes pH7.5 10mM $MgCl_2$ 1mM DTT and 0.1mM ATP 1μCi γ-^{32}P ATP). Each assay was initiated with an access of G actin relative to YpkA and then incubated at room temperature for 30 minutes. The reactions were then visualized by SDS-PAGE and exposure on a Storage Phosphor Screen / Typhoon Scanner (Molecular Dyanmics, Inc.).

Inhibitor Assay. The 45 top hits from virtual screening were tested for inhibitory activity against full length YpkA initially at 450 μM in the presence of G actin (see Radiological assays). The results of each trial were visualized using the Bio-blot apparatus (Biorad) and exposure on a Storage Phosphor Screen / Typhoon Scanner (Molecular Dynamics). The IC_{50} of the best inhibitors was determined by creating serial dilutions of inhibitor from 0 μM to 181 μM (final reaction concentration) spread over 5 to 6 independent points measured in triplicate. A total of 200 ng of YpkA 1-732 was used per 16 μL reaction volume. After visualization (as above) the intensity of each band was intergraded by the program ImageQuant (Molecular Dynamics). This data with the program, LSW Data Analysis Tool Box version 1.1.1 in MicroSoft Excel, was used to generate curves for IC_{50} calculation. MAP kinase 1 (MAPK) and Protein kinase Cα (PKC) were obtained from Calbiochem and the same protocol followed as YpkA to determine inhibitor IC_{50} values. Each was added in the same molar ratio as YpkA in the final reaction conditions. All small molecules were obtained from Chembridge. The compound IDs are as follows in order, 5362253, 5221100, 5890215, and 7469777.