Using Metabolomic and Transportomic Modeling and Machine Learning to Identify Putative Novel Therapeutic Targets for Antibiotic Resistant Pseudomonad Infections

Peter E. Larsen, Frank R. Collart, and Yang Dai

Abstract— **Hospital acquired infections sicken or kill tens of thousands of patients every year. These infections are difficult to treat due to a growing prevalence of resistance to many antibiotics. Among these hospital acquired infections, bacteria of the genus Pseudomonas are among the most common opportunistic pathogens. Computational methods for predicting potential novel antimicrobial therapies for hospital acquired Pseudomonad infections, as well as other hospital acquired infectious pathogens, are desperately needed. Using data generated from sequenced Pseudomonad genomes and metabolomic and transportomic computational approaches developed in our laboratory, we present a support vector machine learning method for identifying the most predictive molecular mechanisms that distinguish pathogenic from non-pathogenic Pseudomonads. Predictions were highly accurate, yielding F-scores between 0.84 and 0.98 in leave one out cross validations. These mechanisms are high-value targets for the development of new antimicrobial therapies.**

I. Introduction

Hospital Acquired Infections, or HAIs, cause or contribute to approximately 100,000 deaths a year in the United States [1]. These infections are particularly dangerous as their hospital environments predispose these pathogenic populations to acquire multiple antibiotic resistances [2]. To combat this rising threat of antibiotic-resistant HAIs, we developed computational techniques for identifying molecular targets for new antimicrobial therapies [3]. This approach is supported by enhanced capabilities to completely sequence the genomes of previously uncharacterized bacterial from clinical, as well as environmental, isolates provides an opportunity to investigate the molecular mechanisms of pathogenicity in HAIs

Pseudomonads, a genus in the class of gram negative Gammaproteobacteria, are nearly ubiquitous in soil and aquatic environments, and some Pseudomonads are opportunistic human pathogens [4]. For example, *P. aeruginosa* is the second most common cause of hospital acquired pneumonia and the most common cause of pneumonia in intensive care units [5, 6]. Pseudomonads have a broad capacity to synthesize a variety of secondary metabolites [7] and pathogenicity in these organisms is multi-factorial and combinatorial [8]. At the time of this study's inception, there were 44 completely sequenced and annotated Pseudomonads in the NCBI databases, although that number continued to grow. The pathogenicity of some Pseudomonads coupled with their metabolic versatility, adaptations to diverse environments, and the number of available fully sequenced and annotated genomes, make this genus an excellent subject for developing computational methods to identify the molecular mechanisms that distinguish pathogenic from non-pathogenic Pseudomonads.

Here, we present a supervised machine learning approach to identify specific molecular signatures that underlie pathogenicity-related phenotypes using inputs from genomic, metabolomic, and transportomic predictions. We generated metabolomic and transcriptomic predictions for Pseudomonads through application of computational biology approaches previously developed in our lab. For this study, metabolomic models identify all possible metabolites which are capable of interacting with proteins for all unique enzyme functions identified in the set of Pseudomonad genomes. Transportomic models integrated with metabolic pathways and functional assignments can be used to identify the complete set of ligands for which there are transmembrane transporters capable of mediating ligand exchange across the cell membrane. Using Support Vector Machines (SVMs), we identified the key metabolomic and transportomic signatures and molecular mechanisms most predictive for pathogenicity. Results indicate that SVMs can predict a Pseudomonad's pathogenicity from computational models with both high accuracy and high precision (Leave-one-out F-scores averaged 0.84 and the best F-score observed was 0.98). Transportomic models were found to be most predictive of pathogenicity, indicating that the key adaptation for pathogenicity is in a Pseudomonad's ability to detect and manipulate its environment.

Combining SVM predictions and the comparative metabolomic and transportomic models generates a system scale understanding of how specific molecular mechanisms relate to pathogenic phenotype and infers possible molecular pathways that may be vulnerable to targeted therapeutic agents. Our results indicate that combining genomic sequence data, comparative metabolomic and transportomic models, and SVM provide potentially powerful tools for combating HAI-related illness and death.

This research has been funded by the U.S. Department of Energy, Office of Biological and Environmental Research, under Contract DE-AC02-06CH11357.

P. E. Larsen is with the Biosciences Division, Argonne National Laboratory, Argonne IL 60439 (phone: 630-252-3984; email: plarsen@anl.gov).

F. R. Collart is with the Biosciences Division, Argonne National Laboratory, Argonne IL 60439 (email: fcollart@anl.gov).

Y. Dai is with the Department of Bioengineering, University of Illinois at Chicago, 60612 (e-mail: yangdai@uic.edu).

978-1-4244-7929-0/14/$26.00 ©2014 European Union

II. Materials and Methods

A. Characterized Pseudomonad Genomic Sequences

There were 44 fully sequenced Pseudomonas strains, annotated in The Kyoto Encyclopedia of Genes and Genomes (KEGG) [9, 10] and available from NCBI (ftp://ftp.ncbi.nih.gov/genomes/) at the time the calculations in this manuscript were performed. All subsequent analysis used the gene models and predicted coded proteins (.faa files in NCBI genomic sequence database). Pathogenicity is described as a function of Pseudomonad species [4]. *P. aeruginoa*, *P. mendocina, P. putida, and P. syringae* are annotated as pathogens, representing 33% of sequenced Pseudomonad species (4 out of 12 species with at least one example of a sequenced genome) and 61% of total available sequenced genomes (27 out of the 44 genomes).

B. Re-annotation of Pseudomonad Predicted Proteins

As the genomes in the current study have been sequenced over a span of 14 years, all Pseudomonas genomes were re-annotated using a specific set of well characterized protein sequences. This approach insured that all functional assignments for predicted proteins from genomic sequence data were derived using uniform conditions and consistent functional nomenclature. KEGG was used as a source of annotated protein sequences of metabolic enzymes and transmembrane transporter activities that use a common nomenclature for metabolites and ligands. KEGG annotations are compatible with metabolomic and transportomic modeling approaches described below. For enzyme functions annotations, Enzyme Commission (EC) annotation numbers [11] were used. A database of bacterial enzymes annotated with EC numbers and associated with specific reactions in KEGG metabolic pathways (downloaded May 16, 2011) was collected. This set contains 754,066 protein sequences, annotated with 2,605 unique EC number enzyme function annotations. For transmembrane transporter function KEGG Orthology (KO) annotations were used. The specific annotations considered were annotated as membrane transport (Ko02000), ABC transporters (Ko02010), Major Facilitator Superfamily (MFS), Phosphotransferase System (PTS) (Ko02060), and 'Other' (pores/ion channels, electrochemical driven transporters, transmembrane electron carriers, and 'other'). The set of transmembrane transporters is comprised of 164,321 protein sequences, annotated with 891 unique transporter/sensor functions, and are associated with the transport of 272 unique ligands. It is possible for a single protein sequence to be present in both the set of enzymes and the set of transmembrane transporters. Protein annotations were assigned to single best BLAST-P hit with e-values < $1x10^{-10}$ (NCBI-Blast 2.2.23+). Enzyme function profiles for Pseudomonads were generated as lists of all possible enzyme or transmembrane transporter annotations and the number of genes in each Pseudomonad for that function.

C. Predicted Relative Metabolite Turnover (PRMT)

Predicted Relative Metabolic Turnover (PRMT) uses enzyme function profiles for quantifying the relative metabolic turnover between two metabolomes [12]. PRMT scores do not indicate rates of reaction or predict quantities or concentrations of compounds in a metabolome. PRMT scores were used to identify the differential metabolic capacities that distinguishes pathogenic from non-pathogenic Pseudomonads

.

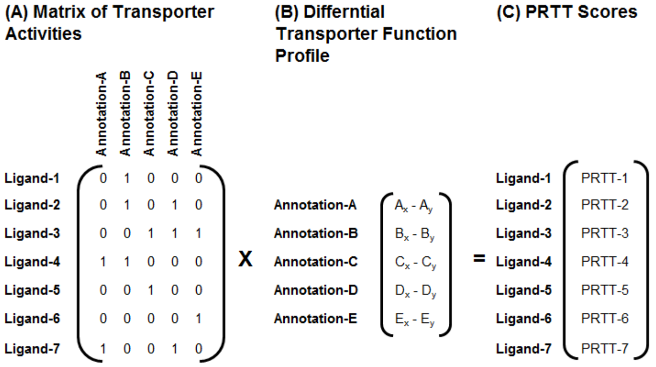

Fig 1. Predicted Relative Transmembrane Transport (PRTT) score calculation summary. (A) is a matrix of specific ligands (Ligands 1-7) associated with specific transporter annotations. Each annotation can be associated with one or more ligands. The matrix is constructed such that a '0' indicates that a ligand is not transported and a '1' indicates that a ligand is transported by a transporter of a given annotation. For example, in the cartoon above, ligands 4 and 7 are transported by a transporter annotated with function 'A'. (B) is an array of the difference between the number of times each annotation appears in genomes '*x*' and '*y*'. (C) is an array of PRTT-scores calculated by taking the product of matrix in (A) and array in (B). PRTT score is a quantification of the relative ability for an organism with genome *x* relative to an organism with genome *y* to transport ligand *n*.

Two sets of PRMT models were generated. The first used the complete set of enzyme functions identified in the set of the 44 Pseudomonad genomes. The second set was restricted to the subset of enzyme activities that is also present in KEGG Biosynthesis of Secondary Metabolites pathway (KEGG map01110). Both sets were calculated using the average enzyme function count across all Pseudomonads. In this analysis, the reference genome *y* is always calculated as the average unique enzyme function counts of all Pseudomonad genomes.

D. Predicted Relative Transmembrane Transport (PRTT)

Predicted Relative Transmembrane Transport (PRTT) is a system scale metric that quantifies relative ability of an organism to transport specific metabolites across the cellular membrane. PRTT-scores are calculated as a special case of PRMT-scores. The application is described here for the first time and is summarized in Fig. 1.

Required input for PRTT is vectors of length *ko* of the $\log_2$-transformed number of transmembrane transport function representation in genomes, where *ko* is the number transporter function annotations in the transportomic model. Also required is transporter ligand specificity matrix $\boldsymbol{T}$ of size *l* by *ko*, where *l* is the total number of ligands present in transporter ligand specificity matrix.

The PRMT score vector between transportomes encoded by genomes *x*, and *y* is given by:

$$\overrightarrow{PRTT_{x,y}} = \boldsymbol{T}(\overrightarrow{k^x} - \overrightarrow{k^y}) \tag{1}$$

The resulting set of values, $\overrightarrow{PRTT_{x,y}}$, is a vector of PRTT-scores of length *l* for the comparison of predicted relative transmembrane transport of each ligand in ***T*** for transportome encoded by genome *x* relative to genome *y*. A positive PRTT score indicates an increased relative capacity for transmembrane transport of a specific ligand in the transportome in genome *x* relative to genome *y*. A negative PRTT score indicates a decreased relative capacity of transmembrane transport of a ligand. PRTT scores do not indicate absolute rates or directionality of transmembrane transport activity. PRTT scores were used to identify what capacities to sense or manipulate the environment via transmembrane transport distinguish pathogenic from non-pathogenic Pseudomonads. As with PRMT scores, all PRTT scores were calculated using reference genome *y* calculated as the average transmembrane transport function counts of all Pseudomonad genomes.

E. *Support Vector Machine Learning*

SVMs are supervised learning methods that analyze data and recognize patterns. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

Enzyme function profiles, PRMT scores, secondary metabolism PRMT scores, or PRTT scores used as features in training SVMs were non-zero in more than half of the genomes and had a standard deviations greater than 0.2 (i.e. features were present in most Pseudomonas genomes and there is variation in feature values).

Accuracy of SVM models was calculated using Leave One Out Validation (LOOV). In LOOV, a single genome is used as a validation set and the model is trained on the remaining data with a 10-fold cross-validation procedure and linear kernels. The process is repeated until each of the 44 Pseudomonas genomes was used as validation set once. For generation of SVM, R-project (http://www.r-project.org/) and package 'e1071' v1.6-1 (August 29, 2013, http://cran.r-project.org/web/packages/e1071/index.html) were used. To identify the most predictive features in SVM for each feature type, enzyme function count, PRMT, and PRTT were ranked by the weight in the corresponding trained model. SVMs were retrained using 10%-90% of the highest ranked features.

To quantitate the predictive power of SVMs, F-score was used. F-score is a metric that combines precision (specificity) and recall (sensitivity) of predictions and is calculated as follows:

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2}$$

Where,

$$Precision = \frac{tp}{tp+fp}\ ,\ Recall = \frac{tp}{tp+fn} \tag{3,4}$$

In precision and recall, *tp* is the number of true positives, *fp* is the number of false positives, and *fn* is the number of false negatives in predictions.

III. Results

A. *Pseudomonad Metabolomic and Transportomic Models*

The re-annotated genomes identify 1092 unique enzyme activities and 195 transmembrane transport annotations that were present in at least one Pseudomonad. 606 of the enzyme functions and 169 transporter functions were present and showed variation in representation in over half of the re-annotated genomes. Metabolic and transportomic models for Pseudomonads were constructed using PRMT and PRTT based on the re-annotated Pseudomonad genomes. The complete metabolomic model is comprised of 6642 enzymatic transformation interactions of 3688 metabolites, of which 2143 metabolite PRMT scores were non-zero and showed variation in over half of pseudomonad models. The secondary metabolism model is comprised of 1649 enzymatic transformations between 1494 metabolites carried out by 301 unique enzyme functions. There are 714 secondary metabolite PRMT scores that were non-zero and showed variation in over half of models. The non-zero feature values with standard deviations over 0.2 were selected for training SVMs. There were 606 enzyme functions, 2143 metabolite PRMT scores, 714 secondary metabolite PRMT scores, and 169 ligand PRTT scores that fit these criteria.

B. *Prediction F-scores*

Using 100% of input features, transportomic model was the most predictive of Pseudomonad pathogenicity, followed by secondary metabolism model, metabolomic model, and enzyme function profiles (Fig. 2A). This trend persisted across the majority of conditions for which nodes were restricted to high-weight SVM features, excepting a region for which enzyme function profile outperformed total metabolome (between top 30-60% of features) and a single point at which secondary metabolism outperformed transportomic model (top 10% of features) (Fig. 2B). Reducing the number of features to those with highest SVM feature weight improved prediction of SVM for every input type. There were 121 unique enzyme functions, 214 metabolites, 71 secondary metabolites, and 40 transmembrane transported ligands that are most predictive of Pseudomonad pathogenicity identified by SVM model analysis. The complete list of predictive features can be downloaded from http://www.bio.anl.gov/molecular_and_systems_biology/S1_PathogenFeatures.txt.

IV. Discussion and Conclusion

Our analysis suggests genomic features of bacteria such as enzyme function profile, metabolomic models, and transportomic model data can be used to accurately distinguish pathogenic from non-pathogenic Pseudomonads. Molecular characteristics supportive of transportomic models are the most informative feature for predicting pathogenicity in Pseudomonads and enzyme function profile is the least predictive.

These results have significant biological implications for experimental identification of specific targets as possible novel therapeutic targets. While experimental validation of

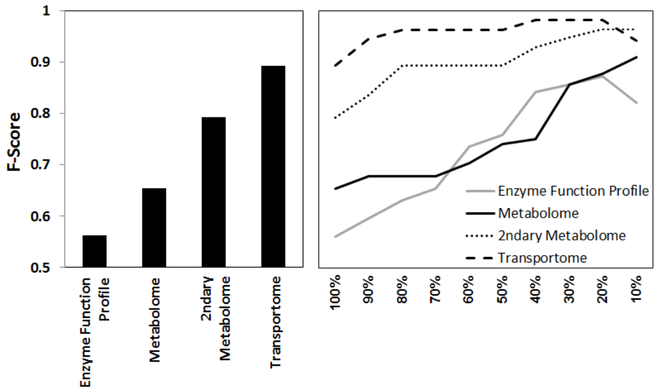

Figure 2. F-scores of SVM predictions. (A) F-score for predictions using complete set of features for each input type. (B) F-scores for predictions using only the high-weight features identified in (A) as percentages of total features for each input type.

computational results is beyond the scope of this study, a number of transported ligands predictive for pathogenicity are suggestive of putative targets for therapeutic intervention in Pseudomonad infections. Many examples of pathogenesis-specific transporter functions in Pseudomonads obtained from our SVM models have been previously identified as important to pathogenicity in similar systems. Colicins, for example, are released into the environment to reduce competition from other bacterial strains [13]. Interfering with an infectious Pseudomonad's ability to compete with native gut flora is a potential approach to treating Pseudomonad infections. Homoserine is a signaling compound implicated in the formation of biofilms and a contributor to resistance to antibiotics [14]. Drugs that interfere with Pseudomonad's ability to form biofilms could substantially reduce Pseudomonads native antibiotic resistances [15]. Transmembrane sugar transporters are identified by this method and sugar metabolism has been previously identified as a virulence factor in enterobacteria [16]. Zinc, Mg^{2+}, and K^{+} cation transport abilities are highly predictive of Pseudomonad pathogenicity and novel mechanisms involving direct metal intoxication of microorganisms have begun to be understood as important components of the immune system [17]. The agreement between predictions and previously reported mechanisms of pathogenicity lends credence to the hypothesis that some of the predictions that are not supported by previous literature may be novel discoveries and valuable targets for drug development.

This approach using SVM models trained on features from metabolomic and transportomic models can be generalized to any bacterial pathogen for which a body of sequences and annotated genomes are available. The ability to generate metabolic and transportomic models directly from sequenced bacterial genomes is a powerful tool to computationally analyze new or uncharacterized pathogens and devise effective therapies using genomic data from clinical isolates. Computational methods that can leverage bacterial pathogen genomic data to propose new patient therapies, such as the machine learning method we have described here, are needed to reduce the risk of illness and death from HAIs and to counter the rising risk of multiple-resistant bacterial pathogens.

V. Acknowledgment

This contribution originates in part from the "Environment Sensing and Response" Scientific Focus Area (SFA) program at Argonne National Laboratory. This research was supported by the U.S. Department of Energy, Office of Biological and Environmental Research (BER), as part of BER's Genomic Science Program. This research has been funded by the U.S. Department of Energy, Office of Biological and Environmental Research, under Contract DE-AC02-06CH11357. Authors would like to thank Danielle Larsen for a critical review of this manuscript.

References

[1] Klevens, R.M., et al., Estimating health care-associated infections and deaths in U.S. hospitals, 2002. Public Health Rep, 2007. 122(2): p. 160-6.

[2] Menichetti, F. and E. Tagliaferri, Antimicrobial resistance in internal medicine wards. Intern Emerg Med. 7 Suppl 3: p. S271-81.

[3] Talbot, G.H., et al., Bad bugs need drugs: an update on the development pipeline from the Antimicrobial Availability Task Force of the Infectious Diseases Society of America. Clin Infect Dis, 2006. 42(5): p. 657-68.

[4] Silby, M.W., et al., Pseudomonas genomes: diverse and adaptable. FEMS Microbiol Rev. 35(4): p. 652-80.

[5] Govan, J.R.W. and V. Deretic, Microbial pathogenesis in cystic fibrosis: Mucoid Pseudomonas aeruginosa and Burkholderia cepacia. Microbiological Reviews, 1996. 60(3): p. 539-+.

[6] Lister, P.D., D.J. Wolter, and N.D. Hanson, Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. Clin Microbiol Rev, 2009. 22(4): p. 582-610.

[7] Gross, H. and J.E. Loper, Genomics of secondary metabolite production by Pseudomonas spp. Nat Prod Rep, 2009. 26(11): p. 1408-46.

[8] de Bentzmann, S. and P. Plesiat, The Pseudomonas aeruginosa opportunistic pathogen and human infections. Environ Microbiol. 13(7): p. 1655-65.

[9] Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, 2012. 40(Database issue): p. D109-14.

[10] Ogata, H., et al., KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research, 1999. 27(1): p. 29-34.

[11] International Union of Biochemistry and Molecular Biology. Nomenclature Committee. and E.C. Webb, Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. 1992, San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press. xiii, 862 p.

[12] Larsen, P.E., et al., Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. Microbial Informatics and Experimentation, 2011. 1(4).

[13] Hibbing, M.E., et al., Bacterial competition: surviving and thriving in the microbial jungle. Nature Reviews Microbiology, 2010. 8(1): p. 15-25.

[14] Davies, D.G., et al., The involvement of cell-to-cell signals in the development of a bacterial biofilm. Science, 1998. 280(5361): p. 295-8.

[15] Bordi, C. and S. de Bentzmann, Hacking into bacterial biofilms: a new therapeutic challenge. Ann Intensive Care. 1(1): p. 19.

[16] Le Bouguenec, C. and C. Schouler, Sugar metabolism, an additional virulence factor in enterobacteria. Int J Med Microbiol. 301(1): p. 1-6.

[17] Botella, H., et al., Metallobiology of host-pathogen interactions: an intoxicating new insight. Trends in Microbiology, 2012. 20(3): p. 106-112.