Bioinformatics, 31(8), 2015, 1267–1273
doi: 10.1093/bioinformatics/btu792
Advance Access Publication Date: 4 December 2014
Original Paper

OXFORD

Data and text mining

Retrieval of *Enterobacteriaceae* drug targets using singular value decomposition

Rita Silvério-Machado[1,*], Bráulio R. G. M. Couto[2] and Marcos A. dos Santos[1]

[1]Institute of Biological Sciences and Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil and [2]Centro Universitário de Belo Horizonte/UNI-BH, Belo Horizonte, MG 30455-610, Brazil

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 29, 2014; revised on November 4, 2014; accepted on November 23, 2014

Abstract

Motivation: The identification of potential drug target proteins in bacteria is important in pharmaceutical research for the development of new antibiotics to combat bacterial agents that cause diseases.

Results: A new model that combines the singular value decomposition (SVD) technique with biological filters composed of a set of protein properties associated with bacterial drug targets and similarity to protein-coding essential genes of *Escherichia coli* (strain K12) has been created to predict potential antibiotic drug targets in the *Enterobacteriaceae* family. This model identified 99 potential drug target proteins in the studied family, which exhibit eight different functions and are protein-coding essential genes or similar to protein-coding essential genes of *E.coli* (strain K12), indicating that the disruption of the activities of these proteins is critical for cells. Proteins from bacteria with described drug resistance were found among the retrieved candidates. These candidates have no similarity to the human proteome, therefore exhibiting the advantage of causing no adverse effects or at least no known adverse effects on humans.

Contact: rita_silverio@hotmail.com.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Members of the *Enterobacteriaceae* family are globally ubiquitous organisms found in soil, water and vegetation and are part of the normal gut microbiota of many animals, including humans. These gram-negative bacteria include a large number of genera (e.g. *Klebsiella*, *Proteus*, *Serratia*, *Enterobacter* and *Escherichia*), are among the most abundant commensal microorganisms in humans, are the most frequent cause of invasive diseases in patients of all ages and acquire drug resistance at a concerning rate (Guerrero *et al.*, 2014). Their ubiquity and frequent acquisition of mobile genetic elements that give them a selective advantage drives human hosts to be regularly exposed to new bacteria strains with novel genetic repertoires including antibiotic resistance. Resistant gram-negative bacteria are responsible for many nosocomial infections and healthy non-hospitalized patient infections and are a major concern due to the severity of the infections they cause, the ease of developing multi-resistance and the absence of new therapeutic agents active against this group of pathogens (CDC, 2013; WHO, 2014). The development of antibiotic resistance in bacteria is a growing phenomenon in the twenty-first century and is one of the world's most pressing public health concerns (Martins *et al.*, 2013). Recent reports have described the spread of carbapenem-resistant or carbapenemase-producing *Enterobacteriaceae* as being associated with high mortality rates and having the potential to spread widely. The first report of a carbapenem-resistant strain of *Klebsiella pneumoniae* was in 2001 in the United States; in Europe, the first case of infection, which was of US origin, was reported in 2005 in France, and in 2006, a case was first reported in Colombia and then in Brazil and

Argentina (Cuzon *et al.*, 2010; Monteiro *et al.*, 2009; Yigit *et al.*, 2001).

New antibiotics are needed as drug resistance continues to grow, but little progress has been observed in the development of new antibiotics to combat bacterial infections, mainly because many pharmaceutical companies are withdrawing from antibacterial research and development likely due to the scientific challenges associated with manufacturing a novel antibacterial drug, the low return on investment that currently affects most pharmaceuticals and the difficulties associated with the discovery of new drug targets in bacteria (Lewis, 2013). Typically, a drug target is a molecular structure, such as a protein or a nucleic acid, through which a drug modulates its therapeutic activity, and it can be a molecule in the human body that causes disease for some reason or a molecule from a disease-causing microorganism (Imming *et al.*, 2006; Overington *et al.*, 2006). In this study, we will focus only on drug targets that are proteins from species of the *Enterobacteriaceae* family.

Different approaches can be applied for selection of a good candidate to therapeutic target like models based on sequence to function (Geyer *et al.*, 2005), comparative genomics (Abadio *et al.*, 2011), metabolic pathways (Huthmacher *et al.*, 2010), structure-to-function (Darapaneni *et al.*, 2009) and data mining (Bakheet and Doig, 2010; Chanumolu *et al.*, 2012). The most common approach to assist new drug target identification has been based on the determination of high sequence similarity using the BLAST algorithm which is a good initial starting point, but structure similarities are also important along with other explicit or hidden correlations even for proteins with low sequence similarities (Haupt and Schroeder, 2011). New drug targets can also be identified by establishing correlations between known protein signatures with contiguous patterns of 10– to 50-residue-long amino acids associated with a particular structure or function in proteins (Sheridan and Venkataraghavan, 1992), as annotated on public resources such as InterPro (Hunter *et al.*, 2012). The model that we are proposing here in one of the only few methods that is able to predict the existence of potential drug targets based on the biological function, the domain and functional sites of a protein (Bender *et al.*, 2009; Santos *et al.*, 2013). Hopkins and Groom (2002) identified 130 InterPro entries to be sufficient for the prediction of all of the human 'druggable' proteome.

Herein, we propose a model to identify potential antibiotic drug targets of the *Enterobacteriaceae* family by exploring semantic similarities across InterPro entries annotated to known *Enterobacteriaceae* drug targets. In this model, we first specified a drug target vector space model (VSM) using the known *Enterobacteriaceae* drug targets and their associated InterPro entries and applied the singular value decomposition (SVD) technique to this dataset to create a reduced drug target space that reveals latent correlations between the known drug targets that are impossible to perceive by strictly direct queries onto relational databases. Through the projection of the *Enterobacteriaceae* proteome vectors onto the reduced drug target space and the application of the cosine metric to the vectors' angles, we were able to retrieve a set of potential antibiotic drug targets that incorporate the latent structural and functional correlations between the known drug targets. The final set of *Enterobacteriaceae*-specific antibiotic drug target candidates was defined by filtering some of the protein properties associated with bacterial drug targets and determining the similarity to protein-coding essential genes of *Escherichia coli* (strain K12) and with the human proteome.

2 Methods

2.1 Data collection and generation

The protein drug targets for all of the bacteria were downloaded from DrugBank, and the Enterobacteriaceae complete proteome was downloaded from UniProtKB/Swiss-Prot (Magrane and Consortium, 2011); both of these datasets were downloaded in April 2014. A MySQL relational database was built using the large downloaded XML files, all of the data were integrated, and a dataset of 414 *Enterobacteriaceae* antibiotic drug targets with 1173 unique InterPro entries (Hunter *et al.*, 2012) and another dataset of 33 500 *Enterobacteriaceae* query proteins (*Enterobacteriaceae* complete proteome, excluding the known *Enterobacteriaceae* antibiotic drug targets) with 5839 unique InterPro entries, were obtained.

The InterPro entries were collected from the database 'Cross-references' lines in UniProtKB/Swiss-Prot. From the 'Ontologies' section of UniProtKB/Swiss-Prot, we collected the list of manually annotated keywords associated with each protein in our datasets that we used for knowledge of some sequence properties, such as subcellular location and molecular functions. The enzyme commission numbers (EC number) were collected from the 'Description' lines in UniProtKB/Swiss-Prot, where the enzyme classes for primary EC numbers 1-6 correspond to oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases, respectively.

The protein-coding essential genes were downloaded from the database of essential genes (DEG; Luo *et al.*, 2014). The database contains 609 protein-coding essential genes of *E.coli* (strain K12).

Each *Enterobacteriaceae*-specific antibiotic drug target candidate was subjected to BLASTp analysis against each *E.coli* (strain K12) protein-coded essential gene and the human UniProtKB/Swiss-Prot database, using the NCBI BLAST web server (http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi).

2.2 Singular value decomposition

For the establishment of the model, the first step was to convert each protein from the dataset of *Enterobacteriaceae* antibiotic drug targets and the dataset of *Enterobacteriaceae* query proteins to a multidimensional binary feature vector, in which each descriptor represented an InterPro entry. Each drug target was represented in a VSM, which is one of the most commonly used models of information retrieval in real-life systems (Salton, 1988). This algebraic model was originally created for the representation of text documents as vectors in a V-dimensional vector space, where V represents the number of words or terms in the document, i.e. the number of axes in the space and the documents points or vectors in this space. Here, the drug targets are treated as documents, and the InterPro entries are treated as terms. One of the important properties of these column vectors for our model is that they are very sparse vectors because one drug target is associated with only one or a few InterPro entries, which insinuates the application of latent semantic indexing (LSI), an effective method in the text mining domain (Eldén, 2006) and most recently in biological studies (Santos *et al.*, 2011) revealing the implicit semantic structure in the usage of words in a document, returning query results that are conceptually similar in meaning even if the documents do not share a specific word or words with the search criteria (Deerwester *et al.*, 1990; Dumais and Nielsen, 1992). LSI uses the mathematical technique SVD, a commonly used technique for the analysis of multivariate data that rearranges the vector space by eliminating the noise from the original matrix in such a way that the more relevant associative relationships become more visible and thereby establishing a

non-evident relationship between the clustered elements (Berry *et al.*, 1995; Chen *et al.*, 2008; Eldén, 2006).

A close analysis of each vector revealed that many *Enterobacteriaceae* antibiotic drug targets contained only one feature that would not even be repeated for another drug target. For this reason, a matrix (M) with dimensions 243-by-290 was generated using the drug target multidimensional binary feature vectors associated with more than one InterPro entry, and the SVD of the matrix was computed using MATLAB. SVD factorizes M as $M = USV^{T}$, where U is the orthogonal matrix with the left singular vectors of M, representing M's columns (drug targets), V is the orthogonal matrix with the right singular vectors of M, representing M's lines (InterPro entries) and S is the diagonal matrix of the same dimension as M, with $r > 0$ diagonal elements, which are the singular values of M, sorted by convention with the highest singular value in the upper left index of the S matrix, where r is the rank of M (the number of linearly independent columns or rows of M). The highest singular values are directly associated with more significant characteristics within the drug targets (Eldén, 2006). If we use the notation $M = U(SV^{T})$, M stands out as a linear combination of U, where this U matrix is the base that represents the columns of M and SV^{T} represents the relationship between the columns of M, denoting the relationship between the drug targets (Eldén, 2006). Considering only the k most significant singular values of M, where $k < r$, matrix M can be approximated by a low-dimensional matrix (M_k) given by $M_k = U_k S_k V_k^{T}$. The dataset represented by a smaller number of singular values shows a tendency to cluster elements that would not be grouped if the original dataset was used.

3 Results

A reduced space of *Enterobacteriaceae* antibiotic drug targets was defined using the $S_k(V_k)^{T}$ components of the M_k matrix with a rank corresponding to the k largest singular values of M, where $k < r$. The drug target matrix SVD was computed, and its dimensionality was reduced to the 65 largest singular values of this matrix, a value that is sufficiently high to fit all the data in the real structure and to remove the sampling error and unimportant data. The choice of the number of singular values that are to be used for the construction of M_k after SVD is critical and normally empirically decided. The singular value spectrum is easily visualized in a 1D plot, and the height of any singular value is indicative of its importance for explaining the data. We drew a 1D plot for the relative singular value spectrum (Fig. 1) and applied Cattell's graphical method to help decide on the significant values (Cattell, 1966).

Cattell's principle assumes that if the original variables are linear combinations of a determined number of underlying variables, the plot will tend to decrease sharply for the singular values associated with the core variables and then much more slowly for the remaining ones; thus, the significant singular values to be retained are those corresponding to the sharp dropdown line (Cattell, 1966). Everitt and Dunn (2001) proposed an alternate heuristic approach for deciding the significant singular values that recommends the analysis of the relative variances of each singular value and ignoring the singular values with a relative variance that is lower than $0.7/n$, where n is the number of proteins in the original matrix. We combined the two heuristics to obtain the k value of 65. By representing the data into this SVD subspace, we created a semantic VSM, in which the known drug targets are represented by vectors that express the linear combination of the relevant descriptors, revealing the hidden structure of the data (Fig. 2).

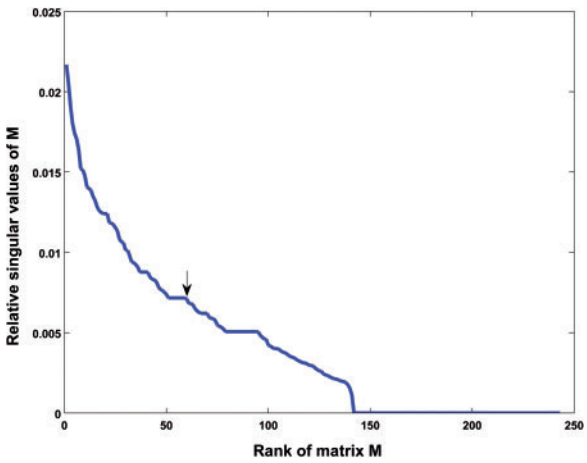

Fig. 1. One-dimensional plot of the relative singular value spectrum of matrix *M*. The arrow indicates the 65 largest singular values of this matrix

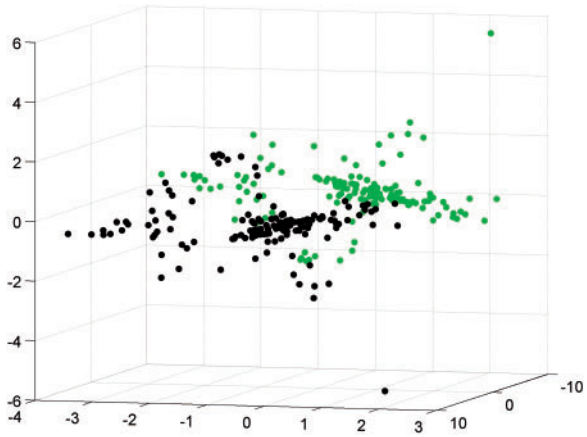

Fig. 2. Visualization of *Enterobacteriaceae* antibiotic drug targets (dark circles) and drug target candidates (light circles) in 3D. The vectors in the space with dimension 65 were projected in space IR^3 using the method described by Marcolino *et al.* (2010)

Using the U_k of the truncated matrix M_k, each *Enterobacteriaceae* query protein (q) was projected onto the reduced *Enterobacteriaceae* antibiotic drug target space to obtain the equivalent vector q_k using the equation $q_k = q^{T}U_k$, which was proposed by Eldén (2006). The pairwise semantic similarity was determined by calculating the cosine between each pair of query protein and drug target. The numeric similarity between two vectors can be measured by the cosine of the angle that they form, with the cosine of the angle yielding a value in the real range $[-1.0, +1.0]$, where $+1.0$ is associated with identical vectors or similar proteins, 0.0 represents orthogonal vectors or dissimilar proteins, and a negative value close to -1.0 indicates vectors in the opposite direction (Singhal, 2001). For determining the dataset of *Enterobacteriaceae* potential drug targets, the 0.2 cosine similarity between each pair was considered, a value that aims to retrieve the topologically closely related descriptors with no evident feature relationship. By applying the described model and using the referred cosine cut-off value,

we recovered 13 092 *Enterobacteriaceae*-specific antibiotic drug target candidates.

Latent semantic relations are the consequence of InterPro entries sharing among proteins. Within this framework, and to make such relations more explicit, the InterPro entries that didn't repeat and their associated drug targets were removed from the original matrix. For this reason, this model was developed with 290 drug targets, out of a total of 415. The robustness of this model was thoroughly evaluated and sensitivity and specificity values were determined taking into consideration the context of the model and the choices that were done while modelling the problem. After removal of several test sets and their subsequent projection in the reduced spaced created by the remaining drug targets, the model was able to classify the test sets as drug targets (the cosine between any pair of test protein and its most similar drug target present in the model was >0.9). As it is known beforehand that the model fails to classify 125 (415-290) drug targets because their projection in the reduced vector space is the null vector, the sensitivity of this model is 70% (290/415). The specificity is 100% because we are considering a non-target to be any protein that doesn't share any InterPro entry with the 290 drug targets of the model.

In 2010, Bakheet and Doig (2010) identified many protein properties associated with bacterial drug targets. We refined our list of *Enterobacteriaceae*-specific antibiotic drug target candidates by applying some of these properties (Fig. 3). The first chosen property that was filtered was the existence of catalytic activity because drug targets are most likely to be enzymes (Bakheet and Doig, 2010), and this filtering resulted in 7989 candidates. It has also been revealed that the preferred subcellular location for bacterial drug targets are the cytoplasm and periplasm, which is a logical approach because the highest percentage of known bacterial drug targets have these preferred locations (Bakheet and Doig, 2010). Gathering the candidates based on this former filter resulted in 1127 *Enterobacteriaceae* antibiotic drug target candidates. After applying a set of keywords related to antibiotic response, cellular processes of transcription, translation, replication and bacterial cell biosynthesis, a list shaped based on the properties identified by Bakheet and Doig (2010), no more candidates were excluded, which means that the model developed preserved the most relevant drug target properties in the *Enterobacteriaceae* family.

Antibiotic drug targets are ideally protein-coding sequences essential to the pathogen, have a unique function in the pathogen, are present only in the pathogen and are able to be inhibited by a small molecule. The importance of a unique protein as a candidate to drug target is determined on the basis of its functional importance, for example, if it is essential in cell survival or if it is an enzyme that catalyses a reaction in which either a substrate or product is uniquely consumed or produced. Even if a small drug molecule can modulate an identified candidate, if it is not involved in critical functions in the pathogen, this modulation will have no serious consequence on their survival or growth. Targeting an essential protein for a parasite may provide an effective way to control infection; thus, proteins encoded from pathogen essential genes serve as potential drug targets. Essential genes are the minimal set of genes that code for central cellular processes required for the viability or fertility of an organism so they are indispensable for the survival of an organism, and their functions are, therefore, considered a foundation of life (Chen *et al.*, 2012; Gerdes *et al.*, 2006). A sequence similarity search of the 1127 *Enterobacteriaceae* antibiotic drug target candidates against the *E.coli* (strain K12) protein-coding essential genes using the BLASTp tool, returned 860 candidates similar to a protein-coding essential gene (identity >70%), with 19 *E.coli* (strain K12) protein-coding essential genes and 620 proteins in the *Enterobacteriaceae* family with a known human host, as described in HAMAP (Pedruzzi *et al.*, 2013), similar to protein-coding essential genes. For similarity determination, only the sequences with a query cover higher than 70% and *E*-values lower than 10^{-20} were considered.

Enterobacteriaceae antibiotic drug target candidates that are essential gene products or that are similar to protein-coding essential genes were checked for similarity to the human proteome to identify potential targets that may cause adverse effects in patients. A list of 99 proteins were found not to be similar to any human protein (identity <20%), indicating that when modulated by an antibiotic drug, they have little chance of causing known adverse events in their human hosts or at least no known adverse events. This list is available in Supplementary Material. Out of these *Enterobacteriaceae* antibiotic drug target candidates with a human host, some proteins are from eight bacteria with described drug-resistance in HAMAP: *E.coli* O17:K52:H18 (strain UMN026/ExPEC), *Klebsiella pneumonia* subsp. *pneumoniae* (strain ATCC 700721/MGH 78578), *Salmonella choleraesuis* (strain SC-B67), *Salmonella agona* (strain SL483), *Salmonella newport* (strain SL254), *Salmonella paratyphi* A (strain AKU_12601), *Salmonella heidelberg* (strain SL476) and *Salmonella schwarzengrund* (strain CVM19633). The *Enterobacteriaceae* antibiotic drug target candidates in bacteria with a human host hold eight different functions: 3-oxoacyl-[acyl-carrier-protein] synthase 3 (KAS III), UDP-*N*-acetylglucosamine acyltransferase (LpxA), D-alanine-D-alanine ligase B, serine acetyltransferase (SATase), 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate *N*-succinyltransferase (THP succinyltransferase), ribosomal-proteinalanine acetyltransferase, ribosomal-protein-serine acetyltransferase and thioredoxin-2 (Trx2).

Fatty acids are a main source of energy in bacteria and play an important role in the formation of cell membranes (Heath *et al.*, 2001), making the fatty acid biosynthesis pathway an attractive but still largely unexploited target for the development of new antibacterial agents (Lee *et al.*, 2012) with successful validated antibiotic targets, such as KAS III from *E.coli*, which is inhibited by the approved drug Cerulenin (http://www.drugbank.ca/drugs/DB01034) that irreversibly binds to it (Heath and Rock, 2004; Wright and Reynolds, 2007). This enzyme is essential for bacterial fatty acid synthesis in both gram-negative and gram-positive bacteria (Lai and Cronan, 2003). There is no reference in the literature of inhibition studies in this enzyme for the *Enterobacteriaceae*, family such as *Klebsiella pneumoniae* strains, with the exception of *E.coli*. This is a encouraging potential antibiotic drug target for this bacterial family.

LpxA catalyses the first step of the biosynthesis of Lipid A, a phosphorylated glucosamine disaccharide with a long-chain saturated fatty acid that serves as a hydrophobic anchor for the lipopolysaccharide (LPS) in the outer membrane of gram-negative bacteria (Lee *et al.*, 2013). The outer membrane of gram-negative bacteria is an effective permeability barrier against external pernicious agents, and in the *Enterobacteriaceae* family, LPS molecules occupy the outer leaflet of the outer membrane, creating a highly ordered monolayer with low fluidity, which is a structure that is shown to create a poor partition of hydrophobic molecules into the hydrophobic interior portion of isolated LPS (Vaara and Nurminen, 1999). *E.coli* lpxA gene mutants have severe defects in the biosynthesis of lipid A, making them extremely susceptible to hydrophobic antibiotics, probably due to the lack of a continuous LPS layer in the outer leaflet and the resultant compensatory presence of glycerol-phospholipids patches in this leaflet or even transient ruptures,

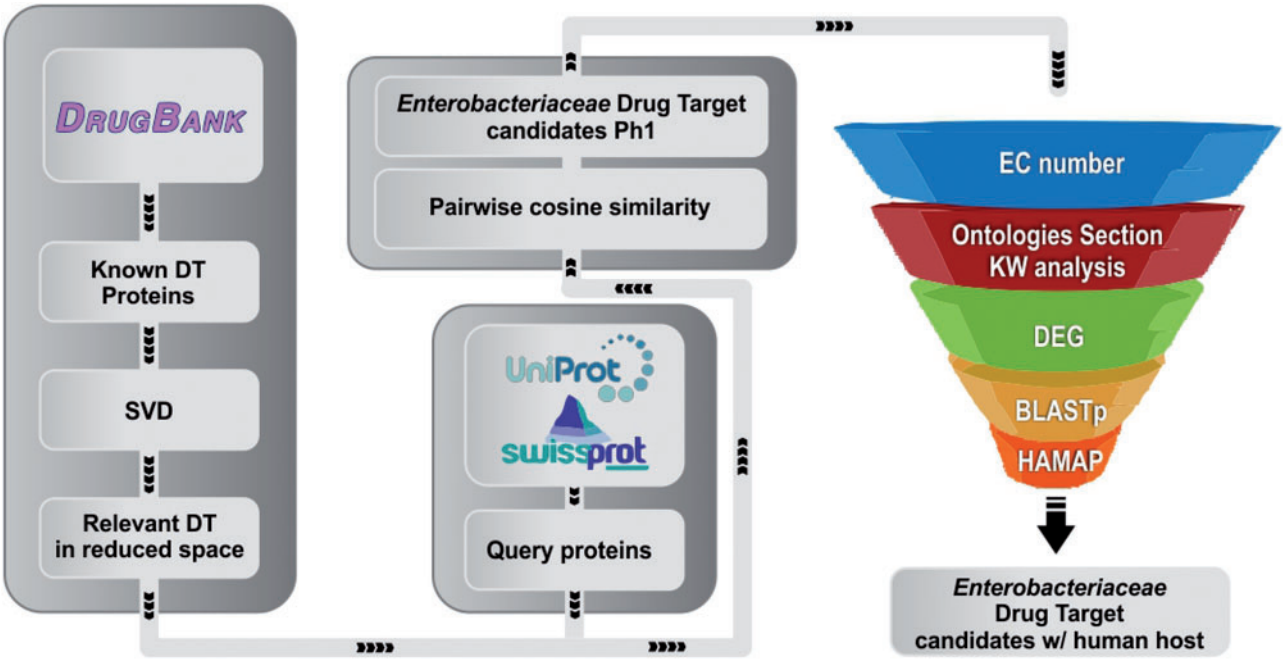

Fig. 3. Diagram of the proposed model to identify potential antibiotic drug targets (DT) in the *Enterobacteriaceae* family. A drug target vector space was specified using the known *Enterobacteriaceae* drug target proteins and their associated InterPro entries and SVD was applied to this dataset to create a reduced drug target space that reveals latent correlations between the known drug targets. Through the projection of the *Enterobacteriaceae* proteome vectors (query proteins) onto the reduced drug target space and the application of the cosine metric to the vectors' angles, we were able to retrieve a set of potential antibiotic drug targets (*Enterobacteriaceae* drug target candidates Ph1) that incorporate the latent correlations between the known drug targets. The final set of *Enterobacteriaceae*-specific antibiotic drug target candidates (*Enterobacteriaceae* drug target candidates w/human host) was defined by application of a set of biological filter

allowing the diffusion of hydrophobic molecules (Vaara and Nurminen, 1999). The use of *E.coli* mutants defective in the first step of lipid A biosynthesis has shown that the lipid A domain of LPS is essential to the growth and viability of cells (Galloway and Raetz, 1990; Heath *et al.*, 2001). Consequently, LpxA, the enzyme catalysing the first step of the biosynthesis of lipid A, is a noteworthy target for therapeutic intervention in gram-negative bacteria that cause diseases.

D-Alanine-D-alanine ligase B is one of two D-alanine-D-alanine ligases that exist in *E.coli*, is produced by the essential gene ddlB and is an isoenzyme of the encoding ddlA gene product (Zawadzke *et al.*, 1991), which synthesizes D-alanyl-D-alanine, a dipeptide that is then added to the precursor subunit of peptidoglycan and that provides the site of peptidoglycan framework cross-linkage (Ellsworth *et al.*, 1996). Peptidoglycan is a structure that is unique to bacterial cells, is essential to the maintenance of the structural integrity of the cell and is a structure that most bacterial cells contain and other organisms do not have. This combination of major uniqueness and cell survival essentiality makes peptidoglycan an excellent target for antibiotics. In fact, D-alanine-D-alanine ligase is an antibacterial drug target, and *E.coli* D-alanine-D-alanine ligase A is a long-known declared target of cycloserine (http://www.drugbank.ca/drugs/DB00260), which inhibits both D-alanine-D-alanine ligases but exhibits high toxicity to the human host (Batson *et al.*, 2010). This ddlB essential gene is preserved in over 80% of diverse bacterial genomes (Gerdes *et al.*, 2003), highlighting this gene's product, which is ubiquitous in prokaryotes possessing peptidoglycan, as a good drug target.

SATase carries out the first step in the pathway of cysteine biosynthesis, catalysing the transference of an acetyl group to *L*-serine into *O*-acetyl-*L*-serine, which will later be converted into *L*-cysteine (Kredich and Tomkins, 1966). *L*-cysteine is required for cell growth in some bacteria, such as *E.coli*, and also operates as a sulphur donor for the synthesis of other cellular molecules, such as *L*-methionine, glutathione, coenzyme A and molybdenum cofactor (Sekowska *et al.*, 2000). Because SATase is one of the key enzymes for the biosynthesis of cysteine in *E.coli*, encoded by the essential gene cysE (Gerdes *et al.*, 2003), and is not crucial in humans as *L*-cysteine is the produced in several reactions in these organisms, it may serve as a potential novel target for developing new antibiotic drugs.

The biosynthesis of lysine in bacteria occurs via the diaminopimelate-lysine pathway with THP succinyltransferase acting as a key enzyme in this *E.coli* lysine biosynthetic pathway (Simms *et al.*, 1984). This pathway generates an important intermediate, namely diaminopimelate (Danks *et al.*, 2013), which is a building block of peptidoglycan in the cell wall of many gram-negative bacteria (Hutton *et al.*, 2007). THP succinyltransferase is of special interest as a bacteria-specific drug target because it is the product of an essential gene, is involved in cell wall synthesis and is absent in mammalian cells because mammals lack the ability to biosynthesize lysine that is therefore one of the nine essential amino acids that must be provided through diet. This is a potential drug target that has yet to be exploited.

Acetylation is a post-translational modification that is less prevalent in bacteria than in eukaryotes (Cain *et al.*, 2014). In *E.coli*, the essential genes rimI and rimL encode two *N*-acetyltransferases, namely ribosomal-protein-alanine acetyltransferase and ribosomal-protein-serine acetyltransferase, respectively, with the first being specific for ribosomal protein S18 and the latest being specific for ribosomal protein L7/L12 (Cain *et al.*, 2014; Isono and Isono, 1981; Tanaka *et al.*, 1989; Yoshikawa *et al.*, 1987). The biological

function of this post-translational modification remains to be determined; however, because one of the roles established for *N*-acetylation in eukaryotes is protein stability, the same possibility has been assumed for bacteria (Cain *et al.*, 2014). Ribosomal proteins are the protein components of the ribosomal subunits that are involved in protein biosynthesis, and assuming that the raised possibility is a true fact, an inhibition of these *N*-acetyl-transferases would cause major damage to bacteria, making these excellent potential drug targets. In addition, the lack of protein similarity to the human proteome makes the antibacterial drugs to be used to inhibit these good candidates to drug targets and causative of no known adverse effects.

Protection from oxidative stress and efficient redox regulation are essential for bacteria that can grow in aerobic environments. Thioredoxins are small ubiquitous redox proteins that are found in nearly all organisms and that comprise the thioredoxin system, an antioxidant system acting against oxidative stress through its disulfide reductase activity that regulates the protein dithiol/disulfide balance, and in *E.coli*, it constitutes a critical system for the maintenance of cellular protein disulfide/dithiol redox control (Lu and Holmgren, 2014; Stewart *et al.*, 1998). Disulfide bonds in proteins play numerous important roles, namely stabilization of the proteins, or are formed transiently as part of a catalytic or regulatory cycle (Ritz and Beckwith, 2001). Trx2 is the product of the trxC essential gene in *E.coli* (Miranda-Vizuete *et al.*, 1997), is an oxidative stress-induced protein and play a role in the cellular defence against oxidative stress (Collet *et al.*, 2003). The important redox roles of Trx2, the identified essentiality of its coded gene and the lack of similarity to the human proteome make this protein a noticeable candidate as an antibiotic drug target.

4 Conclusions

There is a race against time to develop new antibiotics to combat bacterial agents causing diseases. This rushes the need to discover new antibiotic drug targets. We propose a model for the identification of potential antibiotic drug targets that incorporate the latent structural and functional correlations between the known drug targets in the *Enterobacteriaceae* family, using the SVD technique followed by fine-tuning of the results with a set of most relevant protein properties associated with bacterial drug targets and similarity to *E.coli* (strain K12) protein-coding essential genes. The application of this model to the *Enterobacteriaceae* family proteome returned 99 proteins candidates that have no similarity to the human proteome and holding eight different functions. Of these, some proteins are from eight bacteria with described drug resistance.

It is now a known fact that drugs almost invariably influence more than one target to some extent as either a consequence of structural similarities between the intended target and other proteins, through allosteric effects on other proteins or due to genuine multivalent target binding and if this occurs the result is usually a series of unwanted side effects in addition to the desired therapeutic effect. We adopted the criterion that if two proteins share <20% sequence identity, they are considered not to be similar and so no adverse events to the human host would arise from a potential inhibitor against a proposed candidate to bacteria drug targets. One concern is that two proteins that fall in this cut-off value may have similar binding sites; regions in the proteins' structures deeply related to their pharmacology (Durrant *et al.*, 2010; Ho Sui *et al.*, 2012). This is a limitation of the current method and it is important to compare the various protein molecules in both the pathogen and the host at the binding site level to remove target candidates that may have very high similarity with the human proteome.

Predicting individual drug targets candidates is already of value but this single protein strategy may flaw due to drug resistance development from a single amino acid mutation in the target protein (Hopkins, 2008). But one must not forget that the essential nature of a drug is its polypharmacology, therefore a drug that might be developed or repositioned for one of these target candidates will potentially exhibit bioactivity against a series of other relevant targets in a pathogen, preventing or delaying drug resistance. The findings from this model should now be addressed with a structure-based systems biology approach to identify off-targets related to the proposed target candidates on a proteome-wide scale (Ng *et al.*, 2014; Pei *et al.*, 2014; Xie *et al.*, 2014).

This model paves the way for the discovery of novel bacterial drug targets for therapeutic intervention and has allowed the identification of a set of, still to be exploited, drug target candidates, incorporating the latent structural and functional correlations between the known drug targets, a distinguishing feature among other models for drug target retrieval.

Funding

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Conflict of Interest: none declared.

References

Abadio,A.K. *et al.* (2011) Comparative genomics allowed the identification of drug targets against human fungal pathogens. *BMC Genomics*, **12**, 75.

Bakheet,T.M. and Doig,A.J. (2010) Properties and identification of antibiotic drug targets. *BMC Bioinformatics*, **11**, 195.

Batson,S. *et al.* (2010) Crystallization and preliminary X-ray analysis of a D-alanyl-D-alanine ligase (EcDdlB) from Escherichia coli. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **66**, 405–408.

Bender,A. *et al.* (2009) Use of ligand based models for protein domains to predict novel molecular targets and applications to triage affinity chromatography data. *J. Proteome Res.*, **8**, 2575–2585.

Berry,M.W. *et al.* (1995) Using linear algebra for intelligent information retrieval. *SIAM Rev.*, **37**, 573–595.

Cain,J.A. *et al.* (2014) Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J. Proteomics*, **97**, 265–286.

Cattell,R.B. (1966) The scree test for the number of factors. *Multivariate Behav. Res.*, **1**, 245–276.

CDC. (2013) Vital signs: carbapenem-resistant Enterobacteriaceae. *MMWR. Morbidity and mortality weekly report*. Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Atlanta, pp. 165–170.

Chanumolu,S.K. *et al.* (2012) UniDrug-target: a computational tool to identify unique drug targets in pathogenic bacteria. *PLoS One*, **7**, e32833.

Chen,M.-C. *et al.* (2008) An information granulation based data mining approach for classifying imbalanced data. *Inform. Sci.*, **178**, 3214–3227.

Chen,W.H. *et al.* (2012) OGEE: an online gene essentiality database. *Nucleic Acids Res.*, **40**, D901–D906.

Collet,J.F. *et al.* (2003) Thioredoxin 2, an oxidative stress-induced protein, contains a high affinity zinc binding site. *J. Biol. Chem.*, **278**, 45325–45332.

Cuzon,G. *et al.* (2010) Worldwide diversity of Klebsiella pneumoniae that produce beta-lactamase blaKPC-2 gene. *Emerg. Infect. Dis.*, **16**, 1349–1356.

Danks,G. *et al.* (2013) OikoBase: a genomics and developmental transcriptomics resource for the urochordate Oikopleura dioica. *Nucleic Acids Res.*, **41**, D845–D853.

Darapaneni,V. *et al.* (2009) Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *J. Gen. Virol.*, **90**, 2124–2133.

Deerwester,S. *et al.* (1990) Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.*, **41**, 391–407.

Dumais,S.T. and Nielsen,J. (1992) Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Copenhagen, Denmark, pp. 233–244.

Durrant,J.D. *et al.* (2010) A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput. Biol.*, **6**, e1000648.

Eldén,L. (2006) Numerical linear algebra in data mining. *Acta Numerica*, **15**, 327–384.

Ellsworth,B.A. *et al.* (1996) Synthesis and evaluation of inhibitors of bacterial D-alanine:D-alanine ligases. *Chem. Biol.*, **3**, 37–44.

Everitt,B.S. and Dunn,G. (2001) *Applied Multivariate Data Analysis*, 2nd edn. Arnold, London, UK.

Galloway,S.M. and Raetz,C.R. (1990) A mutant of *Escherichia coli* defective in the first step of endotoxin biosynthesis. *J. Biol. Chem.*, **265**, 6394–6402.

Gerdes,S. *et al.* (2006) Essential genes on metabolic maps. *Curr. Opin. Biotechnol.*, **17**, 448–456.

Gerdes,S.Y. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.

Geyer,J.A. *et al.* (2005) Targeting malaria with specific CDK inhibitors. *Biochimica et Biophysica Acta*, **1754**, 160–170.

Guerrero,P.P. *et al.* (2014) Infecciones por enterobacterias. *Medicine Programa de Formación Médica Continuada Acreditado*, **11**, 3276–3282.

Haupt,V.J. and Schroeder,M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief. Bioinform.*, **12**, 312–326.

Heath,R.J. and Rock,C.O. (2004) Fatty acid biosynthesis as a target for novel antibacterials. *Curr. Opin. Invest. Drugs*, **5**, 146–153.

Heath,R.J. *et al.* (2001) Lipid biosynthesis as a target for antibacterial agents. *Prog. Lipid Res.*, **40**, 467–497.

Ho Sui,S.J. *et al.* (2012) Raloxifene attenuates *Pseudomonas aeruginosa* pyocyanin production and virulence. *Int. J. Antimicrob. Agents*, **40**, 246-251.

Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.

Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Huthmacher,C. *et al.* (2010) Antimalarial drug targets in plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC Syst. Biol.*, **4**, 120.

Hutton,C.A. *et al.* (2007) Inhibition of lysine biosynthesis: an evolving antibiotic strategy. *Mol. Biosyst.*, **3**, 458–465.

Imming,P. *et al.* (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Dis.*, **5**, 821–834.

Isono,S. and Isono,K. (1981) Ribosomal protein modification in Escherichia coli. III. Studies of mutants lacking an acetylase activity specific for protein L12. *Mol. Gen. Genetics MGG*, **183**, 473–477.

Kredich,N.M. and Tomkins,G.M. (1966) The enzymic synthesis of L-cysteine in *Escherichia coli* and *Salmonella typhimurium*. *J. Biol. Chem.*, **241**, 4955–4965.

Lai,C.Y. and Cronan,J.E. (2003) Beta-ketoacyl-acyl carrier protein synthase III (FabH) is essential for bacterial fatty acid synthesis. *J. Biol. Chem.*, **278**, 51494–51503.

Lee,C.R. *et al.* (2013) Lipid a biosynthesis of multidrug-resistant pathogens: a novel drug target. *Curr. Pharm. Des.*, **19**, 6534–6550.

Lee,J.Y. *et al.* (2012) Discovery of novel selective inhibitors of *Staphylococcus aureus* beta-ketoacyl acyl carrier protein synthase III. *Eur. J. Med. Chem.*, **47**, 261–269.

Lewis,K. (2013) Platforms for antibiotic discovery. *Nat. Rev. Drug Dis.*, **12**, 371–387.

Lu,J. and Holmgren,A. (2014) The thioredoxin antioxidant system. *Free Radical Biol. Med.*, **66**, 75–87.

Luo,H. *et al.* (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.

Magrane,M. and Consortium,U. (2011) UniProt knowledgebase: a hub of integrated protein data. *Database J. Biol. Databases Curation*, **2011**, bar009.

Marcolino,L.S. *et al.* (2010) Genome visualization in space. In *Proceedings of IWPACBB*, Springer Berlin Heidelberg, pp. 225–232.

Martins,A. *et al.* (2013) Mechanisms of resistance in bacteria: an evolutionary approach. *Open Microbiol. J.*, **7**, 53–58.

Miranda-Vizuete,A. *et al.* (1997) Cloning, expression, and characterization of a novel *Escherichia coli* thioredoxin. *J. Biol. Chem.*, **272**, 30841–30847.

Monteiro,J. *et al.* (2009) First report of KPC-2-producing Klebsiella pneumoniae strains in Brazil. *Antimicrob. Agents Chemother.*, **53**, 333–334.

Ng,C. *et al.* (2014) Anti-infectious drug repurposing using an integrated chemical genomics and structural systems biology approach. In *Pacific Symposium on Biocomputing*. Pacific Symposium on Biocomputing, World Scientific Publishing Company, pp. 136–147.

Overington,J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Dis.*, **5**, 993–996.

Pedruzzi,I. *et al.* (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res.*, **41**, D584–D589.

Pei,J. *et al.* (2014) Systems biology brings new dimensions for structure-based drug design. *J. Am. Chem. Soc.*, **136**, 11556–11565.

Ritz,D. and Beckwith,J. (2001) Roles of thiol-redox pathways in bacteria. *Annu. Rev. Microbiol.*, **55**, 21–48.

Salton,G. (1988) Automatic text indexing using complex identifiers. In ACM (ed.) *DOCPROCS '88 Proceedings of the ACM Conference on Document Processing Systems*, Association for Computing Machinery (ACM), New York, pp. 135–144

Santos,A.R. *et al.* (2011) A singular value decomposition approach for improved taxonomic classification of biological sequences. *BMC Genomics*, **12** (Suppl 4), S11.

Santos,E.C. *et al.* (2013) A semantic-based similarity measure for human druggable target proteins. In *BIOTECHNO 2013: The Fifth International Conference on Bioinformatics*. Biocomputational Systems and Biotechnologies, BIOTECHNO 2013 Editors in Lisbon, Portugal, pp. 9–14.

Sekowska,A. *et al.* (2000) Sulfur metabolism in *Escherichia coli* and related bacteria: facts and fiction. *J. Mol. Microbiol. Biotechnol.*, **2**, 145–177.

Sheridan,R.P. and Venkataraghavan,R. (1992) A systematic search for protein signature sequences. *Proteins*, **14**, 16–28.

Simms,S.A. *et al.* (1984) Purification and characterization of succinyl-CoA: tetrahydrodipicolinate N-succinyltransferase from *Escherichia coli*. *J. Biol. Chem.*, **259**, 2734–2741.

Singhal,A. (2001) Modern information retrieval: a brief overview. *Bull. IEEE Comp. Soc. Tech. Committ. Data Eng.*, **24**, 35–42.

Stewart,E.J. *et al.* (1998) Disulfide bond formation in the *Escherichia coli* cytoplasm: an in vivo role reversal for the thioredoxins. *EMBO J.*, **17**, 5543–5550.

Tanaka,S. *et al.* (1989) Cloning and molecular characterization of the gene rimL which encodes an enzyme acetylating ribosomal protein L12 of *Escherichia coli* K12. *Mol. Gen. Genetics MGG*, **217**, 289–293.

Vaara,M. and Nurminen,M. (1999) Outer membrane permeability barrier in *Escherichia coli* mutants that are defective in the late acyltransferases of lipid A biosynthesis. *Antimicrob. Agents Chemother.*, **43**, 1459–1462.

WHO. (2014) Antimicrobial resistance global report on surveillance 2014, World Health Organization, pp. 257.

Wright,H.T. and Reynolds,K.A. (2007) Antibacterial targets in fatty acid biosynthesis. *Curr. Opin. Microbiol.*, **10**, 447–453.

Xie,L. *et al.* (2014) Towards structural systems pharmacology to study complex diseases and personalized medicine. *PLoS Comput. Biol.*, **10**, e1003554.

Yigit,H. *et al.* (2001) Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. *Antimicrob. Agents Chemother.*, **45**, 1151–1161.

Yoshikawa,A. *et al.* (1987) Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylating ribosomal proteins S18 and S5 of *Escherichia coli* K12. *Mol. Gen. Genetics MGG*, **209**, 481–488.

Zawadzke,L.E. *et al.* (1991) Existence of two D-alanine:D-alanine ligases in Escherichia coli: cloning and sequencing of the ddlA gene and purification and characterization of the DdlA and DdlB enzymes. *Biochemistry*, **30**, 1673–1682.