BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–16

Using Comparative Genomics and Virtual Screening for Antibiotic Drug Discovery

Peter Hebden

School of Computing Science, Newcastle University

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

The application of a combination of methods from comparative genomics, sequence analysis in particular, and virtual screening to predict new drug targets and the chemical compounds that bind most strongly to those targets is presented.

Motivation: The evolution and spread of antibiotic resistant pathogenic bacteria has been rapid and often lethal, while the pipeline for new antibiotics has remained virtually bone dry. We face an urgent need for new antibiotics and more cost effective methods to support drug discovery; greater efficiency may be achieved by prioritizing *in vitro* testing.

Results We found that methods based on comparative genomics and virtual screening are both time and cost effective and, more to the point, help to predict suitable drug targets, their active sites, and prioritize chemical compounds for *in vitro* testing of their antimicrobial properties.

Availability Software and data are available upon request.

Contact phebden@gmail.com

Supplementary Information Supplementary tables and figures are available with the online version.

1 INTRODUCTION

In this thesis we report on our application of methods from comparative genomics and virtual screening to discover new bacterial drug targets and chemical compounds that bind to those targets. We hypothesized that given the growing abundance of genomic sequence data and computing power, the combination of sequence analysis and virtual screening provide an efficient way to identify probable active sites on essential proteins and the chemical compounds that bind to them with the greatest affinity. Our hope is that this research will contribute the antibiotic drug discovery process and that new antibiotics will be developed and, eventually, be used to promote health and save lives.

The specific objectives of this research include the following:

- Acquire the complete proteomes of *Bacillus subtilis* and 10 to 100 diverse bacterial proteomes.
- Identify the cell wall associated proteins in *Bacillus subtilis* that are the most broadly conserved across bacterial species, both Gram-positive and Gram-negative, and use sequence analysis to determine probable active sites.
- Determine target proteins: broadly conserved proteins have a known 3D structure.
- Virtually screen chemical compounds against probable active sites on target proteins.
- Identify the "best" compounds for *in vitro* testing against pathogenic bacteria.
- Develop software and techniques, especially those that use high throughput distributed computing, that scale to very large data sets.

The treatment of infectious disease with antibiotics has been one of the major successes of the 20th century and has led to large increases in life expectancy. However, microorganisms are becoming increasingly resistant to multiple classes of antibiotics and there is an urgent need for new compounds (Mills, 2006). For example, *Staphylococcus aureus* (*S. aureus*) is the leading cause of bacterial infections of the skin, soft tissue, bloodstream, and lower respiratory tract in developed countries (DeLeo and Chambers, 2009). Although penicillin was effective initially, resistant strains emerged in the mid-1940s, methicillin resistant *S. aureus* (MRSA) was discovered in the 1960s (DeLeo and Chambers, 2009), and strains with some degree of vancomycin resistance have recently emerged worldwide (Howe *et al.*, 2004). The implication is clear. If one of the greatest successes of the 20th century is lost, we face the possibility of a commensurate decrease in life expectancy.

In an effort to meet this challenge, the first major part of our research applies sequence analysis to recently determined genomic data. The availability of bacterial genome sequences was expected have a large impact on antibiotic discovery by enabling new genomics-based approaches for identifying new molecular targets (Knowles, 1997) but, it should be emphasized, genomics does not significantly alter the time frame for the drug discovery process, which can take up to 12 years (McDevitt and Rosenberg, 2001). In practice, genomics has delivered novel essential drug targets for target based discovery efforts (McDevitt and Rosenberg, 2001; Mills, 2003).

The second major part of our research uses the power of distributed computing for virtual screening of compounds. This application of virtual screening was motivated by several considerations. First, computational methods have advanced to become a crucial component of many drug discovery programs, and virtual screening techniques are now widely used (Kitchen *et al.*, 2004). Second, the size of compound libraries has been growing exponentially from the output of robotic "combinatorial chemistry" (Hogan, 1996). Third, the number of known 3D structures continues to grow, with thousands being added each year (Warren *et al.*, 2006). Fourth, small organic ligands are integral parts of many protein structures (Lesk,

2001, p. 2) and, more importantly, they usually participate directly in the function of a protein and, consequently, may be one of the most interesting components of a structure to study (Lesk, 2001, p. 103). And finally, we noted that virtual screening to discover new inhibitors is becoming a common practice in modern drug discovery (Shoichet, 2004).

Although pathogenic bacteria have proven to be remarkably adaptable, the cell wall remains a good target for antibiotics (Silver, 2003, 2006; Schneider and Sahl, 2010) and more than 100 genes involved in bacterial cell wall biosynthesis and division have been identified. Starting with these genes, we used a comparative genomics approach (Arigoni *et al.*, 1998) to identify a set of homologous genes/proteins in multiple bacterial species.

We used BLAST (Altschul *et al.*, 1997) for pairwise sequence alignment and MUSCLE (Edgar, 2004) for multiple sequence alignment (MSA), Jalview (Waterhouse *et al.*, 2009) for visualization of the MSA and calculation of conservation scores for each amino acid in the *B. subtilis* query protein, and PyMOL (Delano, 2002) for 3D molecular visualization.

We downloaded ligand files representing $\approx$ 100,000 compounds from the ZINC database (Irwin and Shoichet, 2005); almost 90,000 were from the Natural Products Database (NPD) meta data set. Given a promising set of candidate proteins, probable active sites, and ligands we used AutoDockTools (ADT) (Sanner, 1999) to prepare receptor and ligand input files and AutoDock Vina (Trott and Olson, 2010) molecular docking software for virtual screening to identify small molecule inhibitors (Mills, 2006).

We found that amino acids with high conservation scores corresponded well with known active sites, i.e., organic ligand binding sites according to the Protein Data Bank (PDB). We also found that by the application of molecular evolutionary principles, sequence analysis and virtual screening, we were able to prioritize bacterial species, their proteins and matching compounds for wet lab experiments. In summary, we considered 123 *B. subtilis* proteins and $\approx$ 100,000 chemical compounds. Four proteins, FtsZ, CoaD, YwtF, and RacE were identified as candidates for virtual screening; FtsZ and YwtF and a small number of specific compounds were selected for testing which, based on empirical evidence, had the highest probability of exhibiting antimicrobial properties.

This thesis is organized as follows. Section 2 provides background information on bacteria, antibiotics, and the motivations underlying our research. Section 3 provides a conceptual introduction to the methods used in our research and why they were chosen. Section 4 describes how our methods were applied. Section 5 presents our results and interpretations. Section 6 provides a brief discussion of our findings. In addition, the Supplementary Material provides detailed tables of data and figures.

2 BACTERIA AND ANTIBIOTICS

2.1 Background

The 'golden era' of antibiotic research occurred from the late 1940s to the late 1960s (McDevitt and Rosenberg, 2001). In fact, nearly all classes of antibiotics (acting on the same target) currently in use were developed prior to 1970 (Knowles, 1997). Since the 1980s, there has been a decline in the number of new antibiotics brought into clinical practice, and even these new antibiotics bind to the same target molecules as their predecessors. And now, as pan-resistant strains of pathogenic bacteria have become a clinical reality, the pipeline of new drugs is "virtually bone dry" (Projan, 2008).

This has been due to a number of factors including the high expense of developing new drugs, the relatively low profit potential of antibiotics, industry concentration through mergers and acquisition, and higher regulatory hurdles (Projan, 2003, 2008; Payne *et al.*, 2007). Consequently, new classes of antibiotics are desperately needed to combat the global emergence of bacterial pathogens – this has been described as one of the most important challenges facing the pharmaceutical industry (Brown and Warren, 1998).

Rational drug design has been based on detailed structural information of the drug target. However, the success of this paradigm, which emerged in the late 1960s, has been limited by the small number of well defined molecular targets and progress has been limited to a few chemical classes, e.g., β-lactamase inhibitors, carbapenems and fluroquinolones (Knowles, 1997). This limitation and the lack of new antibiotic classes since 1970 is shown in Figure 1 where the graph flattens out in the late 1960s.

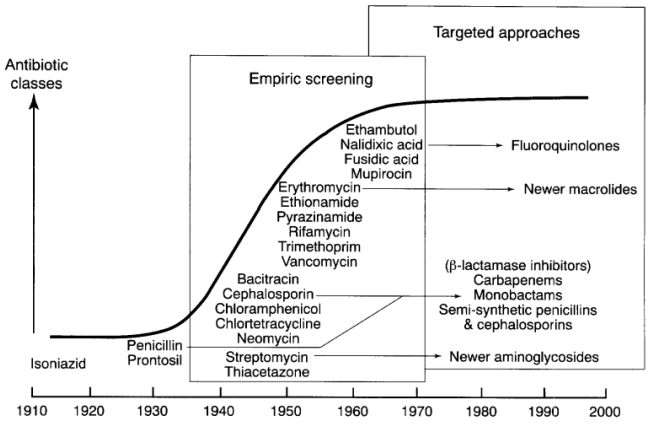


Fig. 1: Discovery of new classes of antibiotics: Empiric screening has generated antibiotic classes, while targeted approaches (i.e., rational drug design) have generated antibiotic agents that act by related mechanisms. Some representative antibiotics are shown here. Image: (Knowles, 1997).

2.2 Antibiotic Resistance

Pathogens such as *Staphylococcus aureus (S. aureus)*, *Streptococcus pneumoniae (S. pneumoniae)* and *Enterococcus faecalis (E. faecalis)* have become increasingly resistant to multiple antibiotics (McDevitt and Rosenberg, 2001). Methicillin-resistant *S. aureus* (MRSA) and vancomycin-resistant *Enterococci* (VRE) infections tend to occur in hospitals whereas multi-drug-resistant *pneumococci* are more likely to occur in the community (Archer, 1998; Pfaller *et al.*, 1998; Jones *et al.*, 1999). If current trends continue, even vancomycin, a last line of defense for hospitalized patients, will no longer be effective (McDevitt and Rosenberg, 2001).

For example, *S. aureus USA300* is resistant to methicillin, ampicillin, penicillin, mupirocin, macrolides, lincosamides, tetracycline,

ciprofloxacin and streptogramin. Antibiotic resistance genes are present on strain USA300's chromosome and plasmid, they include *mecA* on the chromosome, *ermC* and *ileS* on both the chromosome and plasmid, and gene *blaZ* on the plasmid (Diep *et al.*, 2006). *S. aureus USA300* sensitive to vancomycin (Diep *et al.*, 2006). Vancomycin resistant *S. aureus* (VRSA) is a rare strain of *S. aureus* that was identified in 2002 (Weigel *et al.*, 2003); it has become resistant to the glycopeptide antibiotic vancomycin (DeLeo and Chambers, 2009).

The main mechanisms by which microorganisms exhibit resistance include drug inactivation or modification through the production of β-lactamases, alteration of a target site, alteration of a metabolic pathway, and reduced drug accumulation by decreasing permeability and/or increasing active efflux of the drugs across the cell surface (Hawkey, 1998). However, the acquisition of drug resistance is not necessarily the product of a slow evolutionary process.

Nonresistant bacteria may rapidly acquire resistance from resistant bacteria by horizontal gene transfer using one of three mechanisms: transformation, transduction and conjugation (Ochman *et al.*, 2000). "Unlike eukaryotes, which evolve principally through the modification of existing genetic information, bacteria have obtained a significant proportion of their genetic diversity through the acquisition of sequences from distantly related organisms. Horizontal gene transfer produces extremely dynamic genomes in which substantial amounts of DNA are introduced into and deleted from the chromosome" (Ochman *et al.*, 2000). The significance of horizontal gene transfer for bacterial evolution was recognized when multidrug resistance patterns emerged on a worldwide scale (Davies, 1996).

2.3 Broad Spectrum Antimicrobial Agents

The range of different microbes against which an antimicrobial agent acts is called call its *spectrum of activity*, and agents that are effective against bacteria from a wide range of taxonomic groups, including Gram-positive and Gram-negative, are called *broad spectrum*. Broad spectrum drugs are appropriate where a patient is seriously ill with an infection but the organism is not known (Black, 2008), i.e., such a drug will have a higher probability of success. Consequently, in this thesis, broadly conserved cell wall related proteins, especially those conserved in both Gram positive and Gram negative bacteria, are analyzed for their drug target potential.

2.4 The Cell Wall and Membrane as Targets

The cell wall is an essential structure for virtually all bacteria as it protects the cell from damage and osmotic lysis; it is the target of our best antibiotics (Leaver *et al.*, 2009). In general, antimicrobial drugs act on important structures or functions that exist in bacteria (prokaryotic cells) but not in humans (eukaryotic cells). For example, Gram-positive bacteria have high osmotic pressure and would burst without a sturdy cell wall. Penicillin and cephalosporin contain a *β lactam ring* which attaches to enzymes that cross-link peptidoglycans. Interference with the cross-linking of tetrapeptides prevents cell wall synthesis (Black, 2008).

Although all cell membranes are similar, those of bacteria are significantly different from animal cells. For example, some polypeptide antibiotics such as polymyxins act as detergents and distort bacterial cell membranes – they are especially effective against Gram-negative bacteria which have an outer membrane rich in phospholipids (Black, 2008).

3 COMPARATIVE GENOMICS AND VIRTUAL SCREENING

In this section we provide general background information on comparative genomics and virtual screening.

The application of comparative genomics for target identification is this paper is similar to the application in (Arigoni *et al.*, 1998) where conserved proteins were identified, determined to be essential and, therefore, potential drug targets. We chose comparative genomics to identify antibacterial targets because, while not a flood (Projan, 2003), an unprecedented number of novel targets have been discovered via this approach (Payne *et al.*, 2004).

Given broadly conserved proteins and their sequences, amino acids (and their locations) in those sequences that are important functionally and for binding tend to be conserved across species during evolution (Zvelebil and Baum, 2008, p586). Given a 3D structure, an *active site* is a localized combination of amino acids (which may be far apart in the amino acid sequence) that can interact with a chemically specific substrate and provide the protein with biological activity; as a result, proteins with very different amino acid sequences may fold into structures that produce the same active site (Mount, 2004, p415). Consequently, *in silico* screening of chemical compound libraries forms the second major part of our approach. Once an active site or probable binding site has been identified, the binding of small molecule ligands can be modeled. Molecular modeling is a high throughput way of rapidly investigating the binding potential of large numbers of small molecule drugs (Zvelebil and Baum, 2008, p587).

While the design, analysis and enhancement of ligands are also important steps in the drug discovery process (Lesk, 2008), they are outside the scope of this thesis.

3.1 Comparative Genomics

The first insights into genomic drug discovery approaches were made possible by the comparative analysis of the complete genome sequences of 10 bacterial pathogens (Galperin and Koonin, 1999). Broadly conserved genes conserved tend to be essential, which makes them attractive targets for new broad-spectrum antibiotics (Galperin and Koonin, 1999). In (Arigoni *et al.*, 1998), researchers used *Escherichia coli* (*E. coli*) as a reference organism. They identified four novel genes (*ygjD, ycfB, yihA, yjeQ*, and their respective orthologs). These genes were attractive targets, i.e., they were novel, essential, broadly conserved (homologs in a wide range of bacteria including *H. influenza, S. pneumoniae, H. pylori, and B. burgdorferi*), and low toxicity for higher organisms.

In other research, it was reported that many cell division proteins are conserved across Gram-positive and Gram-negative bacteria (Lock and Harry, 2008). Since essential bacterial genes tend to be more conserved that nonessential genes over short and long evolutionary time scales (Jordan *et al.*, 2002), the most highly conserved genes associated with cell wall biosynthesis should be the best targets for new broad spectrum antibiotics.

B. subtilis has approximately 4,100 genes; 192 were shown to be indispensable, 79 were predicted to be essential – with about 44 of these involved in the synthesis of cell envelope and the determination of cell shape and division (Kobayashi *et al.*, 2003). As of 2000, about 150 essential genes were identified in *S. aureus* (Ji *et al.*, 2001).

3.2 Sequence Analysis

3.2.1 Foundation The fundamental basis for our approach is the well established notion that similar sequences tend to have similar structures and, consequently, similar functions. Burkhard Rost analyzed more than a million alignments of pairs of protein sequences to determine a threshold for sequence identity (Rost, 1999), i.e., the minimum level of sequence identity required for an alignment to provide a reliable measure of homology. He found that 90% of sequence pairs with at least 30% identity over their whole length were structurally similar. Below 25% sequence identity, only 10% of the aligned pairs were structurally similar.

As a result, 30% sequence identity is often used to justify an initial presumption of homology (Zvelebil and Baum, 2008). The region between 30% and 20% has been called the *twilight zone* – where homology may exist but cannot be assumed without additional evidence – and below 20% lies the *midnight zone* (Zvelebil and Baum, 2008).

3.2.2 Pairwise Sequence Alignment The Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997) may be applied as a first step in sequence analysis. BLAST uses heuristics to find high scoring pairs and the computationally expensive Smith-Waterman algorithm (Smith and Waterman, 1981) to generate the final alignment (Altschul *et al.*, 1997). In general, each query is "blasted" against a database of proteomes to find similar proteins in other organisms. This generates one set of similar sequences for each query protein.

3.2.3 Complexity Alignment algorithms such as Needleman-Wunsch (Needleman and Wunsch, 1970) and Smith-Waterman (Smith and Waterman, 1981) are based on dynamic programming: pairwise sequence alignments have $O(n^2)$ time and space complexity; multiple sequence alignments of k sequences have $O(n^k)$ time and space complexity and are not feasible for alignments of more than three sequences (Koonin and Galperin, 2003). Hence, all available methods are approximations, i.e., hierarchical clustering of pairwise alignments roughly approximates the phylogenetic tree and guides the multiple alignment (Koonin and Galperin, 2003). This hierarchical approach to progressive sequence alignment, an idea presented in (Feng and Doolittle, 1987), reduces the $O(n^k)$ multiple alignment problem to a series of $O(n^2)$ problems (Koonin and Galperin, 2003).

3.2.4 Multiple Sequence Alignment (MSA) MSA may be viewed as a extension of pairwise sequence alignment that may reveal small active sites in otherwise dissimilar sequences. CLUSTAL is a multiple sequence alignment algorithm introduced in (Thompson *et al.*, 1994). It uses heuristics, as do all available MSA methods (Koonin and Galperin, 2003), to generate an alignment that is not guaranteed to be optimal for a given scoring scheme. CLUSTALW (Higgins *et al.*, 1996) is an improved version CLUSTAL (Higgins and Sharp, 1988) – the W stands for weighted. Although CLUSTALW is still the most widely used multiple sequence alignment tool, no significant improvements have been made to the algorithm since its introduction and several more recently developed algorithms perform better in terms of accuracy, speed or both (Edgar and Batzoglou, 2006).

In a recent review, MUSCLE (Edgar, 2004) and T-COFFEE (Notredame *et al.*, 2000) were listed as the most accurate programs (Edgar and Batzoglou, 2006). On nearly all benchmarks, these newer programs outperformed CLUSTALW in terms of average accuracy. Given a typical desktop PC, the authors recommended MUSCLE for aligning over 100 sequences that are approximately globally alignable (Edgar and Batzoglou, 2006). Since we expected to use a typical desktop PC for MSA of large numbers of sequences that are at least 30% identical, we chose MUSCLE generating for generating all multiple sequence alignment and Jalview (Waterhouse *et al.*, 2009) for MSA visualization – an example is shown in Figure 2.

3.3 Virtual Screening

Virtual screening is an *in silico* structure-based design strategy where small molecules are "docked" into the structures of target molecules and scored for their complementarity to binding sites. This process models similar events that occur in nature. For example, Figures 3a and 3b show two proteins, FtsZ and CoaD. An organic ligand (in green) bound to an active site on the surface of each protein; the figures also show polar contacts as defined in the Protein Data Bank (pdb) file for each protein.

Fortunately, accurate energy calculations and scores are not necessarily required for meaningful compound selection and, in a typical docking study, a large compound database may be reduced to a short list of $\sim$100 preferred candidates (Kitchen *et al.*, 2004). Docking and scoring are important steps in the filter between a total potential library and testing at the bench (Lesk, 2008) and, by some measures, its major achievement has been to eliminate inactive compounds from further consideration (Bhöm and Schneider, 2000). Indeed, such computational approaches have been widely used for hit identification and lead optimization (Kitchen *et al.*, 2004).

A central problem in drug discovery is the identification of a compound that will bind tightly and specifically to a target protein. Tight binding promotes efficacy at low concentrations, and specificity helps to minimize side effects (Lesk, 2008). Although it is difficult to estimate absolute affinities, comparative docking can indicate relative affinities. This means that given a scoring function that can rank ligands in approximate order of affinity, we can select compounds and set priorities for experimental testing (Lesk, 2008).

However, the probability of finding a novel antibacterial compound with broad spectrum activity depends on the diversity of biologically relevant compounds available for screening. Chemical diversity is an extremely important factor when searching for new drugs (Payne *et al.*, 2007), and ZINC provides a large database of millions of commercially available compounds for virtual screening (Irwin and Shoichet, 2005).

Docking *in silico* of large sets of ligands is computationally expensive and, consequently, high-throughput computing clusters and algorithms are widely used for virtual screening (Prakhov *et al.*, 2010). We found that virtual screening is well suited to distributed computing. Screening one compound does not depend on the output from screening other compounds. As a result, many compounds can be screened simultaneously on many CPUs located anywhere on a network.

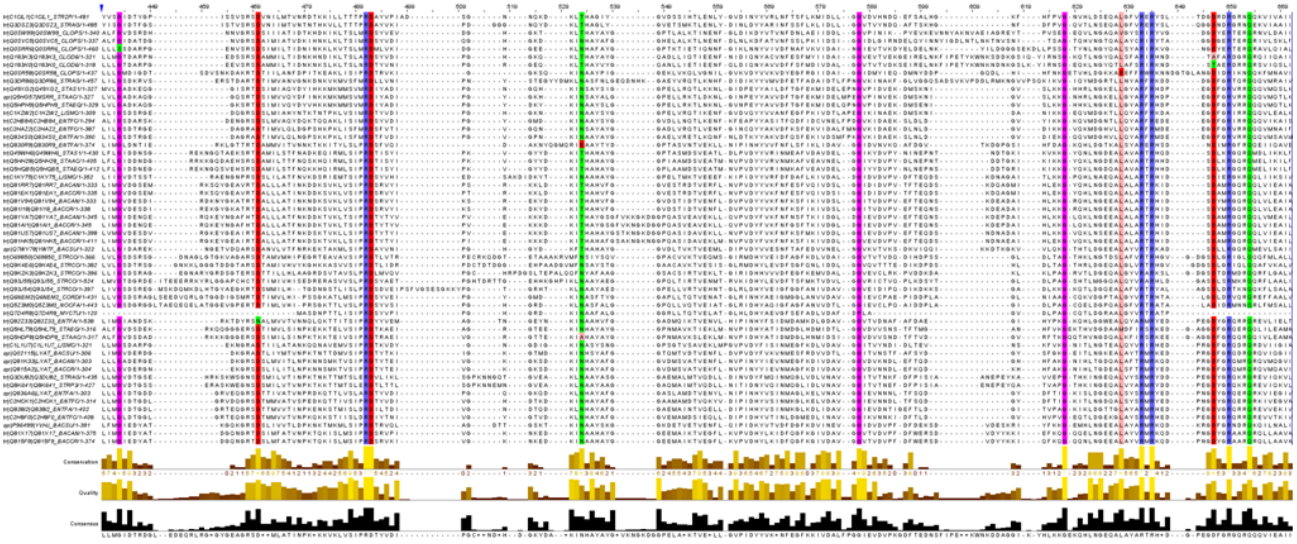
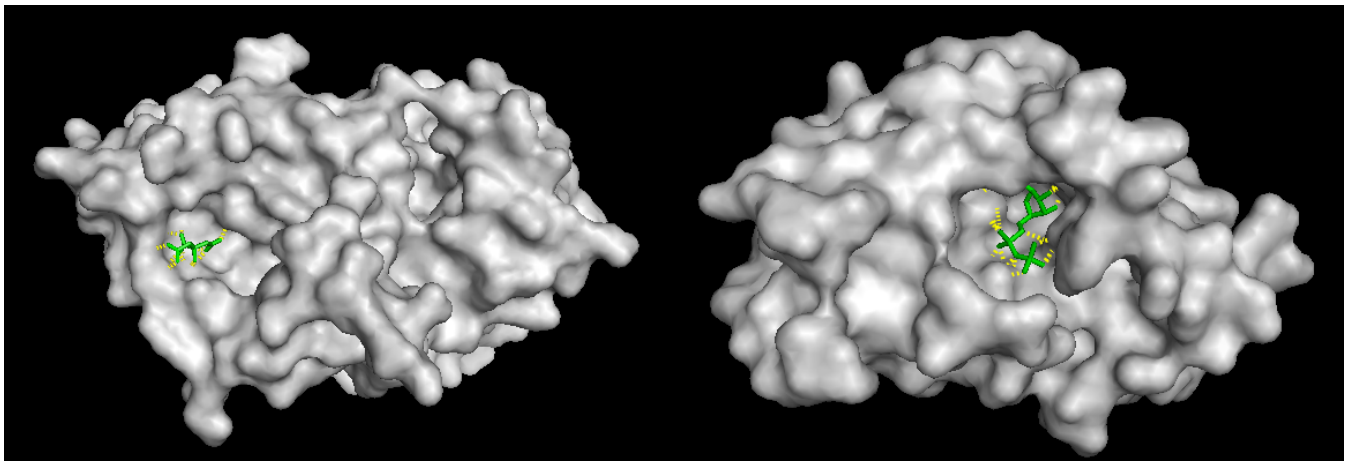


Fig. 2: Multiple sequence alignment for sequences similar to *B. subtilis* (columns with 100% conservation are colored).



(a) FtsZ 2VXY.pdb. (b) CoaD 1O6B.pdb.

Fig. 3: Pymol images of FtsZ and CoaD with organic ligand using Protein Data Bank files.

4 METHODS

In this section we describe the application of our methods and algorithm parameters.

A list of 123 *B. subtilis* proteins was provided by Dr. Nick Allenby of Demuris Ltd., Newcastle Upon Tyne, United Kingdom. We downloaded these *B. subtilis* proteins from the UniProt website. Complete protein lists may be found in the Supplementary Material, Tables **??** and **??**. We used the Database of Essential Genes (Zhang *et al.*, 2004; Zhang and Lin, 2009) to confirm whether or not broadly conserved genes are essential.

Next we downloaded 60 complete bacterial proteomes in fasta format from the Integr8 sequence database (Kersey *et al.*, 2005). The 60 proteomes were selected for pathogenicity and diversity. They included 25 Gram positive, 25 Gram negative, and 10 proteomes, such as Mycoplasmas, that do not have a cell wall (Glass *et al.*, 2006). Complete proteome lists may be found in the Supplementary Material, Table **??**.

As our first computational task, we compared each of the 123 *B. subtilis* proteins with every protein in each of the 60 proteomes in order to identify potential drug targets, i.e., we used BLAST (Altschul *et al.*, 1997) for pairwise sequence alignment to identify broadly conserved cell wall proteins. Then we used MUSCLE (Edgar, 2004) for multiple sequence alignment to identify highly conserved residues in each query sequence and mapped conservation scores for those residues to the query protein's 3D surface to indicate probable active sites. The workflow for the these tasks is shown in Figure 4.

We also used BLAST to identify proteins conserved in the 25 Gram positive and the 25 Gram negative proteomes; and to gauge the redundancy of

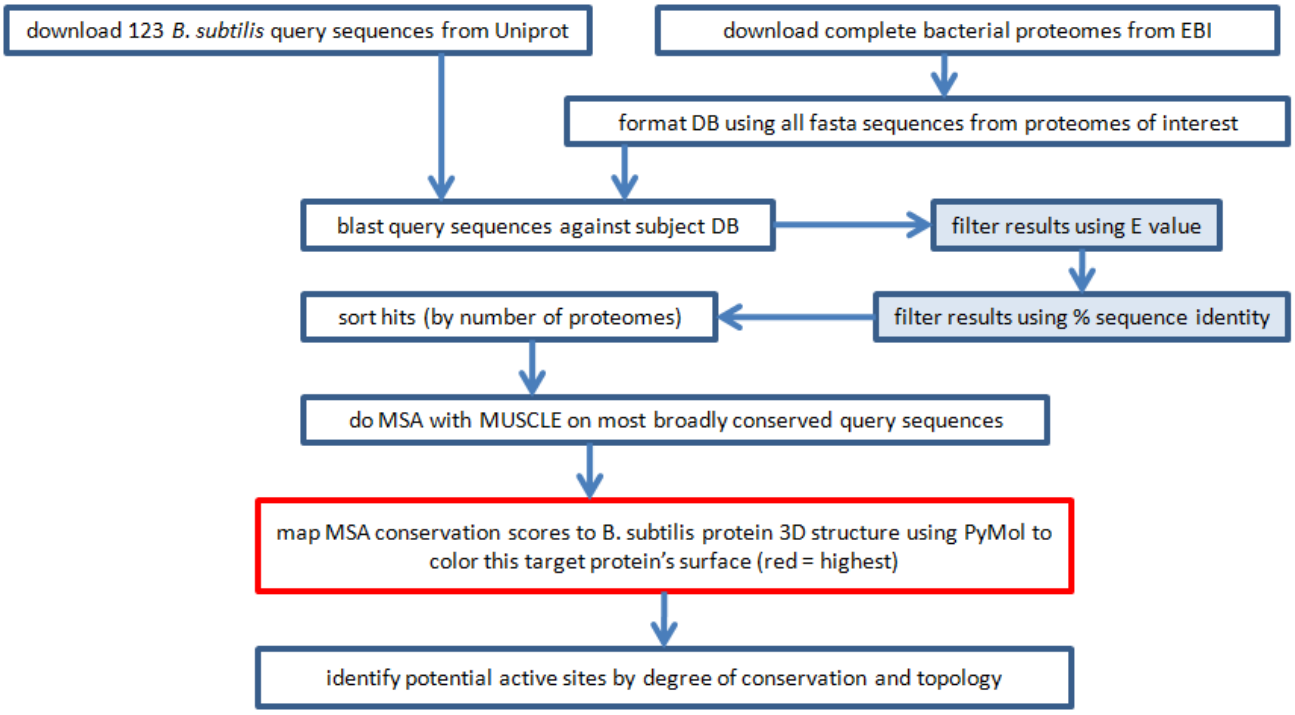


Fig. 4: Workflow 1: pairwise and multiple sequence alignment; mapping conservation scores by color to surface of *B. subtilis* 3D structure.

broadly conserved proteins in the *B. subtilis* proteome, i.e., we assumed that relatively unique proteins are more likely to be essential.

After identifying potential targets, we docked NPD ligands from the online ZINC database with each potential target protein *in silico*, i.e., we used Vina (Trott and Olson, 2010) to perform virtual screening to identify chemical compounds for *in vitro* testing. The search space that included the most promising, highly conserved region of the receptor molecule – this region either included a known active site or a predicted active site. And finally, top scoring ligands were selected for *in vitro* testing. The workflow for these tasks is shown in Figure 5.

4.1 Pairwise Sequence Alignment

We used the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997) to identify the set of proteins that have homologs in many species of pathogenic bacteria. BLAST uses heuristics in an attempt to align pairs of sequences such that the resulting alignment's similarity score S is maximized and, consequently, its E-value, which is a a measure of the reliability of the S score, is minimized (Korf *et al.*, 2003).

$$E = kmne^{-\lambda S} \tag{1}$$

Equation 1 states that E, the number of alignments expected by chance, is a function of several parameters: (1) the size of the search space $m*n$; (2) the normalized score λS, and (3) a minor constant k. In a database search, the size of the search space is the product of the number of letters in the query m and the number of letters in the database n. As equation 1 indicates, a given increase in S results in an exponential decrease in the expected number of alignments E with a similarity score at least as good as S; therefore, the smaller the E value, the more significant the alignment (Korf *et al.*, 2003).

From Equation 1: A lower E value reduces false positives at the expense of an increase in false negatives. In fact, very short sequence alignments (i.e., small m) with very high S scores will be filtered out because, all else being equal, a shorter sequence will result in a higher e-value and, therefore, deemed less statistically significant. Consequently, alignments with very short *B. subtilis* query proteins were filtered out and not analyzed.

> The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. ... The lower the E value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance. [http://blast.ncbi.nlm.nih.gov]

The following pseudo code represents our use of BLAST (Altschul *et al.*, 1997) to identify one or more homologous proteins for each *B. subtilis* query protein. **Input**: a query fasta file of 123 protein sequences, 60 subject proteomes, and a set of thresholds T: a BLAST E value of 0.00001 (a smaller E value will filter out more alignments than a larger one) and a minimum percent identity of 30%.

```
foreach protein P in the B_subtilis fasta file
  blast P against the proteomes
    for each hit > T1
      record hit

foreach protein P in the B_subtilis fasta file
  if P had >= 1 hit in >= T2 proteomes
    save set of hit sequences to file
```

Output: a list of *B. subtilis* proteins that are probably homologous to proteins in subject species ranked in order of the number of subject species with one or more probable homologs (hits).

Time complexity: blasting 123 *B. subtilis* proteins against 60 proteomes is relatively fast: ≈ 162 seconds, or 2.7 seconds per bacterial proteome. This step of our research was greatly facilitated by software that we implemented in the Java programming language, shown in Figure **??** of the Supplementary Material.

Although outside the scope of this thesis, we note that additional work may be needed for aligning subsequences that constitute active sites, i.e., multiple protein sequences may lack global similarity but possess critical local similarity.

4.2 Multiple Sequence Alignment

After we determined which sequences were suitable, based on E value, sequence identity, and whether or not the top scoring *B. subtilis* proteins had a known 3D structure, we generated multiple sequence alignments using MUSCLE (Edgar, 2004). Each MSA included the *B. subtilis* sequence and, typically, about 50 very similar sequences from at least 40 of the 60 proteomes in our local database.

The MSAs produced by MUSCLE were visualized with Jalview (Waterhouse *et al.*, 2009) and conservation scores were extracted with a Perl script, mapped to each amino acid in the *B. subtilis* protein of interest, saved to a pml file, and visualized with PyMOL (Delano, 2002).

4.3 Virtual Screening

The virtual screening method applied in this research is a receptor based screen Kitchen *et al.* (2004) where we attempted to "dock" members of a chemical library against a given protein structure *in silico* and software predicted the conformation and binding affinity of NPD ligands.

4.3.1 Chemical Compounds used as Ligands Natural Products Database (NPD) of commercially available natural products and natural product derivatives was downloaded from the ZINC database (Irwin and Shoichet, 2005). The 2008/5 version of the NPD contains 89,425 entries for commercially available compounds from seven vendors that advertise their compounds as being of natural origin, either pure natural products, or chemical derivatives of natural products. We also downloaded the "usual" data sets of 11 random micro vendors, $approx$ 11,000 from the ZINC database. The vendors are listed in Tables **??** and **??** of the Supplementary Material.

4.3.2 Docking Software For virtual screening, we considered AutoDock 4 (Morris *et al.*, 1998; Huey *et al.*, 2007) and AutoDock Vina (Vina) (Trott and Olson, 2010). Both have been found to be are equally capable in ranking smaller molecules with few rotatable bonds, and both exhibit a size-related bias in scoring (Chang *et al.*, 2010). However, AutoDock Vina has an improved local search algorithm and the ability to detect and utilize multiple CPUs. As a result, Vina executes more quickly than AutoDock 4 and ranks larger molecules more accurately, "researchers should look to it first when undertaking a virtual screen" (Chang *et al.*, 2010).

Hence, given a set of target proteins and the Natural Products Database (NPD) ligands from the ZINC database (Irwin and Shoichet, 2005), we used Vina (Trott and Olson, 2010) to screen these ligands versus our target proteins, i.e., we used Vina to predict how well our drug candidates will bind *in vitro* to probable active sites (identified above by sequence analysis) of known 3D structure. Note: to achieve high throughput, depending on availability, up to 88 receptor-ligand docking operations were performed in parallel on a cluster of computers with a total of 88 CPUs (cisbclust at Newcastle University). This reduced expected time complexity from months to days, i.e., from ≈ 30 to 60 seconds to less than 1 second per docking.

4.3.3 Ligand and Receptor Preparation for Vina The "usual" set NPD ligands were downloaded from ZINC in the mol2 file format (p0.0, p0.1, p0.2, p0.3, p0.4, p1.0). Each file, which contained thousands of ligands, was split into single ligand files in mol2 format. Each single ligand file was prepared for docking and converted to pdbqt format using Perl and the prepare_ligand4.py script from AutoDockTools (Sanner, 1999). Each protein receptor was prepared and converted to pdbqt format using the AutoDockTools graphical user interface (Sanner, 1999): polar hydrogens were added before conversion to pdbqt format.

4.3.4 Search Space for Vina AutoDockTools (Sanner, 1999) was also used calculate the x, y, z coordinates of the search space, Table 1, to be explored by Vina (Trott and Olson, 2010); spacing was set to 1.000 Angstrom. The probable active sites identified from our sequence analysis helped to limit this search space and reduce computation time. We noted that the quality of the starting coordinates for each receptor, in effect, put a limit on the accuracy of our docked results (Morris *et al.*, 2009).

Table 1. Vina Search Space Parameters (Angstroms).

receptor	center_x	center_y	center_z	size_x	size_y	size_z
FtsZ	-17.047	31.454	24.275	22	22	22
CoaD	126.100	69.632	117.280	20	20	20
YwtF	24.077	58.595	85.242	22	22	22
CoaD	-.0972	12.775	18.182	34	24	28

4.3.5 Vina Parameters We used Vina's default parameters except for exhaustiveness, which we set to 6 instead of 8. This allowed us to uniformly screen the ligands against more receptor proteins in the time available and, time permitting, subsequently screen a small number of high scoring ligands with exhaustiveness set to at a higher level to make a final selection for *in vitro* testing.

5 RESULTS

Our BLAST results indicated that four *B. subtilis* proteins with known 3D structure were highly conserved; FtsZ and CoaD in both Gram positive and Gram negative bacteria, Table 2; and YwtF and RacE in Gram positive bacteria only, Table 3. Complete tables, including information indicating the existence of a known 3D structure for matching proteins on other proteomes, may be found in the Supplementary Material, Tables **??**, **??** and **??**. We confirmed that two are listed as essential genes for multiple species of bacteria, at the Database of Essential Genes (DEG) web site (Zhang *et al.*, 2004; Zhang and Lin, 2009). This database listed 12 species for FtsZ, 6 species for CoaD, 0 species for YwtF (function is unknown), and only *B. subtilis* was listed for RacE. However, our results suggest that all four may serve as viable drug targets.

Our multiple sequence alignment results showed that highly conserved regions of each protein's 3D structure was consistent with known or plausible active sites. We also found that virtual screening by Vina predicted binding sites for the NPD ligands that were consistent with sites characterized by high conservation, pocket topologies, and experimentally verified organic ligand binding sites as recorded in the FtsZ (2VXY) and CoaD (1O6B) pdb files from the Protein Data Bank; and the positions of high scoring ligands docked with YwtF (3MEJ) and RacE (1ZUW) pdb files were consistent with regions found by our sequence analysis to be highly conserved.

5.1 BLAST

The following four *B. subtilis* proteins, FtsZ, CoaD, YwtF, and RacE, have a known 3D structure and were found to be broadly conserved in our BLAST results, Tables 2 and Table 3. While this

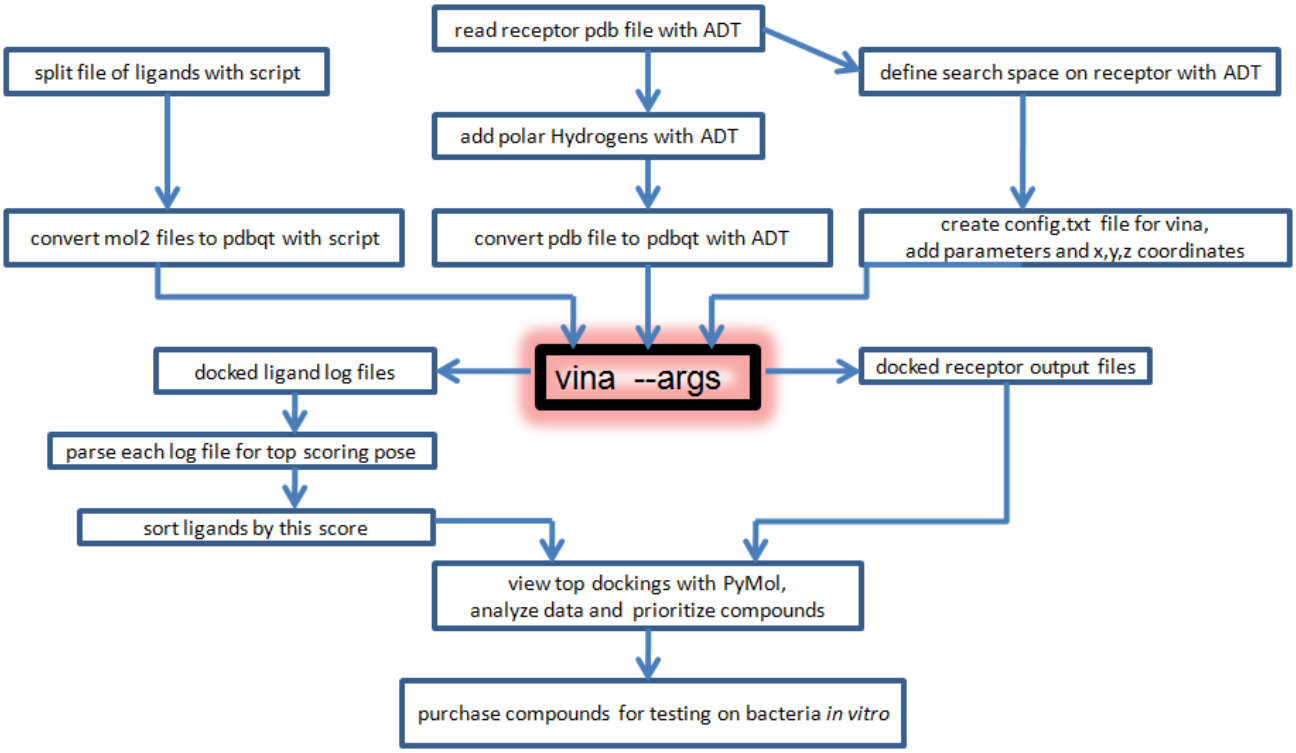


Fig. 5: Workflow 2: virtual screening and ligand selection.

alone suggests that they might serve as suitable drug targets, FtsZ and CoaD are unique in the *B. subtilis* proteome – as Table 5 shows, FtsZ and CoaD had no matches with other proteins in their own proteome and YwtF and RacE had only one match other than with themselves. The complete table may be found in the Supplementary Material, Table **??**.

5.1.1 Four proteins: FtsZ, CoaD, YwtF, RacE In this section we provide a concise summary of our four proteins of interest, but provide somewhat more detail for YwtF and RacE because their drug target potential may be less obvious. **FtsZ** may be the most critical component of bacterial cell division machinery (Errington *et al.*, 2003). It is an essential bacterial protein (Kobayashi *et al.*, 2003), guanosine triphosphatase, and a homolog of mammalian beta-tubulin that polymerizes and assembles into a ring to initiate cell division (Haydon *et al.*, 2008). **CoaD** was found to be essential for several species of bacteria, including *Streptococcus pneumoniae* (Thanassi *et al.*, 2002; Song *et al.*, 2005; Song and Ko, 2008). The structure of CoaD was reported in (Badger *et al.*, 2005). **YwtF**: Although the 3D structure for this protein has been determined its function is unknown [http://dx.doi.org/ 10.2210/pdb3mej/pdb]. Specifically, this entry describes a domain of unknown function that is found in the predicted extracellular domain of a number of putative membrane-bound proteins, which includes protein psr: described as a penicillin binding protein 5 (PDP-5) synthesis repressor. Another putative membrane-bound protein is Bacillus subtilis LytR. And a third protein, CpsA, is described as a putative regulatory protein involved in exocellular polysaccharide biosynthesis [InterPro entry IPR004474]. **RacE**: D-glutamate is an essential building block of the peptidoglycan layer in bacterial cell walls and can be synthesized from L-glutamate by glutamate racemase (RacE). The structure of a complex of *B. subtilis* RacE (Kobayashi *et al.*, 2003) with D-glutamate reveals that the glutamate is buried in a deep pocket. This structure provides new insights into the RacE mechanism and an explanation for the potency of a family of RacE inhibitors (Ruzheinikov *et al.*, 2005).

5.2 MSA and active sites

Based on our BLAST results for *B. subtilis* versus 60 proteomes, we used MUSCLE to generate MSAs using multiple sequences: 54 for FtsZ, 50 for CoaD, 57 for YwtF, and 38 for RacE. The four proteins vary in length: FtsZ 382, CoaD 161, YwtF 322, and RacE 272 amino acids; and various numbers of gaps were inserted into their MSAs. Figure 6 shows an overview of the MSA for FtsZ, CoaD, YwtF and RacE respectively. These visualizations were generated by Jalview such that colored bars appear only at positions where residues are 100% conserved.

The MSA for FtsZ in Figure 6a shows the greatest degree of conservation, and this is even more apparent in Figure 7a where the conservation scores calculated by Jalview, ranging from 0 (lowest) to 11 (highest), have been mapped to the surface of FtsZ as colors: gray {0..5}, blue {6,7}, yellow {8}, orange {9}, and red {10,11}. Perhaps the most striking feature of this FtsZ 3D image is the large highly conserved area, colored red, and a deep pocket on the left side of Figure 7a. In contrast, the opposite side, Figure 7b, shows less conservation and lacks a similar pocket.

Table 2. *B. subtilis* query proteins versus 60 Proteomes

Proteomes	Hits	mean id.	UniProt #	GN	3D
54	54	55%	P17865	*ftsZ*	yes
51	70	50%	P70965	*murAA*	no
51	70	47%	P19670	*murAB*	no
51	61	46%	O31751	*uppS*	no
50	54	39%	Q03523	*murE*	no
49	49	46%	O34797	*coaD*	yes
48	49	46%	P14192	*glmU*	no
48	60	38%	P40778	*murC*	no
47	77	34%	P31114	*hepT*	no
46	80	39%	P54383	*ispA*	no
46	66	47%	O31822	*yngB*	no
46	64	51%	Q05852	*gtaB*	no
45	45	36%	P96613	*murF*	no
44	57	39%	P96612	*ddl*	no
43	46	37%	Q06755	*ispD*	no
42	43	37%	Q03522	*murD*	no
41	55	43%	P54523	*dxs*	no

Table 3. *B. subtilis* query proteins versus 25 Gram Positive Proteomes

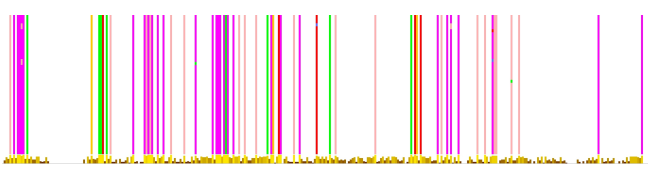
Proteomes	Hits	mean id.	UniProt #	GN	3D
25	25	63%	P17865	*ftsZ*	yes
25	25	51%	P14192	*glmU*	no
25	25	50%	O34797	*coaD*	yes
25	34	48%	O31751	*uppS*	no
24	27	41%	Q03523	*murE*	no
23	29	43%	P96612	*ddl*	no
21	24	47%	P94556	*racE*	yes
21	58	34%	Q7WY78	*ywtF*	yes
19	20	34%	O05412	*yrpC*	no
16	18	51%	O31753	*dxr*	no
16	33	37%	P39581	*dltA*	yes
15	16	60%	P54482	*ispG*	no

Table 4. *B. subtilis* query proteins versus 25 Gram Negative Proteomes

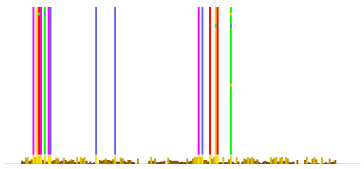
Proteomes	Hits	mean id.	UniProt #	GN	3D
25	25	47%	P17865	*ftsZ*	yes
25	26	46%	P70965	*murAA*	no
25	25	45%	O31751	*uppS*	no
25	26	42%	P19670	*murAB*	no
25	27	36%	Q03523	*murE*	no
24	24	42%	O34797	*coaD*	yes
23	25	40%	P14192	*glmU*	no
23	43	37%	P54383	*ispA*	no
23	39	32%	P31114	*hepT*	no
23	34	32%	P40778	*murC*	no
22	22	56%	Q06756	*ispF*	no
22	28	42%	P54523	*dxs*	no
22	22	33%	P96613	*murF*	no
21	27	46%	Q05852	*gtaB*	no
21	31	43%	O31822	*yngB*	no
21	29	36%	P96612	*ddl*	no
21	43	34%	P38422	*dacF*	no
21	22	32%	Q03522	*murD*	no
20	20	36%	Q06755	*ispD*	no
19	20	37%	P54473	*ispH*	no
19	19	34%	O05412	*yrpC*	no
19	40	33%	P35150	*dacB*	no
18	18	41%	O31753	*dxr*	no
17	95	37%	P96740	*pgdS*	no

The MSA for CoaD in Figure 6b shows less conservation than FtsZ and its highly conserved residues are grouped closer together. Figure 7c shows a more compact highly conserved region (mostly red) and a substantial pocket; in stark contrast, the opposite side, shown in Figure 7d, shows very little conservation or concave areas.

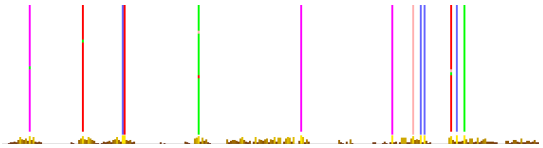
The MSA for YwtF in Figure 6c shows much less conservation than FtsZ. However, as Figure 8a shows, this protein's highly conserved residues map to one highly conserved surface region



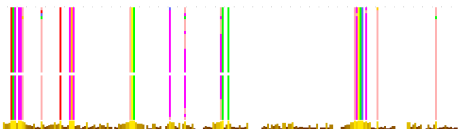
(a) FtsZ



(b) CoaD

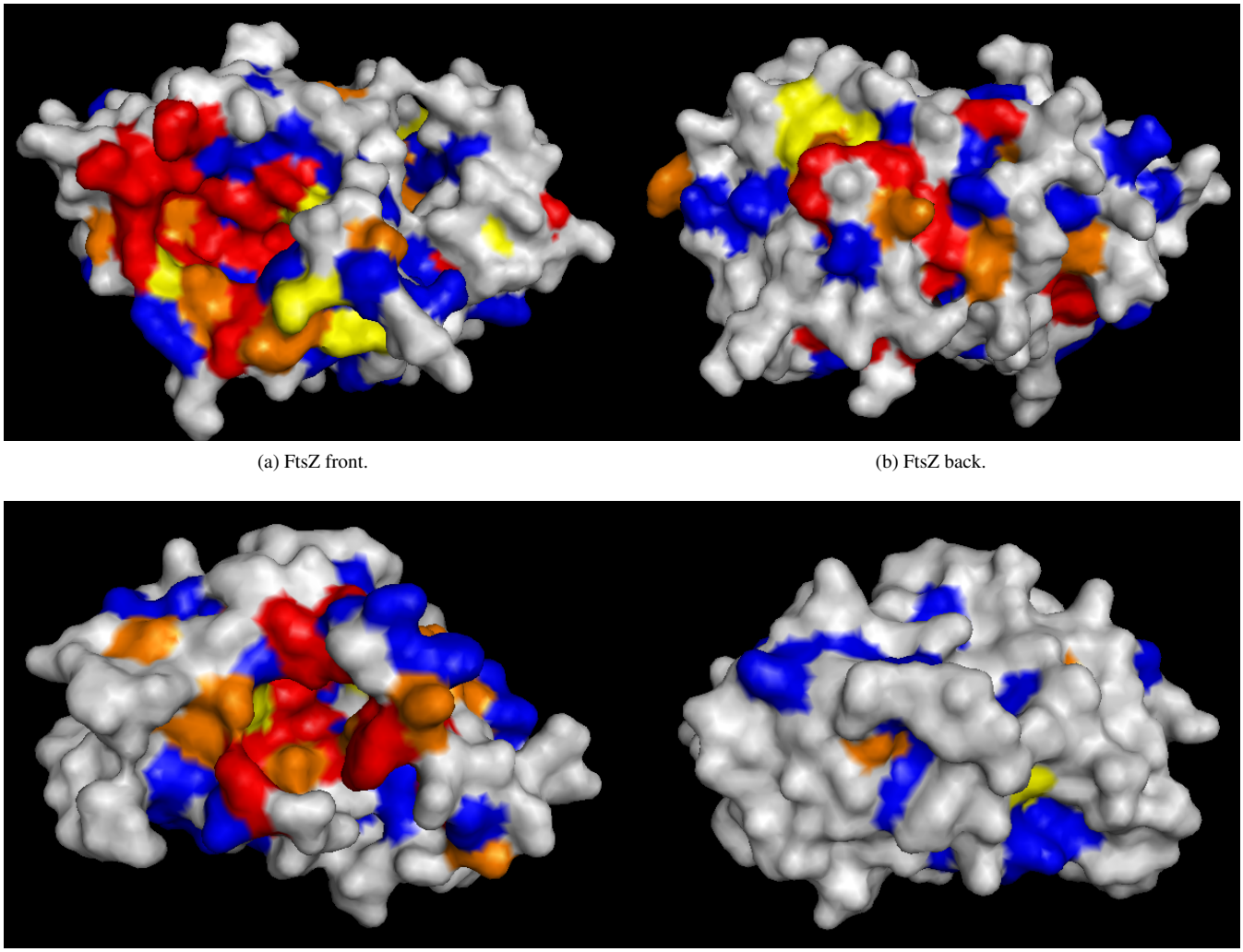


(c) YwtF



(d) RacE

Fig. 6: Multiple sequence alignments: columns are in color where residues are 100% conserved. The FtsZ sequence shows the greatest degree of conservation.



(a) FtsZ front. (b) FtsZ back.

(c) CoaD front (d) CoaD back

Fig. 7: FtsZ and CoaD with surface colored by conservation score: gray {0..5}, blue {6,7}, yellow {8}, orange {9}, red {10,11}.

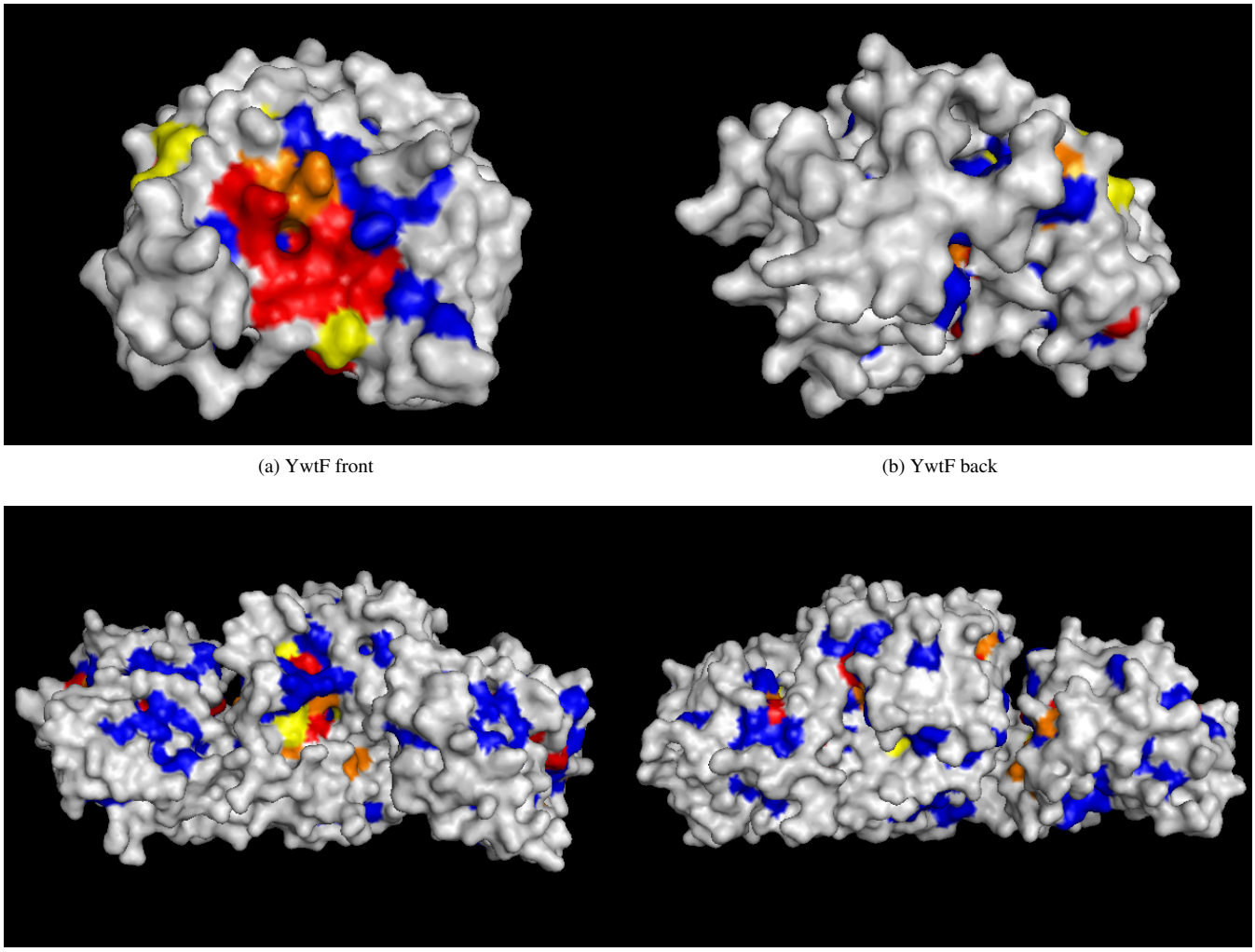
(mostly red) which lacks a substantial pocket but nonetheless presents a region that probably plays an important role. In contrast, the opposite side, shown in Figure 8b, shows very little conservation.

The MSA for RacE in Figure 6d appears to show more conservation than the MSA for YwtF, but this greater conservation is not evident from the surface colors for RacE in Figures 8c and 8d. The implication is that many of RacE's highly conserved residues are located in the interior. This was confirmed by rendering RacE as lines in PyMOL, shown in Figure **??** of the Supplementary Material. Consequently, unlike the other three proteins, RacE does not present an obvious target for docking ligands – where highly conserved surface regions, especially those characterized by a substantial pocket, are probably good drug targets – but its highly conserved residues are internal, the topology of RacE is complex and, given the potency of a family of RacE inhibitors developed as novel antibiotics (Ruzheinikov *et al.*, 2005), merits further investigation.

5.3 Virtual Screening

A concise illustration of our results from screening ≈ 90,000 ligands against the four proteins identified from our BLAST results is shown in Figure 9. Scores were relatively consistent between proteins, i.e., NPD ligands docked with FtsZ tended to score the best (lowest binding energy as measure by kcal/mol), followed by CoaD, YwtF, and RacE; Table 6. In addition, NPD ligands scored consistently higher than the micro vendor ligands, Table **??** of the Supplementary Material.

Docking scores indicate that FtsZ contains the most promising drug target, while CoaD which has a similar appearance (Figure 7c), also scores relatively well. Figures 10a, 10b, 10c, and 10d show the top scoring ligand for each protein docked with FtsZ, CoaD, YwtF, and RacE respectively. We noted that the top scoring ligands for FtsZ tend to have a bend similar to the one shown in Figure 10a such that they conform closely to the contours of the pocket, which



(a) YwtF front (b) YwtF back

(c) RacE front (d) RacE back

Fig. 8: FtsZ and CoaD with surface colored by conservation score: gray {0..5}, blue {6,7}, yellow {8}, orange {9}, red {10,11}.

Table 5. Redundancy of 123 *B. subtilis* proteins in the *B. subtilis* Proteome. Using BLAST with an E value of 0.1 and filtered with a sequence identity threshold of $>= 30\%$.

Proteomes	Hits	mean identity	UniProt #	GN	3D
1	1	100%	P17865	*ftsZ*	yes
1	1	100%	O34797	*coaD*	yes
1	1	100%	P39844	*dacC*	yes
1	1	100%	P50740	*fni*	yes
1	1	100%	P39131	*mnaA*	yes
1	1	100%	P27623	*tagD*	yes
1	2	66%	O05412	*yrpC*	no
1	2	66%	P94556	*racE*	yes
1	3	59%	P96499	*yvhJ*	no
1	3	59%	Q02115	*lytR*	no
1	3	58%	Q7WY78	*ywtF*	yes

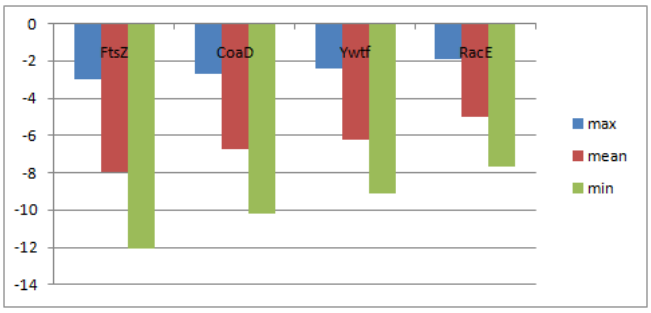


Fig. 9: Docking scores in kcal/mol for NPD ligands. A more negative value indicates a lower binding energy.

Table 6. Top scoring NPD ligands using predicted site

Protein Name	UniProt #	pdb id	Top Scoring Ligand	Score kcal/mol
FtsZ	P17865	2VXY	ZINC04236084	-12.10
CoaD	O34797	1O6B	ZINC03846570	-10.20
YwtF	Q7WY78	3MEJ	ZINC08791231	-9.10
RacE	P94556	1ZUW	ZINC03844349	-7.70

Table 7. Docking Scores for NPD p0.4 ligands: search areas include predicted active site (a) versus predicted inactive site (b).

receptor	$mean_a$	$mean_b$	min_a	min_b	max_a	max_b
FtsZ	-8.24	-5.89	-11.90	-8.40	-3.40	-2.50
CoaD	-6.97	-4.61	-9.60	-6.60	-2.80	-1.90
YwtF	-6.41	-6.75	-9.00	-11.40	-2.60	-2.70
RacE	-5.21	-5.64	-7.60	-8.30	-2.10	-2.60

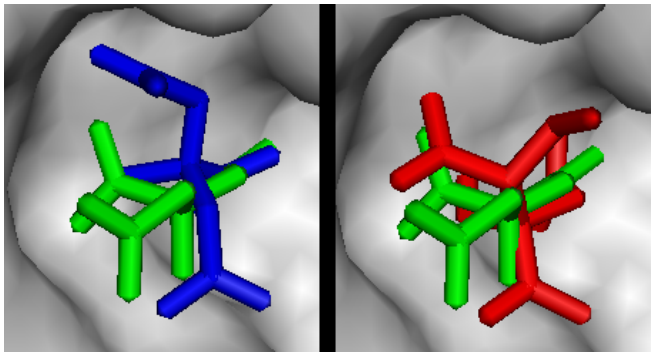
is also a characteristic of high scoring ligands for the other proteins. This is consistent with other research, which has found that in most cases the largest pocket/cavity is the active site (Liang *et al.*, 1998).

5.4 Validation

We validated our *in silico* results using against the data contained in the Protein Data Bank pdb files and ZINC mol2 ligand files.

(1) Highly conserved regions in FtsZ and CoaD were compared with their known active sites. We found that the locations of highly conserved residues corresponded very closely to known active sites.

(2) Although Vina's accuracy was confirmed in (Chang *et al.*, 2010), we compared the predicted binding site of ligands similar to the organic FtsZ ligand, CIT, with CIT's experimentally determined binding site. As the images in Figure 11 indicate, the organic ligand's known position closely matches the position predicted *in silico* for a virtually identical ligand.



(a) Modified ligand (blue) (b) ZINC1726452 ligand (red)

Fig. 11: FtZ with organic ligand (green) in known position versus docked ligands (a) and (b).

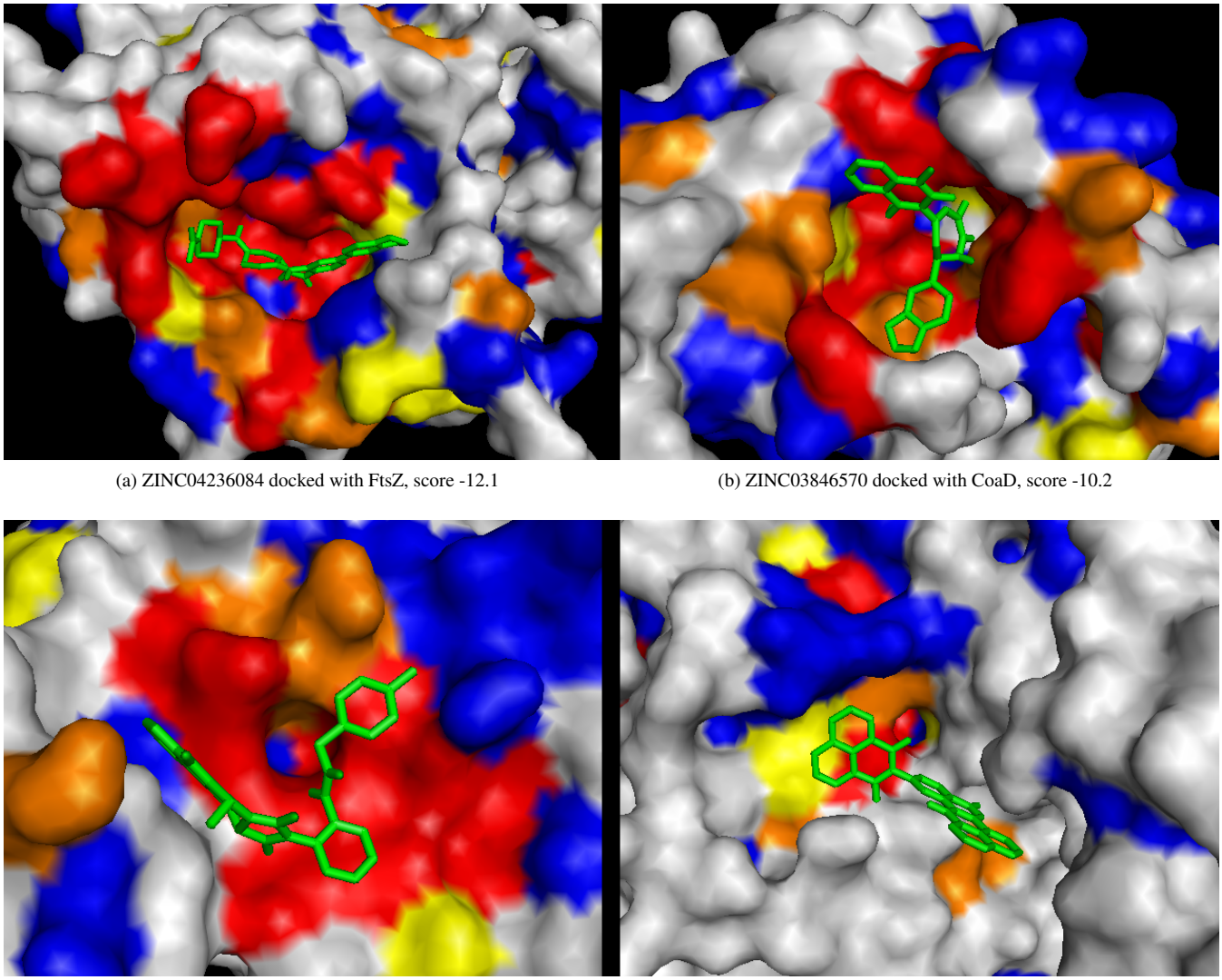
(3) We compared the docking scores of a set of 10777 NPD ligands (p0.4) when docked within a search area centered on our predicted active sites versus docking scores for the same set of ligands in a search space of the same size but centered on a poorly conserved region on the opposite side of each protein. Although results for FtsZ and CoaD were as expected – the ligands docked on the "back" side did not score as well – ligands docked on the "back" side of YwtF and CoaD scored better, Table 7. This was quite interesting. However, Figure 12c shows that the top scoring ligand for YwtF in this case is docked deep in an elaborate pocket characterized by some conservation (blue ={6,7} on the scale from 0 to 11), whereas the "front" side of YwtF presents a large highly conserved area, Figure 10c, lacks a substantial pocket. In the case of RacE, Figure 12d shows the top scoring ligand docked in a small pocket in a relatively conserved area which, in hindsight, is not obviously a less likely binding site than the one shown in Figure10d.

6 DISCUSSION

In this thesis we considered 123 genes associated with bacterial cell wall biosynthesis and cell division from the *B. subtilis* genome. From these genes we used a comparative genomics approach to identify essential proteins in multiple bacterial species because the proteins most conserved across species are most likely to be essential for the survival of more species of bacteria (Arigoni *et al.*, 1998) and, therefore, good potential drug targets (Galperin and Koonin, 1999; Pucci, 2006). We found that the conserved residues in each protein coincided with either known actives sites, or sites where the best scoring ligands docked with the lowest binding energy. However, while our results indicate that the combination of very highly conserved residues and a substantial pocket clearly indicate a potential drug target, less favorable conservation scores and surface topology indicate the need for further investigation and probably a larger search space for the docking software.

While the images of proteins presented in this paper may appear to be precisely colored according to conservation and, therefore, functional importance, we can only say that based on the data and algorithms used, some surface regions are probably more important than others. First, the sequences used for MSA depended not only on the set of input sequences but on BLAST (which uses hueristics) and its parameters. Second, the MSAs produced by MUSCLE may differ from those produced by other widely used algorithms and, of course, all are probabilistic. Different MSAs may yield differ conservation scores and different 3D images. Third, blasting against a different set of proteomes may also lead to different conservation scores. Fourth, the results from virtual screening depended on multiple factors including the set of ligands and the search parameters used by AutoDock Vina. Another cause for concern is that the 3D structure of the vast majority of proteins is not known, which means our investigation has in fact been limited to a small minority of *B. subtilis* proteins.

The scope for future work includes many areas. First, this work could be extended to consider the 3D structure of proteins similar to essential *B. subtilis* proteins. Second, additional methods and resources could be combined with ours to form an ensemble with greater



(a) ZINC04236084 docked with FtsZ, score -12.1

(b) ZINC03846570 docked with CoaD, score -10.2

(c) ZINC08791231 docked with YwtF, score -9.1

(d) ZINC04015296 docked with RacE, score -7.7

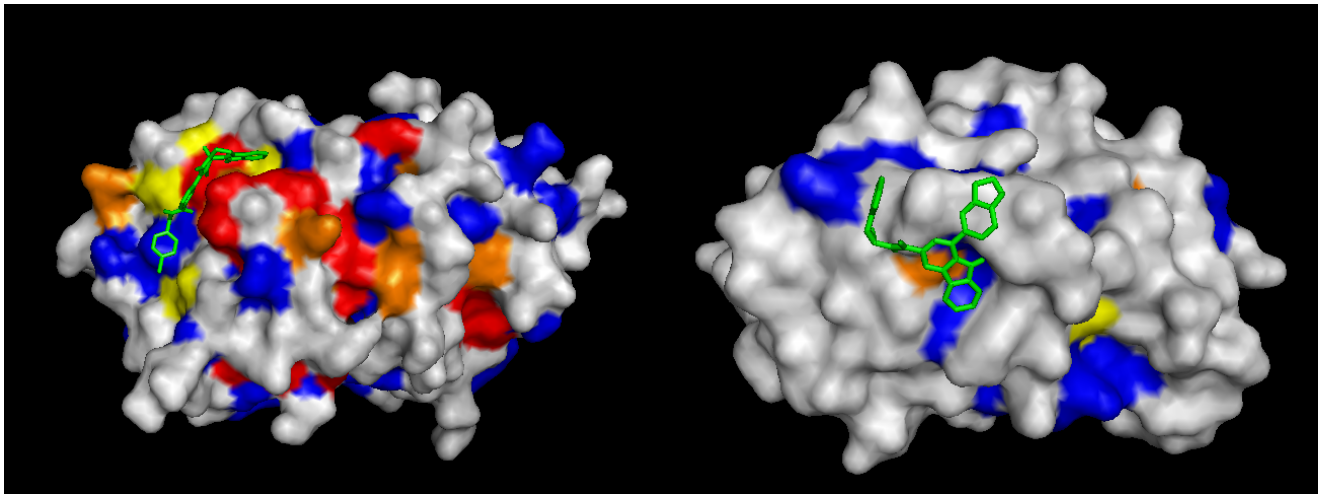
Fig. 10: Top scoring NPD ligands (complete set) docked in search areas that include **predicted active sites**.

accuracy. Third, the data analyzed in this paper should be amenable to machine learning techniques. For example, which features of a chemical compound are the best predictors binding energy may depend on multiple factors and the receptor protein, but this could be learned from the data produced by our methods and data from online databases such as ZINC, UniProt, PDB, and many others.

In conclusion, we found that the methods presented provide a principled and rational way to reduce costs and manage complexity. Given an extremely large number of possible protein ligand combinations, our results were used to decide which proteins to target, which ligands to purchase, and thereby prioritize *in vitro* experiments designed to test compounds for antimicrobial properties.

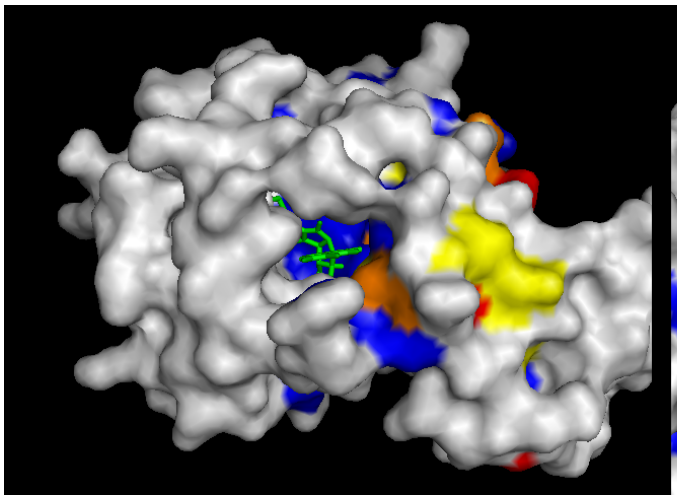
Table 8. Index to Web Sites Referenced

AD	autodock.scripps.edu
ADT	mgltools.scripps.edu/packages/adt
DEG	www.essentialgene.org
EBI	www.ebi.ac.uk
Integr8	www.ebi.ac.uk/integr8
Java	http://java.sun.com/reference/api
PDB	www.pdb.org
Perl	http://www.perl.org
NPD	wiki.compbio.ucsf.edu/wiki
UniProt	www.uniprot.org
Vina	vina.scripps.edu
ZINC	zinc.docking.org

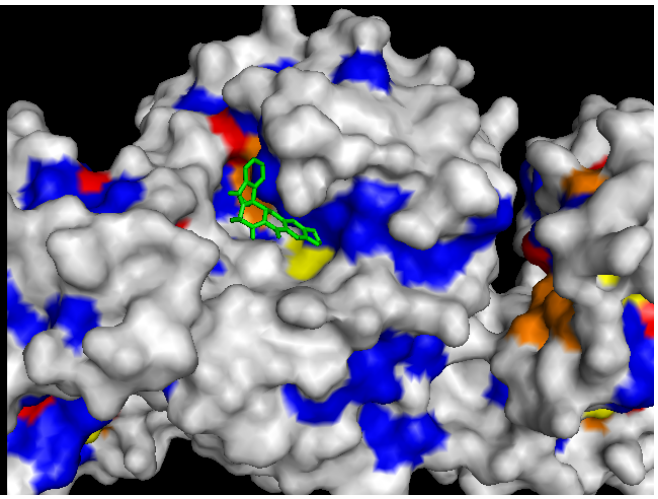


(a) ZINC12865513 docked with FtsZ, score -8.40

(b) ZINC02127995 docked with CoaD, score -6.60



(c) ZINC02161101 docked with YwtF, score -11.4



(d) ZINC12886695 docked with RacE, score -8.30

Fig. 12: Top scoring NPD p0.4 ligands docked in search areas that include **predicted inactive sites** (using 10777 ligands instead of ≈ 90000).

ACKNOWLEDGMENT

First and foremost, I would like to thank Dr. Nick Allenby, Demuris Ltd., for supervising this research; and Dr. Richard Daniel, The Centre for Bacterial Cell Biology, Newcastle University, for his expert opinion. I would also like to thank the following people for their excellent technical support. Dr. Daniel Swan and Dr. Simon Cockell, Bioinformatics Support Unit, Newcastle University; Professor Anil Wipat and Dr. Keith Flanagan, School of Computing Science, Newcastle University.

Funding: Biotechnology and Biological Sciences Research Council (BBSRC).

REFERENCES

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Archer, G. L. (1998). Staphylococcus aureus: a well-armed pathogen. *Clin Infect Dis*, **26**(5), 1179–1181.

Arigoni, F., Talabot, F., Peitsch, M., Edgerton, M. D., Meldrum, E., Allet, E., Fish, R., Jamotte, T., Curchod, M. L., and Loferer, H. (1998). A genome-based approach for the identification of essential bacterial genes. *Nat Biotechnol*, **16**(9), 851–856.

Badger, J., Sauder, J. M., Adams, J. M., Antonysamy, S., Bain, K., Bergseid, M. G., Buchanan, S. G., Buchanan, M. D., Batiyenko, Y., Christopher, J. A., Emtage, S., Eroshkina, A., Feil, I., Furlong, E. B., Gajiwala, K. S., Gao, X., He, D., Hendle, J., Huber, A., Hoda, K., Kearins, P., Kissinger, C., Laubert, B., Lewis, H. A., Lin, J., Loomis, K., Lorimer, D., Louie, G., Maletic, M., Marsh, C. D., Miller, I., Molinari, J., Muller-Dieckmann, H. J., Newman, J. M., Noland, B. W., Pagarigan, B., Park, F., Peat, T. S., Post, K. W., Radojicic, S., Ramos, A., Romero, R., Rutter, M. E., Sanderson, W. E., Schwinn, K. D., Tresser, J., Winhoven, J., Wright, T. A., Wu, L., Xu, J., and Harris, T. J. R. (2005). Structural analysis of a set of proteins resulting from a bacterial genomics project. *Proteins*, **60**(4), 787–796.

Bhöm, H.-J. and Schneider, G., editors (2000). *Virtual Screening for Bioactive Molecules*. Wiley. Weiheim, Germany.

Black, J. G. (2008). *Microbiology*. John Wiley & Sons (Asia) Pte Ltd., 7th edition.

Brown, J. R. and Warren, P. V. (1998). Antibiotic discovery: is it all in the genes? *Drug Discovery Today*, **3**, 564–566.

Chang, M., Ayeni, C., Breuer, S., and Torbett, B. (2010). Virtual screening for hiv protease inhibitors: A comparison of autodock 4 and vina. *PLoS ONE*, **5(8)**, e11955.

Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia*, **12**(1), 9–16.

Delano, W. (2002). The pymol molecular graphics system.

DeLeo, F. R. and Chambers, H. F. (2009). Reemergence of antibiotic-resistant staphylococcus aureus in the genomics era. *J Clin Invest*, **119**(9), 2464–2474.

Diep, B. A., Gill, S. R., Chang, R. F., Phan, T. H., Chen, J. H., Davidson, M. G., Lin, F., Lin, J., Carleton, H. A., Mongodin, E. F., Sensabaugh, G. F., and Perdreau-Remington, F. (2006). Complete genome sequence of usa300, an epidemic clone of community-acquired meticillin-resistant staphylococcus aureus. *Lancet*, **367**(9512), 731–739.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5), 1792–1797.

Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, **16**(3), 368–373.

Errington, J., Daniel, R. A., and Scheffers, D.-J. (2003). Cytokinesis in bacteria. *Microbiol Mol Biol Rev*, **67**(1), 52–65, table of contents.

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, **25**(4), 351–360.

Galperin, M. Y. and Koonin, E. V. (1999). Searching for drug targets in microbial genomes. *Curr Opin Biotechnol*, **10**(6), 571–578.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., Smith, H. O., and Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A*, **103**(2), 425–430.

Hawkey, P. M. (1998). The origins and molecular basis of antibiotic resistance. *BMJ*, **317**(7159), 657–660.

Haydon, D. J., Stokes, N. R., Ure, R., Galbraith, G., Bennett, J. M., Brown, D. R., Baker, P. J., Barynin, V. V., Rice, D. W., Sedelnikova, S. E., Heal, J. R., Sheridan, J. M., Aiwale, S. T., Chauhan, P. K., Srivastava, A., Taneja, A., Collins, I., Errington, J., and Czaplewski, L. G. (2008). An inhibitor of ftsz with potent and selective anti-staphylococcal activity. *Science*, **321**(5896), 1673–1675.

Higgins, D. G. and Sharp, P. M. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**(1), 237–244.

Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using clustal for multiple sequence alignments. *Methods Enzymol*, **266**, 383–402.

Hogan, J. C. (1996). Directed combinatorial chemistry. *Nature*, **384**(6604 Suppl), 17–19.

Howe, R. A., Monk, A., Wootton, M., Walsh, T. R., and Enright, M. C. (2004). Vancomycin susceptibility within methicillin-resistant staphylococcus aureus lineages. *Emerg Infect Dis*, **10**(5), 855–857.

Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*, **28**(6), 1145–1152.

Irwin, J. J. and Shoichet, B. K. (2005). Zinc–a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, **45**(1), 177–182.

Ji, Y., Zhang, B., Van, S. F., Horn, Warren, P., Woodnutt, G., Burnham, M. K., and Rosenberg, M. (2001). Identification of critical staphylococcal genes using conditional phenotypes generated by antisense rna. *Science*, **293**(5538), 2266–2269.

Jones, R. N., Low, D. E., and Pfaller, M. A. (1999). Epidemiologic trends in nosocomial and community-acquired infections due to antibiotic-resistant gram-positive bacteria: the role of streptogramins and other newer compounds. *Diagn Microbiol Infect Dis*, **33**(2), 101–112.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*, **12**(6), 962–968.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., and Apweiler, R. (2005). Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*, **33**(Database issue), D297–D302.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, **3**(11), 935–949.

Knowles, D. J. (1997). New strategies for antibacterial drug design. *Trends Microbiol*, **5**(10), 379–383.

Kobayashi, K., Ehrlich, S. D., and A. Albertini, *et al.*. (2003). Essential bacillus subtilis genes. *Proc Natl Acad Sci U S A*, **100**(8), 4678–4683.

Koonin, E. V. and Galperin, M. Y. (2003). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers.

Korf, I., Yandell, M., and Bedell, J. (2003). *BLAST*. O'Reilly Media.

Leaver, M., Domnguez-Cuevas, P., Coxhead, J. M., Daniel, R. A., and Errington, J. (2009). Life without a wall or division machine in bacillus subtilis. *Nature*, **457**(7231), 849–853.

Lesk, A. M. (2001). *Introduction to Protein Architecture*. Oxford University Press.

Lesk, A. M. (2008). *Introduction to Bioinformatics*. Oxford University Press, 3rd edition.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, **7**(9), 1884–1897.

Lock, R. L. and Harry, E. J. (2008). Cell-division inhibitors: new insights for future antibiotics. *Nat Rev Drug Discov*, **7**(4), 324–338.

McDevitt, D. and Rosenberg, M. (2001). Exploiting genomics to discover new antibiotics. *Trends Microbiol*, **9**(12), 611–617.

Mills, S. D. (2003). The role of genomics in antimicrobial discovery. *J Antimicrob Chemother*, **51**(4), 749–752.

Mills, S. D. (2006). When will the genomics investment pay off for antibacterial discovery? *Biochem Pharmacol*, **71**(7), 1096–1102.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. O., and J., A. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, **19**, 1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J Comput Chem*, **30**(16), 2785–2791.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Press.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3), 443–453.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304.

Payne, D. J., Gwynn, M. N., Holmes, D. J., and Rosenberg, M. (2004). Genomic approaches to antibacterial discovery. *Methods Mol Biol*, **266**, 231–259.

Payne, D. J., Gwynn, M. N., Holmes, D. J., and Pompliano, D. L. (2007). Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov*, **6**(1), 29–40.

Pfaller, M. A., Jones, R. N., Doern, G. V., and Kugler, K. (1998). Bacterial pathogens isolated from patients with bloodstream infection: frequencies of occurrence and antimicrobial susceptibility patterns from the sentry antimicrobial surveillance program (united states and canada, 1997). *Antimicrob Agents Chemother*, **42**(7), 1762–1770.

Prakhov, N. D., Chernorudskiy, A. L., and Gainullin, M. R. (2010). Vsdocker: a tool for parallel high-throughput virtual screening using autodock on windows-based computer clusters. *Bioinformatics*, **26**(10), 1374–1375.

Projan, S. J. (2003). Why is big pharma getting out of antibacterial drug discovery? *Curr Opin Microbiol*, **6**(5), 427–430.

Projan, S. J. (2008). Whither antibacterial drug discovery? *Drug Discov Today*, **13**(7-8), 279–280.

Pucci, M. J. (2006). Use of genomics to select antibacterial targets. *Biochem Pharmacol*, **71**(7), 1066–1072.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, **12**(2), 85–94.

Ruzheinikov, S. N., Taal, M. A., Sedelnikova, S. E., Baker, P. J., and Rice, D. W. (2005). Substrate-induced conformational changes in bacillus subtilis glutamate racemase and their implications for drug discovery. *Structure*, **13**(11), 1707–1713.

Sanner, M. F. (1999). Python: a programming language for software integration and development. *J Mol Graph Model*, **17**(1), 57–61.

Schneider, T. and Sahl, H.-G. (2010). An oldie but a goodie - cell wall biosynthesis as antibiotic target pathway. *Int J Med Microbiol*, **300**(2-3), 161–169.

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*, **432**(7019), 862–865.

Silver, L. L. (2003). Novel inhibitors of bacterial cell wall synthesis. *Curr Opin Microbiol*, **6**(5), 431–438.

Silver, L. L. (2006). Does the cell wall of bacteria remain a viable source of targets for novel antibiotics? *Biochem Pharmacol*, **71**(7), 996–1005.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1), 195–197.

Song, J.-H. and Ko, K. S. (2008). Detection of essential genes in streptococcus pneumoniae using bioinformatics and allelic replacement mutagenesis. *Methods Mol Biol*, **416**, 401–408.

Song, J.-H., Ko, K. S., Lee, J.-Y., Baek, J. Y., Oh, W. S., Yoon, H. S., Jeong, J.-Y., and Chun, J. (2005). Identification of essential genes in streptococcus pneumoniae by allelic replacement mutagenesis. *Mol Cells*, **19**(3), 365–374.

Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., and Pucci, M. J. (2002). Identification of 113 conserved essential genes using a high-throughput gene disruption system in streptococcus pneumoniae. *Nucleic Acids Res*, **30**(14), 3152–3162.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22), 4673–4680.

Trott, O. and Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, **31**(2), 455–461.

Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *J Med Chem*, **49**(20), 5912–5931.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.

Weigel, L. M., Clewell, D. B., Gill, S. R., Clark, N. C., McDougal, L. K., Flannagan, S. E., Kolonay, J. F., Shetty, J., Killgore, G. E., and Tenover, F. C. (2003). Genetic analysis of a high-level vancomycin-resistant isolate of staphylococcus aureus. *Science*, **302**(5650), 1569–1571.

Zhang, R. and Lin, Y. (2009). Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res*, **37**(Database issue), D455–D458.

Zhang, R., Ou, H.-Y., and Zhang, C.-T. (2004). Deg: a database of essential genes. *Nucleic Acids Res*, **32**(Database issue), D271–D272.

Zvelebil, M. J. and Baum, J. O. (2008). *Understanding Bioinformatics*. Garland Science.