Improved Annotations of 23 Differentially Expressed Hypothetical Proteins in Methicillin Resistant S. aureus

Jessica Marklevitz¹ and Laura K. Harris¹,²∗

¹Department of Science, Davenport University, Lansing, Michigan, United States of America; ²Department of Health Informatics, Rutgers School of Health Professions, Newark, New Jersey, United States of America; Laura K. Harris - E-mail: laura.harris@davenport.edu; ∗Corresponding author

Received March 21, 2017; Revised April 12, 2017; Accepted April 12, 2017; Published April 30, 2017

Abstract:
Antibiotic resistant Staphylococcus aureus is a major public health concern effecting millions of people annually. Medical science has documented completely untreatable S. aureus infections. These strains are appearing in the community with increasing frequency. New diagnostic and therapeutic options are needed to combat this deadly infection. Interestingly, around 50% of the proteins in S. aureus are annotated as hypothetical. Methods to select hypothetical proteins related to antibiotic resistance have been inadequate. This study uses differential gene expression to identify hypothetical proteins related to antibiotic resistant phenotype strain variations. We apply computational tools to predict physiochemical properties, cellular location, sequence-based homologs, domains, 3D modeling, active site features, and binding partners. Nine of 23 hypothetical proteins were <100 residues, unlikely to be functional proteins based on size. Of the 14 differentially expressed hypothetical proteins examined, confident predictions on function could not be made. Most identified domains had unknown functions. Six hypothetical protein models had >50% confidence over >20% residues. These findings indicate the method of hypothetical protein identification is sufficient; however, current scientific knowledge is inadequate to properly annotate these proteins. This process should be repeated regularly until entire genomes are clearly and accurately annotated.

Keywords: Annotations, Hypothetical proteins, Methicillin, S. aureus

Background:
Antibiotic therapy has been the marvel of modern medicine since the advent of Penicillin in the 1920s. Over seventy billion doses of antibiotics are consumed globally each year [1]. Antibiotics are a low-cost resource to treat food-borne and other sanitation-related infections that commonly affect poor people. Among wealthier countries, antibiotics play a pivotal role as a prophylactic, controlling infections associated with medical practices such as surgery [1, 2]. Unfortunately, this usage exposes normal microbial flora to anti-bacterial drugs, allowing them to develop resistances so the drugs lose effectiveness. Medical science has been unable to cultivate new antibiotics as fast as resistances to current therapies are rising [2, 3]. Infectious organisms that are resistant to every antibiotic developed have been reported. This antibiotic resistance crisis is a critical challenge for humanity’s medical future.
including children [5]. While there are no S. aureus strains currently resistant to all antibiotics, completely resistant strains of other infectious organisms have emerged so the same outcome will likely befall S. aureus soon.

A challenge to developing new antibiotic therapies is genome annotation. Around 50% of proteins identified in the S. aureus genome are annotated as hypothetical [6, 7]. At annotation, hypothetical proteins are predicted by sequence only and lack homology to known proteins. Researchers further define hypothetical proteins by their larger than 100 amino acids size, since smaller sequences likely represent other macromolecular structures such as short interfering RNA (siRNA) rather than functional proteins [8]. True hypothetical proteins have similar features to other hypothetical proteins due to lack of experimental evidence to predict function for the protein family, though frequently hypothetical proteins found in databases represent old genome annotations in need of update. Several studies have used various methods to identify hypothetical proteins related to antibiotic resistance in S. aureus. Early studies randomly selected hypothetical proteins for characterization [6, 7, 9, 10]. While this approach developed and demonstrated computational procedures that contribute to hypothetical protein characterization, it is limited in its ability to identify hypothetical proteins specifically connected to antibiotic resistance. To improve the selection process, we formerly developed cross-species approach that used proteins with experimentally established structures from the major facilitator superfamily; a large, highly conserved protein family associated with antibiotic resistance [7]. This approach worked because of the large percentage of hypothetical proteins in the S. aureus genome, but it becomes inadequate if a hypothetical protein related to resistance has no well-characterized homolog in another species, a common challenge for hypothetical proteins. Better methods for identifying antibiotic resistant-related hypothetical proteins are needed.

Microarray and other forms of publicly accessible gene expression data can provide an excellent repository for targeted identification of resistance linked hypothetical proteins in S. aureus. For example, Ham and colleagues examined mRNA expression between antibiotic resistant (MRSA; ATCC 33591, shown to be susceptible only to Vancomycin and Kanamycin) and sensitive (MSSA; ATCC 25923) strains using Affymetrix GeneChip® technology [11]. They statistically compared mRNA expression levels between the strains to uncover potential mechanisms of resistance, but did not consider hypothetical proteins. This presents an opportunity to characterize hypothetical proteins whose differential expression constitutes a drug-resistant genomic background.

This study uses computational procedures to characterize statistically significant differentially expressed hypothetical proteins from the microarray data generated by Ham and associates. By comparing natural gene expression between antibiotic sensitive and resistant strains, new insight into strain background differences is gained. These variations could uncover new resistance mechanisms, further developing into a useful diagnostic tool or potential antibiotic therapeutic target. This would improve outcomes for patients infected with MRSA strains through faster and more effective treatment options.

Methodology:

Normalized mRNA expression data from Ham’s study is available at the National Center for Biotechnology Information’s (NCBI) Gene Expression Omnibus (GEO; Dataset Record GDS4242; GEO accession GSE18289) [11]. Data consisted of 7774 entries, each with probe name and six samples representing triplicates of both MSSA (ATCC 25923) and MRSA (ATCC 33591) strains. Probe names were converted to gene names and descriptions per Affymetrix chip platform and non-hypothetical proteins were removed. Excel calculated T-scores and p-values based on Student’s T-test two-tailed, equal variance formulas. The study rejected hypothetical proteins with a p>0.05 as these were not differentially expressed. The National Center for Biotechnology Information (NCBI) and UniProt databases confirmed hypothetical protein annotation.

This study used numerous algorithms to characterize these hypothetical proteins and default program settings were used for all analyses. ExFASy’s ProtParam server calculated physiochemical properties including number of amino acids, molecular weight, positively and negatively charged residues, theorectically isoelectric point (pl), extinction coefficient, aliphatic index (Al), instability index (II), and the grand average hydropathy (GRAVY) [12]. By hypothetical protein definition, those identified through differential expression yet smaller than 100 amino acids were excluded from further study.

PSortB and SOSUI servers predicted each hypothetical protein’s cellular location. PSortB predicted between cytoplasm, cytoplasmic membrane, cell wall, or extracellular locations [13]. SOSUI calculated transmembrane regions and solubility indices, a valuable confirmation of PSortB predictions [14]. These complementary algorithms provide confidence for cellular localization estimates.

Sequence similarity and domain identification projected functional features of hypothetical proteins. The Position-Specific Iterative (PSI) Basic Local Alignment Search Tool (BLAST) identified potential homologs from the NCBI database based on protein sequence similarities. Further, both Conserved Domain Database (CDD) BLAST and Pfam algorithms predicted potential domains within each hypothetical protein. CDD-BLAST uses a PSI-BLAST variation to identify domains by comparison of the protein sequence’s position specific scoring matrix to those in the NCBI database [15]. Alternatively, Pfam is a separately curated database of Hidden Markov Models and multiple sequence alignments representing protein domain families [16]. These complementary approaches provide a level of validation to this study’s findings.

For model development and characterization, we used the integrated Phyre2 and 3DLigandSite servers. Phyre2 produced a
tertiary structure model, predicted ligand-binding sites, and analyzed the effect of amino acid variants through automatic homology detection methods [17]. Phyre2's model advanced to 3DLigandSite for active site characterization and docking predictions. 3DLigandSite identifies homologous structures with bound ligands by searching a structural library then superimposing those ligands onto the Phyre2's protein structure [18]. Together, Phyre2 and 3DLigandSite servers modeled the protein and characterized its binding sites.

The Search Tool for Interactions of Chemicals (STITCH) database predicted potential ligand interactions for each hypothetical protein. STITCH draws upon scientific literature and several databases, including the formerly separate Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database, which houses high-throughput experiment and conserved co-expression data, to calculate drug-target interactions, binding affinities, and biological pathways [19]. STITCH is a useful tool to predict protein and chemical binding partners.

Results:
The mRNA expression dataset, GSE18289, was downloaded from GEO and Excel calculated the T-score value for each protein. Twenty-seven proteins labeled as hypothetical in NCBI, 16 and 11 up- and down regulated in MRSA, respectively, had <0.05 p-values. Four of these proteins had predicted functions in UniProt, an endotoxin (SACOL0468, up regulated, T-score 9.00), exotoxin (SACOL1178, up regulated, T-score 10.17), phosphate dikinase regulatory protein (SACOL1620, down regulated, T-score -9.80), and a lipoprotein (SACOL1531, up regulated, T-score 7.89). Since these proteins had predicted identities, they were excluded from further study. The remaining 23 proteins are listed by T-score in Table 1.

![Figure 1: Venn diagram illustrating overlap of study evaluations.](image)

Table 1: Differential expression T-scores and physicochemical properties of 23 hypothetical proteins

| Protein     | T-score | # AA | MW     | pI   | # neg | # pos | EC   | II  | AI  | GRAVY |
|-------------|---------|------|--------|------|-------|-------|------|-----|-----|-------|
| SACOL091918.77 | 45      | 520  | 9.03   | 4    | 6     | 2980  | 17.29| 136.22| 0.560 |
| SACOL185912.23 | 106     | 1206 | 5.70   | 147  | 130   | 16360 | 37.50| 95.75 | -0.415 |
| SACOL134610.83 | 64      | 7573 | 4.10   | 16   | 6     | 5960  | 21.26| 92.81 | -0.420 |
| SACOL10356.29 | 78      | 8726 | 4.32   | 17   | 7     | 5960  | 41.83| 86.28 | -0.529 |
| SACOL30267.48 | 74      | 8841 | 4.54   | 18   | 7     | 11460 | 52.27| 77.70 | -0.938 |
| SACOL12373.52 | 102     | 11944| 7.91   | 17   | 18    | 9970  | 33.41| 89.80 | -0.762 |
| SACOL10196.83 | 135     | 15123| 4.45   | 14   | 8     | 26930 | 39.15| 132.89| 0.757 |
| SACOL0876.62  | 35      | 4172 | 6.00   | 5    | 5     | 4470  | 25.62| 94.57 | -0.149 |
| SACOL0736.04  | 200     | 2262 | 9.55   | 9    | 20    | 31860 | 42.51| 121.35| 0.665 |
| SACOL11044.35 | 209     | 24690| 9.55   | 14   | 25    | 43890 | 29.04| 125.48| 0.448 |
| SACOL3503.80  | 118     | 13923| 10.08  | 14   | 28    | 12950 | 26.44| 76.02 | -0.804 |
| SACOL03263.73 | 66      | 7068 | 8.03   | 8    | 9     | 9970  | 37.99| 125.45| 0.185 |
| SACOL24813.23 | 121     | 14067| 4.54   | 21   | 11    | 5960  | 35.52| 118.43| -0.098 |
| SACOL0835.26  | 209     | 20707| 9.07   | 31   | 36    | 8940  | 64.82| 34.16 | -1.974 |
| SACOL2241.6.45| 129     | 14638| 9.73   | 3    | 8     | 18450 | 26.53| 155.74| 1.209 |
| SACOL2123.6.59 | 223    | 2956 | 4.74   | 43   | 28    | 28550 | 41.80| 93.00  | -0.289 |
| SACOL2491.8.97| 63      | 7221 | 4.60   | 10   | 6     | 7450  | 25.52| 97.46 | -0.146 |
| SACOL2571.9.80 | 63    | 7266 | 5.44   | 9    | 7     | 1490  | 17.06| 103.65| -0.233 |
| SACOL2076.10.78| 45     | 5070 | 10.46  | 3    | 10    | 0[^1] | 45.35| 114.67| -0.424 |
| SACOL1956.14.64| 176    | 20513| 9.25   | 10   | 15    | 21555 | 39.83| 132.95| 0.747 |
| SACOL267.15.31| 507    | 57978| 8.02   | 93   | 95    | 36790 | 23.76| 74.48 | -0.906 |
| SACOL088.24.11 | 107    | 13458| 5.23   | 26   | 21    | 15930 | 65.54| 59.25  | -1.693 |
| SACOL1701.25.02| 165   | 19009| 5.08   | 27   | 17    | 10430 | 33.94| 100.48| -0.181 |

[^1]: T-score excluded from further study.

# AA, number of amino acids; MW, molecular weight; pI, theoretical isoelectric point; # neg, total number of negatively charged residues (Asp + Glu); # pos, total number of positively charged residues (Arg + Lys); EC, extinction coefficient assuming all pairs of
Cys residues form cystines; II, instability index; AI, aliphatic index; GRAVY, grand average hydropathy. As there are no Trp, Tyr, or Cys in the region considered, protein should not be visible by UV spectrophotometry.

Few hypothetical proteins had well defined homologs in the NCBI database as identified by PSI-BLAST (Table 4). Most top homologs came from S. aureus and were vaguely annotated or had low sequence similarity to the hypothetical protein. Three proteins, SACOL0323, SACOL2481, and SACOL0710, had homologs from other species, *Muclagenibacter*, *Helicobacter mustelae*, and *Bacillus cereus*, respectively. Interestingly, four hypothetical proteins had membrane protein for their top homolog. PSortB and SOSUI confirm that SACOL0109, SACOL0075, and SACOL2241 are likely membrane proteins too (Tables 2 and 3, respectively). However, according to these algorithms, SACOL2481 is a soluble, cytoplasmic protein. Further, PSortB predicted SACOL0488 to reside in the cytoplasm, which PSI-BLAST’s top homolog confirmed, though PSortB was unable to confirm extracellular locations for SACOL0835 and SACOL0267 where top homologs are exported proteins. Interestingly, for SACOL2123, PSI-BLAST identified its top homolog as a PF11042 family member. This matched CDD-BLAST domain identification of pfam11042, showing the interconnectivity of these computational tools, and was confirmed by Pfam itself.

Phyre2 and 3DlLigand servers performed hypothetical protein modeling and active site characterization. Similarity measurements of the hypothetical protein target to its experimental structure template are in Table 7. These findings represent Phyre2 running in normal mode. Hypothetical proteins with coverage >25% in normal mode were re-run under Phyre2’s intensive mode with the results show in Figure 2. Remarkably, under this mode, SACOL1859 and SACOL0710 models had 88% and 89% residues modeled with >90% confidence. No amino acids from the other four proteins could be modelled with that confidence. Unfortunately, 3DLigand was unable to make a prediction for any hypothetical protein examined in this study due to insufficient homologous structures with ligands bound.

STITCH predicted binding partners for hypothetical proteins. STITCH was unable to predict binding partners for the following hypothetical proteins: SACOL2481, SACOL0835, and SACOL2241. Most top binding partners were fellow hypothetical proteins with confidence scores listed in Table 8. This implies that more database annotation and/or wet bench work are needed to fully understand how these proteins work. SACOL0323, SACOL2123, and SACOL0710 had top matching binding partners that were not hypothetical proteins.

SACOL0323 matched a prophage L54a, Cro-like protein. SACOL2123 had equal scores to a M20/M25/M40 family peptidase (SACOL2125) and a hypothetical protein (SACOL2124). SACOL0710 equally matched a phosphotransferase mannos-specific family component IIa (SACOL0709) and a DAK2 domain-containing protein (SACOL0708). These results did not correlate with the findings from other programs used in this study.

Figure 2: Phyre2’s intensive mode models for hypothetical proteins SACOL1859 (A), SACOL0323 (B), SACOL0109 (C), SACOL1956 (D), SACOL0488 (E), and SACOL0710 (F). Image colored by rainbow N- to C-terminus.
Table 2: PSortB cellular location of 14 hypothetical proteins

| Protein   | Location         | Localization Score |
|-----------|------------------|--------------------|
| SACOL1859 | Unknown          | 2.50               |
| SACOL0323 | Cytoplasm        | 7.50               |
| SACOL0109 | Cytoplasmic membrane | 10.00             |
| SACOL0075 | Cytoplasmic membrane | 10.00             |
| SACOL0644 | Cytoplasmic membrane | 10.00             |
| SACOL0350 | Unknown          | 2.50               |
| SACOL2481 | Cytoplasm        | 7.50               |
| SACOL0835 | Cytoplasmic membrane | 9.55             |
| SACOL2241 | Cytoplasmic membrane | 10.00             |
| SACOL2123 | Cytoplasm        | 7.50               |
| SACOL1956 | Cytoplasmic membrane | 10.00             |
| SACOL0267 | Unknown          | 3.33               |
| SACOL0488 | Cytoplasm        | 7.50               |
| SACOL0710 | Cytoplasm        | 7.50               |

1Equal probability of the protein being located in any cellular structure: cytoplasm, cytoplasmic membrane, cell wall, or extracellular. 2Equal probability of protein being located in cytoplasmic membrane, cell wall, or extracellular.

Table 3: SOSUI results for 7 transmembrane hypothetical proteins

| Protein  | N-terminal Transmembrane Region | C-terminal | Type | Length |
|----------|--------------------------------|------------|------|--------|
| SACOL0109| 53 IGK1AIWIGIVAQYPSVVFVVRM      | 75 PRIMARY | 23   |
| SACOL0075| 7 KIAIWIGIVAQYPSVVFVRMIS       | 29 PRIMARY | 23   |
| SACOL0644| 23 YLLDILDSTWLVYFFFPFIFIP      | 45 SECONDARY | 23   |
| SACOL1956| 16 FIILQLVIALFVIFYKWLAVG       | 38 PRIMARY | 23   |
| SACOL0267| 6 KIIIPLVLLIGGIAGVVFYYAF       | 28 PRIMARY | 23   |

Table 4: Top PSI-BLAST result for 14 hypothetical proteins

| Protein            | PSI-BLAST Match | Query Cover | E-value | Identity |
|--------------------|-----------------|-------------|---------|----------|
| SACOL1859          | NTPase          | 100%        | 0.0     | 100%     |
| SACOL0323          | Metallophosphoesterase | 59% | 1.6 | 31% |
| SACOL0109          | Membrane protein | 100%        | 3e-44   | 59%      |
| SACOL0075          | Membrane spanning protein | 90% | 7e-124 | 98%      |
| SACOL0644          | tandem five-TM protein | 100% | 1e-143 | 99%      |
| SACOL0350          | Phage protein   | 100%        | 5e-80   | 98%      |
| SACOL2481          | Outer membrane protein | 59% | 4.3 | 27% |
| SACOL0835          | Exported protein | 91%         | 8e-128  | 100%     |
| SACOL2241          | Membrane protein | 79%         | 3e-64   | 100%     |
| SACOL2123          | PF11042 family protein | 100% | 1e-91 | 65% |
| SACOL1956          | Permease        | 100%        | 3e-123  | 100%     |
| SACOL0267          | Exported protein | 51%         | 8e-170  | 98%      |
| SACOL0488          | Cytosolic protein | 89%         | 2e-59   | 100%     |
| SACOL0710          | RHS repeat-associated core domain-containing protein | 87% | 1e-09 | 29% |

ISSN 0973-2063 (online) 0973-8894 (print)
Table 5: CDD-BLAST domain data for 7 hypothetical proteins

| Protein  | Domains   | Description                  | E-value  |
|----------|-----------|------------------------------|----------|
| SACOL1859| pfam13401 | AAA                          | 1.61e-04 |
|          | smart00382| ATPase                       | 7.59e-03 |
| SACOL0644| pfam04276 | Protein of unknown function  | 1.05e-37 |
| SACOL0350| pfam07768 | PVL ORF-30-like family       | 8.79e-47 |
| SACOL0835| pfam16228 | Domain of unknown function   | 1.18e-12 |
| SACOL2123| pfam11042 | Protein of unknown function  | 2.38e-20 |
| SACOL1956| pfam11667 | Putative zincin peptidase    | 4.49e-08 |
| SACOL0488| pfam13654 | AAA                          | 5.17e-03 |

Table 6: Pfam domain data for 5 hypothetical proteins

| Protein  | Domain   | Description                  | E-value  |
|----------|----------|------------------------------|----------|
| SACOL0644| DUF443   | Unknown function             | 9.8e-56  |
|          | PVL_ORF30| Panton-Valentine leucocidin ORF-30-like family | 2.8e-45  |
| SACOL0835| DUF4887  | Unknown function             | 1.7e-50  |
| SACOL2123| DUF2750  | Unknown function             | 2.1e-21  |
| SACOL1956| DUF3267  | Putative zincin peptidase    | 1.6e-18  |

Table 7: Phyre2 model data for 14 hypothetical proteins

| Protein  | Template   | Description                                             | Confidence | Coverage |
|----------|------------|---------------------------------------------------------|------------|----------|
| SACOL1859| c4kxfF     | nlr family card domain-containing protein 4             | 99.7%      | 30%      |
| SACOL023 | d1nu9c1    | immunoglobulin/albumin-binding domain-like               | 37.8%      | 25%      |
| SACOL109 | c2x29A     | crystal structure of mouse claudin-19                   | 73.2%      | 45%      |
| SACOL0075| c4xxsD     | virion egress protein ul31                               | 55.6%      | 20%      |
| SACOL0644| c4yjbX     | ATP-dependent clp protease adapter protein               | 30.7%      | 7%       |
| SACOL0350| c2dqA      | talin-1                                                 | 40.7%      | 21%      |
| SACOL2481| c3da0B     | putative phosphatase                                     | 23.3%      | 17%      |
| SACOL0835| c2fimA     | pF1 filamentous bacteriophage                            | 80.3%      | 14%      |
| SACOL2241| c2ap8A     | bombinin h4                                             | 43.2%      | 10%      |
| SACOL2123| c1zcbt     | glycogenin-1                                            | 46.3%      | 12%      |
| SACOL1956| c3b4rB     | putative zinc metalloprotease mj0392                     | 89.3%      | 41%      |
| SACOL2067| c3cuj      | photosystem ii reaction center protein j                 | 50.4%      | 5%       |
| SACOL0488| c4c46B     | general control protein gcn4                             | 80.6%      | 29%      |
| SACOL0701| c1kt0A     | lare fkbp-like protein, fkbp51, involved in steroid2 receptor complexes | 93.6%      | 47%      |

Table 8: Top STITCH predicted binding partners for 11 hypothetical proteins

| Protein  | Substrate | Score |
|----------|-----------|-------|
| SACOL1859| SACOL1860 | 0.651 |
| SACOL0323| SACOL0322 | 0.819 |
| SACOL0109| SACOL0110 | 0.692 |
| SACOL0075| SACOL0076 | 0.462 |
| SACOL0644| SACOL0643 | 0.859 |
| SACOL0350| SACOL0351 | 0.859 |
| SACOL2123| SACOL2125 | 0.422 |
| SACOL2124| SACOL2124 | 0.422 |
| SACOL1956| SACOL2519 | 0.685 |
| SACOL2067| SACOL2066 | 0.694 |
| SACOL0488| SACOL0487 | 0.859 |
| SACOL0710| SACOL0709 | 0.570 |
| SACOL0708| SACOL0708 | 0.570 |
Conclusion:
Antibiotic resistance is a major global health crisis. Infections, like those caused by Methicillin-resistant *S. aureus*, are becoming untreatable, and increasing fatalities from these once curable diseases. Faster techniques to identify drug-resistant organisms and new therapeutics are needed to improve patient outcomes. Characterizing hypothetical proteins, particularly those contributing to resistance, may hold the key to unlock this health predicament. This work provides insight into hypothetical proteins related to antibiotic resistance, potentially leading to improved diagnostic tools and therapeutics against antibiotic resistant *S. aureus*. It characterized differentially expressed hypothetical proteins between Methicillin-sensitive and resistant strains whereas other studies have randomly selected or performed cross-species comparisons to identify hypothetical proteins of interest. Our approach to identify hypothetical proteins related to antibiotic resistance is an improvement over prior methods. However, computational algorithms were unable to confidently predict functions for any of the 14 differentially expressed hypothetical proteins examined. Most programs struggled to identify parameters, such as domains or binding partners. Those that were found usually had unknown functions or little sequence homolog. These results indicate that using statistically significant differential expression from a publically available antibiotic resistant strain comparison microarray study will identify proteins potentially related to antibiotic resistance for which more scientific knowledge is needed.

References:
[1] Woolhouse M, et al. *J of Glob Health* 2106 6(1) [PMID: 27350872]
[2] Ventola CL. *Pharmacy and Therapeutics*. 2015 40(4): 277 [PMID: 25859123]
[3] Nathan C & Cars O. *N Engl J Med* 2014 371(19): 1761 [PMID: 25271470]
[4] Rossolini GM, et al. *Curr Opin Pharmacol* 2014 18: 56 [PMID: 25254623]
[5] https://www.cdc.gov/mrsa/tracking/
[6] School K, et al. *Bioinformation* 2016 12(3): 209 [PMID: 28149057]
[7] Marklevitz J & Harris I.K. *Bioinformation* 2016 12(4):254 [PMID: 28197063]
[8] Yanagihara K, et al. *J Antimicrob Chemother* 2006 57(1): 122 [PMID: 16344286]
[9] Bharat Siva Varma P, et al. *J Infect Public Health* 2015 8(6): 526 [PMID: 26025048]
[10] Mohan R & Venugopal S. *Bioinformation* 2012 8(15): 722 [PMID: 23055618]
[11] Ham JS, et al. *Mol Cells* 2010 30(1): 71 [PMID: 20652498]
[12] Gasteiger E, et al. *Methods Mol Biol* 1999 112: 531 [PMID: 10027275]
[13] Yu NY, et al. *Bioinformatics* 2010 26(13): 1608 [PMID: 20472543]
[14] Hirokawa T, et al *Bioinformatics* 1998 14: 378 [PMID: 9632836]
[15] Marchler-Bauer A, et al. *NAR* 2015 43: D222 [PMID: 25414356]
[16] Sonnhhammer EL, et al. *Proteins* 1997 28: 405 [PMID: 9223186]
[17] Kelley LA, et al. *Nat Protoc* 2015 10: 845 [PMID: 25950237]
[18] Wass MN, et al. *NAR* 2010 38 Suppl: W469 [PMID: 20513649]
[19] Szklarczy D, et al. *NAR* 2016 44(D1): D380 [PMID: 26590256]
[20] http://bioinformaticss.psb.ugent.be/webtools/Venn/

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License

Edited by P Kangueane