Supplementary Figures

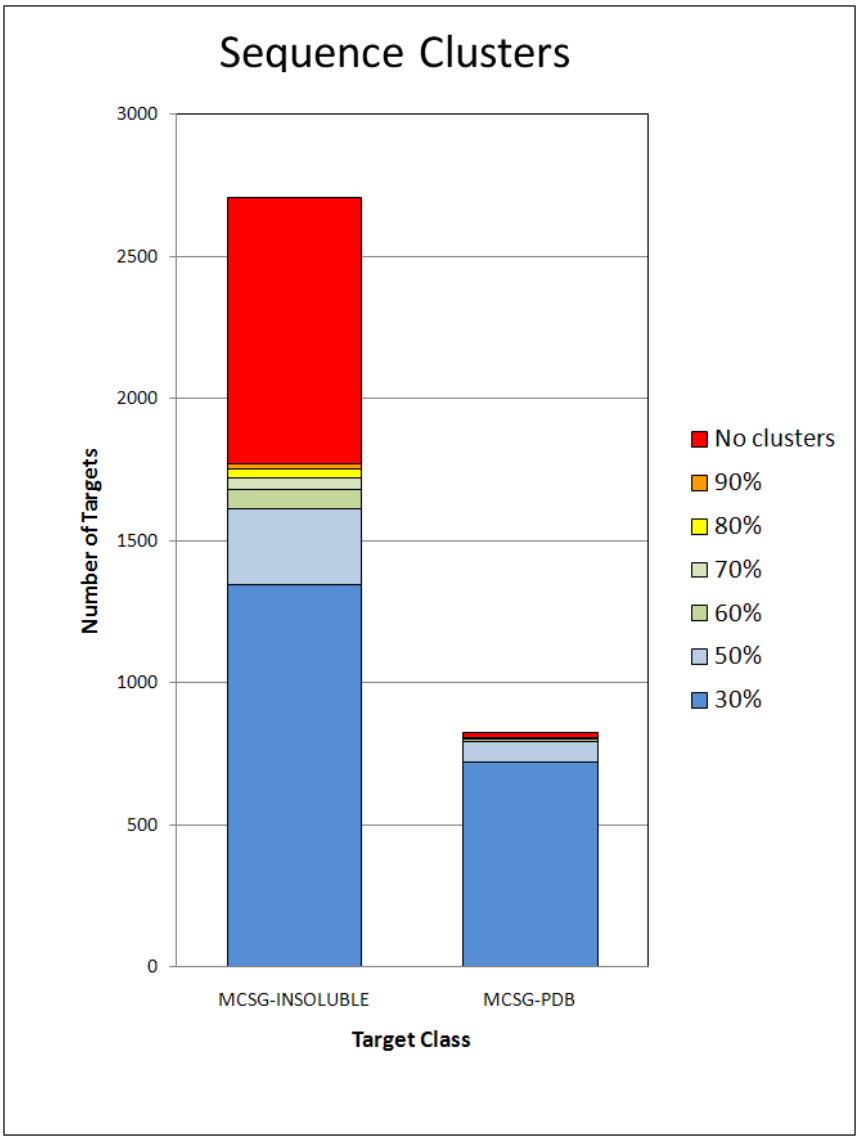

Figure S1. Sequence clusters used in analysis. The protein targets in the MCSG pipeline were divided into two classes: a) the MCSG-INSOLUBLE class represents targets that express at high level, but are insoluble and b) the MCSG-PDB class represents those targets deposited into PDB. In order to increase the success rate in the MCSG pipeline, orthologous targets are used, generating sequence classes with high similarity, while towards the end of the pipeline similar sequences are eliminated in order to increase sequence uniqueness. The classes were clustered using cd-hit [16] and the 60% similarity classes were used in the analysis.

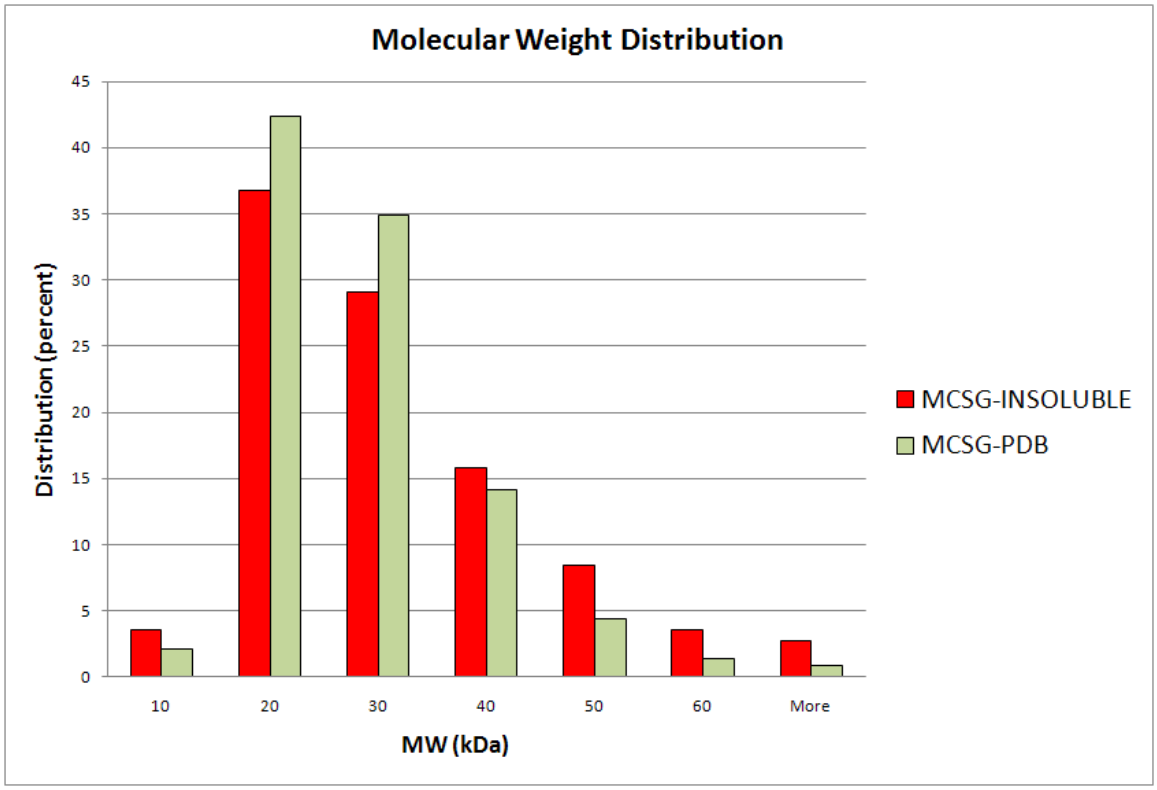

Figure S2. The Distribution of Molecular Weight of Targets Used in Study. The molecular weight of protein targets in the MCSG-INSOLUBLE and MCSG-PDB classes were binned using a 10kDa bin-size and normalized.

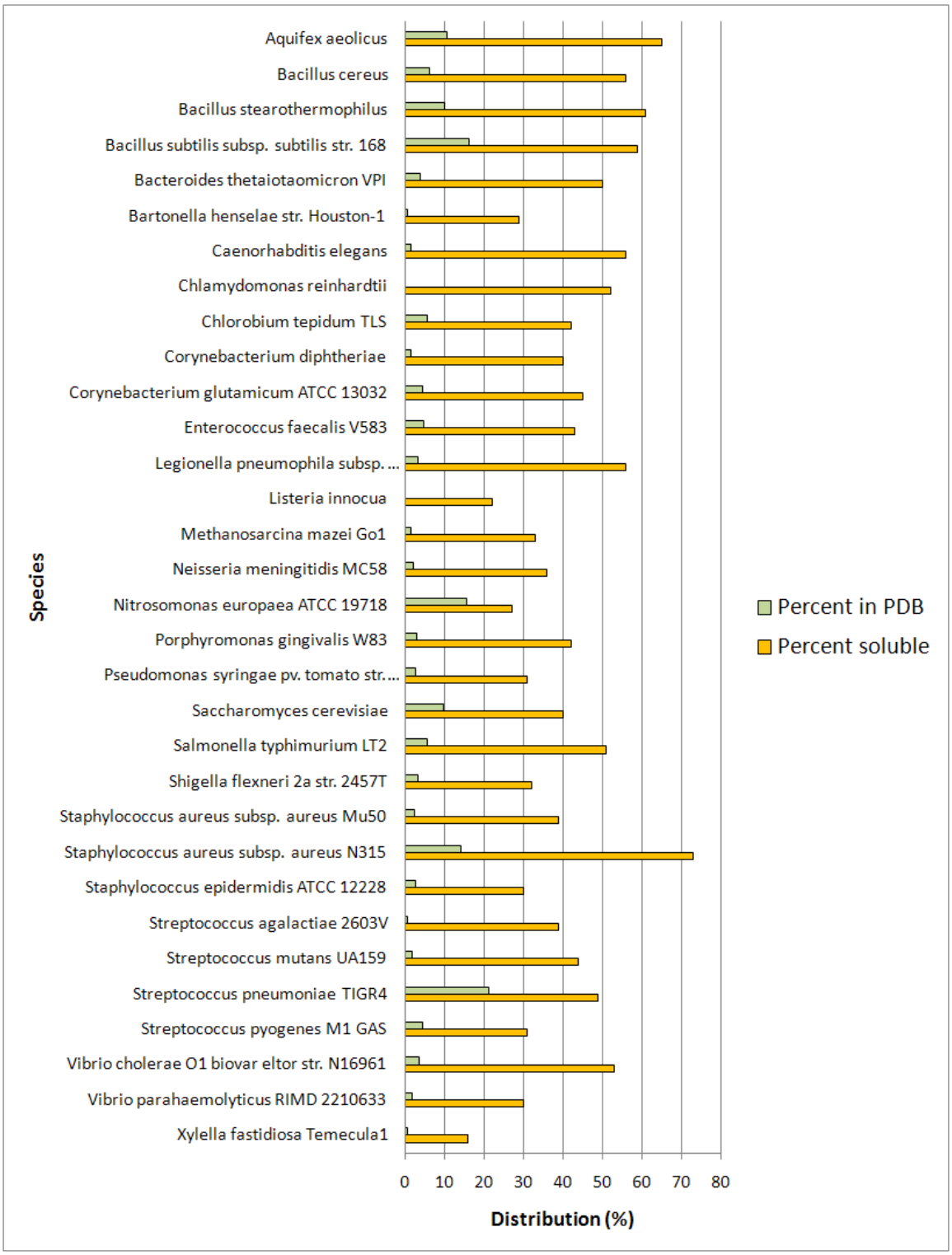
Species
Aquifex aeolicus
Bacillus cereus
Bacillus stearothermophilus
Bacillus subtilis subsp. subtilis str. 168
Bacteroides thetaiotaomicron VPI
Bartonella henselae str. Houston-1
Caenorhabditis elegans
Chlamydomonas reinhardtii
Chlorobium tepidum TLS
Corynebacterium diphtheriae
Corynebacterium glutamicum ATCC 13032
Enterococcus faecalis V583
Legionella pneumophila subsp. ...
Listeria innocua
Methanosarcina mazei Go1
Neisseria meningitidis MC58
Nitrosomonas europaea ATCC 19718
Porphyromonas gingivalis W83
Pseudomonas syringae pv. tomato str. ...
Saccharomyces cerevisiae
Salmonella typhimurium LT2
Shigella flexneri 2a str. 2457T
Staphylococcus aureus subsp. aureus Mu50
Staphylococcus aureus subsp. aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae 2603V
Streptococcus mutans UA159
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes M1 GAS
Vibrio cholerae O1 biovar eltor str. N16961
Vibrio parahaemolyticus RIMD 2210633
Xylella fastidiosa Temecula1
0
10
20
30
40
50
60
70
80
Distribution (%)
Percent in PDB
Percent soluble

Figure S3. The Species Distribution of Targets. The small-scale expression and solubility data was grouped by species as a ratio of low and high solubility clones (yellow bars). The ratio of PDB structures versus clones processed is shown in green.

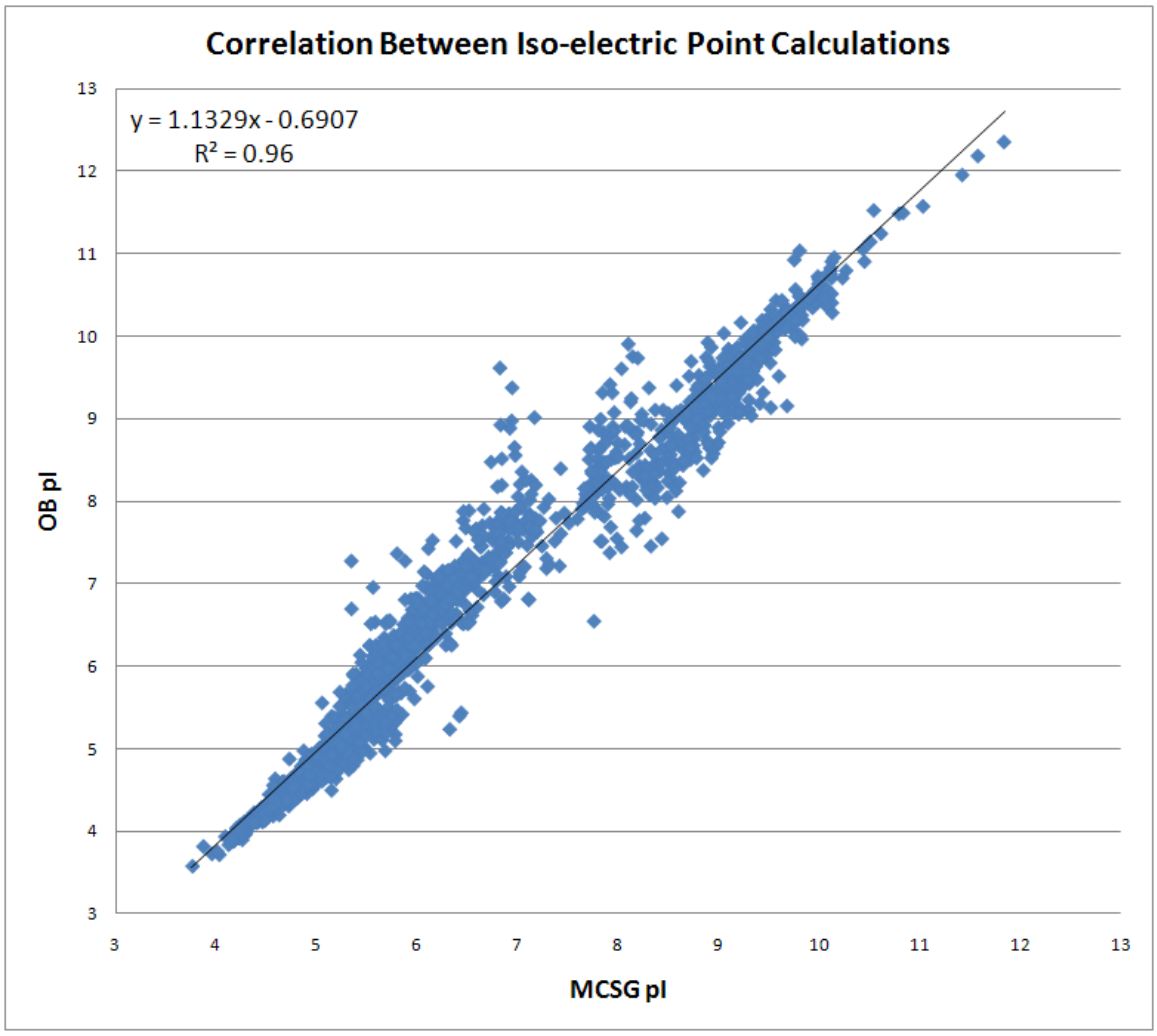

Figure S4. Correlation Between the Iso-electric Points Calculations. The pI was calculated for the datasets using the OB_score program and MCSG method. The pI values calculated by the two methods correlated well in the acidic and basic regions, but showed poor correlation around neutral pI.

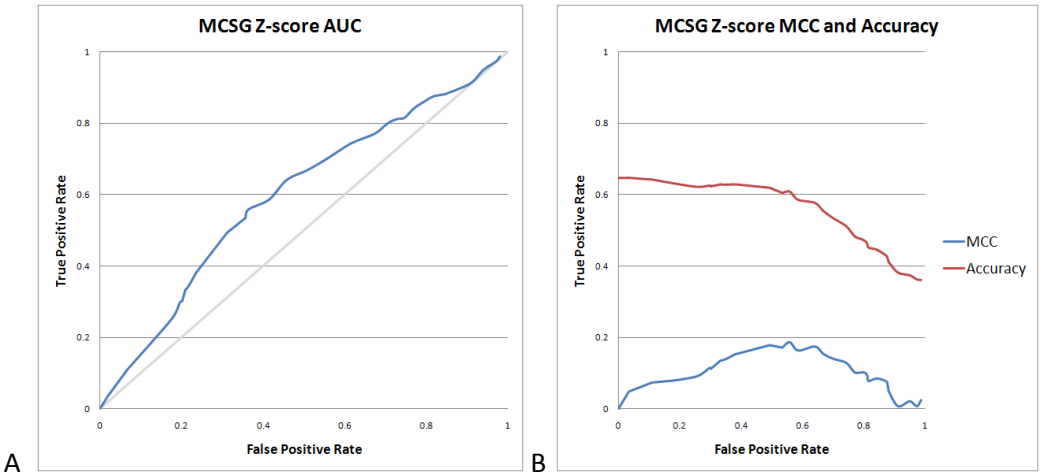

Figure S5. The Receiver Operator Curve of the MCSG Z-score Analysis. The MCSG-INSOLUBLE and MCSG-PDB datasets were combined and a repeated random sub-sampling validation was performed. 60% of the data was used to generate a Z-score matrix from the pI and GRAVY values. The generated Z-score matrix was used to calculate Z-score for the testing dataset (40%). The true positive and false positive rate is shown in A with a 58% area under the ROC. The MCC was calculated and displayed along the accuracy in relation to the true positive rate (B).

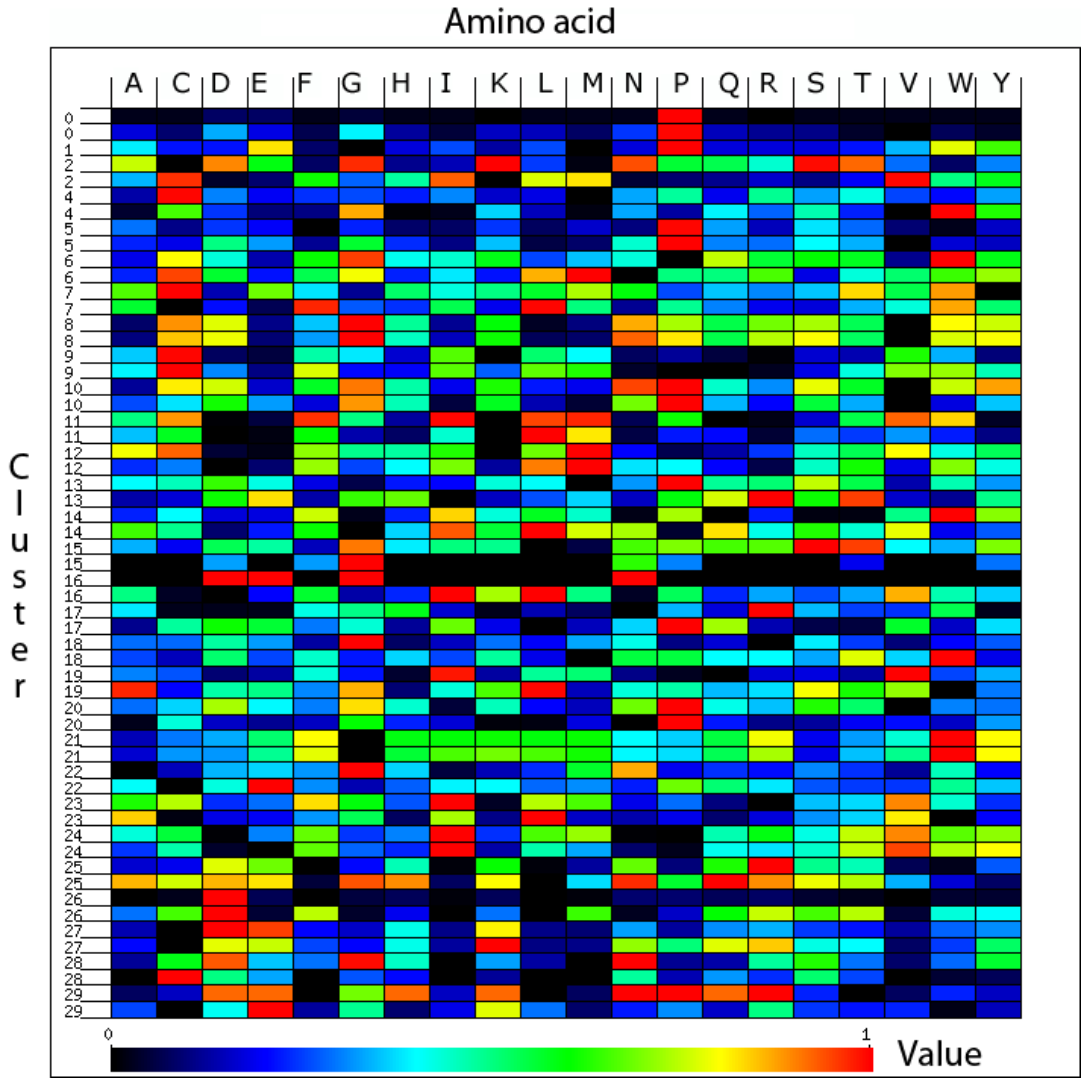

Figure S6. Clustering of AAindex1 attributes. The AAIndex database entries [33] were normalized, and clustered into 30 classes using Kohonen Maps. Two representative attributes from each class were selected in this study. The X-axis denotes the amino acid, while the Y-axis denotes the cluster identifiers. The color scale denotes the normalized value of the attributes.