International Journal of Clinical Biochemistry

ISSN: xxxx-xxxx Vol. 1 (1),
pp. 031-036, December, 2013. ©
Global Science Research Journals

Full Length Research

Virulence prediction model (virprob) using amino acid and dipeptide composition for human pathogens

S. B. Muley, V. Bastikar*, S. Bothe, A. Meshram and N. Roy

Department of Biotechnology and Bioinformatics, Padmashree Dr. D. Y. Patil University, Navi Mumbai, Maharashtra, India. – 400614.

Accepted 26 March, 2013

Pathogenic bacteria that cause infectious disease are operated by various virulence mechanisms. Hence, it is important to develop a reliable system for predicting bacterial virulent proteins aiming at discovering novel drug/vaccine and for understanding virulence mechanisms in pathogens. On the basis of features like amino acid and dipeptide composition, it tried to identify the virulence potential in the given biological protein sequence of bacteria using statistical methods like regression analysis, which is of great use in the prediction strategies of the virulence protein. In this work a bacterial virulent protein prediction model, virprob, is proposed based on classifiers, where the features are extracted directly from the amino acid sequence of a given protein. It is a probabilistic model which predicts the virulence potential of the corresponding human pathogenic bacterial protein. An extensive evaluation according to a blind testing protocol, where the parameters of the system are calculated using the training set and the system is validated in independent dataset, has demonstrated the validity of virprob with 53.6% of accuracy. The statistical analysis method may increase the prediction accuracy when combined with machine learning techniques. The results of this analysis might help in rapidly advancing knowledge of infectious agents.

Key words: Virulence, proteins, virprob, pathogen.

INTRODUCTION

A pathogen is a biological agent that causes disease or illness to its host invading through several substrates and pathways. The majority of pathogens are harmless and sometimes beneficial; a few can cause infectious diseases such as the bacterium *Yersinia pestis*, which caused the Black Plague. Pathogenicity or virulence of a pathogen is its ability to cause disease. Virulence is the result of complex interplay between parasite and host. Only a small portion of the total population of microorganisms contains the attribute while the rest are harmless or even beneficial to humans and other animals. The pathogenic mechanisms of microorganisms causing diseases other than bacteria have been probed at the molecular level but many bacterial diseases are poorly understood. Discovering bacterial virulence factors is important in understanding bacterial pathogenesis and their interactions with the host, which may also serve as novel targets for drug and vaccine development (Hsing et al., 2008).

Bioinformatics approaches and bacterial genomic data are being used to find new mechanisms of virulence, and eventually, targets for novel antimicrobials (Weinstock, 2000). Bioinformatics utilizes large databases of biological information with specific *in silico* tools to complement traditional wet laboratory-based biology (Murray, 1994). Detailed knowledge about the complete sequences of pathogen genomes provides wealth of information about the determinants of bacterial virulence. A large number of predicted proteins in these genomes are yet to be assigned any function and some of them could be virulent proteins. Due to diversity and complexity of virulence proteins, the computational tools for their interpretation, identification and characterization are still limited. Availability of accurate prediction methods for virulent proteins will enhance knowledge about bacterial virulence, annotations of (novel) virulent genes and development of novel antimicrobial targets. Similarity search methods like BLAST (Altschul et al., 1990) distinguish between virulent

*Corresponding author. E-mail: vabastikar@yahoo.co.in

and non-virulent proteins with reasonable accuracy, but reasonable enough in cases where virulent proteins are evolutionarily distant and do not have significant sequence similarity to known virulent protein sequences. Several computational strategies have been proposed to deal with the problems of finding sequences with remote similarity and homology. PSIBLAST is one such algorithm, which aids in identification of remotely similar proteins (Altschul et al., 1997). Another reasonable method to overcome this limitation is the machine learning algorithms. Statistical methods like 'regression analysis' are also of great use in the prediction strategies of the virulence protein, which when combined with machine learning techniques may increase the prediction accuracy.

METHODOLOGY

Extraction of protein sequences

Virprob uses UNIPROT (Apweiler et al., 2004) and VFDB (Chen et al., 2005) as source of human pathogenic bacteria virulent proteins sequences. The sequence entries annotated as "Probable", "Putative", "By similarity", "Fragments" "Hypothetical", "Unknown" and "Possible" were removed. It yielded 406 annotated virulent protein sequences of human pathogenic bacteria referred as 'positive dataset'. For training with non-virulent protein sequences, we selected 350 annotated protein sequences of bacterial enzymes and other non-virulent proteins from UNIPROT database referred as 'negative dataset'. The non-virulent dataset sequences were mainly chosen from the bacterial proteomes, the virulent protein sequences of which are included in the positive dataset.

Reducing redundancy of sequences

For the refinement of dataset, we reduced similarity present between sequences. We used CD-HIT (Li and Godzik, 2006) to scale the redundancy in positive and negative dataset sequences so that no two sequences were more than 40% similar. CD-HIT yielded a non-redundant dataset of sequences, out of which 338 sequences were found to be virulent (positive dataset) sequences. Out of 338, we selected 296 sequences belonging to 12 different organisms: *Vibrio cholera, Staphylococcus epidermidis, Streptococcus pneumonia, Neisseria meningitides, Mycobacterium tuberculosis, Listeria monocytogenes, Helicobacter pylori, Haemophilus influenza, Escherichia coli, Clostridium perfringens, Brucella abortus* and *Bacillus anthracis* to make the training set. Hence, we used 296 positive sequences and 258 negative sequences from bacterial proteomes to complete our final non-redundant training dataset (554 sequences).

Generation of dataset for blind test

Sequences of a few organisms were excluded from the positive non-redundant training dataset to constitute a positive independent dataset. This was done to gauge the classifier prediction efficiency for the sequences of the organisms, which were not represented in the training dataset. Similarly, random non-virulent sequences from these organisms were included in the negative independent dataset. The dataset consists of 50 virulent and 50 non-virulent sequences from the following bacterial pathogens: *Bordetella pertusis*, *Legionella pneumophila, Mycoplasma pneumoniae* and *Yersinia pestis.*

Protein features

Amino acid composition

Amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated using the following equation:

$$fraction\ of\ amino\ acid\ i = \frac{total\ number\ of\ amino\ acid\ i}{total\ number\ of\ amino\ acids\ in\ protein} \quad \text{……… (1)}$$

i = 1 to 20 .

Dipeptide composition

Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20 × 20). This representation encompassed the information about amino acid composition along local order of amino acid. The fraction of each dipeptide was calculated using following equation:

$$fraction\ of\ dep(i) = \frac{total\ number\ of\ dep(i)}{total\ number\ of\ all\ possible\ dipeptides} \quad \text{……… (2)}$$

i = 1 to 400.

Binary logistic regression

Logistic regression model can be used for prediction of dependent variable on the basis of scale and/or categorical independents to rank the relative importance of independents, assess interaction effects and understand the impact of covariate variables. Binary logistic regression is a type of logistic regression model which is used when the dependent variable is of dichotomous type and the independent variables of any type.

The non-redundant dataset of 554 sequences was analyzed and the statistical model was generated using binary logistic regression. In logistic regression, we predict the probability of outcome variable *Y* occurring in a given known values of predictor variables (*X*i). The logistic regression equation from which the probability of *Y* is predicted is given by equation:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon_i)}} \quad \text{……… …… (3)}$$

where, P*(Y)* is the probability of *Y* occurring, *e* is the mathematical constant e = 2.7182 and $b_0,b_1,..,b_n$ are the regression coefficient and ε_i is the error term.

The values of regression coefficient are estimated on the basis of dependent and independent variables which are used to predict the value of Y when X is given. The resulting value from the equation is a probability value ranging from 0 - 1. The value close to 0 indicates that *Y* is very unlikely to have occurred, and a value close to 1 indicates that *Y* is very likely to have occurred.

The values are estimated using maximum-likelihood estimation, which selects coefficients that make the observed values most likely to have occurred. So, we tried to fit a model to data that allows us to estimate values of the outcome variable from known values of the predictor variable or variables. In the present study, the dependent variable is of dichotomous type, the value 1 indicates virulent protein and 0 indicates non-virulent protein. The independent variables comprise 20 variables for amino acid composition and 400 variables for dipeptide composition.

The data is analyzed using Forward LR method of Logistic Regression in SPSS. There are several methods that can be used in logistic regression. Stepwise methods are defensible when used in situations in which no previous research exists to base hypotheses for testing, and in situations in which causality is not of interest and the researcher merely wishes to find a model to fit his data (Andey,????). In the present hypothesis, there is no prior research present to defend. For this analysis Forward: LR method of regression was used in which the computer begins with a model that includes only a constant and then adds single predictors into the model based on the value of the score statistic. The variable with the most significant score statistic is added to the model and continues until none of the remaining predictors have a significant score statistic (the cut-off point for significance being .05). At each step, the computer also examines the variables in the model to see whether any should be removed. In the Forward: LR method, the current model is compared to the model when that predictor is removed. If this removal makes a significant difference as to how well the model fits the observed data, then the computer retains that predictor (because the model is better when predictor is included). However, when the predictor removal makes little difference to the model, the computer rejects that predictor (Andey, 2005).

RESULTS AND DISCUSSION

Variables used in the equation

The same table is divided into three tables. In the table, first is the constant and rest are Amino acid or Dipeptide composition (comp). For example, ALcomp indicates - Alanine Leucine Composition. “Standard amino acid abbreviations are used.”

The results in Table 1, gives the estimates for the values of regression coefficients, that is, in column B, Wald statistic and other statistics of the desired Logistic Regression Model. The B statistic and the corresponding significance p-value, test the significance of each of the independent variable in the model. If the p-value is less than 0.05 then the independent variable is significant in the model. The independent variables included in the equation of the final model have statistical significance (p-value) less than 0.05.

The B-values are the coefficient values that we would replace in logistic regression equation (3) to establish the probability of the protein virulence potential. The Exp (B) statistic gives us the change in odds. If the value is greater than, then it indicates that as the predictor increases, the odds of the outcome occurring also increases. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. The Wald statistic tells us whether a variable is a significant predictor of the outcome or not (Andey, 2005).

The classification in Table 2 shows that, for the current data set, the Model accuracy for the classification is almost 83%. Virprob was validated with the independent dataset. The accuracy of prediction achieved is 53.6%. The dataset consists of 50 virulent and 50 non-virulent sequences from the following bacterial pathogens: *B. pertusis*, *L. pneumophila, M. pneumoniae* and *Y. pestis*.

Conclusion

Discovering virulence factors is important in understanding bacterial pathogenesis and their interactions with the host, which may also serve as novel targets in drug and vaccine development. On the basis of features like amino acid and dipeptide composition, we tried to identify the virulence potential in the given biological protein sequence. In Forward: LR method, logistic regression will move forward while dropping non-significant variables. Same approach can be used for other organism specific bacteria. Statistical method is a pro-step for machine learning technique and the output of this method can be used for machine learning.

In summary, probabilistic models for biological sequences (DNA and proteins) are frequently used in bioinformatics. We describe statistical tests designed to detect the order of dependency among elements of the sequence and to select the most appropriate probabilistic model for an experimental biological sequence. We demonstrate how one can estimate virulence potential in a particular protein that is human pathogenic bacteria by applying binary logistic regression analysis using SPSS. This is one of the approach through which we can define the probability of the particular protein being virulent.

Description of virprob

The programs runs on Windows operating system with Framework. NET 2.0 or higher. To demonstrate the functionality of virprob, we take protein sequences from the blind test dataset. First, we take raw virulent protein sequence of *M. pneumonia* (Dallo et al., 1990) from positive independent dataset and enter it into the text-field and click PREDICT. The output is result of two step process. In first step, it calculates the Amino Acid and Dipeptide Composition and substitutes them with corresponding regression coefficients values in the equation 3. P(Y) is calculated. In final step, P(Y) value is compared with the fixed threshold values. Since the P(Y) value of the entered protein exceeds the threshold of Strong Virulent Potential, the output is displayed as “PROTEIN HAS STRONG VIRULENCE POTENTIAL”, as shown in Figure 1. The protein is responsible for the pathogenecity of the organism.

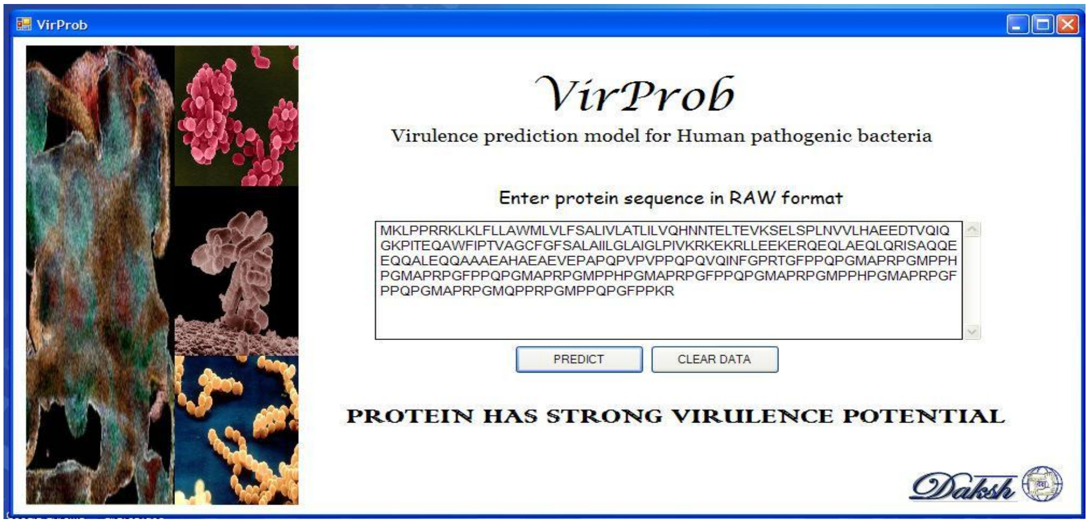
Now we take another raw virulent protein sequence of *M. pneumonia* (Himmelreich et al., 1996) from positive independent dataset. Click CLEAR DATA, it will clear the earlier data from the program. Enter the virulent raw

Table 1. Estimates for the values of regression coefficients.

Variables	Regression coefficients(B)	Standard error	Wald statistics	Significance Value (P value)	Exp (B)
Constant	7.376	1.703	18.764	0.000	0.001597
ALcomp	-0.752	0.231	10.6:09	0.001	0.472
APcomp	1.011	0.390	6.725	0.10	2.749
AVcomp	0.556	0.331	2.822	0.93	1.743
CAcomp	- 2.365	0.916	6.665	0.010	0.094
CGcomp	-2. 508	0.767	10.693	0.001	0.061
DAcomp	-1.189	0.322	13.672	0.000	0.304
DGcomp	-1.002	0.344	8.488	0.004	0.367
DPcomp	- 1.335	0.474	7.928	0.005	0.263
EQcomp	1. 518	0.396	14.791	0.000	4.565
FCcomp	2.380	1.107	4.623	0.032	10.804
FLcomp	1.741	0.440	15.666	0.000	5.706
FTcomp	1.994	0.474	17.728	0.000	7.346
FYcomp	-2.233	0.637	12.307	0.000	0.107
GQcomp	-1.339	0.416	10.367	0.001	0.262
HQcomp	- 2.765	0.787	12.342	0.000	0.063
IDcomp	-0.958	0.390	6.021	0.014	0.384
ILcomp	1. 085	0.325	11.118	0.001	2.960
KRcomp	-0.938	0.395	5.626	0.018	0.391
KVcomp	-0.888	0.321	7.316	0.007	0.420
KVcomp	-2.795	0.999	7.829	0.005	0.061
LFcomp	-1.947	0.379	26.413	0.000	0.143
MDcomp	-2.157	0.662	10.941	0.001	0.116
MRcomp	-2.414	0.691	16.672	0.000	0.089
NKcomp	-1.286	0.402	10.237	0.001	0.276
NWcomp	2.706	1.150	5.536	0.019	14.965
PScomp	-1.399	0.471	8.826	0.003	0.247
QCcomp	-4.945	1.129	19.185	0.000	0.007
QFcomp	1. 631	0.561	8.456	0.004	0.196
QMcomp	-2.049	0.722	8.058	0.005	0.129
QNcomp	-1. 027	0.626	3.814	0.051	0.366
QRcomp	-1.369	0.641	6.309	0.012	0.267
QTcomp	-2.293	0.490	21.896	0.000	0.101
RLcomp	0. 846	0.276	9.424	0.002	2.330
RVcomp	1.309	0.397	10.876	0.001	0.270
SNcomp	-1.204	0.486	6.133	0.013	0.300
THcomp	-2.135	0.843	11.043	0.001	0.118
TQcomp	1. 477	0.499	8.739	0.033	4.376
TScomp	-0.921	0.385	5.721	0.017	0.398
TWcomp	2.723	1.294	4.428	0.035	16.229
VAcomp	-1. 808	0.334	29.316	0.000	0.164
VTcomp	1.187	0.342	12.021	0.001	3.276
WNcomp	-2.285	1.042	4.810	0.028	0.102
WScomp	-3.900	1.021	14.607	0.000	0.020
YLcomp	-1.685	0.614	7.537	0.006	0.185
E	-0.430	0.077	31.174	0.000	0.651
G	-0.259	0.077	11.360	0.001	0.772
N	0. 217	0.092	5.533	0.019	1.242
S	0. 466	0.088	28.244	0.000	1.594
W	0.547	0.204	7.166	0.007	1.727

Table 2. Classification of data set.

	Predicted		
	Non Virulent	**Virulent**	**Percentage**
Non virulent	215	43	83.3
Virulent	48	248	83.8
Overall percentage	-	-	83.6



VirProb

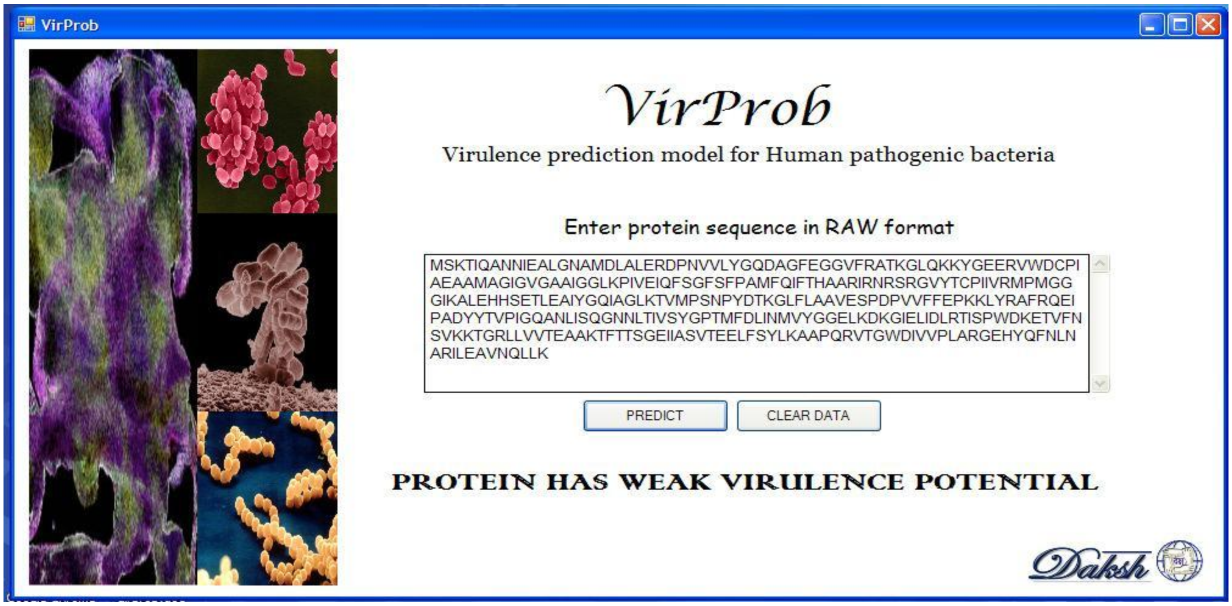
Virulence prediction model for Human pathogenic bacteria

Enter protein sequence in RAW format

```
MKLPPRRKLKLFLLAWMLVLFSALIVLATLILVQHNNTELTEVKSELSPLNVVLHAEEDTVQIQ
GKPITEQAWFIPTVAGCFGFSALAIILGLAIGLPIVKRKEKRLLEEKERQEQLAEQLQRISAQQE
EQQALEQQAAAEAHAEAEVEPAPQPVPVPPQPQVQINFGPRTGFPPQPGMAPRPGMPPH
PGMAPRPGFPPQPGMAPRPGMPPHPGMAPRPGFPPQPGMAPRPGMPPHPGMAPRPGF
PPQPGMAPRPGMQPPRPGMPPQPGFPPKR
```

PROTEIN HAS STRONG VIRULENCE POTENTIAL

Figure 1. First raw virulent protein sequence of *M. pneumonia.*



VirProb

Virulence prediction model for Human pathogenic bacteria

Enter protein sequence in RAW format

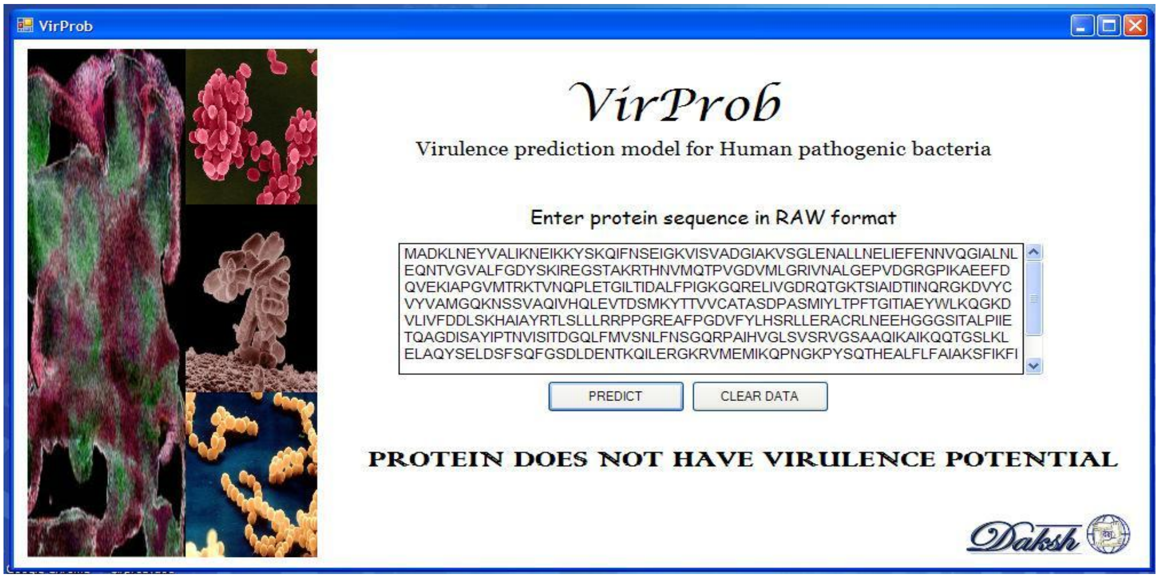
```
MSKTIQANNIEALGNAMDLALERDPNVVLYGQDAGFEGGVFRATKGLQKKYGEERVWDCPI
AEAAMAGIGVGAAIGGLKPIVEIQFSGFSFPAMFQIFTHAARIRNRSRGVYTCPIIVRMPMGG
GIKALEHHSETLEAIYGQIAGLKTVMPSNPYDTKGLFLAAVESPDPVVFFEPKKLYRAFRQEI
PADYYTVPIGQANLISQGNNLTIVSYGPTMFDLINMVYGGELKDKGIELIDLRTISPWDKETVFN
SVKKTGRLLVVTEAAKTFTTSGEIIASVTEELFSYLKAAPQRVTGWDIVVPLARGEHYQFNLN
ARILEAVNQLLK
```

PROTEIN HAS WEAK VIRULENCE POTENTIAL

Figure 2. The second raw virulent protein sequence of *M. pneumonia.*

sequence and click PREDICT. The output is “PROTEIN HAS WEAK VIRULENCE POTENTIAL” (Figure 2). The reason is that the P(Y) value for this sequence is greater than the threshold of Virulence Potential but less than that of Strong Virulent Potential. The protein somewhere helps in the virulence but not completely a virulent protein.

Now we take raw non-virulent protein sequence of *M.*



VirProb

Virulence prediction model for Human pathogenic bacteria

Enter protein sequence in RAW format

```
MADKLNEYVALIKNEIKKYSKQIFNSEIGKVISVADGIAKVSGLENALLNELIEFENNVQGIALNL
EQNTVGVALFGDYSKIREGSTAKRTHNVMQTPVGDVMLGRIVNALGEPVDGRGPIKAEEFD
QVEKIAPGVMTRKTVNQPLETGILTIDALFPIGKGQRELIVGDRQTGKTSIAIDTIINQRGKDVYC
VYVAMGQKNSSVAQIVHQLEVTDSMKYTTVVCATASDPASMIYLTPFTGITIAEYWLKQGKD
VLIVFDDLSKHAIAYRTLSLLLRRPPGREAFPGDVFYLHSRLLERACRLNEEHGGGSITALPIIE
TQAGDISAYIPTNVISITDGQLFMVSNLFNSGQRPAIHVGLSVSRVGSAAQIKAIKQQTGSLKL
ELAQYSELDSFSQFGSDLDENTKQILERGKRVMEMIKQPNGKPYSQTHEALFLFAIAKSFIKFI
```

PROTEIN DOES NOT HAVE VIRULENCE POTENTIAL

Figure 3. Raw non-virulent protein sequence of *M. pneumonia.*
The threshold values are as follows:
No virulence potential = less than 0.5
Virulence Potential = 0.5 to less than 0.8
Strong Virulence Potential = greater than 0.8

pneumonia (Himmelreich et al., 1996) from negative independent dataset. Click CLEAR DATA, it will clear the earlier data from the program. Enter the non-virulent raw sequence and click PREDICT. The output is “PROTEIN DOES NOT HAVE VIRULENCE POTENTIAL” (Figure 3). The reason is the P(Y) value of this sequence does not exceed the threshold value of Virulence Potential and is part of the organism’s normal biological function.

REFERENCE

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool, J. Mol. Biol. 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25: 3389-402.

Andey Field (2005). “Discovering statistics using SPSS” Second Edition Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q (2005). VFDB: a reference database for bacterial virulence factors. Nucleic Acids Research, 33: D325-D328

Dallo SF, Chavaoya A, Baseman JB (1990). Characterization of the gene for a 30-kilodalton adhesion-related protein of Mycoplasma pneumoniae.

Himmelreich R, Hilbert H, Plagens H, Herrmann R (1996). Sequence analysis of 56 kb from the genome of the bacterium Mycoplasma pneumoniae comprising the dnaA region, the atp operon and a cluster of ribosomal protein genes.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, LiB C, Herrmann R (1996). Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae.

Hsing JU. Andrew WU, Wan HJ, Michael PJ (2008). Discovery of virulence factors of pathogenic bacteria. Curr Opin Chem Biol. 12(1): 93-101.

Murray RP (1994). Bioinformatics and drug discovery, Curr. Opin. Biotechnol., 5: 648-653.

Rolf A, Amos B, Cathy H, Wu, Winona CB (2004). Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O’Donovan, Nicole Redaschi and Lai-Su L. Yeh: UniProt: the universal protein knowledgebase Nucleic acids research, 2004 - Oxford University Press.

Soccaronan M (2004). 1 Contact Information, K. Koscaronmelj2, N. Mariniccaron-Fiscaroner3 and L. Vidmar: A prediction model for community-acquired Chlamydia pneumoniae pneumonia in hospitalized patients. Infection, Aug., 32(4): 204-9.

Weinstock GM (2000). Genomics and bacterial pathogenesis, Emer. Infect. Dis., 6: 496-504.

Weizhong Li, Adam G (2006). CD-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13): 1658-1659.