Theoretical Peptide Mass Distribution in the Non-Redundant Protein Database of the NCBI

Dajeong Lim[1], Hee-Seok Oh[2] and Heebal Kim[1]*

[1]Department of Food and Animal Biotechnology, Seoul National University, Seoul 151-742, Korea, [2]Department of Statistics, Seoul National University, Seoul 151-742, Korea

Abstract

Peptide mass mapping is the matching of experimentally generated peptides masses with the predicted masses of digested proteins contained in a database. To identify proteins by matching their constituent fragment masses to the theoretical peptide masses generated from a protein database, the peptide mass fingerprinting technique is used for the protein identification. Thus, it is important to know the theoretical mass distribution of the database. However, few researches have reported the peptide mass distribution of a database. We analyzed the peptide mass distribution of non-redundant protein sequence database in the NCBI after digestion with 15 different types of enzymes. In order to characterize the peptide mass distribution with different digestion enzymes, a power law distribution (Zipf's law) was applied to the distribution. After constructing simulated digestion of a protein database, rank-frequency plot of peptide fragments was applied to generalize a Zipf's law curve for all enzymes. As a result, our data appear to fit Zipf's law with statistically significant parameter values.

Keywords: peptide mass, non-redundant protein database, Zipf's law

Introduction

Proteins are responsible for an organism's phenotype and function. Proteins can be identified using electrophoresis that separates proteins based on their isoelectric points and molecular weights on polyacrylamide gels, like two-dimensional polyacrylamide-gel electrophoresis (2D PAGE) (O'Farrell, 1975). For efficient identification of proteins, mass spectrometry (MS) has become the powerful method for the rapid speed and characterization of post-translational modification (Blackstock and Weir, 1999). Matrix-assisted laser-desorption-ionization-time-of-flight (MALDI-TOF) mass spectrometry is usually used for peptide mass fingerprinting (PMF) (Henzel *et al.*, 1993). The masses of peptides that came from an in-gel proteolytic digestion are searched against the protein sequence databases by used enzyme with cleavage rule. The protein sequence database was precisely digested by several kinds of enzymes and calculated by mass values of amino acids. Trypsin is usually the default enzyme and some proteomic programs do not include any enzyme. Experimental or Computational peptide values are 'average' or 'monoisotopic' mass as ratio of isotope.

Zipf's law (Zipf, 1949) is the word usage in natural languages, which follows a power-law function as a rank-frequency distribution. If each word of a language has a frequency in a large corpus, and them the list words in order of their frequency of occurrence, we can observe the relationship between the frequency of a word $f$and its position in the list as rank r(Manning and Schutze, 1999). There is a constant k such as $f \cdot r = k$ Zipf's plot was exhibited linguistic, social, economic data should be a straight line with slope -1. The abundances of expressed gene follow a power-law distribution with an exponent close to -1 in the SAGE (Serial Analysis of Gene Expression) data (Furusawa and Kaneko, 2003). Protein domains were ranked by the frequency of their connectivity and then found that the curve is similar to a generalized Zipf's law curve in the Prosite, ProDom and Pfam domain databases (Wuchty, 2001). It was found that highly connected Inter Pro domains were observed in the five organisms using topology of domain networks. Mantegna (Mantegna, *et al.*, 1994) reported that the noncoding regions are more similar to natural languages that the coding region of DNA. Distribution of RNA secondary structures for several lengths also good fitted into the power-law distribution.

Zipf's law derived from the word usage in natural language and was applied to cultural sciences, but biological data also has been followed rank-frequency distribution by several studies. We performed that peptide fragments were obtained from non-redundant protein sequence database in NCBI and digested protein sequences with 15 enzymes. After constructing simulated digestion of a protein database, rank-frequency

*Corresponding author: E-mail heebal@snu.ac.kr, Tel +82-2-880-4803, Fax +82-2-883-8812

Accepted 9 May 2006

plot of peptide fragments was able to generalize the Zipf's law curve for all enzymes.

Materials and Methods

Non-redundant protein database were downloaded from NCBI (National Center for Biotechnology Information) website (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/). All 3,154,491 fasta sequences were subjected to digest using enzymes. Enzymes and cleavage rules were listed up in Table 1. Fifteen digested peptide fragment database by enzymes were assigned ranks and frequencies by mass of fragments. We can explore the relationship between the frequency of a domain f and its position in the list, known as its rank r. $f \cdot r = k$, where k is constant. The relationship is described by a power-law that f means the frequency of size of peptide fragment and x is rank of determined from this frequency in the equation $f = ax^{-b}$ or $\log f = \log a - b\log x$. Rank-frequency plot usually shows a straight line on doubly logarithmic axes. We examined fitness of power-law distribution using linear regression model because equation of $\log f = \log a - b\log x$ is same as typical form of regression model ($y = \alpha\chi + \beta + \epsilon$). This is easily achieved using regression model on the $\log_e$ frequency versus $\log_e$ rank transformed data. The term ϵ represents the unpredicted or unexplained variation in the response variable; it is conventionally called the "error term" whether it is really a measurement error or not. The error term is conventionally assumed to have expected value equal to zero, as a nonzero expected value could be absorbed into α. We estimated α, β parameter values in the plot of Zipf's law using the least square method and examined whether parameter values have statistical significance. It is also performed to examine of the residuals (the deviations from the fitted line to the observed values of $\log f$) using residual plots for fitting a group of data. Statistical analysis used the R package (http://www.r-project.org/).

Results and Discussion

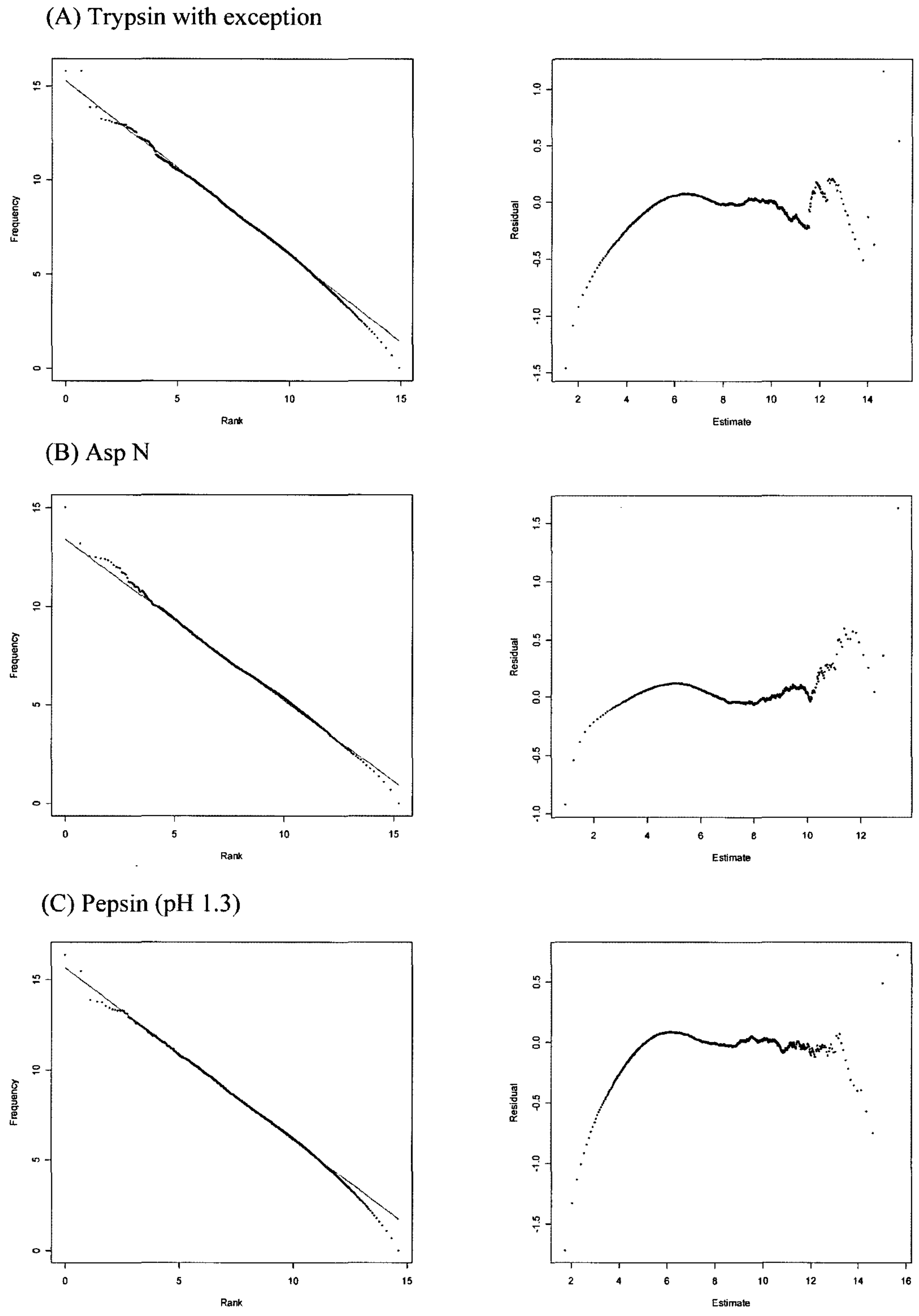
The rank-frequency plots in all the 15 peptide sequence

Table 1. Enzymes and cleavage rules

Enzyme	Cleavage site	Exception
Trypsin	C-terminal side of K or R	If P is C-term to K or R
Trypsin (without exception)	C-terminal side of K or R	
Lys C	C-terminal side of K	
CNBr	C-terminal side of M	
Arg C	C-terminal side of R	If P is C-term to R
Asp N	N-terminal side of D	
Asp N + N-terminal Glu	N-terminal side of D or E	
Glu C (bicarbonate)	C-terminal side of E	If P is C-term to E, or if E is C-term to D or E
Glu C (phosphate)	C-terminal side of D or E	If P is C-term to D or E, or if E is C-term to D or E
Chymotrypsin (Low)	C-terminal side of F, L, M, W, Y	If P is C-term to F, L, M, W, Y, if P is N-term to Y
Chymotrypsin (High)	C-terminal side of F, Y, W	If P is C-term to F, Y, W, if P in N-term to Y
Trypsin/Chymotrypsin	C-terminal side of K, R, F, Y, W	If P is C-term to K, R, F, Y, W, if P in N-term to Y
Pepsin (pH 1.3)	C-terminal side of F, L	
Pepsin (pH > 2)	C-terminal side of F, L, W, Y, A, E, Q	
Proteinase K	C-terminal side of A, C, G, M, F, S, Y, W	

Table 2. Slopes, intercept values, R^2 value and p-values of 15 rank-frequency plots.

Enzyme	Slope	intercept	R2 value	P-value
Trypsin	-0.93	15.3	0.9975	< 2e-16
Trypsin (without exception)	-0.95	15.55	0.998	< 2e-16
Lys C	-0.88	14.04	0.998	< 2e-16
CNBr	-0.79	12.09	0.9957	< 2e-16
Arg C	-0.86	13.93	0.9986	< 2e-16
Asp N	-0.92	13.4	0.9968	< 2e-16
Asp N + N-terminal Glu	-0.93	15.4	0.9976	< 2e-16
Glu C (bicarbonate)	-0.69	5.22	0.9684	< 2e-16
Glu C (phosphate)	-0.91	15.21	0.9977	< 2e-16
Chymotrypsin (Low)	-1.02	16.79	0.9948	< 2e-16
Chymotrypsin (High)	-0.88	14.68	0.9973	< 2e-16
Trypsin/Chymotrypsin	-1.03	16.85	0.9952	< 2e-16
Pepsin (pH 1.3)	-0.95	15.66	0.9976	< 2e-16
Pepsin (pH > 2)	-1.23	18.67	0.996	< 2e-16
Proteinase K	-1.23	18.67	0.995	< 2e-16



(A) Trypsin with exception

(B) Asp N

(C) Pepsin (pH 1.3)

Fig. 1. Rank-Frequency distribution and the residual plot. (A) Trypsin with exception, (B) Asp N, (C) Pepsin (pH 1.3). 15 rank-frequency plots on log-log graph have closely similar patterns. The left side of plot is rank-frequency plots on log-log graph with regression line (red line). The right side of plot is the residual plot of regression. Since all plots described nonrandom form, zipf's law distribution doesn't determine the linearity of regression model

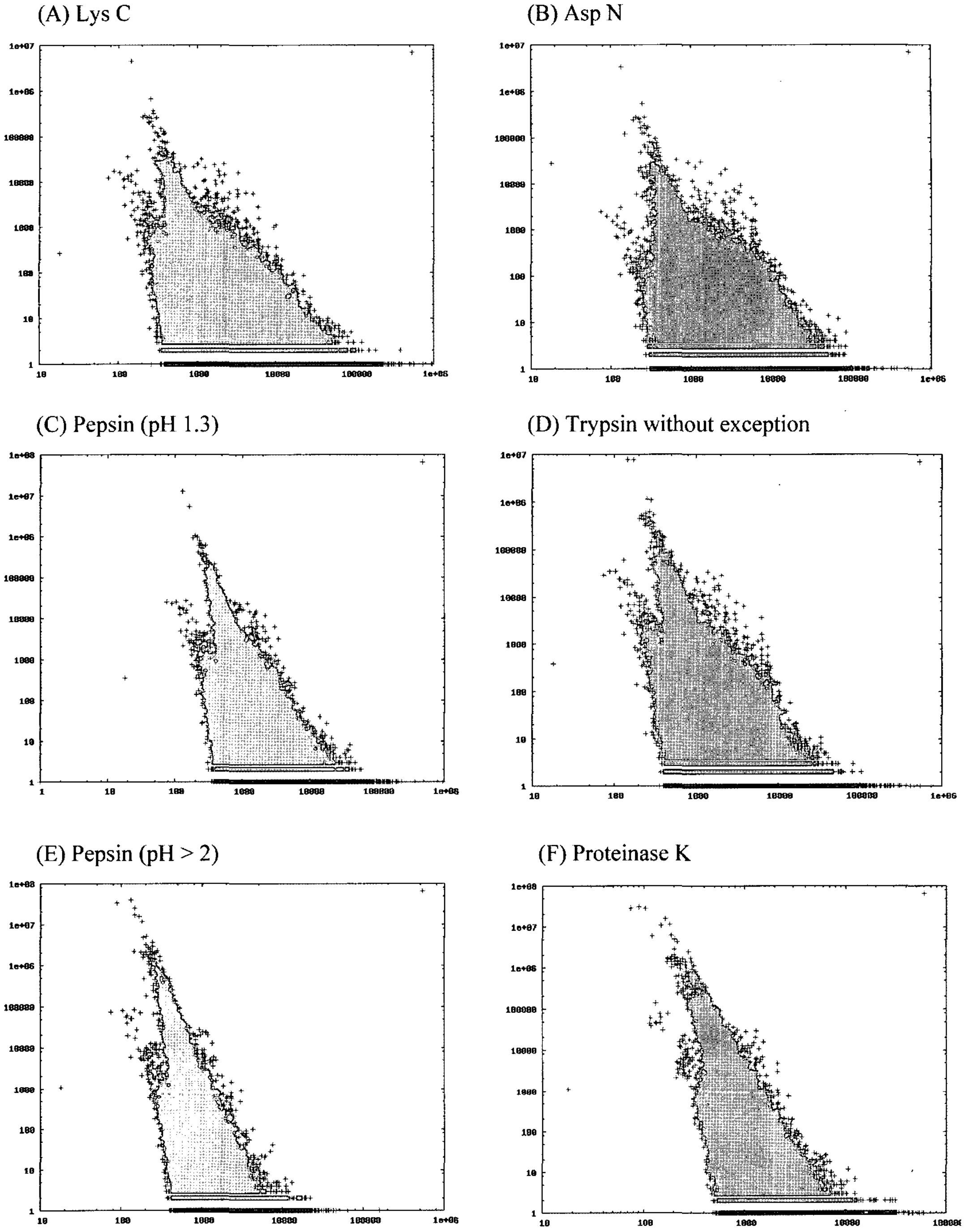


Fig. 2. Size-Frequency distribution of the three enzymes. After determined size-frequency distribution from 15 enzymes digestion, the patterns have nearly similar shape. However, the width of distributions was the difference between 15 enzymes. The more cleavage sites digested by an enzyme, the narrower range obtained. (A) Lys C, (B) Asp N, (C) Pepsin (pH 1.3), (D) Trypsin without exception, (E) Pepsin (pH > 2), (F) Proteinase K.

databases followed the famous Zipf's law curve. The distribution of Zipf's law typically followed a power law function with an exponent close to -1 (Zipf, 1949). The most common method of verifying conformity to Zipf's law is a linear regression on the $\log_e$-$\log_e$ transformed data set (Lu *et al.*, 2005). The distribution has a similar linear appearance when plot on a log-log graph, where -b defines the slope in the equation of $\log f = \log a - b\log x$. Table 2 showed the parameter values and the p-values of 15 rank-frequency plots. As illustrated in Fig. 1, our regression showed a good fit, with significant parameter values. The Zipf's plots of 15 peptide fragment databases showed the power-law distribution with the exponent in the range from -1.23 ~ -0.86. Luscombe *et al.* reported that power-law function provides the best description among linear, exponential, double-exponential, triple-exponential, stretched-exponential and lognormal of a wide group of properties associated with genomes (Luscombe *et al.*, 2002). We found that the overall distributions are very similar against each peptide fragment database. The high- and low-ranking masses of peptide fragment get out of the regression line in Fig. 1. A residual plot is a scatter plot of the residuals and is easier to see nonlinearity. If the regression model represents the data correctly, the residuals are randomly distributed around the line with zero mean (Moore and McCabe, 2003). However, for nonlinearity related data, the residual plot shows a systematic pattern, curve form. Our 15 residual plots showed similar curve forms. In other words, parameters were explained the Zipf's law in the regression model but Zipf's law distribution doesn't determine linearity of regression model. The Zipf's law has been not exactly conformed to the linear model because the line does not have linear forms in the data of high and low ranks. Manning and Schutze noted that the line is often a bad fit, especially for low and high ranks and derives a more general relationship between rank and frequency (Manning and Schutze, 1999). Kalda also reported that the distribution of low variability periods in the activity of human heart rate typically followed a multi-scaling Zipf's law and presened a non-linear time-series method (Kalda *et al.*, 2001). It has been studying statistical model of Zipf's law.

Size distribution and Zipf's law plot show nearly similar patterns. The size distribution has a trend which guides cleavage sites to more frequencies and smaller sizes (Fig. 2). These size-frequency distributions have majority of which are covered a wide range relative to the majority of distribution with only one cleavage site. For example, Pepsin (pH > 2) and Proteinase K cleave the C-terminal side of F, L, W, T, A, E, Q and that of A, C, G, M, F, S, Y, W, respectively. Pepsin has different cleavage sites by pH; Pepsin (pH 1.3) and pepsin (pH >2) cleave C-terminal side of F, L and that of F, L, W, Y, A, E, Q without exception. The shape of size-frequency distribution of pepsin (pH > 2) has more narrow than that of pepsin (pH 1.3). Fig. 2 showed that the more cleavage sites digested by an enzyme, the smaller range obtained.

Both high-ranking (common) and low-ranking (rare) words are not good candidates for keywords in database search and information retrieval (Luhn, 1957). Using this theory, Zipf's law normalization is also a useful tool than existing published normalization methods in microarray (Lu *et al.*, 2005). Several analyses have studied the relationship of power law and biological or medical datasets. We determined the applicability of Zipf's law using frequencies of mass of peptide fragment by 15 enzymes. Our results also followed a power law distribution and were described by a simple mathematical model.

Acknowledgements

This work is supported by the grant (#20050301034487) from BioGreen 21 Program of the Korean Rural Development Administration.

References

Blackstock, W.P. and Weir, M.P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17, 121-127.

Furusawa, C. and Kaneko, K. (2003). Zipf's law in gene expression. *Phys. Rev. Lett.* 90, 88-102.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA.* 90, 5011-5015.

Kalda, J., Sakki, M., Vainu, M., and Laan, M. (2001). Zipf's law in human heartbeat dynamics. *Physics*, 1-4.

Lu, T., Costello, C.M., Croucher, P.J., Hasler, R., Deuschl, G., and Schreiber, S. (2005). Can Zipf's law be adapted to normalize microarrays? *BMC Bioinformatics* 6, 37.

Luhn, H.P. (1957). A statistical approach to mechanized encoding and search of literature information. *IBM J. Res. Develop.* 2, 159-165.

Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T., and Gerstein, M. (2002). The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.* 3, RESEARCH 0040.1-0040.7.

Manning, C.D. and Schutze, H. (1999). *Statistical natural Language processing.* (Cambridge: MIT Press).

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., and Stanley, H.E. (1994). Linguistic Features of Noncoding DNA Sequences. *Phys. Rev. Lett.* 73, 3179-3172.

Moore, D.S. and McCabe, G.P. (2003). *Introduction to the practice of statistics* W.H. Freeman, ed.(New York)

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250, 4007-4021.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* 18, 1694-1702.

Zipf, G.K. (1949). Human behavior and the principle of least *effort: an introduction to human ecology*. (Cambridge: Addison-Wesley Press).