Characterizing Human Protein Mass Density Distributions

Gil Alterovitz[1], Isaac S. Kohane[2,3], and Marco F. Ramoni[2,3]
[1] Health Science and Technology/Electrical Engineering & Computer Science,
Massachusetts Institute of Technology, Cambridge, 02139, USA.
[2] Children's Hospital Informatics Program, Harvard Medical School, Boston, 02115, USA.
[3] Harvard Partners Center for Genetics and Genomics, Boston, 02115, USA.

***Abstract*— Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI or SELDI-TOF MS) has yielded predictive protein profiles for a number of clinically relevant diseases. Yet, rather than identifying specific proteins, such studies have provided diagnostic information solely based on "black box" predictors that look at differential patterns of mass spectrometry peaks.**

This paper analyzes the number of proteins that could be represented by mass spectrometry peaks associated with corresponding masses. It proposes and compares three models to fit the probability density function (PDF) of such a distribution. These include the gamma, Poisson, and negative binomial models.

The results yielded a nonuniform distribution of protein masses- particularly apparent near masses where proteins involved in somatic recombination are prevalent. This may be useful to consider when using protein databases for protein identification near such mass regions. In terms of PDF models, the distribution surprisingly does not follow a simple Poisson distribution of counts. Instead, it follows a negative binomial distribution.

I. Introduction

SELDI-based mass spectrometry is turning out to be one of the high growth areas in proteomics research in recent years [1]. Using statistical and signal processing techniques, early studies [2], [3] were the first to predict pathological states in their respective domains (e.g. ovarian cancer), solely using serum proteins. However, SELDI studies are struggling with actual protein identification, often providing no more than a pattern-based predictor model. Purification, isolation, and manual identification of proteins can take months. However, new computational approaches may help to provide other alternatives.

SELDI works by subjecting protein samples (within an energy absorbing matrix) to a laser pulse. This ionizes the proteins and sends them flying into a chamber. The mass-to-charge ratio of the protein can be calculated from the time of flight (i.e. the time it takes the protein to travel before hitting the detector at the end of the compartment).

II. Methods

Since the mass-to-charge ratio is known for SELDI mass spec peaks, and most peaks represent univalent ions, one can use the effective resolution of the mass spec to try to ascertain the mass of the protein. The problem is that many proteins have similar masses, so it is hard to uniquely identify a protein based solely on mass (even assuming single ionic charge, z=1).

As a first step, this paper looks at the protein "mass density"- namely the number of proteins that could theoretically exist (based on a database) per given mass window (e.g. 1 Dalton). In this paper, human proteins present in the mass range of 700 - 12,000 Daltons (Da) are focused on. This is the same range used in a recent high resolution (in terms of dynamic range) SELDI mass spec instrument reported [4]. Even with an ideal intra-machine mass drift of approximately 100 ppm (parts per million), this aforementioned SELDI-based instrument cannot be more accurate than +/-1.2 Da at 12000 Da. In that study, bins of 400 ppm were actually used for analysis to accommodate inter- and intra-assay variance and drift. Other constraints to mass spectrometry-based proteomics include the probabilities associated with finding particular peptides extracellularly- and within the appropriate tissue or body fluid sampled.

To generate a human protein mass density plot, the Entrez Protein database was searched for all human proteins with molecular weight along the 700 - 12,000 Da mass range within 1 Da windows. This search included cleavage products of proteins and protein precursors (based on annotation features) for a more accurate picture of potential proteins that might be found upon mass spectrometry.

As the Entrez Protein database contains redundant entries, SeqHound [5] was used to determine the non-redundant entries via remote Java Application Protocol Interface (API) calls implemented in Matlab. Finally, each database sequence was examined for similarity directly, further reducing the number of non-redundant proteins.

It would be useful to be able to model the number of proteins per mass unit for probabilistic calculations. Often, the Poisson distribution is used to model situations involving counts or arrivals during an interval of time. In this case, at each mass unit, λ proteins are modeled as being expected to 'arrive' during this interval.

A second model explored is the negative binomial distribution. It has been used modeling count data as well. For example, one common application is modeling daily road accidents at certain highway locations [6]. When parameter r is not restricted in integer values, the more general expression

for the negative binomial distribution is (where Γ(x) is the gamma function) as shown in Equation (1):

$$f(x|r,p) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} p^r (1-p)^x \quad (1)$$

Lastly, the gamma distribution, was examined as a model for the PDF in question. The gamma distribution is useful in measuring failure times and is a superset of the exponential distribution since it allows for an additional dependence on the "age" of the item. In this case, the failure times would be the count of proteins that are confined to a certain mass before the next mass window starts.

The above models assume independent, identically distributed (IID) counts of proteins per mass unit (not perfect satisfied, as alluded to above). Each of the aforementioned distributions' parameters were then estimated via maximum likelihood estimation (MLE).

III. Results

The number of non-redundant human proteins found in Entrez within the 700-12,000 Da range (determined as described above) was 36,024. Thus, it is expected that there would be around 3 proteins per mass unit (Da). However, while the average number of proteins per mass unit is 3.19, the distribution across the mass range clearly shows a pattern, rather than uniformity- with peaks at approximately 2,300 Da and 11,600 Da among others.

Exploration of the first peak (2,300 Da) suggested the commonality was in slightly different proteins dealing with T-cell receptor beta/delta chains. For the second peak, different forms of immunoglobulin heavy/light chain variable regions were seen.

The resulting models for the PDF of the human protein mass density are shown in Figure 1. The negative binomial model (with MLE estimates of r=6.19 and p=0.660) was the best fit by several measures. A Monte Carlo simulation of the Wilcoxon rank sum test at the 5% significance level was performed to test concordance of the models with the data. The negative binomial distribution was the best fitting model. It also had the best log likelihood score at: -2.39×10^4.

IV. Discussion and Conclusion

The nonuniformities noted for the human protein mass density plot involved elements associated with somatic recombination. Since immunology demands a great diversity of T-cell receptor chains and immunoglobulin variable regions, knowing the masses where these molecules are concentrated could potentially help in tuning protein identification algorithms as well as yield insights into the relevant biology. In fact, T-cell receptor loci and immunoglobulin loci both have gene segments with variable regions that are rearranged by exactly the same enzymes [7]. Yet, these regions are typically not seen in isolated form extracellularly. Sequest, Mascot, ProFound, and other applications (which depend on databases like NCBI-nr [8] for accurate protein identifications) are all potentially susceptible to such database biases.

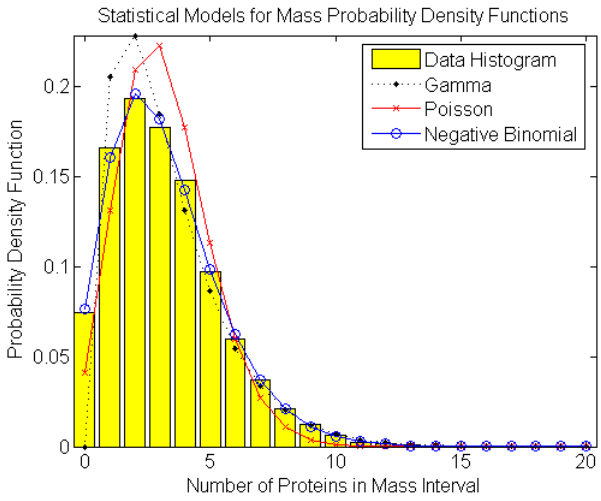

Fig. 1. The probability density function of the data appears to follow a negative binomial distribution.

The Poisson distribution, often used for count data, might be expected to be a good model for the PDF of human protein mass density. However, the Poisson model has only one parameter (λ) which is equal to the mean and variance. Yet, since the variance is 4.82 (compared to 3.19 for mean), the Poisson model is not a good fit in this case (see Figure 1). In such cases, the negative binomial distribution can be used as it has two parameters (p and r). As there can be high variances in this scenario (analogous to daily accidents dependent on the day's weather conditions in the aforementioned highway study), a negative binomial model is more suitable here compared to Poisson.

By exploring statistical models for mapping mass spectrometry peaks to actual proteins, the strengths and weaknesses of existing technologies can be measured and new methodologies explored for better protein identification.

References

[1] G. Alterovitz, E. Afkhami, and M. Ramoni,"Robotics, Automation, and Statistical Learning for Proteomics," in *Focus on Robotics and Intelligent Systems Research*, vol. 1, F. Columbus, Ed. New York: Nova Science Publishers, Inc., 2005 (In press).

[2] E. F. Petricoin, Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A., "Clinical proteomics: translating benchside promise into bedside reality," *Nature Rev. Drug Discovery*, vol. 1, pp. 683-695, 2002.

[3] G. Alterovitz, M. Aivado, D. Spentzos, et al., "Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry," *Proceedings of IEEE EMBS*, San Francisco, CA, USA, 2004.

[4] T. P. Conrads, V. A. Fusaro, S. Ross, et al., "High-resolution serum proteomic features for ovarian cancer detection," *Endocr Relat Cancer*, vol. 11, pp. 163-78, 2004.

[5] K. Michalickova, G. D. Bader, M. Dumontier, et al., "SeqHound: biological sequence and structure database as a platform for bioinformatics research," *BMC Bioinformatics*, vol. 3, pp. 32, 2002.

[6] S. P. Miaou and H. Lum, "Modeling Vehicle Accidents and Highway Geometric Design Relationships," *Accident Analysis and Prevention*, vol. 25, pp. 689-709, 1993.

[7] C. A. Janeway, P. Travers, M. Walport, et al., *Immunobiology: The Immune System in Health and Disease*, 6th ed. London: Garland Science Publishing, 2004.

[8] J. P. Quadroni M, "Proteomics and automation,"*Electrophoresis*, vol. 20, pp. 664-77, 1999.