Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

MODELING PROTEIN TANDEM MASS SPECTROMETRY DATA WITH AN EXTENDED LINEAR REGRESSION STRATEGY

Han Liu[1], *Anthony J.Bonner*[1], *Andrew Emili*[2]

[1]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[2]Department of Medical Genetics & Microbiology, University of Toronto, Toronto, Canada

ABSTRACT

Tandem mass spectrometry(MS/MS) has emerged as a cornerstone of proteomics owing in part to robust spectral interpretation algorithm. The intensity patterns presented in mass spectra are useful information for identification of peptides and proteins. However, widely used algorithms can not predicate the peak intensity patterns exactly. In this paper, we have developed a systematic analytical approach based on a family of extended regression models, which permits routine, large scale protein expression profile modeling. By proving an important technical result that the regression coefficient vector is just the eigenvector corresponding to the least eigenvalue of a space transformed version of the original data, this extended regression problem can be reduced to a SVD decomposition problem, thus gain the robustness and efficiency. To evaluate the performance of our model, from 60,960 spectra, we chose 2,859 with high confidence, non redundant matches as training data. based on this specific problem, we derived some measurements of goodness of fit to show that our modeling method is reasonable. The issues of overfitting and underfitting are also discussed. This extended regression strategy therefore offers an effective and efficient framework for in-depth investigation of complex mammalian proteomes.

keywords− **tandem mass spectrometry, regression, proteomes, protein expression profile, goodness of fit.**

1. INTRODUCTION

Proteomics–the direct analysis of the expressed protein components of a cell, is critical to our understanding of cellular biological processes. Key insights into the action and effects of a disease can be gotten by comparison of the expression of the expressed proteins in normal versus diseased tissue[1]. Tandem mass spectrometry(MS/MS) of peptides is a central technology for proteomics, enabling the identification of thousands of peptides from a complex mixture[2][3]. In tandem mass spectrometry(MS/MS), many peptides are ionized with one or more units of charge, and one chosen for fragmentation by *collision-induced dissociation*(CID). Fragments retaining the ionizing charge after CID have their mass-charge ratio measured. Since peptides typically breadk a peptide-bond when they fragmented by CID, the resulting spectrum contains information about the constituent amino-acids of the peptide[4]. One of the major bottlenecks in the efficiency for such a method is the spectrometry intensity of different peptides. The difference between an experimental result and a theoretically calculated one is quite large. Usually, in an ideal experiment, there would be no loss during the sample preparation procedure and all the digested peptides should be present on the MS/MS spectrum. Therefore, an ideal spectrum would contain as many peaks as there are peptides resulting from the current cleavage of the protein. Furthermore, the MS/MS apparatus would perfectly ionize all the peptides, and as there is theoretically the same quantity of each peptide, the measured intensity of each peak should be the same. However, there are usually only a few theoretical peptides that match well-definded and easily identifiable MS/MS spectrum peaks. The majority of theoretical peptides match only small peaks or no peak at all. Numerous reports exist in the literature that address the question of the factors influencing the quality of a MS/MS experiment[6][7]. An obvious factor influencing peak intensities is the concentration of the peptides in the sample. However, it is not the sole factor. Such factors involve, among others, the sample preparation methods, pH and composition of the solution, the characteristics of the MS/MS apparatus and characteristics of the analyzed sample[5]. All these factors affect the precise predication of the peak intensities, unfortunately, no model exists that predicts accurately the peak intensities of a MS/MS experiment.

In this work, we used a family of extended linear regression models as data minning tools, by taking advantage of the huge amount of data obtained from MS/MS experiment to find clues enabling us to improve the bioinformatics part of the mass spectrometry method, *ie.*, the prediction of the peak intensity of a specific peptide from a protein. Under some given assumptions, our modeling strategy is simple

This work collaborated with Emili Lab in University of Toronto; Han is supported by the scholarship from University of Toronto.

and robust. We evaluated our method on some real data, the result is quite promising. This paper is organized as follows: the next part will introduce our modeling strategy ; part III illustrated our main technique result−an elegant solution for the extended linear regression model and statistic measurements for goodness of fit ; part IV is an experiment on real data and the analysis, also the discussion and analysis. part V gives conclusion and some directions of future work.

2. METHODOLOGY

A theoretical framework combing both regression and singular value decomposition(SVD) was developed in this part, which is purposed on modeling protein tandem mass spectrometry data. Our model takes advantage from both regression and SVD, apparently, it's a supervised parametric model, which has similar form to regression model but has a more powerful expression capability; Essentially, it can be reduced to a SVD decomposition problem, but computationally more efficient.

In this paper, the working hypothesis was that the peak intensity is reproducible for a given peptide in the same experimental conditions. Hence, the peak intensity is only dependent on peptides characteristics and a correlation between these properties and the peak intensity can be found. Another important assumption concerns the suppression effect[8],which is the phenomena of chemical interactions between peptides during the MS/MS experiment that can lead to the absence of MS/MS peaks or to new peaks corresponding to unexpected peptides. Because the regression model requests independence between different observations−the peak intensity of a peptide does not influence the peak intensity of another peptide−we will assume that there are no such interactions between peptides in the MS experiments.

Assume that p_{i1}, p_{i2},...,p_{ik_i} are different peptides fragmented from the same protein P_i in the MS/MS experiment. $y_{i1}, y_{i2}, ...y_{ik_i}$ are the peak intensities for different peptides which were represented by occurrence count number. the true input number in_i for protein P_i is an unknown latent variable, we can give out a model for the relationship between y_{ij}, in_i and ie_{ij} as:

$$f(y_{ij}) = g(in_i \cdot ie_{ij}) + \varepsilon_{ij}, j = 1, ...k_i \qquad (1)$$

Here, $f(\cdot)$ and $g(\cdot)$ are two transformation functions,while ie_{ij} represents the ionization efficiency for the jth peptide p_{ij} fragmented from protein P_i. $j = 1, ..., k_i$(there are altogether k_i peptides from protein P_i), ε_{ij} is the systematic error. Assume that the label-attached peptide is depicted as $LABEL - R_1 - R_2 - R_3 - R_4 - R_5 - R_6 - .....$ with the n^{th} amino acid residue as measured from the label-attached end of the peptide referred to symbolically as R_n,here R_i is one of the 20 amino-acids.

The key issue now is how to model the ionization efficiency ie_{ij}, to this purpose, we must first convert the amino acid subsequences in the peptides into some kind of number on which machine learning technique can be applied directly. Under our assumption, we assume the distribution of the ie_{ij} satisfies a family of linear regression models from the occurrence frequency of different amino acids subsequences. define a predicator matrix X, where an element x_{ij} represents the occurrent frequency of amino acids subsequce s_{ij} for the jth peptide from protien P_i, the rows of X represent the sequences of different peptide fragments in the experiment. There're different kinds of models here, for example, for a 0-order linear model, only subsequences with length 1 are considered while the order information was ignored; For 1-order model, we will consider the occurrence frequency of subsequences with length no more than 2; Similarly, for 2-order model, besides considering the 0 order and 1 order terms, we still need to consider the occurrent frequency of subsequences with length 3; What's more, we have some hybrid model−0-order hybrid model, 1-order hybrid mendel− for which not only the order information, but also the quadratic and cubic terms are considered. Based on this method, an amino acid sequence can be changed to one of the rows of a matrix X. Based on this definition of X, a linear regression model for the ionization efficiency ie_{ij} is:

$$ie_{ij} = [1, x_{ij}]\beta + \varepsilon'_{ij}, s.t.||\beta|^2| = 1, j = 1, ...k_i \qquad (2)$$

In equation (2), the "1" represents the bias term, ε'_{ij} doesn't like ε_{ij} as stated in equation(1), ε'_{ij} means the error generated by the underlying chemical and biological principles. We have a restriction that $||beta||^2 = 1$, this is because in our model, there exists a latent variable in_i, thus, β can be scaled to any constant factor by adjusting the relative value of in_i, thus this assumption is reasonable. By assuming that $f(x) = x$ and $g(x) = x$, through substituting the variable ie_{ij} in equation (1) with equation (2), we can get

$$y_{ij} = in_i \cdot [1, x_{ij}]\beta + \varepsilon''_{ij}, s.t.||\beta||^2 = 1, j = 1, ...k_i \quad (3)$$

here, ε''_{ij} is the combination of both ε_{ij} and ε'_{ij}, normally, we assume that $\varepsilon_{ij} << \varepsilon'_{ij}$, so, $\varepsilon''_{ij} = in_i \cdot \varepsilon'_{ij}$. Equation (3) is a family of extended linear regression models to model the MS/MS data, based on different hypothesis on X, we have different models, besides 0-order,1-order,...,n-order, hybrid models, we also derived some nonlinear models such as log models and squared root models, detailed discussions for this issue could be found in [9]. We say this method is an "extended" regression model because of the existence of the latent variable in_i, classical regression model can not handle this. In the following, we will derive an elegant solution for this problem. by which, the regression coefficient vector β can be easily trained, thus accomplish the modeling task.

3. THE ELEGANT SOLUTION FOR EXTENDED LINEAR REGRESSION MODEL

To find a solution of the column vector β, we must first deal with the latent variable in_i. The trick here is by dividing different equations for peptides from the same protein, to get rid of the latent variable.*ie.* For the peptides from a given protein P_i, we have:

$$\begin{cases} y_{i1} = in_i \cdot [1, x_{i1}]\beta \\ y_{i2} = in_i \cdot [1, x_{i2}]\beta \\ \vdots\vdots\vdots \\ y_{ik_1} = in_i \cdot [1, x_{ik_1}]\beta \end{cases} \quad s.t. ||\beta||^2 = 1 \tag{4}$$

From (4), we can get that $\frac{y_{i,j-1}}{y_{i,j}} = \frac{in_i \cdot [1, x_{i,j-1}]\beta}{in_i \cdot [1, x_{i,j}]\beta} = \frac{[1, x_{i,j-1}]\beta}{[1, x_{i,j}]\beta}$,while $||\beta||^2 = 1$ and $j = 2, ..., k_i$. By defining

$$Y_{ij} = y_{i,j-1} \cdot [1, x_{ij}] - y_{i,j} \cdot [1, x_{i,j-1}], j = 2, ...k_i \tag{5}$$

and Y is the matrix for all the responses Y_{ij}, $i = 1, ...N$(N is the valid protein number− we say a protein P_i is valid if and only if at least two peptides from this protein were observed in the MS/MS experiment) and $j = 1, ...k_i$. We define a matrix$Y$ which rows are Y_{ij}:

$$Y^T = (Y_{11}, ..., Y_{1,k_1}, Y_{21}, ..., Y_{2k_2}, ..., Y_{l1}, ..., Y_{lk_l})^T \tag{6}$$

We proved Lemma 1 and Theorem 2 here, which are two main theoretical underpinnings for our efficient and robust solution of extended linear regression model, the key idea is to reduce the problem of finding coefficient vector for regression model into a problem finding solutions to a linear equation groups(LEG), then, find the solution of the LEG efficiently. The followings are detailed mathematical derivations, more details and discussions can be found in the parallel paper[9]:

Lemma 1:*In the sense of least square error, a parameter vector β is the solution for LEG: $Y\beta = 0$ st. $||\beta||^2 = 1$ if and only if β is the regression coefficient vector for the extended linear regression model as shown in (3)*

Proof: For the peptides from the same protein P_i,β is the regression coefficient vector for the extended linear regression model iff. $y_{ij} = in_i \cdot [1, x_{ij}]\beta$, *st.* $||\beta||^2 = 1$,$j = 1, ..., k_i$, iff.$\frac{y_{i,j-1}}{y_{i,j}} = \frac{in_i \cdot [1, x_{i,j-1}]\beta}{in_i \cdot [1, x_{i,j}]\beta} = \frac{[1, x_{i,j-1}]\beta}{[1, x_{i,j}]\beta}$,while $||\beta|| = 1$ and $j = 2, ..., k_i$. iff. $y_{i,j-1} \cdot [1, x_{ij}]\beta - y_{i,j} \cdot [1, x_{ij}]\beta = 0$,$j = 2, ..., k_i$ iff. $Y_{ij}\beta = 0$, $j = 2, ..., k_i$ for a given protein P_i iff. $Y\beta = 0$,for all the proteins. $\diamond$

Also, from Lemma 1, we know that to make $Y\beta = 0$ *st.* $||\beta||^2 = 1 \Leftrightarrow \hat{\beta} = arg\min_\beta ||Y\beta||^2$ *st.* $||\beta||^2 = 1$. Based on which, we will go on to prove one of our main technical results− theorem 2:

Theorem 1:*In the sense of least square error, the solution vector β for the LEG: $Y\beta = 0$ s.t.$||\beta|| = 1$ is equivalent to an eigenvector of the matrix $A = Y^TY$ which corresponds to the smallest eigenvalue λ_{min}*

Proof: To simplify the proof procedure, not loss of generality, we define Y as :

$$Y^T = (Y_1, ..., Y_N)^T \tag{7}$$

for every $Y_i = (y_{i1}, y_{i2}, ..., y_{ip})$, which is a column vector. For the least square error, the solution vector β satisfies $\hat{\beta} = arg\min_\beta ||Y\beta||^2$ *st.* $||\beta||^2 = 1$, with lagrange multiplier, define $F = (Y\beta)^2 - \lambda||\beta||^2 = (Y\beta)^2 - \lambda\beta^T\beta$, the β should satisfy $\frac{\partial F}{\partial \beta} = 0$, for the jth element β_j in the vector β, we can get

$$0 = 2\sum_i (Y_i\beta) \cdot \frac{\partial Y_i\beta}{\partial \beta_j} - 2\lambda\beta_j \tag{8}$$

while $\lambda\beta = \lambda(\beta_1, \beta_2, ...\beta_p)$. From equation(15), we get that $\lambda\beta_j = \sum_i (Y_i y_{ij})\beta$. By which, we get:

$$\begin{cases} \lambda\beta_1 = (Y_1y_{11} + Y_2y_{21} + \cdots + Y_Ny_{N1})\beta \\ \lambda\beta_2 = (Y_2y_{12} + Y_2y_{22} + \cdots + Y_Ny_{N2})\beta \\ \vdots\vdots\vdots \\ \lambda\beta_p = (Y_1y_{1p} + Y_2y_{2p} + \cdots + Y_Ny_{Np})\beta \end{cases} \tag{9}$$

For the matrix vector $A\beta = Y^TY\beta$, it can be written as:

$$\begin{pmatrix} y_{11} & \cdots & y_{N1} \\ y_{12} & \cdots & y_{N2} \\ \vdots & \vdots & \vdots \\ y_{1p} & \cdots & y_{Np} \end{pmatrix} \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots \\ y_{N1} & \cdots & y_{Np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{10}$$

From equation (9) and equation (10), it's obviously that $Y^TY\beta = A\beta = \lambda\beta$, thus we know that β is an eigenvector of the matrix A, while the lagrange multiplier λ is the eigenvalues.

Further, we will go on to prove that λ is also the least squared error: This is because $\lambda\beta = Y^TY\beta$, then, we have $\lambda\beta^T\beta = Y^TY\beta \Rightarrow \lambda||\beta||^2 = \beta^TY^TY\beta$, because $||\beta||^2 = 1$, we know that $\lambda = \beta^TY^TY\beta$=$(Y\beta)^T(Y\beta)$=$LSE$. Thus finish the proof. $\diamond$

Lemma 1 and Theorem 1 above provide with us an efficient shortcut to solve the extended linear regression fitting problem, From the original data, we fist generate the matrix Y, then, using SVD decomposition[10],makes $Y^TY = UDV^T$, the first column of the eigenvector matrix V is our desired solution. Normally, the SVD decomposition algorithm can deal with dimensionality not more than 8,000 easily, when dimensionality is larger than this, to compute the covariance matrix is too time and space consuming. However, in our condition, this is not a problem, because we only need the eigenvector which corresponding to the smallest eigenvalue, we do not need conduct the full SVD decomposition, but only compute the elements for the first column of the matrix V. this is quite computationally efficient. what's more, we will also use some subset selection technique to reduce the dimensionality.

4. EXPERIMENTAL RESULT

Dataset

The data are generated from Emili LAB in University of Toronto, A fragment of the style of the data is shown in table 1, more details can be found in [12]

Table 1. A fragment of the original data file

Protein ID	Peptide Sequence	Count	Charge
Q91VA7	$TRHNLVIIR$	4	2
Q91VA7	$KLDLVVHVK$	3	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$

The first column is the protein accession number−swissprot, the second column is the peptide sequences, we digest the proteins with trypsin, which cuts c-terminal to lysine (K) or arginine (R); it occasionally may cut at other locations or another protease may have nicked the proteins too; hence, we also often see partial tryptic (a K or R at the c-terminus, or flanking the peptide at the N-terminus). for the third column is the "sequence peak count" correlates rather well with the signal intensity (big peaks trigger the MS to fragment the same peptide many times, whereas most peptides are represented by smaller peaks which trigger the MS to fragment them once - hence, the values are all integers, and most are represented by a 1 count match. The last column represents the charge of the peptide ion.

Statistic Measurements and Experiment Design

To evaluate the goodness of fit, we developed two statistic measurement, more statistical details can be found in [9]:

Extended R-Square: which is just the classical R-square measurment[11] extended to be suitable for our extended regression model and is defined as:

$$Ext-R^2 = \frac{var(X\beta)}{var(Y_{ij}\beta/\sqrt{y_{i,j-1}^2+y_{ij}^2}) + var(X\beta)} \quad (11)$$

In which, the term $var(X\beta)$ represents the variance of the original data that can be explained by the model, while the term $var(Y_{ij}\beta/\sqrt{y_{i,j-1}^2+y_{ij}^2})$ represents the variance of the data originate by noise, $j = 2, .., k_i$.This statistic measures how successful the fit is in explaining the variation of the data. Put another way, the Extended R-square is the square of the correlation between the response values and the predicted response values.

Ratio-Least Square Error: which measures the deviation of the predicated peak intensity ratios from the truth value:

$$Ratio - LSE = \sum_{i=1}^{N}\sum_{j=2}^{k_i}(\frac{y_{i,j-1}}{y_{i,j}} - \frac{\hat{\beta}^T X_{i,j-1}}{\hat{\beta}^T X_{i,j}})^2 \quad (12)$$

For our condition, this statistic can reasonability handle the latent variable in_{ij} by ratio. Extrinsic property of a model can be reflected by the $Ratio - LSE$.

There are altogether 60,960 spectra, we chose 2,859 with high confidence($> 90\%$),non redundant matches to evaluate our method. There are two kinds of data in the whole dataset, the first part are those had been observed in MS/MS experiment, we call these "positive data", the other part contains the peptides which had not been observed in MS/MS experiment but generated by computer from the original protein sequence, we call these "negative data", all the high confidence data are positive data. the 2,859 data were divided into two parts− 1,800 as training data on which we fit the model and evaluate the goodness of fit by $Ext - R^2$, another 1,059 as testing data on which the fitted model were applied and the $Ratio - LSE$ is considered. For a comprehensive evaluation, we used 0-order linear model, 1-order linear model, 2-order linear model and 1-order hybrid model as well, for the 2-order model and hybrid model, because of the high dimensionality($> 8,000$), subset selection technique was used to reduce the dimensionality[11].

Implementation Consideration

We implemented our method only on method 1, which is a specific case for our theory. The implementation software is MATLAB, the hardware condition is a Pententium M 1.3GHz CPU and 512MB memory. The computation is quite efficient and all the algorithms can be finished in no more than 1.5 hours. The source code for this algorithm is attached in the appendix.

Result and Analysis

When applying different models, For the training dataset, the statistic measurements are shown in table 2, here, 0-order have 21 dimensions, 1-order have 421 dimensions, for 2-order linear model and 1-order hybrid model, we use subset selection technique, we select 66 dimensions for each model:

Table 2: comparison of statistics for training cases

-	$Ext - R^2$	$Ratio - LSE$
0-order linear model	0.6525	135.5828
1-order linear model	0.9991	5.3128
2-order linear model	1.0000	1.9689e-019
1-order hybrid model	0.9996	0.5312

From table 2, we see that for the 0-order linear model, the result is not quite good, the $Ext - R^2 = 0.6525$ means almost half of the variance generated by the noise and can not be fitted by the model−thus, the model is not flexible enough. While the result for the 1-order linear model is much better, from the $Ext - R^2$ we can see most of the variance are fitted by the model, while the training error

$Ratio-LSE$ remains small.The 2-order linear model gives out amazing fit to the training data, the $Ext-R^2=1.0000$ means all the variance in the original data can be explained perfectly by the model and the variance generated by the noise can be ignored and the $Ratio-LSE$ is almost 0, but in this case, the model is so flexible(dimensionality $>8,000$) that all the noises are also fitted as data, in fact, the 1-order hybrid model has the similar problem. After modeling the data, the fitted model was applied to the part of test data, the statistic measurement $Ratio-LSE$ for test data is listed in table 3:

Table 3: comparison of statistics for testing cases

$Ratio-LSE$ for 0-order linear	2.3958e+005
$Ratio-LSE$ for 1-order linear	137.8327
$Ratio-LSE$ for 2-order linear	2.9638e+024
$Ratio-LSE$ for 1-order hybrid	1.8732e+019

From table 3, we can see that the $Ratio-LSE$ for 0-order linear model is relatively high, and the $Ratio-LSE$ for the 2-order linear model and 1-order hybrid model are surprisingly high: The former is because of underfitting and the latter is because of overfitting. However, the result of 1-order linear model is quite reasonable, for this model, there're altogether 421 predictors which could resist overfitting while guarantee flexibility at the same time.

5. CONCLUSIONS AND FUTURE WORK

The main purpose of this work was to determine if the peak intensity of the MS/MS experiment could be predicated by some simple models. Based on the working hypothesis that peak intensity could be predicted using the amino acid occurrence frequency and order properties, through finding a correlation between these properties and the peak intensity. Elegant statistical approaches based on a family of extended linear regression models for modeling peak intensity distribution have been developed. Our method are evaluated by some statistic measurements on real world dataset, the result is promising.

In conclusion, the extended linear regression strategy provides a new experimental and analytical framework for systematic, in-depth investigation of the proteomes. One of the difficult of our work is, under our assumption,the ionization efficiency value can be negative, which has no biological meaning, to find a more elegant mechanism to guarantee the ionization efficiency value is in the interval$[0,1]$ would be very useful, we strongly believe that *sigmod* function is a good choice. Currently, we are developing a neural net based modeling framework for modeling MS/MS data, which will be the future work for our work.

Acknowledge

The authors would like to thank Dr. Clement Chung's hard work on generating the experimental data, and also Dr. Rafal Kustra for fruitful suggestions.

6. REFERENCES

[1] Pandey, A., Mann,M. *Protemics to study genes and genomes*, Nature 405, pp837-846, 2000

[2] Mann,M., Hendrickson, R.C., Pandey, A. *Analysis of proteomes by mass spectrometry*, Annu.Rew.Biochem.70, pp437-473, 2001

[3] Aebersold,R., Mann,M. *Mass spectrometry-based proteomics*, Nature 422, pp198-207, 2003

[4] Gay,S. ,Binz, P.A., Hochstrasser,D.F., Appel,R.D. *Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra*,Proteomics 2, pp1374-1391, 2002

[5] Eng,J., McCormack,A., Yates,J.R. *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database*,j.Am.Soc.Mass Spectrom.5,, pp976-989, 1994

[6] Kussmann, M., Nordhoff,E., Rahbek-Nielsen, H., Haebel,S.et al., Journal of Mass Spectrom. 32 pp593-601, 1997

[7] Purves, R. W., Gabryelski,W., Li,L.,Rapic Commun.Mass Spectrom.2, pp695-700, 1998

[8] Knochenmuss,R.,Wiesli,U.,Breuker,K.,Zenobi,R.,Rapic Commun.Mass Spectrom.12, pp529-534, 1998

[9] Han L., Anthony J.B., Anderew E., *Bridging the Gap between regression and PCA:an extended regression strategy and it's application in Tandem Mass Spectrometry*, submitted to IEEE/ACM Transactions on Bioinformatics and Computational Biology,2004

[10] Alter O, Brown PO, Botstein D. *singular value decomposition for genome-wide expression data processing and modeling*, Pro Natl Acad Sci USA 97, pp10101-10106 2000

[11] Trevor H., Robert T., Jerome F. *The elements of statistical learning- Data ming, inference and prediction*, Springer 2001.

[12] Thosma K., Khaled R.,Dragan Radulovic., *PRIM, a Generic Large Scale Proteomics investigation Strategy for Mammals*, Molecular& Cellular Proteomics 2.1 2003