Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, **2014**, *15*, 237-243

A Computational Algorithm for Functional Clustering of Proteome Dynamics During Development

Yaqun Wang[1], Ningtao Wang[1], Han Hao[1], Yunqian Guo[2], Yan Zhen[3], Jisen Shi[3,*] and Rongling Wu[1,*]

[1]*Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA;* [2]*Center for Computational Biology, Beijing Forestry University, Beijing 100083, China;* [3]*Key Laboratory of Forest Genetics and Biotechnology, Nanjing Forestry University, Nanjing 210037, China*

Abstract: Phenotypic traits, such as seed development, are a consequence of complex biochemical interactions among genes, proteins and metabolites, but the underlying mechanisms that operate in a coordinated and sequential manner remain elusive. Here, we address this issue by developing a computational algorithm to monitor proteome changes during the course of trait development. The algorithm is built within the mixture-model framework in which each mixture component is modeled by a specific group of proteins that display a similar temporal pattern of expression in trait development. A nonparametric approach based on Legendre orthogonal polynomials was used to fit dynamic changes of protein expression, increasing the power and flexibility of protein clustering. By analyzing a dataset of proteomic dynamics during early embryogenesis of the Chinese fir, the algorithm has successfully identified several distinct types of proteins that coordinate with each other to determine seed development in this forest tree commercially and environmentally important to China. The algorithm will find its immediate applications for the characterization of mechanistic underpinnings for any other biological processes in which protein abundance plays a key role.

Received on: September 30, 2013- Revised on: March 27, 2014- Accepted on: April 05, 2014

Keywords: Functional clustering, Unsupervised analysis, Dynamic proteomics, Seed development, Forest tree.

INTRODUCTION

Under long natural selection, the organism has evolved into a capacity to alter its form and function in facing rapid changes in the environment. At the cellular level, some fundamental aspects of this response are addressed by proteins that directly maintain the function of genes in the form of cellular building blocks via enzymatic catalysis, molecular signaling, and physical interactions. It is common that protein molecules are continuously synthesized and degraded in response to developmental signals during the organism's lifetime [1]. Proper turnovers of proteins have been thought to be essential for the normal development of a phenotypic trait [2, 3]. Several studies have used proteome dynamics, i.e., temporal changes in protein abundance, to understand the etiology of trait formation and development. In an anticancer drug delivery study, Cohen *et al.* discerned a difference in responding to a drug between seemingly identical cells based on dynamic changes of particular proteins [4]. The use of proteome dynamics to study phenotypic traits in plants has increasingly been interesting to plant geneticists [5-7].

The application of proteome dynamics relies critically upon the analysis and clustering of proteins expressed over time [8, 9]. Since the temporal pattern of protein expression profiles conforms to particular biological processes, the function of proteins can be revealed by clustering them into distinct groups. Measured usually as curves, the approaches for clustering and analyzing dynamic proteomics data are challenging, although the underlying statistical theory is not entirely new. In recent years, intensive efforts have been made to develop computational methods for cataloguing dynamic expression data including supervised approaches such as ML-KNN algorithm, [10] fuzzy KNN algorithm, [11-13] covariance discriminant algorithm, [14, 15] SLLE algorithm [16] and Random forest technique [17, 18]. Several unsupervised approaches developed to analyze time-course gene expression data have also been available [19, 20] and can be, in principle, used for statistical analysis of proteome dynamics. Of these approaches, one integrates mathematical aspects of response dynamics into a mixture model, allowing each mixture component to be represented by a cluster of genes [21, 22]. This approach, called functional clustering, translates the discrete measurements at multiple points to a continuous function of biological relevance. For example, Fourier series approximation was used to fit periodic profiles of expression by clock genes and cluster these genes into different groups in terms of their important dynamic features, such as mean magnitude, amplitude and period. An alternative approach for analyzing time-series genes by functional clustering is to consider response dynamics using non-parametric fitting, [23] thereby increasing the flexibility of the model.

Despite its usefulness, the capacity of functional clustering to cluster dynamic proteomes has not been assessed and

*Address correspondence to these authors at the Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA; Tel: (717) 531-2037; Fax: (717) 531-0480; E-mail: rwu@phs.psu.edu and Key Laboratory of Forest Genetics and Biotechnology, Nanjing Forestry University, Nanjing 210037, China; E-mail: jshi@njfu.edu.cn

1875-5488/14 $58.00+.00

validated in depth. Thus far, many researchers in the area of proteomic dynamics have not been aware of any powerful computational model for protein clustering. In addition, protein data have their unique feature, i.e., a protein may appear in a spectrum typically with thousands of peaks, leading to the number of samples largely smaller than the number of protein peaks. A variety of clustering approaches have been developed to handle such high-dimensional complexity of proteomics data [24-28]. The purpose of this article is to implement functional clustering into analysis and modeling of proteome dynamics. The procedure for analyzing and clustering dynamic changes of protein expression in a time course is described. The model was applied to reanalyze a dataset of proteins measured at discrete time points during seed development in a forest tree. By clustering protein profiles, the model allows fundamental and applied questions of proteomics, i.e., how different proteins are expressed in a coordinated and sequential manner to conform to trait development, to be tested and addressed.

MODEL

Likelihood

Suppose there are n proteins each measured at T time points during the growth of an organism. Let $y_i = (y_i(1), \ldots, y_i(T))$ denote the amounts of time-dependent expression for protein i. If these proteins are grouped into J clusters based on different patterns of their biological trajectories over time, this means that any one of proteins (say i) is assumed to arise from one (and only one) of the J possible clusters. Thus, the distribution of protein profile data is expressed as the J-component mixture probability density function, i.e.,

$$y_i \sim f(y_i; \omega, \boldsymbol{\mu}, \Sigma) = \sum_{j=1}^{J} \omega_j f_j(y_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad (1)$$

where $\omega = (\omega_1, \cdots, \omega_J)$ is a vector of mixture proportions which are non-negative and sum to unity; $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_J)$ contains the mean vector of cluster j; and $\boldsymbol{\Sigma}$ contains residual variances and covariances among T time points which are common for all clusters. The probability density function of cluster j, $f_j(y_i; \boldsymbol{\mu}_j, \Sigma)$, is assumed to be multivariate normally distributed with T-dimensional mean vector

$$\boldsymbol{\mu}_j = (\mu_j(1), \ldots, \mu_j(T) \quad (2)$$

and ($T \times T$) covariance matrix $\boldsymbol{\Sigma}$.

The likelihood based on a mixture model containing J clusters can be written as

$$L(\boldsymbol{\Theta}|y) = \prod_{i=1}^{n} \sum_{j=1}^{J} \left[\omega_j f_j(y_i; \boldsymbol{\mu}_j, \Sigma)\right], \quad (3)$$

where $\boldsymbol{\Theta}$ is a vector of unknown parameters including the mixture proportions, cluster-specific mean vectors, and covariance.

Different from traditional treatments, we will incorporate mathematical and statistical models to fit the mean-covariance structures. Thus, instead of estimating all elements in the vectors and covariance, we estimate the mathematical and statistical parameters that model the mean-covariance structures.

Structural Modeling of Mean Vectors and Covariance

The expression levels of many proteins have been found to vary in a time course [1]. For example, the abundance of proteins within the cycle of cell division may alter periodically, coincident with the cell cycle, aimed at sustaining a proper order during cell division or conserving limited resources. The oscillation of cell cycle-regulated proteins can be mathematically described by periodic Fourier functions or other periodic functions. Thus, by estimating the parameters that define the periodic curves for individual proteins, the differences in the temporal pattern of protein expression can be well determined.

For many proteins whose time-varying expression does not obey an explicit mathematical function, nonparametric approaches, such as B-spline, can be used. In this study, we propose a computation-efficient nonparametric approach based on Legendre orthogonal polynomials (LOP). Since the LOP are orthogonal and integrate to 0 in the interval [-1,1], they have been applied to nonparametric regression, [29] the resulting parameter estimates possessing favorable asymptotic properties [30, 31]. The LOP have also been used to model time-varying phenotypic or genetic variation for milk production [29] and plant growth traits [32, 33].

Let $P_r(t^*) = [P_0(t^*), P_1(t^*), \ldots, P_r(t^*)]$ denote a family of LOP with a particular order r derived from a special differential equation, where t^* is a scaled time with a range $[-1,1]$. Let $u_{jr} = [u_{j0}, u_{j1}, \ldots, u_{jr}]$ denote a vector of base values for cluster j. Then, time-varying mean values for cluster j in equation (2) can be expressed as a linear combination of u_{jr} weighted by the family of LOP, i.e.,

$$\mu_j(t^*) = P_r(t^*)u_{jr}. \quad (4)$$

Our task now is to estimate the base vector u_{jr} from the given data.

The longitudinal covariance among different time points has an inherent structure, which should be modeled for parsimonious parameter estimates. There are many different approaches for covariance structure, including stationary, non-stationary, nonparametric and semiparametric models. For a practical data set, there may exist an optimal approach for structural modeling of the covariance. Zimmerman and Núñez-Antón discussed the procedures and criteria for model selection in covariance structure [34]. These can be directly used in our model for functional clustering of gene expression dynamics.

To illustrate how the covariance is structured, we describe the non-stationary structured antedependence (SAD) model proposed by Zimmerman and Núñez-Antón [34]. The residual term for the SAD model can be expressed as

$$e = A\boldsymbol{\varepsilon}$$

where $e = (e(1), \cdots, e(T))'$ is the residual vector and $\varepsilon = (\varepsilon(1), \cdots, \varepsilon(T))'$ is the innovation error vector. For the first-order SAD model, we have

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \phi & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{pmatrix}$$

where ϕ is the antedependence parameter.

The residual variance-covariance matrix for e is then expressed as

$$\Sigma = AD_l A', \quad (\mathbf{5})$$

where D is the innovation variance-covariance matrix and is expressed as

$$D = \begin{pmatrix} \nu^2(1) & 0 & 0 & \cdots & 0 \\ 0 & \nu^2(2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \nu^2(T) \end{pmatrix},$$

which assumes that innovation errors are independent over different time points. The time-varying innovative variance $\nu^2(t)$ can be approached by a polynomial [35] or, for simplicity, is assumed as a constant ν^2. If $\nu^2(t)$ is assumed to be a constant, the residual matrix $\mathbf{\Sigma}$ contains the parameters, (ϕ, ν^2).

If a study includes protein expression data from multiple organs (say L), we can expand the SAD model to a multivariate case in which correlations among different organs are taken into account. Zhao *et al.* provided a general closed form for solving the determinant and inverse of the multivariate longitudinal covariance matrix [36]. This form can be directly incorporated into our dynamic model, facilitating the computing process of parameter estimation.

Estimation and Tests

A hybrid approach of Expectation-Maximization (EM) and simplex algorithms was implemented to estimate the parameters, $\mathbf{\Theta}$, contained in the likelihood (3). The EM algorithm provides a platform for estimating the proportions of different clusters, within which the simplex algorithm is embedded to estimate base vectors for each cluster and the covariance-structuring parameters. This can be described as follows:

In the E step, we define and estimate the posterior probabilities of protein i, with which it belongs to a particular expression pattern j, by

$$\Omega_{j|i} = \frac{\omega_j f_j(y_i;\boldsymbol{\mu}_j,\Sigma)}{\sum_{j'=1}^{J}[\omega_{j'} f_{j'}(y_i;\boldsymbol{\mu}_{j'},\Sigma)]}. \quad (\mathbf{6})$$

In the M step, the proportion of expression pattern j is calculated by

$$\omega_j = \frac{\sum_i^n = 1\, \Omega\ j|i}{n} \quad (\mathbf{7})$$

Mean-covariance structuring parameters in $\mathbf{\Theta}$ are estimated in this step, but no closed forms can be derived for their estimators. The simplex algorithm, which does not depend on explicit equations, is implemented to estimate these parameters.

Since both the actual number of protein expression patterns and an optimal order of LOP for expression pattern-specific mean fitting are unknown, we employ the commonly used model selection methods, Akaike information criterion (AIC) or Bayesian information criterion (BIC), to estimate these two parameters for a specific data set.

After these parameters are determined, we can formulate several biologically meaningful hypothesis tests. First, we need to determine the optimal number of expression patterns. This can be tested by

$$H_0: u_j = u_{j'} \text{ vs. } H_1: u_j \neq u_{j'}, \text{for } j < j' = 1, \ldots, J \quad (\mathbf{8})$$

If the H_0 is accepted for two given patterns, this means that the optimal number of patterns is $J - 1$. This approach allows the identification of the optimal number of expression patterns.

Second, we can test the significance of protein-protein interactions during development. This can be done by testing

$$H_0: u_j = cu_{j'} \text{ vs. } H_1: u_j \neq u_{j'}, \text{for } j < j' = 1, \ldots, J \quad (\mathbf{9})$$

where c is a constant. If the H_0 is accepted for two given patterns of proteins, j and j', this means that they display significant protein-protein interactions over time. This test provides a quantitative way to study the interplay between proteins and development.

WORKED EXAMPLE

Shi *et al.* reported dynamic profiles of proteins expressed in different stages of early seed development in Chinese fir, *Cunninghamia lanceolata* (Lamb.) Hook [37]. A two-dimensional difference gel electrophoresis approach was used to characterize differentially expressed proteins in developing embryos from seeds dissected from immature cones. Six important developmental phases of early embryogenesis are identified: the cleavage polyembryony-stage seed (stages 1, 2, and 3), dominant embryo-stage seed (stages 4), columnar embryo-stage seed (stage 5), and early cotyledonary-stage seed (stage 6). Protein abundance was measured with three replicates at each stage. The authors identified 136 proteins whose expression levels varied significantly ($P < 0.01$) during seed development. Substantial differences in expression dynamics of these proteins were observed over the six stages of early seed development.

We used the dynamic model to cluster these proteins into different groups in terms of biological functions during early embryogenesis using the mean values of three replicates. According to the BIC values calculated under different numbers of mixture components and different orders of LOP (Fig. **1**), we found that four components each fitted by a fourth order of LOP provide an optimal fitness of the dynamic data of proteomes. Four distinct groups of proteins display different temporal patterns of expression in a time course (Fig. **2**). Starting with a low level of expression at early cleavage polyembryony stages 1 and 2, group A increases its expression exponentially from late cleavage polyembryony stage 3 to early cotyledonary stage 6 through dominant embryo stage 4 and columnar embryo stage 5. From stage 5 to 6, this group of proteins displays the maximum amount of expression among all groups. Compared with group A, group B follows a similar temporal pattern, but with a lesser extent of time-dependent change. Although group C is also up-regulated over time, its slope of increase is much lower than groups A and B. Different from the other groups, group D is slightly down-regulated during seed development.

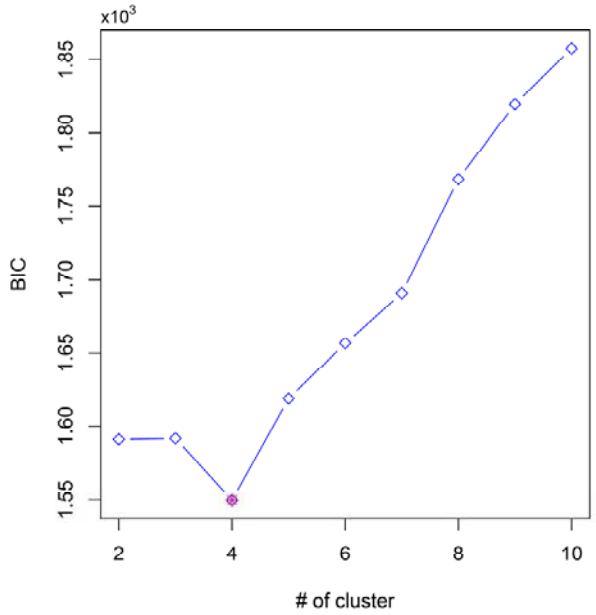

Fig. (1). Bayesian information criterion (BIC) plot that dertemines an optimal number of clusters and an optimal oredr of Legendre orthogonal polymonials by functional clustering.

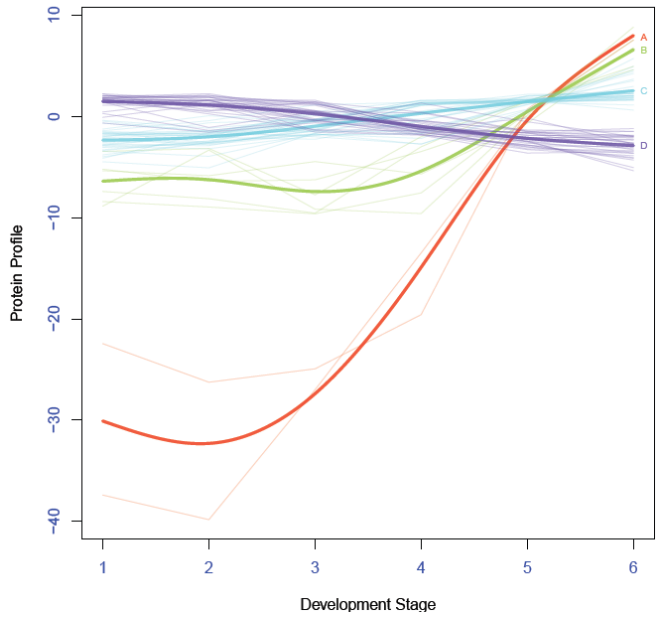

Fig. (2). Expression trajectories of four groups of proteins, labelled by **A**, **B**, **C** and **D**, during early embryogenesis. Light lines within each group are raw data of protein profiles.

In general, all groups of proteins appear to canalize during cleavage polyembryony stages, but alter their expression dramatically in dominant embryo, columnar embryo, and early cotyledonary stages. Because the latter three stages of embryogenesis are of paramount importance in determining variation in seed development, [37] these proteins can be used as biomarkers for explaining phenotypic changes of seeds in the Chinese fir. Especially, group A is highly associated with the rate of seed development from dominant embryo to early cotyledonary stages. There are high interactions among expression profiles of different groups, although the time at which a particular pair of groups triggers an interaction differ from pair to pair. For example, groups C and D generate a crossover in their expression trajectories at stage 3, whereas a crossover between groups A and B occurs at stage 5. From their mutual interactions observed, it is fairly possible that dynamic changes of protein-protein interactions are important determinants of seed development in the Chinese fir.

Specific proteins were identified for each of the four groups (see Supplementary Table **1**). Proteins in each group have a similar function in terms of the dynamic behavior of seed growth. For example, two proteins in group A, legumin-like storage protein and signal transduction-related protein GF14 nu, [37] are up-regulated with a great slope of increase during development, which were also observed in *Arabidopsis thaliana* and *Picea sitchensis* [38, 39]. We used hierarchical clustering to analyze the degree of similarity between different groups of proteins differentially expressed in six stages of seed development (Fig. **3**). Groups C and B are the most similar with each other, and both have a large distance to group D. All these three groups differ tremendously from group A because the latter displays an unusually high slope of increase in a long period of seed development.

DISCUSSION

With the recent advent of global genomic and proteomic approaches, it has been possible to understand important biological processes at a system-wide level. Given that these techniques have mainly focused on analyzing steady-state levels of mRNA or proteins under varying conditions, there is a pressing need to use expression data collected along time to study the intrinsic mechanisms for the dynamic change of phenotypic traits [1, 4, 40]. Here, we provide a computational model to address a fundamental property of temporal data resulting from their directed dependency along time through cluster analysis and functional smoothing. We have shown that the model described can be used to categorize protein profiles into different groups which may correlate with particular biological properties of trait development.

The clustering algorithm derived from the model considers the dependency of temporal observations and allows the number and the members of the clusters to be automatically identified. We implemented a nonparametric approach based on Legendre polynomials for functional smoothing of the dynamic changes of protein expression profiles over time and embedded it into a mixture model framework. A model selection approach is then used to select an optimal number of clusters and their respective members. The main merit of this model lies in its mathematical treatment of time-varying protein expression and the quantitative identification of the intrinsic machinery that govern the expression and degradation of proteins. As a similar application to gene clustering, [21, 23] functional clustering has been validated in terms of its statistical properties through computer simulation. Results from previous simulation and numerical studies suggest that the model is adequately powerful for identifying distinct clusters and characterizing the temporal expression pattern of each group in a time course.

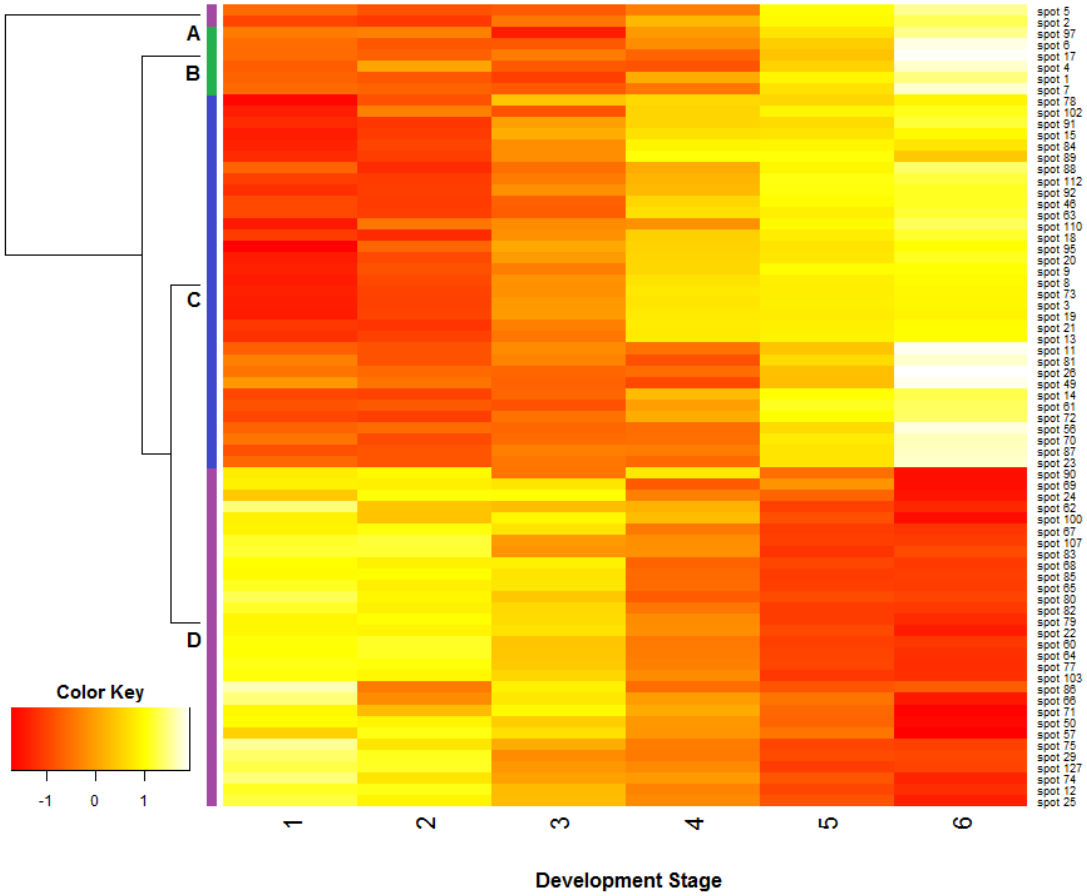

Fig. (3). Comparison of similarity and dissimilarity between four groups of proteins expressed over six stages of seed development. The groups are from our clustering results and indicated by vertical color bars. The names of proteins each labeled by a spot ID are given in (Supplementary Table **2**).

Unlike other dynamic data, proteomic data are often shown as mass spectrometry with a high-dimension. Wavelet-based approaches have proven to be powerful for dimension reduction of high-dimensional data and the extraction of fundamental information from raw data [21, 22]. These dimension reduction approaches can be modified to analyze and cluster high-dimensional protein profiles, although several statistical issues related to curve parameter estimation and longitudinal covariance modeling should be resolved [41]. In particular, when the number of proteins is largely smaller than the number of protein peaks, Bensmail *et al.* proposed an alternative hierarchical clustering algorithm based on a dissimilarity measure combined with a functional data analysis [28]. This alternative can also allow functional smoothing of proteomics expression profiles or spectra.

Further studies that combine our quantitative proteomic strategies with protein-protein and gene-protein coordination will gain new insights into a comprehensive picture of regulatory regulation and pathways involved in the formation of complex phenotypes. Also, there is considerable evidence that genetic variation influences gene and protein expression. In a genome-wide association study, a number of quantitative trait loci (QTLs) were found to influence levels of clinically relevant proteins in human serum and plasma [42]. Thus, by integrating it into a QTL mapping framework, [43] our clustering model will provide a general platform necessary to map the so-called protein QTLs or *p*QTLs that are involved in a sequence of biochemical pathways that cause final phenotypes.

Because of its dynamic features, our model can be modified to provide a vital means of predicting spatiotemporal expression patterns of proteins. The model is also powerful in integrating different types of omics data, allowing key regulatory elements, such as enhancers, to be identified [44]. In particular, a complete view of all of the enhancers that are active in a specific stage of development can be elucidated in a quantitative way.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors, [45] we shall make efforts in our future work to provide a web-server for the method presented in this paper.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This publication was supported by Beijing Forestry University Young Scientist Fund (Grant No. Blx2w8003), National Natural Science Foundation of China (Grant No. 31000287) NSF/IOS-0923975 and UL1 TR000127 from the National Center for Advancing Translational Sciences (NCATS). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

SUPPLEMENTARY MATERIALS

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Price, J.C.; Guan, S.; Burlingame, A.; Prusiner, S. B.; Ghaemmaghami, S. Analysis of proteome dynamics in the mouse brain. *Proc. Natl. Acad. Sci.*, **2010**, *107*, 14508-14513.

[2] Ivan, M.; Kondo, K.; Yang, H.; Kim, W.; Valiando, J.; Ohh, M.; Salic, A.; Asara, J.M.; Lane, W.S.; Kaelin, W.G. Jr. HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: Implications for O2 sensing. *Science*, **2001**, *292*, 464-468.

[3] Pratt, J.M.; Petty, J.; Riba-Garcia, I.; Robertson, D.H.; Gaskell, S.J.; Oliver, S.G.; Beynon, R.J. Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell Proteomics*, **2002**, *1*, 579-591.

[4] Cohen, A.A.; Geva-Zatorsky, N.; Eden, E.; Frenkel-Morgenstern, M.; Issaeva, I.; Sigal, A.; Milo, R.; Cohen-Saidon, C.; Liron, Y.; Kam, Z. Dynamic proteomics of individual cancer cells in response to a drug. *Science*, **2008**, *322*, 1511-1516.

[5] Kleffmann, T. Proteome dynamics during plastid differentiation in rice. *Ann. Rev. Plant Physiol.*, **2007**, *143*, 912.

[6] Elmore, J.M.; Liu, J.; Smith, B.; Phinney, B.; Coaker, G. Quantitative proteomics reveals dynamic changes in the plasma membrane during Arabidopsis immune signaling. *Mol. Cell. Proteomics*, **2012**, *11*(4), M111.014555.

[7] Mastrobuoni, G.; Irgang, S.; Pietzke, M.; Aßmus, H.E.; Wenzel, M.; Schulze, W.X.; Kempa, S. Proteome dynamics and early salt stress response of the photosynthetic organism Chlamydomonas reinhardtii. *BMC Genom.*, **2012**, *13*(1), 215.

[8] Frenkel-Morgenstern, M.; Cohen, A.A.; Geva-Zatorsky, N.; Eden, E.; Prilusky, J.; Issaeva, I.; Sigal, A.; Cohen-Saidon, C.; Liron, Y.; Cohen, L.; Danon, T.; Perzov, N.; Alon, U. Dynamic Proteomics: a database for dynamics and localizations of endogenous fluorescently-tagged proteins in living human cells. *Nucleic Acids Res.*, **2010**, *38*(Database issue), D508-12.

[9] Lee, M.V.; Topper, S.E.; Hubler, S.L.; Hose, J.; Wenger, C.D.; Coon, J.J.; Gasch, A.P. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol. Syst. Biol.*, **2011**, *7*, 514.

[10] Chou, K.C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Mol. Biosys.*, **2013**, *9*, 1092-1100.

[11] Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS ONE*, **2013**, *8*, e72234.

[12] Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theoretical Biol.*, **2013**, *337C*, 71-79.

[13] Xiao, X.; Wang, P.; Lin, W.Z.; Jia, J.H.; Chou, K.C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **2013**, *436*, 168-177.

[14] Chou, K.C.; Elrod D.W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.*, **1998**, *252*, 63-68.

[15] Xiao, X.; Wang, P.; Chou, K.C. iNR-PhysChem: A Sequence-Based Predictor for Identifying Nuclear Receptors and Their Subfamilies via Physical-Chemical Property Matrix. *PLoS ONE*, **2012**, *7*, e30869.

[16] Wang, M.; Yang, J.; Xu, Z.J.; Chou, K.C. SLLE for predicting membrane protein types. *J. Theoretical Biol.*, **2005**, *232*, 7-15.

[17] Kandaswamy, K.K.; Chou, K.C.; Martinetz, T.; Moller, S.; Suganthan, P.N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theoretical Biol.*, **2011**, *270*, 56-62.

[18] Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE*, **2011**, *6*, e24756.

[19] Sivriver, J.; Habib, N.; Friedman, N. An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinform.*, **2011**, *27*(13), i392-400.

[20] Bar-Joseph, Z.; Gitter, A.; Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **2012**, *13*(8), 552-564.

[21] Kim, B-R.; Zhang, L.; Berg, A.; Fan, J.; Wu, R. A computational approach to the functional clustering of periodic gene expression profiles. *Genetics*, **2008**, *180*, 821-834.

[22] Kim, B-R.; McMurry, T.; Zhao, W.; Berg, A.; Wu, R. Wavelet-based functional clustering for high-dimensional dynamic gene expression patterns. *J. Comp. Biol.*, **2010**, *17*, 1067-1080.

[23] Wang, Y.; Xu, M.; Wang, Z.; Tao, M.; Wang, L.; Zhu, J.; Li, R.; Berceli, S.A.; Wu, R. How to cluster gene expression dynamics in response to environmental signals. *Brief. Bioinform.*, **2012**, *13*, 162-174.

[24] Vazquez, A.; Flammini, A.; Maritan, A.; Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, **2003**, *21*(1), 697-700.

[25] Bensmail, H.; Haoudi, A. Postgenomics: proteomics and bioinformatics in cancer research. *J. Biomed. Biotechnol.*, **2003**, *4*, 217-230.

[26] Somorjai, R.L.; Dolenko, B.; Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinform.*, **2003**, *19*, 1484-1491.

[27] Schwartz, S.A.; Weil, R.J.; Johnson, M.D.; Toms, S.A.; Caprioli, R.M. Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin. Cancer. Res.*, **2004**, *10*(3), 981-987.

[28] Bensmail, H.; Aruna, B.; Semmes, O.J.; Haoudi, A. Functional clustering algorithm for high-dimensional proteomics data. *J. Biomed. Biotech.*, **2005**, *2*, 80-86.

[29] Meyer, K. Random regressions to model phenotypic variation in monthly weights of Australian beef cows. *Livest. Prod. Sci.*, **2000**, *65*, 19-38.

[30] Huskova, M.; Sen, P.K. On sequentially adaptive asymptotically efficient rank statistics. *Seq. Anal.* **1985**, *4*, 125-151.

[31] Mackay, M.D. Non-parametric variance based methods for assessing uncertainty importance. *Reliab. Eng. Syst. Saf.*, **1997**, *57*, 267-279.

[32] Cui, Y.; Zhu, J.; Wu, R. Functional mapping for genetic control of programmed cell death. Physiol. *Genomics*, **2006**, *25*, 458-469.

[33] Lin, M.; Wu, R. A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events. *BMC Bioinform.*, **2006**, *7*(1), 138.

[34] Zimmerman, D.; Núñez-Antón. Parametric modelling of growth curve data: an overview (with discussions). *Test*, **2001**, *10*, 1-73.

[35] Pourahmadi, M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **1999**, *86*, 667-690.

[36] Zhao, W.; Chen, Y.Q.; Casella, G.; Cheverud, J.M.; Wu, R. A non-stationary model for functional mapping of complex traits. *Bioinform*, **2005**, *21*, 2469-2477.

[37] Shi, J.; Zhen, Y.; Zheng, R. Proteome profiling of early seed development in Cunninghamia lanceolata (Lamb.) Hook. *J. Exp. Bot.*, **2010**, *61*, 2367-2381.

[38] Bevan, M.; Bancroft, I.; Bent, E. *et al.* Analysis of 19 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. *Nature*, **1998**, *391*, 485-488.

[39] Filonova, L.H.; Bozhkov, P.V.; Brukhin, V.B.; Daniel, G.; Zhivotovsky, B.; von Arnold, S. Two waves of programmed cell death occur during formation and development of somatic embryos in the gymnosperm, Norway spruce. *J. Cell Sci.*, **2000**, *113*, 4399-4411.

[40] Flintoft, L. Gene expression: Predictions across space and time. *Nat. Rev. Genet.*, **2012**, *14*, 78-79. online 27 December.

[41] Yap, J; Fan, J; Wu, R. Nonparametric modeling of covariance structure in functional mapping of quantitative trait loci. *Biometrics,* **2009**, *65*, 1068-1077.

[42] Melzer, D.; Perry, J.R.; Hernandez, D.; Corsi, A.M.; Stevens, K.; Rafferty, I.; Lauretani, F.; Murray, A.; Gibbs, J.R.; Paolisso, G. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.,* **2008**, *4*(5), e1000072

[43] Wu, R.; Ma, C-X.; Casella, G. *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*; Springer-Verlag: New York, **2007**.

[44] Wilczynski, B.; Liu, Y.H.; Yeo, Z. X.; Furlong, Eileen E. M. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput. Biol.,* **2012**, *8*, e1002798.

[45] Chou, K.C.; Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.,* **2009**, *2*, 63-92