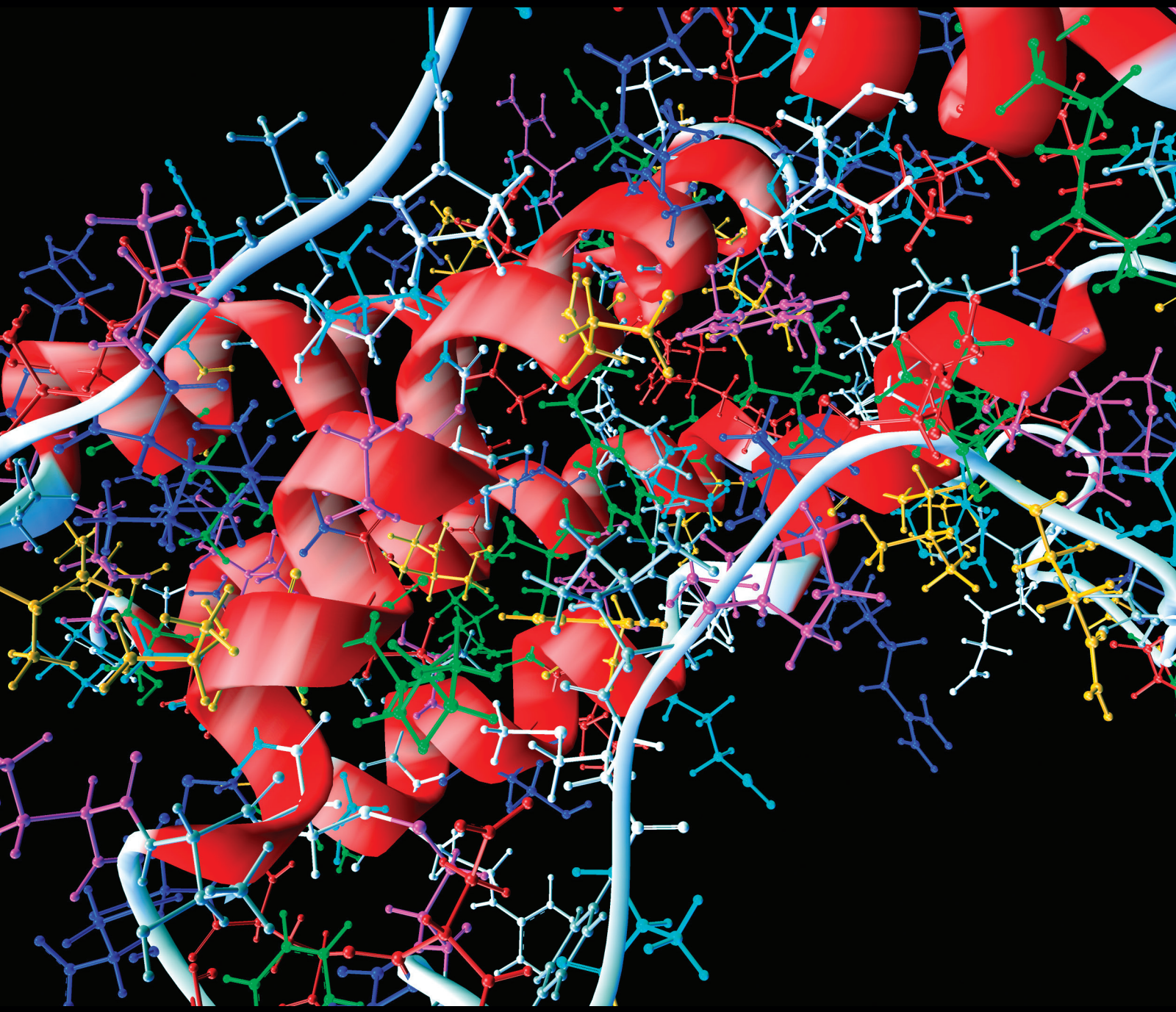
Computational and Mathematical Methods in Medicine

Advances in Computational Methods for Genetic Diseases

Guest Editors: Francesco Camastra, Roberto Amato, Maria Donata Di Taranto, and Antonino Staiano



Advances in Computational Methods for Genetic Diseases

Computational and Mathematical Methods in Medicine

Advances in Computational Methods for Genetic Diseases

Guest Editors: Francesco Camastra, Roberto Amato, Maria Donata Di Taranto, and Antonino Staiano

Editorial Board

Contents

Advances in Computational Methods for Genetic Diseases, Francesco Camastra, Roberto Amato, Maria Donata Di Taranto, and Antonino Staiano
Volume 2015, Article ID 645649, 2 pages

Enhancing the Lasso Approach for Developing a Survival Prediction Model Based on Gene Expression Data, Shuhei Kaneko, Akihiro Hirakawa, and Chikuma Hamada
Volume 2015, Article ID 259474, 7 pages

A New Approach for Mining Order-Preserving Submatrices Based on All Common Subsequences, Yun Xue, Zhengling Liao, Meihang Li, Jie Luo, Qiuhua Kuang, Xiaohui Hu, and Tiechen Li
Volume 2015, Article ID 680434, 11 pages

Statistical and Computational Methods for Genetic Diseases: An Overview, Francesco Camastra, Maria Donata Di Taranto, and Antonino Staiano
Volume 2015, Article ID 954598, 8 pages

Optimization and Corroboration of the Regulatory Pathway of p42.3 Protein in the Pathogenesis of Gastric Carcinoma, Yibin Hao, Tianli Fan, and Kejun Nan
Volume 2015, Article ID 683679, 9 pages

Unified Modeling of Familial Mediterranean Fever and Cryopyrin Associated Periodic Syndromes, Yasemin Bozkurt, Alper Demir, Burak Erman, and Ahmet Gül
Volume 2015, Article ID 893507, 18 pages

Evolutionary Influenced Interaction Pattern as Indicator for the Investigation of Natural Variants Causing Nephrogenic Diabetes Insipidus, Steffen Grunert and Dirk Labudde
Volume 2015, Article ID 641393, 6 pages

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 645649, 2 pages
http://dx.doi.org/10.1155/2015/645649

Editorial

Advances in Computational Methods for Genetic Diseases

Francesco Camastra,[1] Roberto Amato,[2] Maria Donata Di Taranto,[3] and Antonino Staiano[1]

[1]*Department of Science and Technology, University of Naples Parthenope, Centro Direzionale, Isola C4, 80143 Napoli, Italy*
[2]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK*
[3]*IRCCS SDN, Via E. Gianturco 113, 80143 Napoli, Italy*

Correspondence should be addressed to Francesco Camastra; francesco.camastra@uniparthenope.it

Received 10 May 2015; Accepted 10 May 2015

Genetic diseases are a wide group of diseases in which the etiopathogenesis is caused by or related to genetic factors. The role of genetics in the disease development can be more or less relevant depending on the specific characteristics of the disease, and a wide spectrum of complexity exists.

Monogenic diseases, for example, are directly caused by defects in a specific gene whereas complex and polygenic diseases are generally caused by the interactions between multiple genes or between genetic and environmental factors. To the last category belong many forms of cancer, an uncontrolled growth of cells with alterations of the genetic materials.

In the last decade, a large amount of experimental data has become available, so the identification of strategies to process and, most importantly, interpret them is crucial. The massive volume of data, both in terms of quantity and of dimensionality, and their heterogeneity and low signal-to-noise ratio are just some of the most obvious challenges that they present. To give an example, single nucleotide DNA mutations are one of the most common factors analysed in relation to the development of a genetic disease. However, this sometimes translates into dealing with millions of variants measured across thousands of individuals, where only a handful are informative. In fact, other more complex factors, such as gene expression, could play a significant role.

The aim of this special issue is to review the recent advances in computational methods concerned with genetic diseases.

The issue received sixteen submissions; each one was referred by at least two international reviewers that we warmly thank for their time. Six papers have been accepted for the publication.

"A New Approach for Mining Order-Preserving Submatrices Based on all Common Subsequences" by Y. Xue et al. proposes, in the context of gene expression data, a pattern-based subspace clustering or OPSM (order-preserving submatrix model), based on frequent sequential pattern. The approach has been experimentally proven to be able to discover the biological significant OPSMs and deep OPSMs exhaustively.

"Evolutionary Influenced Interaction Pattern as Indicator for the Investigation of Natural Variants Causing Nephrogenic Diabetes Insipidus" by S. Grunert and D. Labudde is devoted to the application of a high-throughput analysis method based on motif conservation among proteins of the same protein family for analysis of interacting sequences. This investigation can help to analyze the pathogenic impact of mutations causing alterations in interacting regions of a protein. This analysis has been applied on membrane proteins, in particular to the aquaporin 2 whose mutants are involved in nephrogenic diabetes insipidus.

"Unified Modeling of Familial Mediterranean Fever and Cryopyrin Associated Periodic Syndromes" by Y. Bozkurt et al. describes a unifying dynamical model for Familial Mediterranean Fever (FMF) and Cryopyrin Associated Periodic Syndromes (CAPS) in the form of coupled nonlinear ordinary differential equations. The authors perform a comprehensive bifurcation analysis of the model and show that it exhibits three modes, capturing the healthy, FMF, and CAPS cases. They present extensive simulation results for the model that match clinical observations.

"Enhancing the Lasso Approach for Developing a Survival Prediction Model Based on Gene Expression Data" by S. Kaneko et al. presents a novel improvement to the lasso

approach, one of the most widely used method to correlate gene expression data with cancer patients' survival. This new algorithm significantly increases the ability to identify "true positives" and its validity is shown on both simulated and real data.

"Statistical and Computational Methods for Genetic Diseases: An Overview" by Francesco Camastra, Maria Donata Di Taranto, and Antonino Staiano gives a survey of statistical and computational methods used to analyse the pathogenic role of sequence variants as well as to identify genetic markers of complex diseases by association studies, meta-analysis, and expression studies.

"Optimization and Corroboration of the Regulatory Pathway of p42.3 Protein in the Pathogenesis of Gastric Carcinoma" by Y. Hao et al. provides important research directions for exploring the mechanism of action of p42.3 protein in gastric cancer. Through a Bayesian network model, the potential important role of p42.3 is verified by both theoretical analysis and preliminary test.

We hope that the readers of this journal will find in the issue interesting papers and that this can encourage and foster further research on computational methods for genetic diseases.

Francesco Camastra
Roberto Amato
Maria Donata Di Taranto
Antonino Staiano

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 259474, 7 pages
http://dx.doi.org/10.1155/2015/259474

Research Article

Enhancing the Lasso Approach for Developing a Survival Prediction Model Based on Gene Expression Data

Shuhei Kaneko,[1] Akihiro Hirakawa,[2] and Chikuma Hamada[1]

[1]*Department of Management Science, Graduate School of Engineering, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan*
[2]*Biostatistics and Bioinformatics Section, Center for Advanced Medicine and Clinical Research, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan*

Correspondence should be addressed to Shuhei Kaneko; skaneko.mobile0724@gmail.com

Received 18 September 2014; Accepted 22 December 2014

Academic Editor: Roberto Amato

In the past decade, researchers in oncology have sought to develop survival prediction models using gene expression data. The least absolute shrinkage and selection operator (lasso) has been widely used to select genes that truly correlated with a patient's survival. The lasso selects genes for prediction by shrinking a large number of coefficients of the candidate genes towards zero based on a tuning parameter that is often determined by a cross-validation (CV). However, this method can pass over (or fail to identify) true positive genes (i.e., it identifies false negatives) in certain instances, because the lasso tends to favor the development of a simple prediction model. Here, we attempt to monitor the identification of false negatives by developing a method for estimating the number of true positive (TP) genes for a series of values of a tuning parameter that assumes a mixture distribution for the lasso estimates. Using our developed method, we performed a simulation study to examine its precision in estimating the number of TP genes. Additionally, we applied our method to a real gene expression dataset and found that it was able to identify genes correlated with survival that a CV method was unable to detect.

1. Introduction

In the past decade, researchers have predicted survival in a cancer patient based on gene expression data [1–4]. Revealing the relationship between gene expression profiles and the time to an event of interest (e.g., overall survival, metastasis-free survival) can improve treatment strategies and establish accurate prognostic markers. The Cox proportional hazard model is the most popular method for relating covariates to survival times [5]. However, due to the high dimensionality of gene expression data (i.e., the number of genes expressed exceeds the number of patients), it is not possible to take an estimation approach based on the Cox log partial likelihood. To overcome this problem, a penalized estimation approach, which includes a shrinkage estimation of coefficients, is frequently taken [6–8].

In penalized estimation approaches, the least absolute shrinkage and selection operator (lasso) [9, 10] is often used because of its attractive ability to simultaneously select the genes correlated with survival and estimate the coefficients in the Cox model. The lasso shrinks most of the coefficients towards zero exactly by adding L_1 norm to the Cox log partial likelihood, and the amount of shrinkage is dependent on the tuning parameter. The value of the tuning parameter is often determined by a cross-validation (CV), which maximizes the out-of-data prediction accuracy [11].

Several researchers have investigated the operating characteristics of the lasso. Goeman [12] used the lasso to analyze a publicly available gene expression dataset, obtained from the articles of van't Veer et al. [2] and van de Vijver et al. [3] in which a 70-gene signature for prediction of metastasis-free survival in breast cancer patients had been established. This data included 295 patients with 4919 genes that were prescreened from 24,885 genes based on the quality criteria in van't Veer et al.'s work [2]. The lasso selected 16 genes with which to develop a prediction model of overall survival when using the tuning parameter that was determined using a CV. Goeman [12] also conducted ridge regression using all 4919

genes to develop a model by adding L_2 norm to the Cox log partial likelihood. The prediction accuracy of the lasso and ridge regression were compared, and the ridge regression with 4919 genes slightly outperformed the lasso with 16 genes. Goeman [12] concluded that the lasso potentially passes over genes that are correlated with survival in order to develop a simple prediction model. Bøvelstad et al. [7] reached the same conclusion in a review of the survival prediction methods available for analyzing breast cancer gene expression datasets. Table 1 summarizes a typical result of gene selection by the lasso.

The CV method determines the value of the tuning parameter by considering the trade-off between the number of true positives (TP) and false positives (FP), and so the possibility of identifying false negatives (FN) cannot be eliminated. One solution for identifying more outcome-predictive genes is to monitor the number of TP in several values of the tuning parameter and, subsequently, determine its final value. In this study, we developed a method for estimating the number of TP for a series of values of the tuning parameter. We assumed a mixture distribution with components of TP and FP for the lasso estimates, and these could be used to estimate the number of TP and FP. It is possible to generate the solution path that includes the lasso estimates for a series of values of the tuning parameter using the methods developed by Goeman [12]. Here, we proposed an algorithm to sequentially fit the mixture distribution for this solution path, and we used a simulation study to test the precision of the algorithm when estimating the number of TP. We further demonstrated the proposed algorithm using a well-known diffuse large B-cell lymphoma (DLBCL) dataset comprising overall survival of 240 DLBCL patients and gene expression data of 7399 genes [1].

2. Materials and Methods

2.1. Lasso in the Cox Proportional Hazard Model. The Cox proportional hazard model is the most popular method for evaluating the relationship between gene expression and time to an event of interest [5]. The hazard function of an event at time t for a patient i ($i = 1, \ldots, n$) with the gene expression levels $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is given by

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a parameter vector and $h_0(t)$ is the baseline hazard, which is the hazard for the respective individual when all variable values are equal to zero. In the general setting where $n > p$, the coefficients are estimated by maximizing Cox log partial likelihood as follows:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log \left\{ \sum_{r \in R(t_i)} \exp\left(\mathbf{x}_r^{\mathrm{T}} \boldsymbol{\beta}\right) \right\} \right], \tag{2}$$

where δ_i is an indicator, which is 1, if the survival time is observed, or 0, if censored. $R(t_i)$ is the risk set of the individuals at t_i.

Table 1: Typical results of gene selection by the lasso.

True condition	The lasso	
	Select	No select
Genes that are not correlated with survival (none-outcome-predictive genes)	False positive (FP)	True negative (TN)
Genes that are truly correlated with survival (outcome-predictive genes)	True positive (TP)	False negative (FN)

In the lasso for the high-dimensional setting where $n < p$, the coefficients are estimated by maximizing the following penalized likelihood function [9, 10]:

$$l_p(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \left|\beta_j\right|, \tag{3}$$

where λ is the tuning parameter, which determines the amount of shrinkage.

2.2. Solution Path of the Lasso Estimates. Goeman [12] introduced a method to calculate the solution path of the lasso estimates as a function of λ, $\widehat{\boldsymbol{\beta}}(\lambda)$, which is based on the algorithm developed by Park and Hastie [13]. The method maximizes $l_p(\boldsymbol{\beta}, \lambda)$ at a fixed λ based on a combination of gradient ascent optimization with the Newton-Raphson algorithm. $\widehat{\boldsymbol{\beta}}(\lambda)$ are calculated for $\lambda_0 > \cdots > \lambda_k > \cdots > \lambda_z > 0$ successively, starting from $\lambda_0 = \max_j \partial l / \partial \beta_j|_{\beta_j = 0}$ (which gives $\widehat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$ because the value has zero gradients). λ_z is chosen arbitrarily but is often set to $0.05 \times \lambda_0$ in analyses of gene expression data [14]. The lasso estimates at a current step are set to initial values for calculation of the subsequent step. Step length $\Delta_k = \lambda_k - \lambda_{k+1}$ is the minimum decrement to change the number of selected genes $m^{(k)}$ ($= \#\{j; \widehat{\beta}_j(\lambda_k) \neq 0\}$); that is, only one gene is newly selected or excluded from λ_k to λ_{k+1}.

2.3. Mixture Distribution for Estimating the Number of TP in the Lasso Estimates. To estimate the number of TP in the lasso estimates at a fixed value of λ, we assumed a mixture distribution developed in our previous study [15]. We introduced the mixture distribution based on the two features of the lasso: (i) the lasso selects at most n genes because of the nature of the convex optimization problem when $n < p$ [16, 17] and (ii) in the Bayesian paradigm the lasso estimates are the posterior mode with the independent Laplace prior distribution $f_L(\beta_j; 0, 1/\tau) = (\tau/2) \exp(-\tau|\beta_j|)$, where $f_L(y; a, b) = 1/2b \exp(-|y - a|/b)$ is the probability density function of Laplace distribution with location parameter a

and scale parameter b [9]. Therefore, the mixture distribution assumed for the lasso estimates at λ was as follows:

$$f\left(\widehat{\beta}_j(\lambda);\pi_0,\pi_c,\tau,\mu_c,\sigma_c\right) = \frac{n}{p}\left\{\pi_0 f_L\left(\widehat{\beta}_j(\lambda);0,\frac{1}{\tau}\right)+\sum_{c=1}^{C}\pi_c f_N\left(\widehat{\beta}_j(\lambda);\mu_c,\sigma_c^2\right)\right\} + \left(1-\frac{n}{p}\right) f_L\left(\widehat{\beta}_j(\lambda);0,\epsilon\right), \tag{4}$$

where π_0 and π_c are mixed proportions ($\pi_0 + \sum_{c=1}^{C}\pi_c = 1$); $f_N(\widehat{\beta}_j(\lambda);\mu_c,\sigma_c^2)$ is the probability density function of the normal distribution with mean μ_c ($\neq 0$) and variance σ_c^2 in component c; C is the number of components, which is determined by model selection criteria; and ϵ is the constant value, which is boundlessly close to 0; for example, $\epsilon = 10^{-8}$. The unknown parameters, π_0, π_c, τ, μ_c, and σ_c, are estimated by maximizing the log-likelihood function of (4) by using the Newton-Raphson method.

The mixture distribution defined in (4) is formulated on the basis of the following concepts: since the lasso selects a maximum of n genes when $p > n$, the coefficients for $p-n$ genes are exactly zero; therefore, (4) consists of 2 terms (n/p term and $1-n/p$ term). In the n/p term, the Laplace distribution with location parameter 0 and scale parameter $1/\tau$ was assumed to be the distribution for the FP on the basis of the lasso feature (ii) discussed above, while the C component normal distribution with location parameter μ_c and scale parameter σ_c^2 was assumed as the distribution for the TP. In the $1-n/p$ term, the Laplace distribution with location parameter 0 and scale parameter ϵ was assumed as the distribution of $p-n$ genes based on the aforementioned lasso feature (i).

The f_L with location parameter 0 and scale parameter $1/\tau$ was assumed to be the distribution for the FP on the basis of lasso feature (i), discussed above. The f_N with location parameter μ_c and scale parameter σ_c^2 was assumed as the distribution for the TP. The f_L of the $(1-n/p)$ term was assumed as the distribution of $p-n$ genes based on the aforementioned lasso feature (ii). Given a cut-off value ζ (>0), the estimated proportions of the FP and TP are the area under the estimated Laplace and normal distribution in the n/p term of (4), respectively, and can be written as follows:

$$\widehat{P}_{\mathrm{FP}} = \widehat{\pi}_0\left[\int_{-\infty}^{-\zeta} f_L\left(u;0,\widehat{\tau}^{-1}\right)du + \int_{\zeta}^{+\infty} f_L\left(u;0,\widehat{\tau}^{-1}\right)du\right],$$

$$\widehat{P}_{\mathrm{TP}} = \sum_{c=1}^{C}\widehat{\pi}_c\left[\int_{-\infty}^{-\zeta} f_N\left(u;\widehat{\mu}_c,\widehat{\sigma}_c^2\right)du + \int_{\zeta}^{+\infty} f_N\left(u;\widehat{\mu}_c,\widehat{\sigma}_c^2\right)du\right]. \tag{5}$$

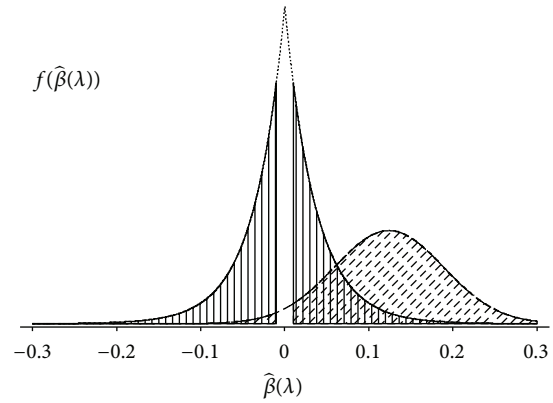


Figure 1: Illustration for estimating the number of FP and TP. The areas denoted by the vertical and diagonal lines are the proportion of FP and TP, respectively.

Figure 1 illustrates the calculation in (5) when the number of components, C, is 1. Using (5), the number of TP and FP was estimated by

$$\widehat{\mathrm{FP}} = \frac{\widehat{P}_{\mathrm{FP}}}{\widehat{P}_{\mathrm{TP}}+\widehat{P}_{\mathrm{FP}}}\times m, \tag{6}$$

$$\widehat{\mathrm{TP}} = \frac{\widehat{P}_{\mathrm{TP}}}{\widehat{P}_{\mathrm{TP}}+\widehat{P}_{\mathrm{FP}}}\times m. \tag{7}$$

2.4. Algorithm for Estimating Number of TP in a Series of Values λ. Here, we propose an algorithm to sequentially fit the mixture distribution in (4) to the solution path of the lasso estimates, which was described in Section 2.2. In this algorithm, we assumed that the number of TP changed when the newly selected or excluded gene from λ_k to λ_{k+1} was truly correlated to survival, based on the maximum log-likelihood of (4). First, we approximated $\widehat{P}_{\mathrm{FP}} \approx \widehat{\pi}_0$ and $\widehat{P}_{\mathrm{TP}} \approx \sum_{c=1}^{C}\widehat{\pi}_c$ in (5) by assuming a suitably small cut-off value ζ (≈ 0). We then obtained $\widehat{\pi}_0 = \widehat{\mathrm{FP}}/m$ and $\widehat{\pi}_c = \widehat{\mathrm{TP}}_c/m$ ($c = 1,\ldots,C$) from (6) and (7), respectively, where $\widehat{\mathrm{TP}}_c$ is an estimate of the number of TP in component c. For $k = 1,\ldots,z$, the proposed algorithm was as follows.

Step 1

Step 1.1. In this step, we assumed that the newly selected or excluded gene from λ_k to λ_{k+1} was FP. π_0 denotes the proportion of FP and is set as

$$\pi_0^{(k+1)} = \begin{cases} \dfrac{\widehat{\mathrm{FP}}^{(k)}+1}{m^{(k+1)}}, & \text{if } m^{(k+1)} = m^{(k)}+1, \\ \dfrac{\widehat{\mathrm{FP}}^{(k)}-1}{m^{(k+1)}}, & \text{if } m^{(k+1)} = m^{(k)}-1. \end{cases} \tag{8}$$

For the other components, c ($c = 1,\ldots,C$), set $\pi_c^{(k+1)} = \widehat{\mathrm{TP}}_c^{(k)}/m^{(k+1)}$.

Step 1.2. Given $\widehat{\boldsymbol{\beta}}(\lambda_{k+1})$ and $\pi_0^{(k+1)}, \ldots, \pi_C^{(k+1)}$, calculate the maximum log-likelihood of (4), $L_0^{(k+1)}$.

Step 2

Step 2.1. Set $c = 1$.

Step 2.2. In this step, we assumed that the newly selected or excluded gene from λ_k to λ_{k+1} was TP. For the component c, set

$$\pi_c^{(k+1)} = \begin{cases} \dfrac{\widehat{\mathrm{TP}}_c^{(k)} + 1}{m^{(k+1)}}, & \text{if } m^{(k+1)} = m^{(k)} + 1, \\ \dfrac{\widehat{\mathrm{TP}}_c^{(k)} - 1}{m^{(k+1)}}, & \text{if } m^{(k+1)} = m^{(k)} - 1. \end{cases} \quad (9)$$

For the other components, set $\pi_0^{(k+1)} = \widehat{\mathrm{FP}}^{(k)}/m^{(k+1)}$ and $\pi_d^{(k+1)} = \widehat{\mathrm{TP}}_d^{(k)}/m^{(k+1)}$ $(d = 1, \ldots, C;\ d \neq c)$.

Step 2.3. Given $\widehat{\boldsymbol{\beta}}(\lambda_{k+1})$ and $\pi_0^{(k+1)}, \ldots, \pi_C^{(k+1)}$, calculate the maximum log-likelihood of (4), $L_c^{(k+1)}$.

Step 2.4. Set $c = c + 1$. Repeat Steps 2.2 and 2.3 until $c = C$.

Step 3. In this step, we determined whether the newly selected or excluded gene from λ_k to λ_{k+1} was TP or FP based on the maximum log-likelihood which was calculated in Steps 1.2 and 2.3. If $L_0^{(k+1)}$ was the largest in $L_c^{(k+1)}$ $(c = 0, \ldots, C)$, we assumed that the newly selected or excluded gene was FP; if not, we assumed that it was TP. Therefore, calculate $C_{\max} = \mathrm{argmax}_{c \in \{0,1,\ldots,C\}} L_c^{(k+1)}$. If $C_{\max} = 0$, update $\widehat{\mathrm{FP}}^{(k)}$ as follows:

$$\widehat{\mathrm{FP}}^{(k+1)} = \begin{cases} \widehat{\mathrm{FP}}^{(k)} + 1, & \text{if } m^{(k+1)} = m^{(k)} + 1, \\ \widehat{\mathrm{FP}}^{(k)} - 1, & \text{if } m^{(k+1)} = m^{(k)} - 1. \end{cases} \quad (10)$$

If $C_{\max} > 0$, update $\widehat{\mathrm{TP}}_{C_{\max}}^{(k)}$ as follows:

$$\widehat{\mathrm{TP}}_{C_{\max}}^{(k+1)} = \begin{cases} \widehat{\mathrm{TP}}_{C_{\max}}^{(k)} + 1, & \text{if } m^{(k+1)} = m^{(k)} + 1, \\ \widehat{\mathrm{TP}}_{C_{\max}}^{(k)} - 1, & \text{if } m^{(k+1)} = m^{(k)} - 1. \end{cases} \quad (11)$$

Here, calculate the estimated TP at $k + 1$ by $\widehat{\mathrm{TP}}^{(k+1)} = \sum_{c=1}^{C} \widehat{\mathrm{TP}}_c^{(k+1)}$.

3. Results

3.1. Simulation Study. We performed a simulation study to examine the precision of our estimated TP. In this study, the number of patients, n, was set to 200. The number of genes, p, was set to 1000, which included the p_1 (=5 or 30) outcome-predictive genes that are randomly chosen from p genes in each simulation. The coefficient for gene j $(j = 1, \ldots, p)$, β_j, was set to 1.5 for the p_1 outcome-predictive genes and 0 for the remaining $p - p_1$ none-outcome-predictive genes. We set λ_z to 5 and the number of components, C, to 1 throughout (although C was determined using a model selection criterion in practice). The gene expression levels for patient i, $\mathbf{x}_i$, were generated from the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ so that the variance was 1 and the correlation $\rho(x_{ik}, x_{il}) = 0$ or $0.5^{|k-l|}$ [18]. The survival time for patient i was generated based on the exponential model $t_i = -\log(U)/\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ where U is the uniform random variable between 0 and 1 [19]. In order to evaluate the precision of the estimated TP for various values of λ, we report a number of selected genes, including true TP, and estimated TP and FP, for λ_k $(k = 5, 10, 50, 100, 150)$.

Table 2 shows the average of λ, a number of selected genes, true TP, and estimated TP and FP, through 1000 repeats. We observed that the precision of estimated TP varied depending on the value of both p_1 and k (see Table 2). When $p_1 = 5$, the precision of the estimates was sufficient for $k = 10, 50, 100$, and 150, while TP was slightly underestimated for $k = 5$. However, when $p_1 = 30$, the precision of the estimates was sufficient for $k = 5$, 10, and 150, while TP was overestimated for $k = 50$ and 100. For example, when $p_1 = 30$, $\rho = 0.5$, and $k = 100$, the average number of true and estimated TP was 29.9 and 35.3, respectively. The values of ρ did not greatly affect the accuracy of the estimated TP.

3.2. Real Data Analysis. To illustrate how our proposed algorithm could be used to determine λ, we applied it to the DLBCL dataset, comprising survival of 240 DLBCL patients and gene expression data from 7399 genes [1]. In the gene expression data from the 240 patients, we identified 434 genes with complete sets of gene expression values; all other genes had missing expression values, with an average of 24.7 missing values per gene. Here, we used 0.0 as the missing expression value for descriptive purposes. Similar to Rosenwald et al. [1], we divided the data into two: training data consisting of 160 patients and validation data consisting of 80 patients.

For the training data, we obtained the solution path of the lasso estimates; $\widehat{\boldsymbol{\beta}}(\lambda_k)$ $(k = 0, 1, \ldots, z)$. $\lambda_0 = 72.5$ was calculated as described in Section 2.2. We set $\lambda_z = 3.625$ $(=0.05 \times \lambda_0)$ according to Simon et al. [14].

We applied our proposed algorithm to the obtained solution path. We assumed three mixture distributions on the lasso estimates with $C = 1$, 2, or 3 and compared their goodness of fit for the $\widehat{\boldsymbol{\beta}}(\lambda_k)$ $(k = 0, 1, \ldots, z)$ by the Akaike information criterion (AIC). As a result, we chose $C = 1$ because it had the best AIC for all λ_k $(k = 0, 1, \ldots, z)$.

Figure 2 shows the estimated number of TP in a series of values of λ. We found that the lasso selected at most 42 TP, with the number of selected genes at 96, when $\lambda = 7.19$ (=0.86 as $\log_{10}$). Therefore, we selected $\lambda = 7.19$ as the optimum λ, and the estimated mixture distribution for the value of λ was as follows:

$$f\left(\widehat{\beta}_j(7.19)\right) = \frac{160}{7399}\left\{0.57 \times f_L\left(\widehat{\beta}_j(7.19); 0, 0.11\right) + 0.43 \times f_N\left(\widehat{\beta}_j(7.19); 0.03, 0.11^2\right)\right\} + \frac{7239}{7399} f_L\left(\widehat{\beta}_j(7.19); 0, 10^{-8}\right). \quad (12)$$

Table 2: Accuracy of the estimated number of true positives (TP) obtained using the proposed algorithm in the simulation study. Average of a tuning parameter (λ), number of selected genes ($\#\{j; \widehat{\beta}_j(\lambda) \neq 0\}$) in the lasso, true number of true positives (True TP), estimated number of TP ($\widehat{\text{TP}}$), and false positives ($\widehat{\text{FP}}$) are reported at λ_k ($k = 5, 10, 50, 100, 150$) of the solution path.

p_1	ρ	k	λ	$\#\{j; \widehat{\beta}_j(\lambda) \neq 0\}$	True TP	$\widehat{\text{TP}}$	$\widehat{\text{FP}}$
30	0	5	47.0	5.0	4.4	2.9	2.2
		10	40.8	10.1	8.0	5.8	4.3
		50	22.9	48.6	25.6	28.5	20.1
		100	12.6	86.7	29.9	32.1	54.7
		150	8.6	124.5	30.0	30.7	93.9
	0.5	5	48.6	5.0	4.1	2.8	2.2
		10	42.1	10.0	7.5	5.8	4.2
		50	23.5	48.1	25.2	31.9	16.3
		100	12.4	84.9	29.9	35.3	49.6
		150	8.4	121.2	30.0	31.6	89.6
5	0	5	66.9	5.0	5.0	3.0	2.0
		10	26.3	10.4	5.0	5.2	5.2
		50	17.2	50.1	5.0	5.2	44.9
		100	12.7	93.9	5.0	5.0	88.9
		150	9.8	128.4	5.0	5.0	123.4
	0.5	5	66.8	5.0	5.0	3.0	2.0
		10	26.5	10.3	5.0	5.2	5.1
		50	16.9	49.5	5.0	5.1	44.4
		100	12.4	92.1	5.0	5.0	87.1
		150	9.6	125.2	5.0	5.0	120.2

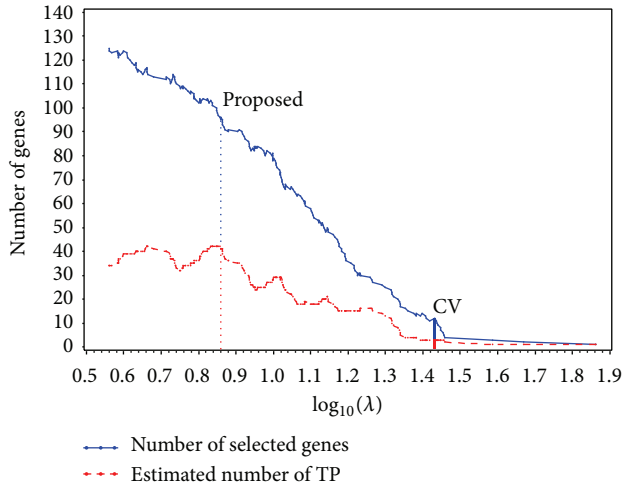


Figure 2: Trace plot of number of selected genes and estimated number of true positives (TP) produced by applying the proposed algorithm to the training data from the diffuse large B-cell lymphoma (DLBCL) dataset. We determined $\lambda = 7.19$ ($\log_{10} = 0.86$) as the optimum λ based on the estimated number of TP. Using cross-validation (CV), we determined $\lambda = 27$ ($\log_{10} = 1.43$) as the optimum λ.

In order to identify the 42 TP from the 96 selected genes, we arranged the 96 in descending order of $|\widehat{\beta}_j|$ and identified the first 42 listed genes with a cut-off value $\zeta = 0.084$. Subsequently, the model that included these 42 genes is identified as the "42 TP-model."

Table 3: GenBank accession numbers and descriptions for 4 genes selected by both CV and the model including the 42 genes identified by the algorithm that we developed.

GenBank accession number	Description
X82240 (AA729003)	T-cell leukemia/lymphoma 1A
AA805575	Thyroxine-binding globulin precursor
LC_29222	—
X59812(H98765)	Cytochrome P450, subfamily XXVIIA polypeptide

In comparison to the 42 TP-model, we performed CV. Briefly, the K-fold CV was given by

$$\text{CV}(\lambda) = \sum_{k=1}^{K} \left\{ l\left(\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda)\right) - l_{(-k)}\left(\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda)\right) \right\}, \quad (13)$$

where $l_{(-k)}(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}}_{(-k)}$ are the log partial likelihood and the lasso estimate with left kth fold out, respectively. The optimal value of λ was obtained by maximizing CV(λ). On the basis of 5-fold CV, 12 genes were selected with $\lambda = 27$ (=1.43 as $\log_{10}$). Subsequently, the model including these 12 genes is identified as the "CV-model." Notably, both the 42 TP-model with 42 genes and the CV-model with 12 genes selected 4 genes in common. Table 3 shows the GenBank accession number and description for each of the 4 genes selected by both models.

We compared the prediction accuracy of the 42 TP-model and the CV-model using validation data consisting

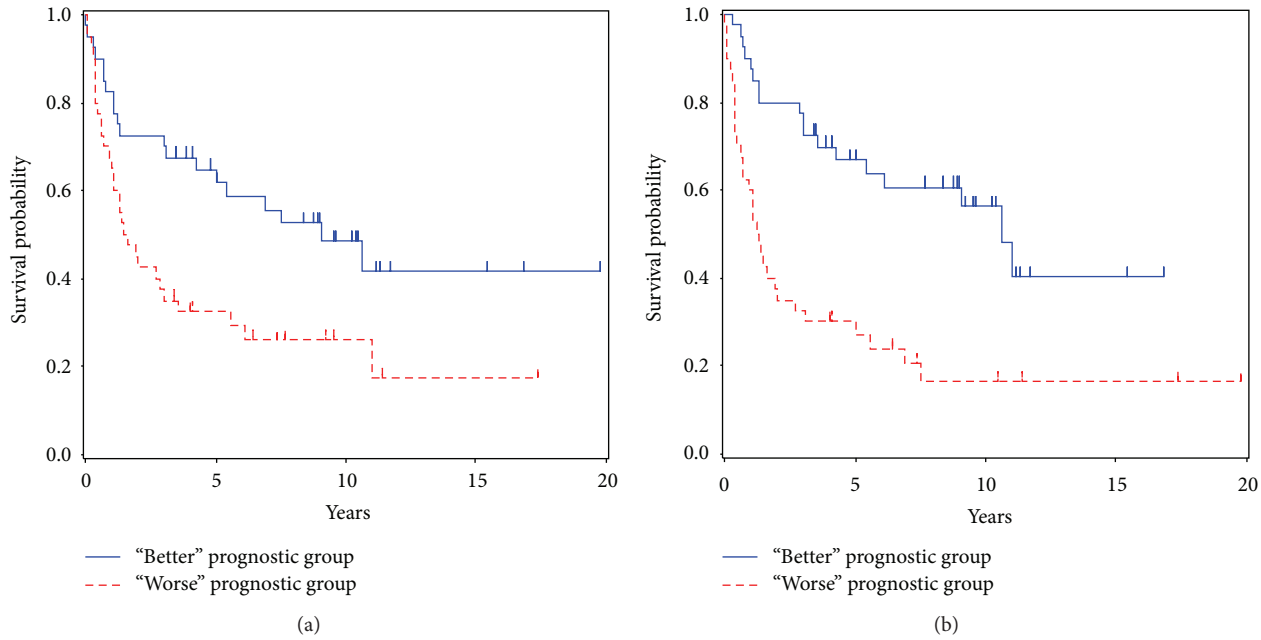


FIGURE 3: Kaplan-Meier curves of overall survival for "better" and "worse" prognostic groups: (a) the model including 12 genes determined by CV (CV-model) and(b) the model including 42 genes identified by the developed method (42 TP-model).

TABLE 4: Values of the comparison criteria for the model including 12 genes determined by CV (CV-model) and the model including the 42 genes identified by our developed algorithm (42 TP-model).

Criteria	CV-model	42 TP-model
P value of the log-rank test	0.007	<0.001
P value for the prognostic index	0.002	<0.001
Deviance	−9.079	−11.297

of 80 patients. For this data, we calculated 3 values that served as comparison criteria: P values for the log-rank test and prognostic index and the deviance. The 80 patients were categorized into 2 groups, the "better" and "worse" prognostic groups, using the boundary of the median of prognostic index $\widehat{\eta}_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. The Kaplan-Meier curves between the 2 groups were compared with a log-rank test. Next, we calculated the P value for the parameter α multiplied by the prognostic index $\widehat{\eta}_i$ in the Cox proportional hazard model $h(t_i \mid \mathbf{x}) = h_0(t)\exp(\alpha\widehat{\eta}_i)$. Finally, the deviance was calculated by $-2\{l^{(\mathrm{validation})}(\widehat{\boldsymbol{\beta}}_{\mathrm{training}}) - l^{(\mathrm{validation})}(\mathbf{0})\}$, where $l^{(\mathrm{validation})}(\widehat{\boldsymbol{\beta}}_{\mathrm{training}})$ and $l^{(\mathrm{validation})}(\mathbf{0})$ are the Cox log partial-likelihood function for the estimated coefficients by using the training data and zero vector $\mathbf{0}$, respectively. For each criterion, the lower value suggested better prediction accuracy.

Table 4 shows the values of the 3 criteria for each model. We found that the values of all 3 criteria for the 42 TP-model were lower than those for the CV-model, suggesting that the model based on the proposed method was more accurate (see Table 4). Additionally, Figure 3 shows that the Kaplan-Meier curves for the 42 TP-model distinguished the "better" and "worse" prognostic groups more definitely than those for the CV-model (42 TP-model, $P < 0.001$; CV-model, $P = 0.007$). Therefore, by using our proposed algorithm, we determined λ and were able to select important genes, likely to be correlated with survival, in which the CV was unable to select.

4. Discussions

In this study, we proposed an algorithm for estimating the number of TP on the solution path of lasso estimates. Monitoring and determining the number of TP for a series of values λ are important because they can increase the probability of uncovering all outcome-predictive genes. The number of TP should be estimated with appropriate accuracy. To confirm the accuracy of our TP, we conducted a simulation study using a typical gene expression dataset. We found that the precision of our algorithm for estimating the number of TP was adequate, although an overestimation occurred with some values of λ. However, the overestimation occurred when the true number of TP was saturated, and so it may not cause a problem by passing over genes that truly correlated with survival. In the simulation study where $p_1 = 30$ and $\rho = 0.5$, the maximum average estimated number of TP was 35.3 at $\lambda = 12.4$ (see Table 2). Using this λ to select TP, an average selection of 29.9 TP within 30 outcome-predictive genes can be made, with the number of TP genes that are passed over being negligible in practice.

The data that have been provided in Table 2 showed that the number of false positives increased, while the number of true positives increased and then plateaued as the tuning parameter decreased. To decrease the number of FP identified while maintaining an adequate number of TP, we should determine the value of λ by monitoring both the number of

TP and the false positive rate (=FP/(TP+FP)) in the proposed method.

Additionally, our proposed algorithm was applied to DLBCL data. We determined the value of the tuning parameter based on the maximum number of estimated TP uncovered by the algorithm. We identified 42 TP genes among 96 selected genes based on the ranking of the absolute values of the lasso estimates. We can also identify TP based on model evaluation criteria such as AIC among all possible combinations of 42 genes from 96, that is, ${}_{96}C_{42}(>10^{27})$ combinations in total; however, calculation of AIC for all possible gene combinations is a distant approach. To evaluate the efficiency of the approach using the ranking of the lasso estimates, we calculated the AIC for 10,000 randomly chosen models among all the possible models and subsequently compared it with the AIC of our approach. From 10,000 models, the AIC of 425 models (4.25%) was better than that of our approach. This result indicated that our ranking-based approach has a satisfactory performance in practice with respect to the identification of 42 genes. Although investigation of all possible gene combinations is ideal, our approach is a good alternative.

In the application to DLBCL data, in comparison to a CV method by which 12 genes were identified, we identified 42 TP genes with our algorithm, and we improved the prediction accuracy of the model. In practice, some researchers might be satisfied with identifying a few promising genes and would not be unduly worried about passing over others. In such a situation, the CV would be preferable because it developed the model to uncover a few genes with just a small loss of prediction accuracy. However, genes that are selected by the lasso are often investigated with greater scrutiny by genetic researchers, and so passing over outcome-predictive genes by the lasso could represent a major problem. Indeed, if the lasso passes over outcome-predictive genes, some genetic research may not take place. Therefore, when identifying all outcome-predictive genes is a priority, our proposed algorithm will be most useful.

5. Conclusions

We developed a method for estimating the number of true positives for a series of values of a tuning parameter in the lasso. We demonstrated the utility of the developed method through a simulation study and an application to a real dataset. Our results indicated that our developed method was useful for determining a value for the tuning parameter in the lasso and reducing the probability of passing over genes that are truly correlated with survival.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. Rosenwald, G. Wright, W. C. Chan et al., "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *The New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, 2002.

[2] L. J. van't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[3] M. J. van de Vijver, Y. D. He, L. J. van't Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[4] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[5] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society B: Methodological*, vol. 34, no. 2, pp. 187–220, 1972.

[6] J. Gui and H. Li, "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Bioinformatics*, vol. 21, no. 13, pp. 3001–3008, 2005.

[7] H. M. Bøvelstad, S. Nygård, H. L. Størvold et al., "Predicting survival from microarray data—a comparative study," *Bioinformatics*, vol. 23, no. 16, pp. 2080–2087, 2007.

[8] W. N. van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix, "Survival prediction using gene expression data: a review and comparison," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1590–1603, 2009.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[10] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[11] P. J. M. Verweij and H. C. van Houwelingen, "Cross-validation in survival analysis," *Statistics in Medicine*, vol. 12, no. 24, pp. 2305–2314, 1993.

[12] J. J. Goeman, "L_1 penalized estimation in the Cox proportional hazards model," *Biometrical Journal*, vol. 52, no. 1, pp. 70–84, 2010.

[13] M. Y. Park and T. Hastie, "L_1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 69, no. 4, pp. 659–677, 2007.

[14] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[15] S. Kaneko, A. Hirakawa, and C. Hamada, "Gene selection using a high-dimensional regression model with microarrays in cancer prognostic studies," *Cancer Informatics*, vol. 11, pp. 29–39, 2012.

[16] B. Efron, T. Hastie, I. Johnstone et al., "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[18] R. J. Tibshirani, "Univariate shrinkage in the Cox model for high dimensional data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 3498–3528, 2009.

[19] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate Cox proportional hazards models," *Statistics in Medicine*, vol. 24, no. 11, pp. 1713–1723, 2005.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 680434, 11 pages
http://dx.doi.org/10.1155/2015/680434

Research Article

A New Approach for Mining Order-Preserving Submatrices Based on All Common Subsequences

Yun Xue, Zhengling Liao, Meihang Li, Jie Luo, Qiuhua Kuang, Xiaohui Hu, and Tiechen Li

Laboratory of Quantum Engineering and Quantum Materials, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China

Correspondence should be addressed to Xiaohui Hu; 1085206157@qq.com

Received 4 October 2014; Revised 31 December 2014; Accepted 24 January 2015

Academic Editor: Antonino Staiano

Order-preserving submatrices (OPSMs) have been applied in many fields, such as DNA microarray data analysis, automatic recommendation systems, and target marketing systems, as an important unsupervised learning model. Unfortunately, most existing methods are heuristic algorithms which are unable to reveal OPSMs entirely in NP-complete problem. In particular, deep OPSMs, corresponding to long patterns with few supporting sequences, incur explosive computational costs and are completely pruned by most popular methods. In this paper, we propose an exact method to discover all OPSMs based on frequent sequential pattern mining. First, an existing algorithm was adjusted to disclose all common subsequence (ACS) between every two row sequences, and therefore all deep OPSMs will not be missed. Then, an improved data structure for prefix tree was used to store and traverse ACS, and Apriori principle was employed to efficiently mine the frequent sequential pattern. Finally, experiments were implemented on gene and synthetic datasets. Results demonstrated the effectiveness and efficiency of this method.

1. Introduction

Recent numerous high-throughput developments in DNA chips generate massive gene expression results, which are represented as matrix D of real numbers with rows (objects) to represent the genes and columns (attributes) to represent the different environmental conditions, different organs, or even different individuals. Each element or entry represents the expression level of a gene under a specific condition.

To analyze the gene expression data, clustering is widely used to gather the objects into different clusters based on similarity. The objects in the same cluster are as similar as possible. Genes in the same cluster may show similar cellular function or expression mode, implying that they are more likely to be involved in the same cellular process. Similarity measurements are mainly based on distance functions, including the Euclidean distance and Manhattan distance. However, these distance functions are not appropriate to measure the object correlation in the gene matrix [1]. Moreover, only a small subset of genes participate in any cellular process of interest, and a cellular process occurs only in a subset of the samples, requiring biclustering or the subspace clustering to capture clusters formed by a subset of genes across a subset of samples [2].

Table 1 shows an example of the original 5×6 data matrix and the corresponding graph is shown in Figure 1(a). If all the rows or columns are considered, then the common mode could not be found. However, if the first five columns are considered, then the 2nd, 3rd, and 4th lines showed the same trend across these five columns as shown in Figure 1(b).

The problem is particularly true for gene expression analysis because the gene expression matrix usually has very high dimension [1]. However, the traditional clustering such as K-means [3] and hierarchical clustering [4] are difficult to use to identify these subsets. Given this observation with respect to the high dimensional data set, those embedded clusters attract wide concern in recent years [5–7], and many biclustering algorithms have been proposed to solve this problem [8–11]. Among them, the pattern-based subspace clustering, which is based on the pattern similarity rather than the distance similarity, has been widely applied in the analysis of gene expression, recommender systems, target sales, and so on.

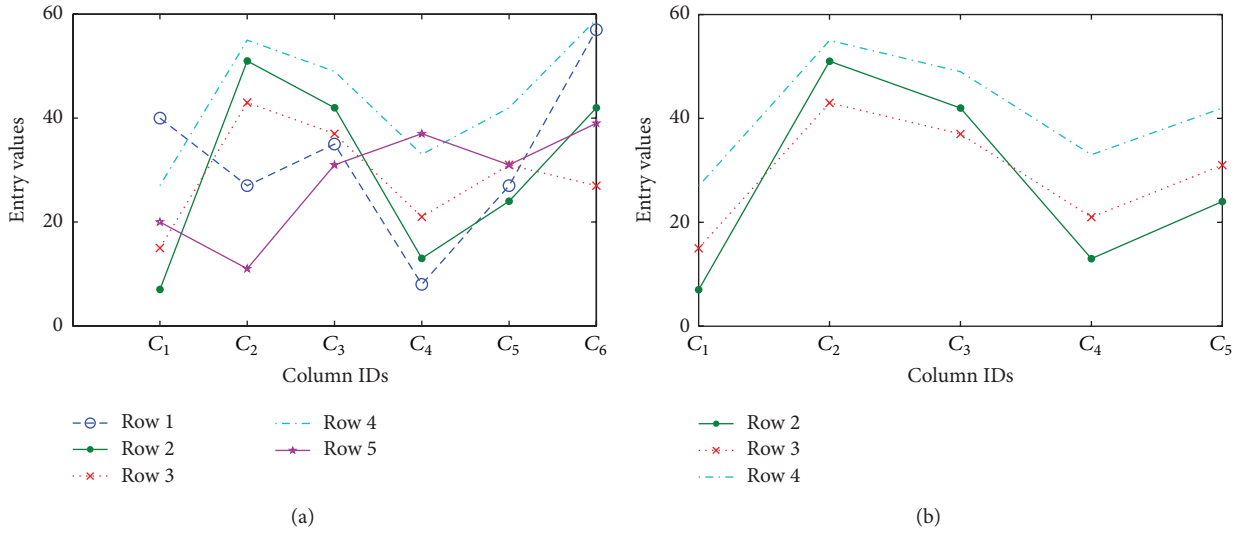


Figure 1: (a) Original data matrix: 5 rows and 6 columns; (b) three rows exhibit a coherent pattern.

Table 1: Raw data matrix.

Rows	Columns					
	1	2	3	4	5	6
Row 1	40	27	35	8	27	57
Row 2	7	51	42	13	24	42
Row 3	15	43	37	21	31	27
Row 4	27	55	49	33	42	59
Row 5	20	11	31	37	31	39

The typical microarray data sometimes has high level noise. Coregulation genes do not necessarily have the same absolute expression level. So to make a comparison of different genes in different experiments, the relative expression levels are more meaningful than their absolute values. Interesting biological knowledge is usually concealed in the genes, which show a similar pattern (rises and falls) in different experimental conditions.

This paper focuses on pattern-based subspace clustering, also known as order-preserving submatrix (OPSM) model. A noncontiguous submatrix is OPSM provided column permutation exists, such that the values in all the rows of the submatrix are strictly monotonically increasing. The tendency among the elements matters more to the model than the actual values. Figure 2 shows that the sequences are monotonically increasing under the new column order given that columns are rearranged. In the field of biology, OPSM model has been accepted as a biologically meaningful cluster model. In addition, the model can also be used in business forecasting. For example, the customers are divided into several categories according to the customer scoring on the telecom tariff packages. Customers who belong to the same class have the same needs such as internet connectivity and surfing speed. The market manager can devise different market strategies for different customer groups based on the results.

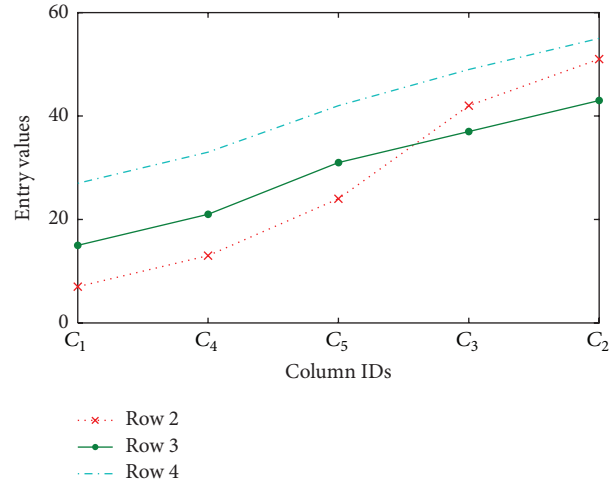


Figure 2: Three rows form a coherent ascending pattern under permutated columns.

If each row vector is sorted in an ascending order with the column indices replacing the original value, then the original matrix is transformed into data set of sequences and OPSM mining problem is simplified as a special case of frequent sequential pattern mining [12]. A frequent sequential pattern is uniquely defined as OPSM with all the supporting sequences as rows. The length of a sequential pattern is the number of columns included. Supporting count is the number of rows containing the sequence. A sequential pattern whose supporting count is beyond a minimum support threshold, min sup, is also known as frequent sequential pattern. Therefore, the problem of mining significant OPSM is equivalent to the search for the complete set of frequent sequential patterns.

Most existing sequential pattern mining methods rely on setting minimum support threshold to narrow the search

Table 2: Transformed sequence data sets.

Rows	Columns					
	1	2	3	4	5	6
Row 1	4	2	5	3	1	6
Row 2	1	4	5	3	6	2
Row 3	1	4	6	5	3	2
Row 4	1	4	5	3	2	6
Row 5	2	1	3	5	4	6

space. Given that the small support threshold will cause the explosive growth of the calculation cost, most of the existing methods improve the efficiency of the algorithm by setting a comparatively larger threshold. However, the large supporting threshold could not find the deep OPSM. The concept of deep OPSM with long patterns and small supporting row count was first proposed by Gao et al. [12]. Deep OPSMs are significant to biologists because they may represent small groups of genes that are tightly coregulated under some conditions. In some important biological processes, such as protein-protein interactions, and biological pathway membership, only a limited number of genes are involved in these processes. However, the general algorithms for frequent sequential pattern mining usually ignore this type of OPSMs.

To solve the above problems, this paper transforms OPSM into frequent sequential pattern first, and then an exact algorithm is proposed to search OPSMs based on frequent common subsequence mining. It can mine all OPSMs embedded in a given matrix and provide flexibility for row and column supports, which allows the discovery of deep OPSMs.

An algorithm calACS proposed by Wang and Lin [13] was improved to determine all common subsequences between two sequences. Then, the Apriori rules were introduced to narrow the search space, and the prefix tree was constructed to store and traverse the sequence modes to reduce time and space complexity. Finally, all OPSMs satisfied the defined threshold. In our algorithm, the computation cost would not increase enormously even if the value of the threshold was very small.

The rest of this paper is organized as follows. In Section 2, we review some related works. Section 3 defines OPSM. Section 4 describes the algorithm and the data structure. Section 5 reports the experimental results. Section 6 concludes the paper.

2. Related Work

Subspace clustering determines the embedded clusters in high dimensional data set. Hartigan [14] first proposed to cluster rows and columns simultaneously. Cheng and Church [15] applied it for knowledge discovery in the expression of gene data. The method overcame the weakness of traditional clustering methods, allowing for the simultaneous clustering of genes and conditions. If the mean square error of submatrix A is less than δ, then submatrix A is a bicluster. A greedy algorithm is proposed to search submatrices with low mean square error in a gene expression matrix, which are consistent biclusters. These submatrices performed well to determine coregulation patterns in genes and attributes [15].

Ben-Dor et al. [16] first proposed OPSM mining model, which pertained to the relative value of entry rather than the actual value. OPSM is essentially a pattern-based subspace clustering. The subset of a matrix is OPSM when the value of each row is strictly increasing or decreasing under column permutation. They proved that the problem is NP-hard and presented a greedy heuristic algorithm for mining OPSM. The algorithm can mine some OPSMs with large row support but cannot guarantee that all OPSMs could be found.

Cheung et al. [1] proposed a maximal OPSM model, converting OPSM problem into a sequential pattern mining problem. To mine all maximal OPSMs with a candidate generation-and-test framework, a new data structure head-tail tree was introduced. However, their algorithm is based on the Apriori principle, and thus the number of maximal OPSMs was affected by supporting row threshold, which increases in proportion with database size.

Gao et al. [12] proposed a new model also known as deep OPSM, referring to long patterns with a few supporting sequences. Deep OPSMs have biological significance. A framework KiWi was proposed to mine deep OPSMs in massive data sets effectively. Two parameters k and w were exploited to bound existing computing resources and determine as many deep OPSMs as possible. However, the algorithm was heuristic, which cannot guarantee the finding of all deep OPSMs.

3. OPSM Problem

In this section, we defined OPSM and detailed the process of transforming OPSM into the problem of mining frequent common subsequences.

Consider $n \times m$ data matrix D, where R is the row set and C is the column set in D. d_{ij} is the entry whose row label is i and column label is j. A cluster $S = (R_S, C_S)$ is a submatrix of D, where R_S is a subset of n rows and C_S is a subset of m columns. The rows and columns do not need to be contiguous in D.

Definition 1. Submatrix S is OPSM if there exists a permutation of C_S. The entries of each row in R_S are strictly monotonically increasing. For example, Table 1 displays a 5×6 matrix. If rows 2, 3, and 4 are increasing from C_1 to C_2, then $(\{2, 3, 4\}, \langle 1, 2\rangle)$ is OPSM. The fundamental goal is to find all the significant OPSMs in a given data matrix D.

In the data preprocess, each row is sorted in an ascending order, and the values are replaced by the original column label. Then, the original matrix is transformed into data set of sequences. The original data matrix of Table 1 is modified into the data set of sequences shown in Table 2. If the values of two entries in a row are the same, then the one that appears earlier is placed in front. A sequential pattern is frequent when the support of the sequence is greater than a predefined minimum support threshold, min sup. Therefore,

OPSM mining problem can be simplified as a special case of frequent sequential pattern mining. A frequent sequential pattern uniquely defines OPSM, in which the sequential pattern is composed of OPSM columns, and the support sequence comprises the rows of OPSM.

Most existing sequential pattern mining methods search OPSMs by finding all the sequences whose support is greater than a given minimum support threshold. The efficiency of the mining algorithm is very sensitive to the minimum support threshold. A larger threshold is adopted to narrow the search space and reduce the complexity of the algorithm because a small threshold results in the high cost of computation. However, this method ignores some statistically and biologically significant OPSMs, deep OPSMs. Deep OPSMs are OPSMs with comparatively more columns and fewer rows that cannot be efficiently discovered by traditional methods [12].

To solve this problem, a new exact algorithm is proposed in this paper. The first step is to determine all common sequences from each two rows in the data set to form the candidate patterns with arbitrary length whose support is at least 2. Then, the database is scanned to calculate the row support for the candidate patterns whose length is 2 to find all the frequent sequential patterns with length 2. The third step is to construct the prefix tree and store the frequent sequential patterns (with length 2). The fourth step is to traverse the prefix tree and insert the node in the branch based on the Apriori principle and calculate the support again to obtain the frequent sequential patterns whose length is 3. The algorithm runs iteratively until all OPSMs satisfying the minimum support threshold could be found. In this process, if larger support threshold is not used to prune, then the results will contain all the deep OPSMs.

4. Algorithm

4.1. All Common Subsequences. All common subsequence (ACS) [17] is a variation from the traditional longest common subsequence (LCS). LCS is a classical problem with a goal to determine LCS from a set of sequences (generally two sequences). Wang [17] proposed this new method to calculate the similarity between two sequences. Different from the previous LCS method, this method calculates the similarity based on the number of all common sequences between the two sequences. calACS [13] is a new method to calculate the number of ACS between sequences A and B. We improved calACS to obtain all common subsequences between two sequences. The pseudocode of the improved calACS algorithm is shown as Algorithm 1.

As shown in the pseudocode, $N_A[i]$ stores the common subsequences whose end is element A_i [line 6]. Provided that any two items in the common sequence remained in the same order in sequences A and B, for any $j < i$, if item A_j in B sequence is arranged before item A_i, the same order in sequence A is retained [line 17]. Hence, the common subsequence ending with A_i must contain the common subsequence ending with A_j. They are combined to form the new common sequences and all common subsequence of A and B is the union of $N_A[i]$ [line 18].

We use a prefix tree to store and traverse all common sequences. Different from the traditional method to solve OPSM problem, frequent common subsequences can be obtained by traversing frequent prefix tree rather than by the columns joint.

The prefix tree, also known as *trie*, is an ordered tree used to store strings or associative arrays, in which the nodes from the root to the leaf form a path. The root node is null corresponding to an empty sequence. The common nodes store the column indices and the leaf nodes retain the row indices, which support the branch (a branch is a common subsequence). The sequence is composed of K nodes known as K sequence as shown in Figure 4. A right path (5, 4) in the tree and the leaf node preserves the number (3, 4, 5). That is, rows 3, 4, and 5 have common subsequence (5, 4).

Suppose a complete set of ACS is obtained, such as $S = (R_{ij}, \langle C_1, C_2, \ldots, C_k\rangle)$. R_{ij} represents the labels of rows i and j. C_i is the element and k is the length of the common subsequence, indicating that $\langle C_1, C_2, \ldots, C_k\rangle$ is ordered. We insert sequence S into the prefix tree whose path is $\langle C_1, C_2, \ldots, C_k\rangle$ and record R_{ij} in leaf nodes, which support the sequence.

The traditional method to construct a prefix tree is described in the subsequent paragraphs.

First, we traverse the prefix tree by preorder. If the first k prefix of length $K + 1$ sequence is the same as length K path in the prefix tree, then $(K + 1)$th node will be added to the path tail before the leaf node. As the length $K + 1$ sequence is different from length K sequence, the corresponding leaf node will be revised, and the rows will be recounted to obtain the support of length $K + 1$ path.

However, if data sets are high dimensional and very dense [12], then the prefix tree will become enormous and occupy a huge space when new sequences are added. Traversing and intersection operations are also time-consuming. Hence, reducing the computational complexity is necessary. In this paper, we develop two kinds of prefix trees, namely, candidate and frequent trees, to save the candidate and frequent sequential patterns, and use the Apriori principle to narrow the search space of patterns.

According to the Apriori principle, if a length K sequence is frequent, then all of its subsequences must be frequent; in other words, if a length K sequence has length $K - 1$ subsequence which is not frequent, then length K sequence must not be frequent either. Thus, if length $K-1$ subsequence which is formed by the first $K - 1$ items of length K sequence is not a branch in the frequent tree, the length K sequence should not be inserted into the candidate tree.

In Section 4.1, calACS is introduced to obtain ACS between any two sequences. The common subsequences with length 2 are employed to generate the 2-candidate prefix tree. The 2-candidate tree is constructed via traversing and inserting with each path retaining the column indices, and all the leaf nodes store the corresponding support row indices, as well as the number of the support rows. Furthermore, we use the number of the support rows to determine whether a branch (i.e., a path) is frequent. If the branch satisfies the support threshold (min sup), it is preserved; otherwise, it is pruned. After all the prune operation is performed, the 2-frequent tree is obtained, which is the first iteration.

Data: Two sequences A and B
Output: acs—the set of all common subsequences of A and B.
(1) * * * * * * * * * * * * * * Begin of Initialization * * * * * * * * * * * * * *
(2) ind[0] = 0, $N_A[0] = \phi$;
(3) acs+ = $N_A[0]$; //acs is the set of ACS between A and B.
(4) *for* ($i = 1$; $i \leq |A|$; i++) *do*
(5) ind[i] = −1;
(6) $N_A[i]$ = null;
(7) *for* ($j = 1$; $j \leq |B|$; j++) *do*
(8) *if* $A_i = B_j$ *then*
(9) ind[i] = j; //ind[i] represents the index of celement i of A in B.
(10) end
(11) end
(12) end
(13) * * * * * * * * * * * * * * End of Initialization * * * * * * * * * * * * * *
(14) *for* ($i = 1$; $i \leq |A|$; i++) *do*
(15) *if* ind[i] ≠ −1 *then*
(16) *for* ($j = 0$; $j < i$; j++) *do*
(17) *if* ind[i] > ind[j] *then* //if i, j stay the same order.
(18) $N_A[i] = N_A[i] \cup (N_A[j] + A_i)$;
(19) end
(20) end
(21) acs = acs ∪ $N_A[i]$;
(22) end

Algorithm 1

The next step is to add the common subsequences with length 3 to the 2-frequent tree. The process is as follows.

Preorder traverses the 2-frequent tree. If the first two prefixes of the length 3 common subsequence are the same as a branch in 2-frequent tree, then the third node is added to the tail of the branch and the leaf node is simultaneously updated, restoring the support row indices and recounting the number of rows.

By contrast, if the first two prefixes of the length 3 common subsequence do not match any path in 2-frequent tree, according to the Apriori principle, then the length 3 common subsequence must not be frequent either and should not be added as a path to the prefix tree. This process reduced the unnecessary traversal and the comparison between sequences, which are very time-consuming in a large prefix tree. Thereafter, we obtain the 3-candidate tree. After pruning infrequent branches, the 3-frequent tree is acquired.

The above process is repeated to generate K-candidate tree from $K-1$ frequent tree. Prune the branches which do not meet the minimum support threshold to obtain K-frequent tree, in which each path or branch is a frequent sequence. The program is not terminated until the common subsequences with the longest length are visited. The final result is a tree with the longest path to satisfy the support. The nodes in each path represent the column indices, and the leaf node of each path stores the corresponding row indices. Thus, all OPSMs can be found.

The flowchart of our algorithm is as Figure 3.

4.2. An Example to Find ACS. Given an original $M \times N$ data matrix D, where d_{ij} represents the expression level of the gene i under the condition j, a matrix is shown in Table 3. When each row in the matrix D is sorted in an ascending order and their values are replaced by the corresponding column indices, the matrix is replaced with a new matrix C as shown in Table 4. ACS could be obtained by applying the improved calACS algorithm for matrix C.

Table 3: A microarray data matrix D.

Rows	Columns				
	1	2	3	4	5
Row 1	120	110	119		100
Row 2	999	128	80	115	810
Row 3	676	300	77	287	264
Row 4	197	107	99	587	101
Row 5	154	78		20	10

Table 4: An example of column permutated matrix C.

Rows	Columns				
	1	2	3	4	5
Row 1	5	2	3	1	
Row 2	3	4	2	5	1
Row 3	3	5	4	2	1
Row 4	3	5	2	1	4
Row 5	5	4	2	1	

Common subsequences from arbitrary two rows are shown in Table 5. However, the relatively large space complexity results in great inconvenience for later traversal,

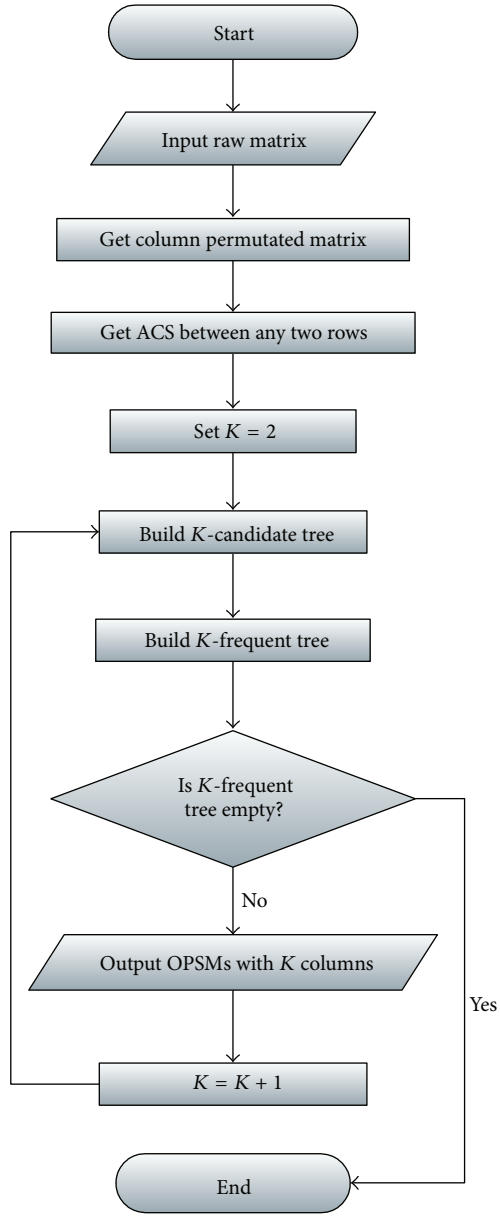


Figure 3: Flowchart of our algorithm.

storage, and support calculations. Hence, a prefix tree is adopted for faster operation to reduce the space complexity.

4.3. Construct ζ-Frequent Prefix Tree. Firstly, ζ-candidate prefix tree would be generated by the candidate ζ-subsequences matrix. Figure 4 illustrates the 2-candidate prefix tree for $\zeta = 2$ after finding ACS operation. The leaf nodes of the prefix tree store the labels of the rows of a common subsequence (a branch). For example, the leaf node of the right branch in Figure 4 records (3, 4, 5), implying that rows 3, 4, and 5 have common subsequence whose column heads are 5 and 4.

The sequences in the leaf nodes of ζ-candidate prefix tree do not necessarily have to be frequent because the ζ-candidate prefix tree would be used to generate the ζ-frequent prefix tree whose leaf nodes are frequent subsequences with length k.

ζ-frequent prefix tree is constructed by deleting the infrequent subsequences that dissatisfy the minimum support δ. Figure 5 is an example of ζ-frequent prefix tree ($\zeta = 2$).

Table 5: Results of ACS discovery of column permutated matrix C.

(a) Candidate 2-subsequences matrix

Sequences	Common 2-subsequences
1, 2	5, 1; 2, 1; 3, 1
1, 3	5, 2; 5, 1; 2, 1; 3, 1
1, 4	5, 2; 5, 1; 2, 1; 3, 1
1, 5	5, 2; 5, 1; 2, 1
2, 3	3, 4; 3, 2; 3, 5; 3, 1; 4, 1; 4, 2; 2, 1; 5, 1
2, 4	3, 4; 3, 2; 3, 5; 3, 1; 2, 1; 5, 1
2, 5	4, 2; 4, 1; 2, 1
3, 4	3, 5; 3, 4; 3, 2; 3, 1; 5, 4; 5, 2; 5, 1; 2, 1
3, 5	5, 4; 5, 2; 5, 1; 4, 2; 4, 1; 2, 1
4, 5	5, 2; 5, 1; 5, 4; 2, 1

(b) Candidate 3-subsequences matrix

Sequences	Common 3-subsequences
1, 3	5, 2, 1
1, 4	5, 2, 1
1, 5	5, 2, 1
2, 3	3, 4, 2; 3, 4, 1; 3, 2, 1; 3, 5, 1; 4, 2, 1
2, 4	3, 5, 1; 3, 2, 1
2, 5	4, 2, 1
3, 4	3, 2, 1; 3, 5, 4; 3, 5, 1; 3, 5, 2; 5, 2, 1
3, 5	5, 4, 2; 5, 4, 1; 5, 2, 1
4, 5	5, 2, 1

(c) Candidate 4-subsequences matrix

Sequences	Common 4-subsequences
2, 3	3, 4, 2, 1
3, 4	3, 5, 2, 1

4.4. Build the (ζ+1)-Frequent Prefix Tree. Specific steps are detailed to construct ($\zeta + 1$)-candidate prefix tree. Based on Apriori principle, if a sequence is frequent, then all of its subsequences must be frequent. Only the frequent sequences can generate the supersequence.

To build ($\zeta + 1$)-frequent prefix tree, first the ($\zeta + 1$)th element of the common ($\zeta + 1$) subsequences is inserted into ($\zeta + 2$)th layer of ζ-frequent prefix tree. At the same time, the leaf nodes of ζ-frequent prefix tree are revised. Second, the infrequent subsequences of ($\zeta + 1$)-candidate prefix tree are rejected, and ($\zeta + 1$)-frequent prefix tree is established in this way. Moreover, ($\zeta + 1$)-candidate prefix tree and ($\zeta + 1$)-frequent prefix tree of the example are shown in Figure 6.

5. Experiments and Results

5.1. The Experiment on Gene Data Set. The algorithm was implemented on the platform of MATLAB R2011b with i3

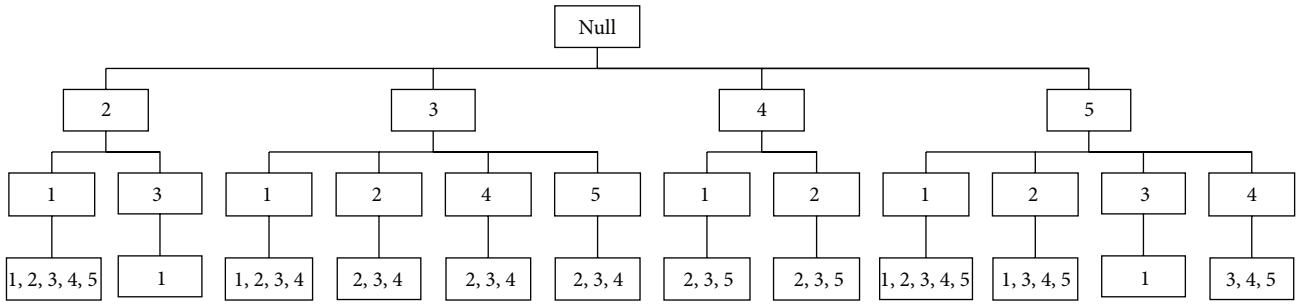


Figure 4: Example of two-candidate prefix tree.

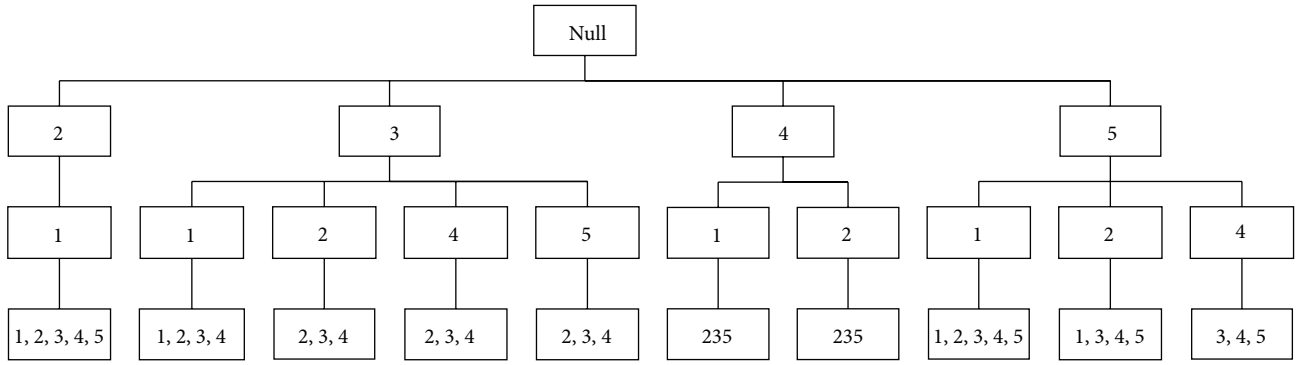


Figure 5: Example of 2-frequent prefix tree with "$\delta = 3$."

380CPU and 4G memory, and the operating system was Windows Server 2007. The real data set was yeast galactose data of [18, 19], which was 205 × 80 real microarray data set obtained from a study of gene response to the knockout of various genes in galactose utilization (GAL) pathway of baker's yeast, with columns corresponding to the knockout conditions and rows corresponding to genes that exhibit responses to the knockouts. The experimental data set is 160× 40 microarray data set by deleting 45 contiguous rows and 40 columns from the original matrix.

5.1.1. Overlap. BicAT software and MATLAB were used in our experiments [20], and overlap is defined as follows [21].

Let G_1, G_2 be two gene sets in biclusters. The overlap of G_1 and G_2 is their intersection divided by their union, and 1 means module identity and 0 means no overlap. Consider

$$S_G(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}. \quad (1)$$

The experimental results are filtered in two steps.

(1) If a bicluster contains another, then the smaller bicluster will be abandoned.

(2) The column threshold is set. For example, if the threshold is six, then the biclusters whose column numbers are less than six will be discarded.

Finally, we obtained all the biclusters corresponding to column threshold six. The total number of OPSMs obtained is shown in Table 6. We can mine all OPSMs that meet the row threshold because our algorithm is exact. The number of OPSMs decreases as the number of row threshold increases.

Table 6: Number of OPSMs of different row thresholds.

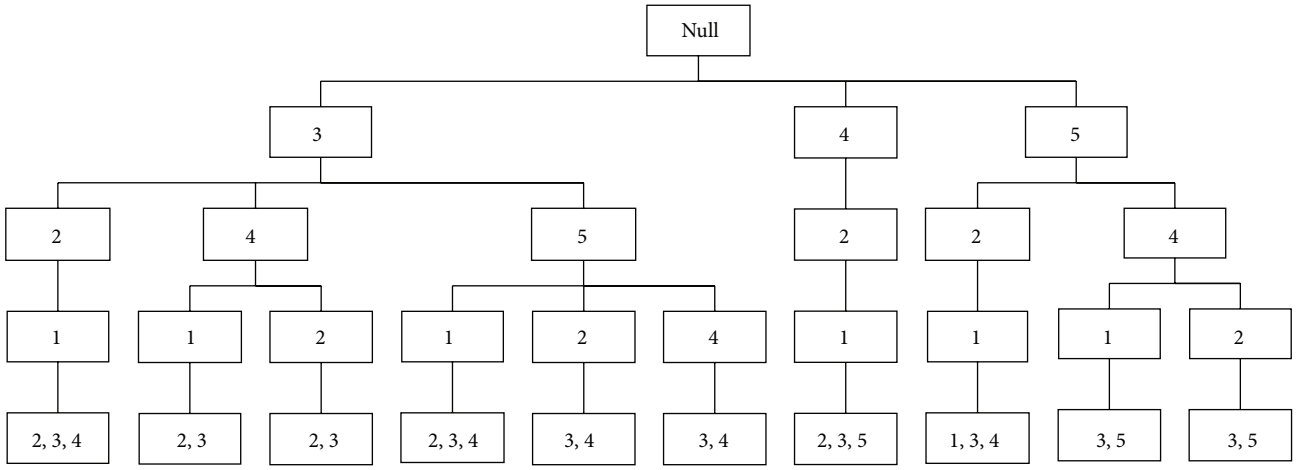
The row threshold	3	5	8	10
Number of biclusters	9248	2791	1350	771

Furthermore, Figure 7 shows the statistical chart on the overlap distribution of 771 biclusters whose row threshold is 10 and column threshold is 6.

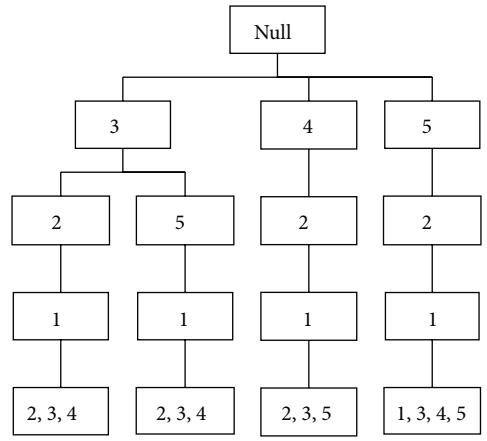
Figure 7 shows that no-overlap biclusters accounted for 60.42% of the total, and the degree of overlap between 0 and 0.1 (excluding 0) accounted for 35.54% of the total. Therefore, the biclusters whose overlap was between 0 and 0.1 (including 0) accounted for 95.96%. That is, the biclusters have no overlap or very small overlap.

5.1.2. An Example of Mined OPSMs. Figures 8(a) and 8(b) show OPSMs that contain the maximal number of columns when the row threshold was set to five and eight. Figure 8(a) shows five genes whose expression values exhibit simultaneous rise and fall across 10 different experiments. Figure 8(b) shows the maximum number of columns that identify OPSM when the row threshold is eight.

5.1.3. Enrichment. The experimental data set is 160 × 40 yeast data set. We first use CC, HCl, *K*-means, OPSM, and xMotif model in the BicAT toolbox to obtain the results. Then, we run our program to obtain the corresponding result.



(a) $(\zeta + 1)$-candidate prefix tree ($\zeta = 2$)



(b) $(\zeta + 1)$-frequent prefix tree ($\zeta = 2$)

Figure 6: Results of $(\zeta + 1)$-frequent prefix tree mining.

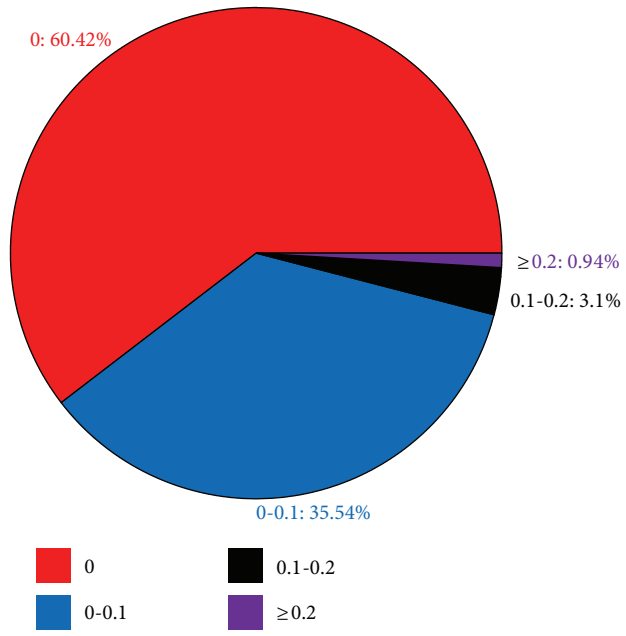


Figure 7: Statistical chart of the overlap distribution.

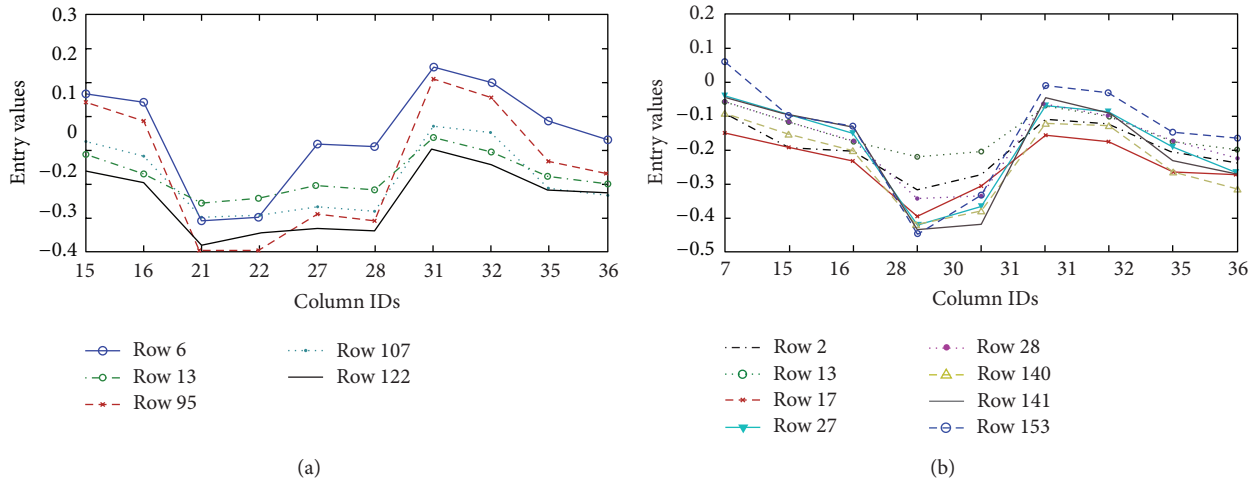


Figure 8: Two examples of mined OPSMs on gene data set.

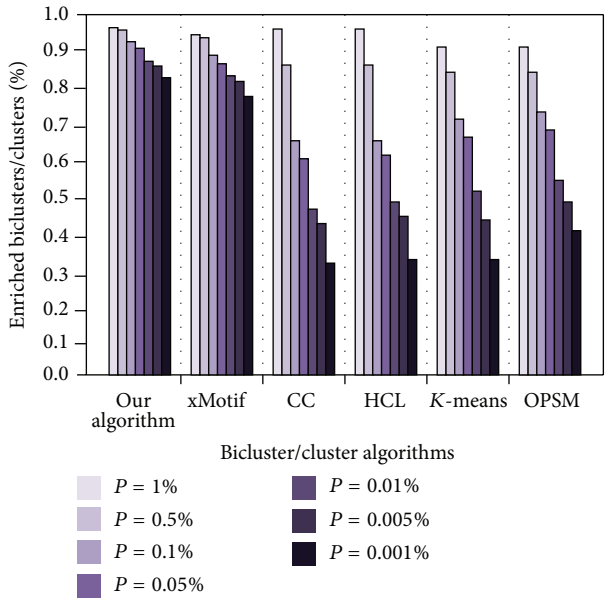


Figure 9: Percentage of significant enriched biclusters/clusters by GO Biological Process category for the five selected biclustering methods and our algorithm at different significance levels P.

The results obtained are packaged, respectively, in GO analysis tool (http://go.princeton.edu/cgi-bin/GOTermFinder) to obtain their P values. Finally, all the results are sorted and analyzed. Figure 9 compares the enrichment results [22, 23].

Figure 9 shows that the enrichment of our algorithm is significantly higher than the enrichment of CC, HCl, K-means, and OPSM. In particular, the smaller P value can show our advantage. The results of xMotif algorithm were close to ours, but slightly less.

5.2. Experiments on Synthetic Data Set

5.2.1. The Influence of Noise. The generation of the simulated data is as follows. First, we generated 30×15 standard normal distribution matrix as the initial matrix with five embedded nonoverlapping 5×3 OPSMs whose row and column sets were recorded. Then, we generated different levels of noise whose means were 0 and variances were 0, 0.002, 0.004, 0.006, 0.008, and 0.01, respectively. The noise will be added to the initial matrix. Finally, we obtained six input matrices with different noise levels.

We introduced match score to evaluate the algorithm [22]. Let M_1 and M_2 be two bicluster sets. Then, the gene match score of M_1 with respect to M_2 is defined as

$$S_G^*(M_1, M_2) = \frac{\sum_{(G_1,C_1)\in M_1} \max_{(G_2,C_2)\in M_2} S_G(G_1, G_2)}{|M_1|}. \quad (2)$$

It shows the average of the maximum gene match scores for all biclusters in M_1 with respect to the biclusters in M_2. An overall match score can be interpreted as $S^*(M_1, M_2) = \sqrt{S_G^*(M_1, M_2) \cdot S_C^*(M_1, M_2)}$, where $S_C^*(M_1, M_2)$ is the corresponding condition match score.

We calculated the match score of different bicluster results, and the comparison was as shown in Figure 10 [22].

The match score of our algorithm is better than others. As the level of noise increases, the total match score decreases slowly.

5.2.2. Overlap. First, we generated 30×15 standard normal distribution matrix with five embedded 5×4 OPSMs whose row and column sets were recorded. Similarly, we obtained five input matrices with different overlap levels. The levels of overlap were L_1, L_2, L_3, L_4, and L_5 corresponding to 0, 0.087, 0.1905, 0.3158, and 0.4706, respectively. The synthetic data were tested by different algorithms, and the match score of all the results was calculated. The performance comparison is shown as in Figure 11 [22].

Figure 11 shows that the match score of our algorithm was better than other algorithms. The parameter settings of other algorithms were based on the best experimental results.

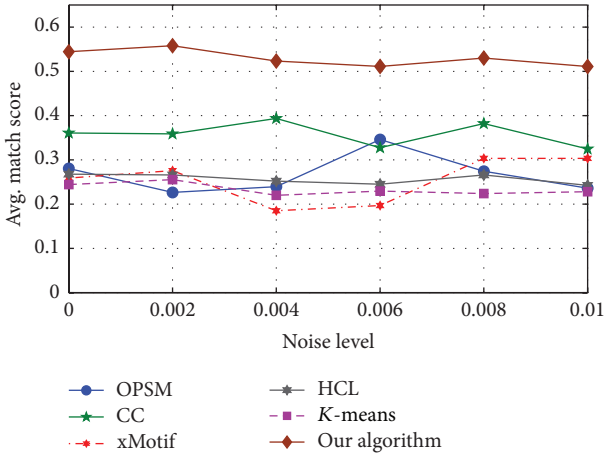


Figure 10: Effect of noise.

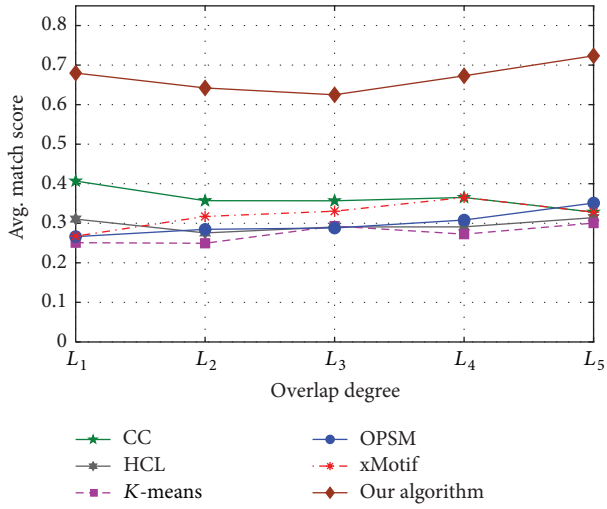


Figure 11: Effect of overlap.

6. Conclusion

OPSMs have been accepted as a biologically meaningful bicluster model. Deep OPSMs consisting of a small number of genes sharing expression patterns over many conditions are very interesting to biologists.

In this paper, an exact algorithm was proposed based on frequent sequential patterns to mine not only all OPSMs, but also the deep OPSMs. The experiment on the gene data set showed that this approach can discover the biological significant OPSMs and deep OPSMs exhaustively. Moreover, the experimental results for synthetic data sets proved that our method can effectively mine the implanted biclusters under different noise and overlap levels.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank their colleagues who were involved in this study and provided valuable technical support. The work is supported by Guangdong Science and Technology Department under Grant no. 2012B091100349, Guangdong Economy & Trade Committee under Grant no. GDEID2010IS034, Guangzhou Yuexiu District Science and Technology Bureau under Grant no. 2012-GX-004, and National Natural Science Foundation of China (Grant nos. 71102146, 3100958).

References

[1] L. Cheung, D. W. Cheung, B. Kao, K. Y. Yip, and M. K. Ng, "On mining micro-array data by order-preserving submatrix," *International Journal of Bioinformatics Research and Applications*, vol. 3, no. 1, pp. 42–64, 2007.

[2] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[3] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, Berkeley, Calif, USA, 1967.

[4] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[5] G. Getz, E. Levine, and E. Domany, "Coupled two-way clustering analysis of gene microarray data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12079–12084, 2000.

[6] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. S136–S144, 2002.

[7] J. Yang, H. X. Wang, W. Wei et al., "Enhanced biclustering on expression data," in *Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*, pp. 321–327, 2003.

[8] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[9] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, "A comparative analysis of biclustering algorithms for gene expression dat," *Briefings in Bioinformatics*, vol. 14, no. 3, pp. 279–292, 2013.

[10] J. A. Nepomuceno, A. Troncoso, and J. S. Aguilar-Ruiz, "Biclustering of gene expression data by correlation-based scatter search," *BioData Mining*, vol. 4, no. 1, article R1, 2011.

[11] J. L. Flores, I. Inza, P. Larrañaga, and B. Calvo, "A new measure for gene expression biclustering based on non-parametric correlation," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 3, pp. 367–397, 2013.

[12] B. J. Gao, O. L. Griffith, M. Ester, H. Xiong, Q. Zhao, and S. J. M. Jones, "On the deep order-preserving submatrix problem: a best effort approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 2, pp. 309–325, 2012.

[13] H. Wang and Z. W. Lin, "A novel algorithm for counting all common subsequences," in *Proceedings of the IEEE International Conference on Granular Computing (GrC '07)*, pp. 502–505, IEEE, Fremont, Calif, USA, November 2007.

[14] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[15] Y. Z. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 93–103, AAAI Press, 2000.

[16] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 373–384, 2003.

[17] H. Wang, "All common subsequences," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 635–640, January 2007.

[18] T. Ideker, V. Thorsson, J. A. Ranish et al., "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science*, vol. 292, no. 5518, pp. 929–934, 2001.

[19] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biology*, vol. 4, no. 5, p. R34, 2003.

[20] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler, "BicAT: a biclustering analysis toolbox," *Bioinformatics*, vol. 22, no. 10, pp. 1282–1283, 2006.

[21] J. Supper, M. Strauch, D. Wanke, K. Harter, and A. Zell, "EDISA: extracting biclusters from multiple time-series of gene expression profiles," *BMC Bioinformatics*, vol. 8, article 334, 2007.

[22] A. Prelić, S. Bleuler, P. Zimmermann et al., "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[23] F. M. Al-Akwaa and Y. M. Kadah, "An automatic gene ontology software tool for bicluster and cluster comparisons," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '09)*, pp. 163–167, IEEE, April 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 954598, 8 pages
http://dx.doi.org/10.1155/2015/954598

Review Article

Statistical and Computational Methods for Genetic Diseases: An Overview

Francesco Camastra,[1] Maria Donata Di Taranto,[2] and Antonino Staiano[1]

[1]*Department of Science and Technology, University of Naples Parthenope, Centro Direzionale Isola C4, 80143 Napoli, Italy*
[2]*IRCCS SDN, Via E. Gianturco 113, 80143 Napoli, Italy*

Correspondence should be addressed to Antonino Staiano; antonino.staiano@uniparthenope.it

Received 16 September 2014; Accepted 23 April 2015

Academic Editor: Abdul Salam Jarrah

The identification of causes of genetic diseases has been carried out by several approaches with increasing complexity. Innovation of genetic methodologies leads to the production of large amounts of data that needs the support of statistical and computational methods to be correctly processed. The aim of the paper is to provide an overview of statistical and computational methods paying attention to methods for the sequence analysis and complex diseases.

1. Introduction

The concept that some disease could be inherited by parents was always present, but only after the discovery of DNA as the genetic material, the research about molecular causes of diseases started. Since the first associations of a disease to a defect in a specific gene, the genetic diagnosis becomes an aim of medical scientists in order to early identify the affected patients and to improve their treatments. For simple monogenic diseases, the conventional way to search for mutations in a gene is the sequencing of amplified fragments corresponding to the gene regions. Innovation in molecular methods together with innovation in computational methods allowed developing new analytical techniques useful to unravel most complicated cases. When the gene responsible for the disease is unknown, in order to identify the genetic defects, the next-generation sequencing could be applied to sequence the whole genome/exome of affected patients, producing then a huge amount of data. In early 2001, during the first assemblies of the human genome, Baldi and Brunak, in their seminal book [1], stressed on the need of statistical and computational supports to the genetic analysis: "[...] *these high throughput technologies are capable of rapidly producing terabytes of data that are too overwhelming for conventional biological approaches. As a result, the need for computer/statistical/machine learning techniques is today stronger rather than weaker*". Today, after fourteen years, the need has become even stronger as the human knowledge of genetic mechanisms still increases, making the research on genetic diseases an amazing adventure as well as difficult and demanding. In case of diseases with a complex etiopathogenesis, for example, those caused by several variants in different genes, more advanced investigations are required. Some examples of methods for association studies are here reported together with methods for meta-analysis of different studies. The study of quantitative traits associated with specific variants is a hot topic in the field of complex diseases, as well as gene expression studies. The presence of a genetic mutation/variant is not the only dysfunction cause of the encoded protein; in fact also alterations in its levels could be responsible for a pathological phenotype. Here we report the example of the combination of both studies, the analysis of expression quantitative trait loci that investigates the association of the quantitative data about gene expression with the presence of specific variants across the genome. Thus the aim of this paper is to provide the reader with an overview of the statistical and computational methodologies, focusing on sequence analysis and complex diseases. Further hot topics, such as methods for next-generation sequencing, gene expression studies, miRNA regulation, and epigenetics, are not discussed merely for sake of space.

The paper is organized as follows: in Section 2, the study of sequence variants is described, while in Section 3 methods for association studies, meta-analysis, and expression quantitative trait loci, specifically targeted to the study of complex diseases, are discussed; finally, some conclusions are drawn in Section 4.

2. Sequencing Analysis

The classical approach for identifying the genetic alteration of a hereditary disease is the sequence of causative genes. Although, in the past, a variant identified in patients and not in control subjects was called pathogenic, currently the definition of pathogenicity should be better demonstrated because some variants have only little effects on the disease [2] and could not be considered the real cause of phenotypic alterations. The only one direct criterion to demonstrate the pathogenicity of a variant is the functional characterization of the protein carrying the variant. If this is difficult to be performed, in silico predictions could help.

Research in databases is the fastest way to retrieve information about a variant and to know if the variant was previously identified. The research in database of mutations (e.g., the Human Gene Mutation Database, HGMG—http://www.hgmd.org/) and single nucleotide polymorphisms (SNP) (e.g., http://www.ncbi.nlm.nih.gov/snp) allows linking to previous papers about the variant or linking to 1000 genome data, for example, the variant frequency.

Some mutation types can be immediately considered pathogenic because they lead to a dramatic change of the encoded protein; these include large deletions and insertions comprising one or more exons and deletion and insertion causing reading frameshift and nucleotide substitutions leading to the formation of a premature stop codon (nonsense mutations). Computational predictions are essential for other mutations with uncertain significance, for example, substitutions leading to an amino acid change (missense), not changing amino acid sequence (synonymous), leading to possible splicing alterations and deletion or insertions without frameshift. Different approaches are utilized to evaluate variant effects depending on the mutation type, as listed below.

(1) *Missense Mutations*. The change of a single amino acid could not be deleterious if the affected amino acid is not included in the functional domains of the protein or if it is not essential in the protein folding. The simplest method utilized to evaluate the relevance of an amino acid is the multiple alignment of the orthologous sequences allowing identification if the mutated amino acid is conserved during evolution. This is the basis of several algorithms created to evaluate the pathogenicity of a missense mutation such as *SIFT* (Sorting Intolerant From Tolerant; http://sift.jcvi.org/) [3] that is solely based on sequence. *PolyPhen-2* (Polymorphism Phenotyping; http://genetics.bwh.harvard.edu/pph2/) [4] evaluates the variant effect using 11 features based on the sequence alignment and on the structure data selected from a wider pool using machine learning methods. Another tool based on both sequence and structure data is *PMut* (http://mmb2.pcb.ub.es:8080/PMut/) that is based on the use of neural networks [5] trained with disease-associated mutations and neutral variants. *Mutation Taster* (http://www.mutationtaster.org/) [6] is useful for different mutation types and uses 3 different models all based on a Bayes Classifier [5] trained with disease-causing mutations and with neutral polymorphisms.

(2) *Synonymous Mutations*. Synonymous mutations are often excluded as causative mutations at the first screening, since they do not cause an apparent change in the protein but they can modify the regulatory mechanisms at the basis of gene expression. Any change in the nucleotide sequence can lead to splicing alterations or to mRNA instability caused by alterations of secondary structure or by altered binding of miRNAs, resulting in decreased protein expression. An additional mechanism of synonymous mutations pathogenicity is due to the alternative codon usage that can increase or decrease the elongation rate depending on the relative abundance of tRNA and influencing the protein folding [7]. Computational approaches to the study of synonymous mutations include the analysis of mRNA structure calculating the ΔG induced by sequence variations [8, 9], of the codon usage [10], of miRNA binding, and of splicing prediction as reported in the next paragraphs.

(3) *Splicing*. An intronic nucleotide change near to the acceptor and donor site is easily presumed to affect splicing mechanisms leading to intron retention or exon skipping. Each intronic variant should be assessed for its potential effects on splicing and recently also exonic variants in the CFTR gene leading to a missense variation have been demonstrated to be more relevant in the splicing process than in the protein alteration due to the amino acid change [11]. Tools to identify alterations at the acceptor/donor sites include, for example, *Human Splicing Finder* that calculate the strength of a nucleotide as splicing site based on position weight matrices [12] and *NNSplice* based on a stochastic grammar inference [13]. *GeneSplicer* improves splice site detection using an algorithm to characterize the nucleotide sequence around the site based on Markov modeling techniques [14]. Other methods are focused on the evaluation of Exonic Splicing Enhancer such as *ESEfinder* [15].

(4) *Deletion or Insertion without Reading Frameshift*. A deletion or an insertion without reading frameshift induces a deletion or an insertion of few amino acids and should be studied with respect to the conservation of involved region and the possible alteration of protein structure. De novo prediction of a protein structure is still a challenge but increasing data of experimentally determined structure allowed creating tools such as *Rosetta* [16] that searches for preexisting structures of fragments with similar

sequence and perform the fragment assembly. An innovative approach to the structure study is its coupling with evolution study of protein sequence that help to identify the most important region of the protein [17].

3. Complex Diseases

Many common diseases, including heart disease, diabetes, hypertension, and schizophrenia, are complex; that is, they are caused by many genes interacting with environmental factors [18, 19], making its study difficult. Complex diseases are due to the presence of a set of gene variants potentially predisposing to the disease that can develop if other nongenetic factors are present, for example, environmental factors. These diseases are also defined as polygenic and/or multifactorial in order to highlight the complexity of their etiopathogenesis. The genetic variants associated with a complex disease are often common polymorphisms that individually have little impact on the phenotype; for example, the presence of a single variant could not cause any alteration, whereas the presence of several variants in specific conditions could be considered the cause of the disease. In order to determine disease mechanisms, disease-associated genes must be identified and analyzed in combination; nonetheless determining how they interact to cause the disease is a challenge.

3.1. Association Studies. First studies on variant associations were conducted by case-control design. In this design, the frequencies of alleles or genotypes at the site of interest are compared in populations of cases and controls; a higher frequency in cases is taken as evidence that allele or genotype is associated with increased risk of disease. The usual conclusion of such studies is that the polymorphism being tested either affects risk of disease directly or is a marker for some nearby genetic variant that affects risk of disease. Due to the modest role of a single variant, the studied population becomes even more large and the number of studied variants increased. Genome-wide association studies (GWAS) have revolutionized human genetics. They have led to the identification of thousands of loci that affect the disease susceptibility and clarified our understanding of the architecture of complex major diseases [20]. In GWAS many common genetic variants in different individuals are analyzed in order to establish if any variant is associated with a phenotypic trait. A *single nucleotide polymorphism*, or SNP, is a single base-pair change in the DNA sequence that occurs with a frequency greater than 1% [21]. Although in the last years a profusion of GWAS for complex human traits was successfully completed [22], even for the simplest analyses there is little general agreement on the most appropriate statistical procedure, including preliminary analyses, that is, Hardy-Weinberg equilibrium testing, inference of phase and missing data, SNP tagging, and single SNP and multipoint tests for association [23]. When a well-defined phenotype has been selected for a study population, and genotypes are collected using well suited techniques, the statistical analysis of genetic data begins. An overview of statistical approaches for genetic association studies is given in [23].

The de facto analysis of genome-wide association data is a series of single locus statistic tests where each SNP is independently examined for association to the phenotype. The usual approach to assess evidence for an association between genetic variants and a phenotype is to compute a p-value for the null hypothesis (H_0), of no association. We recall that the p-value is the probability of obtaining a result of a statistic test identical to the one actually observed when the null hypothesis is true. Some widely used methods for computing p-values are linear regression, logistic regression, Fisher exact test, and χ^2 test [23, 24]. If multiple tests are performed, adjustments of p-values are required. To this aim, several methods are available, for example, Bonferroni, False Detection Rate (FDR), and q-value. We recall that the q-value of an hypothesis is the minimum FDR at which the test is statistically significant. q-values are usually derived from the full distribution of p-values across all tests. However, with p-value only, it is difficult to quantify how much confident one should be that a given SNP is truly associated with a phenotype. Indeed, the same p-value computed at different SNPs or in different studies can have different implications for the plausibility of a true association depending on the factors that affect the power of the test, such as the minor allele frequency of the SNP and the size of the study. This is because the probability that a SNP with a given p-value is truly associated with the phenotype depends not only on how unlikely that p-value is under H_0 but also on how unlikely it is under the alternative hypothesis H_1 (which differs from test to test) [25]. Bayesian methods provide an alternative approach for assessing associations that alleviates the limitations of p-values at the cost of some additional modelling assumptions. As an example, a bayesian analysis requires explicit assumptions about effect sizes at truly associated SNPs. Bayesian methods [5] compute measures of evidence that can be directly compared among SNPs within and across studies, and for combining results across studies, across SNPs in a gene, and across gene pathways. For a comprehensive guide to bayesian methods for genetic association studies, refer to [25]. In general, the discovered genetic variants based on univariate analysis account for only a small proportion of the heritability of complex traits [26, 27]. One possible explanation for the "missing heritability" is that testing for association of the phenotype with each SNP individually is not well suited for detecting multiple variants with small effects [28]. Analyzing SNPs one by one can neglect information on their joint distribution. Therefore, a number of association tests involving multiple SNPs have been applied or developed [23, 29]. The development of a multiple testing procedure involves two steps: ranking the hypotheses and choosing a cutoff (i.e., a threshold value) along the rankings. Different methods use SNPs dependency for choosing the cutoff [30, 31], while [29] uses the dependency of adjacent SNPs, discovered by a *Hidden Markov Model* (HMM) [32], to create more efficient rankings. A gene-based test for association has been, instead, proposed in [33], where a *greedy* [34] bayesian model selection is used

to identify the independent effects within a gene and then combined to generate a stronger statistical signal. A further strategy to uncover the "missing heritability" is to use Gene Set Analysis (GSA) as a way to extract additional information from genome-wide SNP data [35]. GSA has the objective of assessing the overall evidence of variant association in a whole set of genes with a disease status. A gene set is a predefined set of genes based on criteria other than the data being analyzed, for example, genes within a specific biological pathway [22]. Several methods for performing the gene enrichment in GSA are based on Fisher's exact test and the χ^2 test [36]. GSA has the potential to detect subtle effects of multiple SNPs in the same gene set that might be missed when assessed individually [37]. Since numerous genes can be combined into a limited number of gene sets for analysis, the multiple testing burden may be greatly reduced by GSA. Moreover, the incorporation of biological knowledge in the statistical analysis may aid the researchers in the interpretation of the results. For a state-of-the art review of gene set studies the reader can refer to [22], while a thorough review of statistical approaches for "prioritizing" the GWAS results is given in [35]. In [38], instead, the SNPs are grouped into SNP sets on the basis of proximity to genomic features such as gene or haplotype blocks, and then the joint effect of each SNP set is tested. The testing of each SNP set is made via the logistic kernel-machine based test. The latter test provides a statistical framework that allows flexible modeling of epistatic and nonlinear SNP effects ([38] and the references therein). Several further proposals to GWAS come from the machine learning research field [39, 40]. From this perspective, it is argued that methods like *Neural Networks* (NNs) [5], *Support Vector Machine* (SVM) [41], and *Random Forests* (RFs) [42] may more naturally and effectively deal with the high dimensionality of data and the occurrence of multiple polymorphisms with respect to more traditional statistical techniques [40, 43]. A number of applications of NNs and hybrid NN have been developed to study childhood allergic asthma [44], Parkinson's disease [45], Alzheimer's disease [46], and multiple sclerosis [47]. SVM has been applied to Parkinson disease [48] and type 2 diabetes [49], while RFs have been applied to study Crohn disease [50], familial combined hyperlipidemia [51], and colon and ovarian cancers [52].

3.2. Methods for Meta-Analysis. To date, a huge number of association studies identified many genetic variants associated with complex diseases. However, these studies often explain only a small proportion of the disease trait's variability [53, 54]. Genetic effects due to common alleles are small and detecting signal requires larger sample sizes [55]. With this growth in evidence has come an increasing need to collate and summarize the evidences in order to identify true genetic associations among the large volume of false positives ([54] and references therein). Furthermore, replication of findings in independent data sets is now widely regarded as a prerequisite for convincing evidence of association [56]. This is why meta-analysis has become an ever more popular approach for the validation of genetic loci predisposing for common disease and phenotypes. *Meta-analyses* can be defined as the statistical integration of information from multiple independent studies with the aim of obtaining an overall estimator (e.g., significance level, p-value, and odd ratio) of the investigated association [57]. Most genetic risk variants discovered in the past few years have come from large-scale meta-analyses of GWASs and several hundred GWAS meta-analyses have already been published [58, 59]. Most of these meta-analyses had sample sizes in the discovery phase exceeding 10,000 participants [60]. These efforts have dramatically increased the yield of discovered and validated genetic risk loci and large meta-analyses may continue to increase the yield of loci in proportion to the total sample sizes [57]. GWAS meta-analysis can be organized in a number of stages (see references [58, 59] for a more detailed description and reference [57] for a more concise one). However, this overview is focused on the state-of-the-art of statistical models for data synthesis in GWAS meta-analysis and following closely the review given in [57].

One possible approach, that is, the Fisher's approach [57], is based on combining p-values. Here the null hypothesis that the true effect is null in each of the combined data sets is checked against the alternative hypothesis that there is nonnull association in at least one data set. A closely related approach to p-value combination is based on the average of Z-values [61]. Although the two methods are correlated, one advantage of the Z-score approach, over the Fisher method, is that it takes into account the direction of the effect, and it is rather straightforward to introduce the weights for each study. An alternative and popular approach is fixed effects meta-analysis, used for synthesizing GWAS data and resulting to be very effective for prioritizing and discovering phenotype-associated SNPs [62]. Fixed effects meta-analysis assumes that the true effect of each risk allele is the same in each data set. The inverse variance weighting [56] is the most used model for fixed effects meta-analysis, in which each study is weighted according to the inverse of its squared standard error [58]. Cochran-Mantel-Haenszel [63] approach is a further popular used method in genetics which provides similar results to the inverse variance weighting method [61]. A well known estimator of the between-study variance for the random effect approach is the DerSimonian and Laird estimator (see [57] and references therein). However, this method might be less robust with respect to rare variants [64]. Although random effect models are not adopted in discovery efforts, they are suitable when the goal is to estimate the average effect size of the investigated variant and its uncertainty through different populations, for example, as for predictive purposes [65]. In Han and Eskin [66], a novel random effect method has been suggested to improve discovery power when heterogeneity in effect sizes exists across the studies, differently to traditional random effect models. Bayesian techniques have been also used for GWAS meta-analyses. The Bayes factor [67] has been used by the Wellcome Trust Case Control Consortium, while the Coronary Artery Disease Consortium has estimated the posterior probabilities that a given variant is null [68]. Moreover, bayesian methods have been developed to identify the best inheritance model for variants discovered by GWAS meta-analyses [69] and

the polygenic structure of complex diseases [70]. Nevertheless, bayesian models have two main drawbacks. Firstly, they depend on the assumption that the parameters of interest follow a given prior distribution. Secondly, their genome-wide implementation can require a huge computational burden [57].

3.3. Expression Quantitative Trait Loci. Quantitative trait locus (QTL) is a DNA region associated with a quantitative phenomenon. In most genetic diseases, quantitative traits are often a measure of the disease severity, such as the lipid levels in a dyslipidemia. Genetic variants could be studied for its capacity to affect these quantitative traits and then to influence the disease severity. Differences in gene expression levels between patients and controls are now recognized as an additional mechanism influencing the development of a complex disease. We are here reporting an example of QTL study based on gene expression levels, the expression Quantitative Trait Locus (eQTL), for example, the study of the effect of a DNA variant on the gene expression. Experimental data from eQTL mapping are mainly formed by a genetic map, marker genotypes, and microarray data extracted by a set of individuals. After the removal of systematic effects, it can obtain measures of gene expression levels. This section does not deal with statistical issues related to a correct eQTL experimental design. To this purpose the reader can refer to [71] and references therein.

eQTL data were used for the identification of the so-called *hot spots* [72], constructing gene networks [73] and the setup of subclasses of clinical phenotypes [74], and shortening the list of candidate genes [75]. All these studies are based on the generation of a list of transcripts and the respective genomic locations these transcripts correspond to. The methods for the eQTL localization are mainly based on usual QTL mapping techniques. A *logarithms of odd* (LOD) score curve is computed for each transcript. LOD score allows comparing the probability of measuring the observed values if two loci are linked with respect to the probability of observing the same values at random. LOD score curve is obtained computing LOD score for all genomic positions. Several approaches have been proposed to control the FDR based on p-values and q-values [76].

Having said that, in eQTL studies p-values (corresponding to the peaks of LOD score curves from each transcript) are used to yield and to control the FDR for a list of transcripts mapping to one location. Since this approach takes into account only LOD score peaks, it cannot be used for transcript mapping to multiple loci [76]. In order to cope with this problem, statistical methods have been designed to control the overall FDR for single and multiple linkage [77, 78]. In particular, an empirical bayesian method to eQTL mapping has been proposed by Kendziorski et al. [78]. The method shares information across transcripts to estimate the posterior probability that each transcript maps to each marker. The method has two different steps. Firstly, transcripts are identified. Then, multiple eQTL are identified using the posterior probability. The method states a genome linked to a trait if its posterior probability of linkage is in the top $(100 - \alpha)$ percent of all probabilities for the trait. A typical value for α is 5.

After having generated the list of transcripts, the identification of the *hot spots* is usually the next task. Hot spots are genomic regions where there is plenty of transcript maps. The simpler method for identifying the hot spots is the following. For each genomic region, the overall number of mapping transcript is computed. Hot spot candidates are the region whose overall number is ranked among highest ones. Although very simple, the method above can fail if there are several loci with effects whose intensity is not adequately large to be considered statistically significant. A strategy for coping with the problem above has been proposed by Kendziorski et al. [78]. The strategy consists in summing evidence in favor of mapping across every transcript and verifying that the obtained score exceeds a given threshold. Further approaches proposed for the hot spots identification consist in computing profiles averaged across correlated transcripts [79] and profiles from transcripts that are functionally related [72]. After having determined the candidate hot spots, it is necessary to use statistical tests in order to assess the confidence that each spot is hot. Therefore a crucial problem is the identification of the so-called *ghost hot spots*, that is, candidate spots that have been considered erroneously hot. This problem has been partially addressed by a Poisson-based test [80] that can detect ghost spots, by computing the probability that a particular genome region would have at least k transcripts linked to it if there were not any hot spots. Unfortunately, this test cannot be applied when the candidate hot spots are identified by summing the evidence of linkage across all transcripts.

The detection of hot spots yields list of comapping transcripts and involves the inspection of further candidates controlling the whole collection. This is motivated by the observation that comapping is the result of comembership in a biological pathway where functional information is deduced by means of temporally correlated transcripts. Jansen and Nap [81] showed first how spot list could be used to make networks, represented mathematically by *graphs*. A graph is a couple of a set of vertices and a set of edges, connecting couples of vertices. In this case, a vertex represents either a gene or a transcript. An edge connects two vertices when there is some relationship between them; besides, a weight, measured by correlation coefficient, is generally associated to the edge. Pairwise correlations among all transcripts are used to identify *cliques* [82], namely sets of vertices, representing transcripts, completely connected by edges. We have to recall that the clique's identification in a graph is a NP-problem [34]. This implies that it is an intractable problem if the graph of the transcript is not adequately small. Mapping regions common to clique members are studied to identify potential candidates that are likely affecting the pathway.

Other approaches that can permit the identification of potentially causal relationships among transcripts are the ones based on *bayesian networks* [83]. Bayesian networks have the aim of finding the so-called *best model*, namely, the model that optimally describes the data (i.e., the transcript and/or the loci) in some given model space. Finding the best model usually requires the computation of penalized

likelihood that manages the trade-off between the goodness of the fit of the model and the number of model parameters. In order to guarantee that the problem is computationally feasible, the model space has to be moderate. Narrowing down the model space for eQTL mapping is usually performed considering only the transcripts that maps to at least one location [84, 85].

We conclude the section quoting that several software tools for eQTL analysis are currently available [86–88].

4. Conclusions

In the paper an overview of statistical and computational methods focused on sequence analysis and complex diseases has been presented. Among the different techniques discussed in this overview, bayesian techniques seem to be promising in terms of performance in some fields, for example, complex diseases [89]. Since these methods generally require a remarkable computational burden, their application has not been popular in the past. Therefore, the development of new high performing computing platforms makes possible, in the next future, a massive use of bayesian techniques in order to cope with biological problems and in particular with complex disease tasks. Although some biological problems have been solved, new ones, even more complex, arise representing, in this way, novel challenges for either biological or statistical and computational methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors wish to thank the anonymous reviewers for their valuable comments.

References

[1] P. Baldi and S. Brunak, *Bioinformatics—The Machine Learning Approach*, MIT Press, 2nd edition, 2001.

[2] A. Ruotolo, M. D. Di Taranto, M. N. D'Agostino et al., "The novel variant p.Ser465Leu in the PCSK9 gene does not account for the decreased LDLR activity in members of a FH family," *Clinical Chemistry and Laboratory Medicine*, vol. 52, no. 8, pp. e175–e178, 2014.

[3] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[4] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, New York, NY, USA, 2000.

[6] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "Mutation-taster evaluates disease-causing potential of sequence alterations," *Nature Methods*, vol. 7, no. 8, pp. 575–576, 2010.

[7] Z. E. Sauna and C. Kimchi-Sarfaty, "Understanding the contribution of synonymous mutations to human disease," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 683–691, 2011.

[8] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.

[9] A. Xayaphoummine, T. Bucher, and H. Isambert, "Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots," *Nucleic Acids Research*, vol. 33, no. 2, pp. W605–W610, 2005.

[10] N. C. Edwards, Z. A. Hing, A. Perry et al., "Characterization of coding synonymous and non-synonymous variants in ADAMTS13 using ex vivo and in silico approaches," *PLoS ONE*, vol. 7, no. 6, Article ID e38864, 2012.

[11] C. Raynal, D. Baux, C. Theze et al., "A classification model relative to splicing for variants of unknown clinical significance: application to the cftr gene," *Human Mutation*, vol. 34, no. 5, pp. 774–784, 2013.

[12] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, "Human splicing finder: an online bioinformatics tool to predict splicing signals," *Nucleic Acids Research*, vol. 37, no. 9, article e67, 2009.

[13] A. Y. Kashiwabara, D. C. G. Vieira, A. Machado-Lima, and A. M. Durham, "Splice site prediction using stochastic regular grammars," *Genetics and Molecular Research*, vol. 6, no. 1, pp. 105–115, 2007.

[14] M. Pertea, X. Lin, and S. L. Salzberg, "Genesplicer: a new computational method for splice site prediction," *Nucleic Acids Research*, vol. 29, no. 5, pp. 1185–1190, 2001.

[15] L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, and A. R. Krainer, "Esefinder: a web resource to identify exonic splicing enhancers," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3568–3571, 2003.

[16] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[17] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nature Biotechnology*, vol. 30, no. 11, pp. 1072–1080, 2012.

[18] W. S. Bush and J. H. Moore, "Genetic-wide association studies," *PLoS Computational Biology*, vol. 8, no. 12, Article ID e1002822, 2012.

[19] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," *Genetics in Medicine*, vol. 4, no. 2, pp. 45–61, 2002.

[20] G. S. Barsh, G. P. Copenhaver, G. Gibson, and S. M. Williams, "Guidelines for genome-wide association studies," *PLoS Genetics*, vol. 8, no. 7, Article ID e1002812, 2012.

[21] G. P. Consortium, "A map of human genome variation from population scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[22] B. L. Fridley and J. M. Biernacka, "Gene set analysis of SNP data: benefits, challenges, and future directions," *European Journal of Human Genetics*, vol. 19, no. 8, pp. 837–843, 2011.

[23] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[24] M. D. Di Taranto, A. Staiano, M. N. D'Agostino et al., "Association of USF1 and APOA5 polymorphisms with familial combined hyperlipidemia in an Italian population," *Molecular and Cellular Probes*, vol. 29, no. 1, pp. 19–24, 2015.

[25] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 681–690, 2009.

[26] E. E. Eichler, J. Flint, G. Gibson et al., "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[27] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[28] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[29] Z. Wei, W. Sun, K. Wang, and H. Hakonarson, "Multiple testing in genome-wide association studies via hidden Markov models," *Bioinformatics*, vol. 25, no. 21, pp. 2802–2808, 2009.

[30] D. R. Nyholt, "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other," *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 765–769, 2004.

[31] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! Rapid adjustment of *P* values for multiple correlated tests," *American Journal of Human Genetics*, vol. 81, no. 6, pp. 1158–1168, 2007.

[32] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[33] H. Huang, P. Chanda, A. Alonso, J. S. Bader, and D. E. Arking, "Gene-based tests of association," *PLoS genetics*, vol. 7, no. 7, Article ID e1002177, 2011.

[34] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Mass, USA, 2nd edition, 2001.

[35] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing gwas results: a review of statistical methods and recommendations for their application," *The American Journal of Human Genetics*, vol. 86, no. 1, pp. 6–22, 2010.

[36] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[37] P. Holmans, "Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits," *Advances in Genetics*, vol. 72, pp. 141–179, 2010.

[38] M. C. Wu, P. Kraft, M. P. Epstein et al., "Powerful snp-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.

[39] C. C. M. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. MacRossan, "Methods for identifying SNP interactions: a review on variations of logic regression, random forest and Bayesian logistic regression," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1580–1591, 2011.

[40] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh, "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology," *BioMed Research International*, vol. 2013, Article ID 432375, 13 pages, 2013.

[41] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.

[42] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[43] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[44] Y. Tomita, S. Tomida, Y. Hasegawa et al., "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma," *BMC Bioinformatics*, vol. 5, article 120, 2004.

[45] A. A. Motsinger, S. L. Lee, G. Mellick, and M. D. Ritchie, "GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease," *BMC Bioinformatics*, vol. 7, article 39, 10 pages, 2006.

[46] M. D. Ritchie, A. A. Motsinger, W. S. Bush, C. S. Coffey, and J. H. Moore, "Genetic programming neural networks: a powerful bioinformatics tool for human genetics," *Applied Soft Computing Journal*, vol. 7, no. 1, pp. 471–479, 2007.

[47] G. Calcagno, A. Staiano, G. Fortunato et al., "A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients," *Information Sciences*, vol. 180, no. 21, pp. 4153–4163, 2010.

[48] Y. Shen, Z. Liu, and J. Ott, "Detecting gene-gene interactions using support vector machines with L1 penalty," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '10)*, pp. 309–311, December 2010.

[49] H.-J. Ban, J. Y. Heo, K.-S. Oh, and K.-J. Park, "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine," *BMC Genetics*, vol. 2, article 26, 2010.

[50] D. F. Schwarz, I. R. König, and A. Ziegler, "On safari to random Jungle: a fast implementation of random forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, Article ID btq257, pp. 1752–1758, 2010.

[51] A. Staiano, M. D. Di Taranto, E. Bloise et al., "Investigation of single nucleotide polymorphisms associated to familial combined hyperlipidemia with random forests," in *Neural Nets and Surroundings*, vol. 19 of *Smart Innovation, Systems and Technologies*, pp. 169–178, Springer, Berlin, Germany, 2013.

[52] X. Chen and H. Ishwaran, "Pathway hunting by random survival forests," *Bioinformatics*, vol. 29, no. 1, pp. 99–105, 2013.

[53] S. M. Lutz, T. Fingerlin, and D. W. Fardo, "Statistical approaches to combine genetic association data," *Biometrics and Biostatistics*, vol. 4, no. 3, Article ID 1000166, 2013.

[54] G. S. Sagoo, J. Little, and J. P. T. Higgins, "Systematic reviews of genetic association studies," *PLoS Medicine*, vol. 4, no. 3, Article ID e1000028, 2009.

[55] K. Chapman, T. Ferreira, A. Morris, J. Asimit, and E. Zeggini, "Defining the power limits of genome-wide association scan meta-analyses," *Genetic Epidemiology*, vol. 35, no. 8, pp. 781–789, 2011.

[56] F. K. Kavvoura and J. P. A. Ioannidis, "Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls," *Human Genetics*, vol. 123, no. 1, pp. 1–14, 2008.

[57] E. Evangelou and J. P. A. Ioannidis, "Meta-analysis methods for genome-wide association studies and beyond," *Nature Reviews Genetics*, vol. 14, no. 6, pp. 379–389, 2013.

[58] E. Zeggini and J. P. A. Ioannidis, "Meta-analysis in genome-wide association studies," *Pharmacogenomics*, vol. 10, no. 2, pp. 191–201, 2009.

[59] J. R. Thompson, J. Attia, and C. Minelli, "The meta-analysis of genome-wide association studies," *Briefings in Bioinformatics*, vol. 12, no. 3, pp. 259–269, 2011.

[60] O. A. Panagiotou, C. J. Willer, J. N. Hirschhorn, and J. P. A. Ioannidis, "The power of meta-analysis in genome-wide association studies," *Annual Review of Genomics and Human Genetics*, vol. 14, pp. 441–465, 2013.

[61] H. Cooper, L. V. Hedges, and J. C. Valentine, *The Handbook of Research Sybthesis and Meta-Analysis*, Russel Sage Foundation, 2009.

[62] R. M. Pfeiffer, M. H. Gail, and D. Pee, "On combining data from genome-wide association studies to discover disease-associated SNPs," *Statistical Science*, vol. 24, no. 4, pp. 547–560, 2009.

[63] N. Mantel, "Chi-square tests with one degree of freedom, extensions of the Mantel-Haenszel procedure," *Journal of the American Statistical Association*, vol. 58, no. 303, pp. 690–700, 1963.

[64] J. J. Shuster, "Empirical vs natural weighting in random effects meta-analysis," *Statistics in Medicine*, vol. 29, no. 12, pp. 1259–1265, 2010.

[65] T. V. Pereira, N. A. Patsopoulos, G. Salanti, and J. P. A. Ioannidis, "Discovery properties of genome-wide association signals from cumulatively combined data sets," *American Journal of Epidemiology*, vol. 170, no. 10, pp. 1197–1206, 2009.

[66] B. Han and E. Eskin, "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies," *The American Journal of Human Genetics*, vol. 88, no. 5, pp. 586–598, 2011.

[67] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, 2007.

[68] N. J. Samani, J. Erdmann, A. S. Hall et al., "Genome-wide association analysis of coronary artery disease," *The New England Journal of Medicine*, vol. 357, no. 5, pp. 443–453, 2007.

[69] G. Salanti, L. Southam, D. Altshuler et al., "Underlying genetic models of inheritance in established type 2 diabetes associations," *American Journal of Epidemiology*, vol. 170, no. 5, pp. 537–545, 2009.

[70] E. A. Stahl, D. Wegmann, G. Trynka et al., "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis," *Nature Genetics*, vol. 44, no. 5, pp. 483–489, 2012.

[71] C. Kendziorski and P. Wang, "A review of statistical methods for expression quantitative trait loci mapping," *Mammalian Genome*, vol. 17, no. 6, pp. 509–517, 2006.

[72] H. Lan, M. Chen, J. B. Flowers et al., "Combined expression trait correlations and expression quantitative trait locus mapping," *PLoS Genetics*, vol. 2, no. 1, article e6, 2006.

[73] N. Bing and I. Hoeschele, "Genetical genomics analysis of a yeast segregant population for transcription network inference," *Genetics*, vol. 170, no. 2, pp. 533–542, 2005.

[74] L. Bystrykh, E. Weersing, B. Dontje et al., "Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'," *Nature Genetics*, vol. 37, no. 3, pp. 225–232, 2005.

[75] N. Hubner, C. A. Wallace, H. Zimdahl et al., "Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease," *Nature Genetics*, vol. 37, no. 3, pp. 243–253, 2005.

[76] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[77] J. D. Storey, J. M. Akey, and L. Kruglyak, "Multiple locus linkage analysis of genomewide expression in yeast," *PLoS Biology*, vol. 3, no. 8, article e267, 2005.

[78] C. M. Kendziorski, M. Chen, M. Yuan, H. Lan, and A. D. Attie, "Statistical methods for expression quantitative trait loci (eqtl) mapping," *Biometrics*, vol. 62, no. 1, pp. 19–27, 2006.

[79] G. Yvert, R. B. Brem, J. Whittle et al., "Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors," *Nature Genetics*, vol. 35, no. 1, pp. 57–64, 2003.

[80] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak, "Genetic dissection of transcriptional regulation in budding yeast," *Science*, vol. 296, no. 5568, pp. 752–755, 2002.

[81] R. C. Jansen and J. P. Nap, "Genetical genomics: the added value from segregation," *Trends in Genetics*, vol. 17, no. 7, pp. 388–391, 2001.

[82] E. J. Chesler, L. Lu, S. Shou et al., "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function," *Nature Genetics*, vol. 37, no. 3, pp. 233–242, 2005.

[83] T. D. Nielsen and F. Jensen, *Bayesian Networks and Decision Graphs*, Springer, Berlin, Germany, 2009.

[84] H. Li, L. Lu, K. F. Manly et al., "Inferring gene transcriptional modulatory relations: a genetical genomics approach," *Human Molecular Genetics*, vol. 14, no. 9, pp. 1119–1125, 2005.

[85] J. Zhu, P. Y. Lum, J. Lamb et al., "An integrative genomics approach to the reconstruction of gene networks in segregating populations," *Cytogenetic and Genome Research*, vol. 105, no. 2–4, pp. 363–374, 2004.

[86] D. M. Gatti, A. A. Shabalin, T.-C. Lam, F. A. Wright, I. Rusyn, and A. B. Nobel, "FastMap: fast eQTL mapping in homozygous populations," *Bioinformatics*, vol. 25, no. 4, pp. 482–489, 2009.

[87] M. Pérez-Enciso and I. Misztal, "Qxpak.5: old mixed model solutions for new genomics problems," *BMC Bioinformatics*, vol. 12, article 202, 2011.

[88] F. A. Wright, A. A. Shabalin, and I. Rusyn, "Computational tools for discovery and interpretation of expression quantitative trait loci," *Pharmacogenomics*, vol. 13, no. 3, pp. 343–352, 2012.

[89] B. Han, X.-W. Chen, Z. Talebizadeh, and H. Xu, "Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks," *BMC Systems Biology*, vol. 6, supplement 3, article S14, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 683679, 9 pages
http://dx.doi.org/10.1155/2015/683679

Research Article

Optimization and Corroboration of the Regulatory Pathway of p42.3 Protein in the Pathogenesis of Gastric Carcinoma

Yibin Hao,[1] Tianli Fan,[2] and Kejun Nan[3]

[1]*Zhengzhou Central Hospital, Zhengzhou, Henan 450007, China*
[2]*Department of Pharmacology, School of Basic Medicine, Zhengzhou University, Zhengzhou, Henan 450001, China*
[3]*Department of Oncology, The First Affiliated Hospital, College of Medicine of Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China*

Correspondence should be addressed to Kejun Nan; nankejun2013@yeah.net

Received 29 July 2014; Revised 17 October 2014; Accepted 24 October 2014

Academic Editor: Antonino Staiano

Aims. To optimize and verify the regulatory pathway of p42.3 in the pathogenesis of gastric carcinoma (GC) by intelligent algorithm. *Methods*. Bioinformatics methods were used to analyze the features of structural domain in p42.3 protein. Proteins with the same domains and similar functions to p42.3 were screened out for reference. The possible regulatory pathway of p42.3 was established by integrating the acting pathways of these proteins. Then, the similarity between the reference proteins and p42.3 protein was figured out by multiparameter weighted summation method. The calculation result was taken as the prior probability of the initial node in Bayesian network. Besides, the probability of occurrence in different pathways was calculated by conditional probability formula, and the one with the maximum probability was regarded as the most possible pathway of p42.3. Finally, molecular biological experiments were conducted to prove it. *Results*. In Bayesian network of p42.3, probability of the acting pathway "S100A11 → RAGE → P38 → MAPK → Microtubule-associated protein → Spindle protein → Centromere protein → Cell proliferation" was the biggest, and it was also validated by biological experiments. *Conclusions*. The possibly important role of p42.3 in the occurrence of gastric carcinoma was verified by theoretical analysis and preliminary test, helping in studying the relationship between p42.3 and gastric carcinoma.

1. Introduction

The occurrence and development of gastric carcinoma is a multifactor, multistage, and multistep process [1]. A large number of molecules have been involved in it and constituted a complex regulatory network [2]. Finding and identifying the key biomarkers of high-risk warning, early diagnosis, and effective treatment of gastric cancer are a focus of gastric cancer research [3]. So far, studies have confirmed that multiple antioncogenes such as PTEN [4], p16 [5], p21 [3], Smad4 [2], Fas [6], and RECK [3] and oncogenes such as, Ras [7], c-myc [1], and MMPs [8] are associated with the development of gastric carcinoma. p42.3 is a novel gene, cloned by applying synchronization, mRNA differential display and bioinformatics. Researchers have proved that p42.3 may play a vital role in the occurrence and development of gastric carcinoma [9]. Some studies indicate that p42.3 has the characteristic of oncogenes and tumor markers and it may be one of the early molecular events in the development from gastric mucosa lesion to gastric carcinoma [10]. However, these results did not explain systematically the specific function of p42.3 in it.

According to our early study, p42.3 may be involved in the regulatory pathway in the occurrence and development of gastric carcinoma and the regulatory pathway is as follows: Ras → Raf-1 → MEK → MAPK kinase → MAPK → microtubule-associated protein → spindle protein → centromere protein → cell proliferation [11]. Nevertheless, it has not been verified by molecular biological experiments. On the basis of our previous study, through improvement of the similarity algorithm between the reference protein and p42.3 protein, this study investigated the biological features of p42.3 by means of the regulatory network of the reference protein, optimized the regulatory network of p42.3, modulated the maximum possible pathway correspondingly, and verified it by preliminary molecular biological experiments.

2. Materials and Methods

2.1. Materials. Gastric carcinoma cell lines BGC823, MGC803, SGC7901, AGS, N87, and GES1 were provided by Beijing Cancer Hospital (the original sources were from Shanghai Bioleaf Biotech Co., Ltd.) and were cultivated in DMEM culture media with 5% fetal bovine serum in a 5% CO_2 cell culture box at 37°C.

2.2. Methods

2.2.1. Structural Features of p42.3. After obtaining the amino acid sequence of p42.3 (GenBank: NP_848543) from NCBI database, the spatial structure of protein was predicted by the threading prediction tool Phyre[2] (http://www.imperial.ac.uk/phyre/, Imperial College London) [12]. Then, relevance to cell proliferation in terms of function was set as the restrictive condition, based on which the protein with the two structural domains were searched and constituted the data set of reference proteins. Whereby, the possible biological property of p42.3 was studied.

2.2.2. Similarity Calculation of the Reference Protein and p42.3 Protein. Multiparameter weighted sum method was put to use in calculating the similarity of reference protein and p42.3. First select several parameters in which the two proteins have similarity to calculate the degree of similarity of each parameter, and then add the weight trained by artificial neural network. Finally, obtain the degree of similarity after a weighted summation.

2.2.3. Selection of the Parameters. According to the literature data, the following nine parameters of protein similarity were selected: protein spatial structure, the number of atoms inside the molecule, the number of amino acids in each protein, the species of amino acids, the location of element P and element S in the protein molecule, and the proportion of the number of atoms C, N, and O in the protein molecule [13–16].

2.2.4. Similarity Calculation of the Spatial Structure of Protein. Before calculating similarity values, the coordinates of each atom in the protein structure file (pdb file) were determined and Euclidean coordinates were used as spatial coordinates, with the geometrical center of the protein as the origin. The distance from each atom to the origin was then calculated. According to these distances, the protein was divided into layers and the structure similarity of two proteins in corresponding layers was analyzed by stratified analysis. It was found that the distances between most of the atoms of p42.3 protein and the origin were in the range of 0~80 nm and a small portion of the distance were within 80~100 nm, and also, very few of them were above 100 nm. Therefore, based on the length of radius, p42.3 protein was divided into 10 layers from the center to outer edge. The distances of each layer were as follows: the first layer 0~10 nm; the second layer 10~20 nm; the third layer 20~30 nm; the fourth layer 30~40 nm; the fifth layer 40~50 nm; the sixth layer 50~60 nm; the seventh layer 60~70 nm; the eighth layer 70~80 nm; the ninth layer 80~100 nm; and the tenth layer beyond 100 nm. The number of atoms in each layer was counted for each of the proteins being compared and stored in array vector data 1 and data 2, respectively. The similarity in atom numbers in each layer was then compared using the formula: sim = 1 − (|data 1 − data 2|/data 1), wherein sim represents a ten-dimensional vector that has stored the similarity of each layer.

Weights were then added to the similarity of each layer and the overall density similarity was calculated by the weighted summation method. It is reasonable to suppose that the layers that contain the most atoms will be more likely to determine properties of the protein. Based on this assumption, the more atoms the layer owns, the higher the weight of this layer is, so the proportion of the atoms number in each layer determined the weight of this layer. Of course, it is maybe different in every layer for two proteins, so the average would be taken. Hence, each layer was weighted as the following formula: $w_i = ((l_{1i}/n_1) + (l_{2i}/n_2))/2, i = 1, 2, \ldots, 10$, where n_1 is the total number of atoms of the first protein, n_2 is the total number of atoms of the second protein, while l_{1i} and l_{2i} are the number of atoms in the ith layer in protein 1 and protein 2, respectively. Thus, the spatial structure similarity of the two proteins was obtained.

2.2.5. Similarity of the Total Number of Atoms and the Number and Type of Amino Acids. Similarity algorithms of the three parameters were alike. The number of atoms and amino acids, and the number of amino acid types in the two proteins were calculated by textread function in MATLAB software. Then, the number of atoms and the number and type of amino acid can be read from the pdb file of the two proteins. The total number of atoms of the two proteins was recorded as n_1 and n_2, respectively, and then the formula used to calculate the similarity in atom numbers was $\text{sim}_a = 1 - (|n_1 - n_2|/n_1)$. Likewise, the similarity of the number of amino acids and its types could be also obtained.

2.2.6. Similarity of Each Element. This study was mainly to analyze elements C, N, O, P, and S. Firstly, the proportion of the number of C, N, and O to the total number of atoms in each protein was calculated. Then, the similarity was calculated among C, N, and O in accordance with the formula: $\text{sim\_element} = 1 - (|n_1 - n_2|/n_1)$. In addition, in protein molecules, the number of elements P and S was usually small, but they both play crucial roles in the function of protein. While in p42.3 protein, there was only one S atom and no P atoms. Therefore, it is obviously not scientific to calculate the degree of similarity according to the number of atoms of the two elements. Instead, similarity of the location between atoms P and S was set as the criteria for calculation. In this algorithm, it was assumed that if the two elements P and S were in the same layer, the similarity was regarded as 1.0; if they were in adjacent layers, the similarity was 0.8; otherwise, the similarity was 0. Therefore, the similarity parameter of each element in proteins was achieved.

2.2.7. Calculation of the Weight of Each Parameter. Based on the similarity of each parameter of protein that had been

Table 1: Similarity weights for each parameter.

Parameter	Similarity weight
Q1: Protein density	0.3183
Q2: Total number of atoms in each proteins	0.0343
Q3: Number of amino acids	0.0204
Q4: Amino acid type	0.0603
Q5: C	0.0653
Q6: N	0.1062
Q7: O	0.1002
Q8: P	0.1477
Q9: S	0.1480

figured out, the overall similarity was worked out by the weighted summation method. Before this, data of 100 pairs of similar protein pairs had been collected. According to the methods described above, the similarity of each parameter in each pair of proteins had been calculated: $S1$–$S9$. Then, BLASTp was used to search the homology of each pair of proteins, which was regarded as the overall similarity. Therefore, for each pair of proteins, a similarity data vector of $1*10$ can be achieved: $[S1, S2, S3, S4, S5, S6, S7, S8, S9, S]$. Then the similarity data of the 100 pairs of proteins was input to BP (back propagation) artificial neural network for training; thus, the weights of each parameter Qi had been achieved (Table 1).

Therefore, for each pair of proteins, their overall similarity can be calculated by formula: $S = 0.3183S_1 + 0.0343S_2 + 0.0204S_3 + 0.0603S_4 + 0.0653S_5 + 0.1062S_6 + 0.1002S_7 + 0.1477S_8 + 0.1480S_9$. In this formula, S is the overall similarity of the two proteins. S_i represents the similarity of each parameter. $i = 1, 2 \ldots, 9$ were spatial structure (density), number of atoms in the protein, number and type of amino acids, number and proportion of C, N, and O atoms [8], and spatial position of P and S atoms in the protein, respectively. On the basis of this formula, the similarity of the reference protein and p42.3 was thus figured out and the data set of the reference proteins was composed.

2.2.8. Construction and Optimization of a Bayesian Regulatory Network. In condition of cellular proliferation, the reference protein set obtained by the similarity calculation was screened out. Then, with a reference protein as the starting point and cell proliferation as the ending point, the acting pathway and node of each reference protein were collected. There are crosses between different pathways, thus constituting a regulatory network [11]. In the network, "+" indicates a positive role in promoting the regulation; "–" represents a negative role in inhibiting the regulation. The similarity of each reference protein and p42.3 was set as the initially prior probability. By applying knowledge of conditional probability, the probability of occurrence in each node was worked out. The formula is

$$P(E) = P\left(\overline{A}\,\overline{B}C\right) + P\left(\overline{A}\,\overline{B}D\right) + P\left(\overline{A}\,\overline{B}CD\right). \quad (1)$$

Bayesian networks are Directed Acyclic Graphs (DAGs), which describe the joint probability distribution of a finite set of variables $U = \{X_1, X_2, \ldots, X_n\}$. Bayesian networks can be symbolized by the element pair $B = (G, \theta)$, where G is a DAG in which the nodes represent random variables $X_1, X_2, \ldots, X_n$. It can symbolize gene expression vectors in expression profiling data, while θ represents the conditional probability of each variable. DAG showed the independent relation under the following conditions. It was the Markov assumption; each variable X_i was independent of its nonchild node in the prerequisite that it was the parent node in G. Based on the assumption of independence, the Bayesian network G had only one joint probability distribution for set U was $P(X_i, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$, where $\mathrm{Pa}(X_i)$ symbolizes the parent node of X_i. In order to determine the joint probability above, all the conditional probabilities in this formula need to be confirmed.

In the Bayesian network in this paper, after obtaining the probability of occurrence in each node and pathway, the Bayes theorem was used to inverse the probability of protein in acting their roles in each node, thus finding the highest possible regulatory pathway of p42.3 protein.

2.2.9. The Molecular Biological Test of the Optimal Path. After obtaining the highest possible acting pathway of p42.3 protein through calculation and prediction, some basic biological experiments were carried out for initial validation. To begin with, Trizol (invitrogen, America) method was used for extraction of the total mRNA in the six cell lines: BGC823, MGC803, SGC7901, AGS, N87, and GES1. Through reverse transcription, cDNA was synthesized (cDNA reverse transcription kit, Thermo fisher scientific company, United Kingdom). According to the gene sequence of the reference proteins and p42.3, the primers were designed. The sequence of the primers was shown in Table 2. In contrast with β-actin, the RT-PCR was used to amplify, respectively, (PCR amplifier, Eppendorf Company, Germany). After PCR products were detected by agarose gel electrophoresis, the expressions of various proteins in different cell lines were compared.

3. Results

3.1. Structural Features of the EF-Hand and CC-Domain. The spatial structure of protein was predicted by the threading prediction tool Phyre. A three-dimensional ligand-binding model of the characteristic of p42.3 in EF-hand region was predicted by using 3DLigandSite (http://www.sbg.bio.ic.ac.uk/~3dligandsite/, Imperial College London) [17]. The metal ion binding sites of p42.3 were ALA78, SER79, TYR81, and ARG86, as shown in Figure 1. The protein data set that had high structural homology with EF-hand and CC-domain (p42.3 molecule) was searched. Some of them with the same structure of EF-hand were shown in Table 3. Then, proteins relating to cell proliferation functionally were screened out as the reference protein.

3.2. Similarity Calculation of the Reference Protein and p42.3. The similarity algorithm of protein was compared by the similarity of nine parameters mentioned above. The MATLAB software (MathWorks, America) was used for programming.

Table 2: Primers sequence of PCR.

Name	Primer sequence	Amplification length	Renaturation temperature
S100A11	F: 5′-ATCGAGTCCCTGATTGCTGT-3′ R: 5′-AGAAAGGCTGGAAGGAAAGG-3′	331 bp	59°C
S100A2	F: 5′-CGCGAATTCATGTGCAGTTCTCTGGA-3′ R: 5′-CCGGGATCCCTCAGGGTCGGTCTGG-3′	294 bp	56°C
β-actin	F: 5′-TTCTGACCCATACCCACCAT-3′ R: 5′-ATTACAGTGCGTGCTAAAGG-3′	508 bp	56°C

Table 3: Structural data set of EF-hand that is similar to the partial structure of p42.3 molecule.

SCOP code	*E* value	Estimated precision	Fold/PDB descriptor	Superfamily	Family
d1iq3a [18, 19]	0.51	80%	EF hand-like	EF-hand	EH domain
d1s6ja [20, 21]	0.52	80%	EF hand-like	EF-hand	Calmodulin-like
c2pmyB [22, 23]	0.61	80%	PDB header: structural genomics, unknown function	PDB molecule: ras and EF-hand domain-containing protein	PDB title: EF-hand domain of human rasef
d1sw8a [24–26]	0.71	75%	EF hand-like	EF-hand	Calmodulin-like
d1c7va [27, 28]	0.76	75%	EF hand-like	EF-hand	Calmodulin-like
d1hqva [29]	0.85	75%	EF hand-like	EF-hand	Penta-EF-hand proteins
d1tiza [30, 31]	0.89	75%	EF hand-like	EF-hand	Calmodulin-like
d1g33a [32, 33]	0.91	75%	EF hand-like	EF-hand	Parvalbumin
d1fw4a [34]	1	75%	EF hand-like	EF-hand	Calmodulin-like
d1f54a [35]	1	75%	EF hand-like	EF-hand	Calmodulin-like

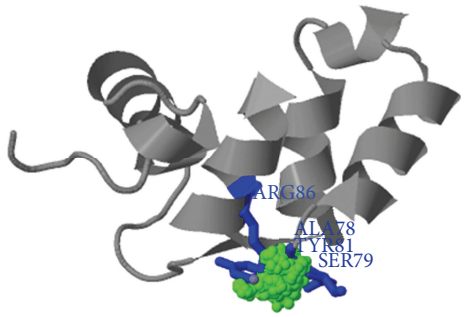


Figure 1: Ligand combination model of EF-hand structural domain in p42.3 molecule.

The similarity of the reference protein and p42.3 was calculated. After screening by "cell proliferation," the results were displayed in Table 4.

3.3. Bayesian Regulatory Network. Cell proliferation was set as the restrictive condition, and the acting pathways and nodes of different reference proteins were worked out by literature collection. With different acting pathways crossing, the regulatory network was thus formed, shown in Figure 2. The round nodes represent the reference proteins and they are the initial nodes. The relation between each node is upstream and downstream regulation. Arrows indicate the direction of action. "+": a positive regulation and "−": a reverse regulation.

The similarity of the reference proteins and p42.3 was treated as prior probability of the initial parent nodes. According to formula (1), the probability of occurrence in each node downstream was calculated until figuring out the final results of cell proliferation. The final one is the probability of occurrence of that pathway. The results were shown in Figure 3. The probability of the path in thick line was 0.9781, higher than that of other pathways. Connected with the results of protein similarity comparison, it can be initially verified that the pathway "S100A11 → RAGE → P38 → MAPK → Microtubule-associated protein → Spindle protein → Centromere protein → Cell proliferation" was with the highest possibility.

3.4. The Molecular Biology Test. Based on the analysis of the spatial structure of p42.3 and Bayesian regulatory network, expressions of S100A11 (the protein with the largest positive maximum weighted value) and S100A2 (the protein with the shortest negative acting path) in gastric carcinoma cell lines were examined, respectively, for preliminary valediction of the correlation of p42.3 and S100A11. The results showed that when p42.3 showed normal expression, both S100A11 and S100A2 had shown expressions. In Figure 4, it was indicated that expression of S100A11 was extremely similar to that of p42.3, while the expression of S100A2 was considerably different from that of p42.3. By referring to the analysis of the protein structure, it could be concluded that the regulatory pathway of p42.3 may be consistent with that of S100A11, or it may be involved in the regulatory pathway.

4. Discussion

The occurrence and development of gastric carcinoma involve changes in the structure and expression of multiple

Table 4: Protein data set gained by using the spherical coordinate space hierarchical similarity algorithm.

Protein name	Number of atoms	Number of amino acids	Type of amino acid	C	N	O	S	P	Spatial structure	Overall similarity
Weight allocation	0.0343	0.020	0.0603	0.0653	0.1062	0.1002	0.1480	0.1477	0.3183	—
S100A11	0.9708	0.9083	0.5294	0.9704	0.8677	0.9226	0.8000	1.0000	0.7384	0.8102
RASEF	0.6557	0.6881	0.9412	0.9929	0.9497	0.9069	0.8000	1.0000	0.6515	0.8068
GCN4	0.6241	0.2752	0.8235	0.9615	0.9271	0.9972	0.8000	1.0000	0.6231	0.7624
FKBP	−0.0255	0.9817	0.7647	0.9655	0.9067	0.9589	1.0000	1.0000	0.5055	0.7334
CENP-B	0.1144	0.4587	0.7647	0.9478	0.9172	0.9488	1.0000	1.0000	0.5535	0.7312
S100A2	0.2032	0.8532	0.5882	0.9611	0.8015	0.8642	0.8000	1.0000	0.5756	0.7046
CIB	0.1034	0.0826	0.8235	0.9497	0.8854	0.9778	0.8000	1.0000	0.5765	0.7026
GPD1	−0.5864	0.6789	0.7647	0.9561	0.8754	0.9505	0.8000	1.0000	0.5905	0.6944
PAK1	0.1521	−0.0275	0.6471	0.9575	0.8543	0.9398	0.8000	1.0000	0.5647	0.6716
ACTN1	−0.2701	−0.3761	0.8235	0.9596	0.9004	0.9721	1.0000	1.0000	0.5902	0.6883
APC	0.9720	0.5046	1.0000	0.9902	0.8994	0.8022	0	1.0000	0.4929	0.6709
GP41	0.4489	0.4128	0.7647	0.9504	0.9527	0.9699	0	1.0000	0.4517	0.6166
S100A12	0.2007	0.8257	0.2941	0.9603	0.8809	0.9300	0	1.0000	0.5668	0.6058
MACF	−0.7944	−0.3578	0.5294	0.9641	0.8395	0.8972	0.8000	1.0000	0.4992	0.5691
MST3	−0.4866	−0.5963	0.2941	0.9407	0.8296	0.9467	0	1.0000	0.5876	0.4997
CHP1	−1.7798	0.1376	0.8824	0.9796	0.9118	0.9021	0	1.0000	0.4568	0.4969
S100A1	−1.5122	0.8532	0.8824	0.5230	0.4119	0.5927	1.0000	1.0000	0.3412	0.4923

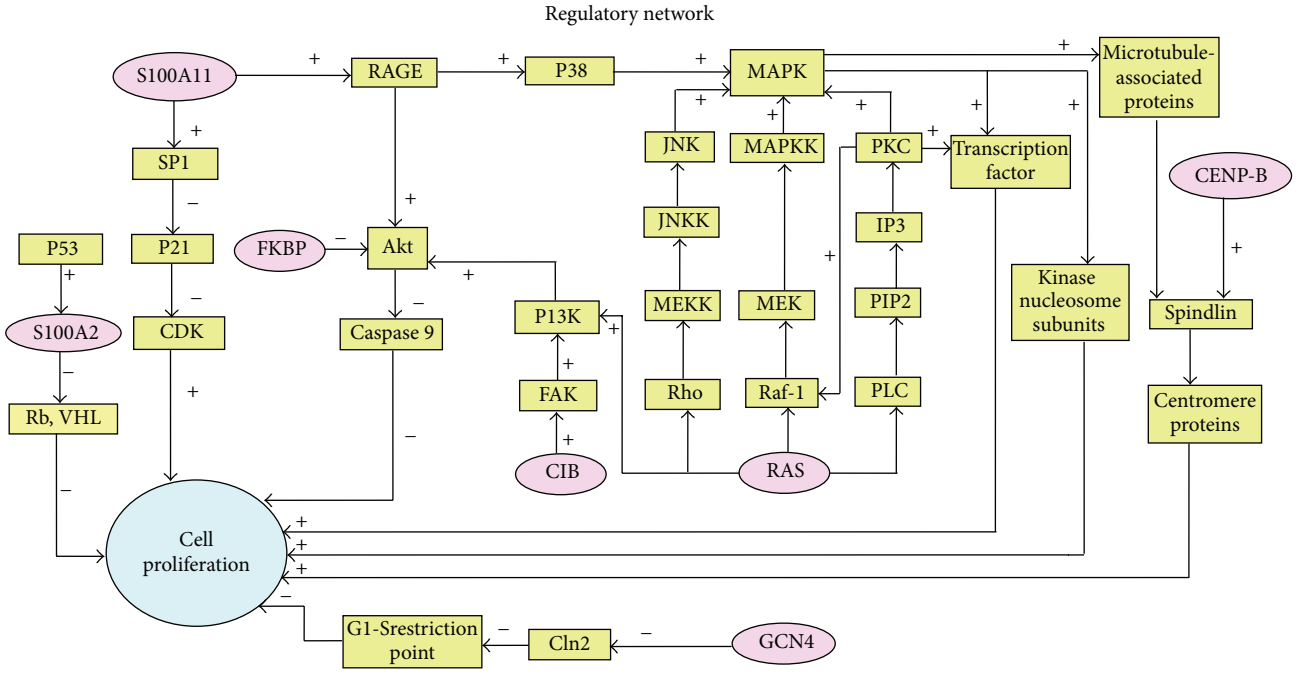


Figure 2: Primary regulatory networks.

related genes [9]. In particular, the activation of oncogenes and inactivation of tumor suppressors play important roles in it [10]. So far, many studies have tried to disclose the molecular regulatory mechanisms of gastric carcinoma in order to find biomarkers for the diagnosis and treatment of gastric cancer, which is expected to be an effective adjuvant therapy of surgery and chemoradiotherapy.

p42.3 expression is dependent on mitosis and is expressed at low levels or not at all in normal gastric mucosa but is highly expressed in gastric carcinoma tissues. It has the effect of promoting cellular proliferation and tumor metastasis [9]. Changes of p42.3 gene expression that occur during the development of gastric carcinoma indicate that p42.3 might be a direction of gastric carcinoma diagnosis and treatment

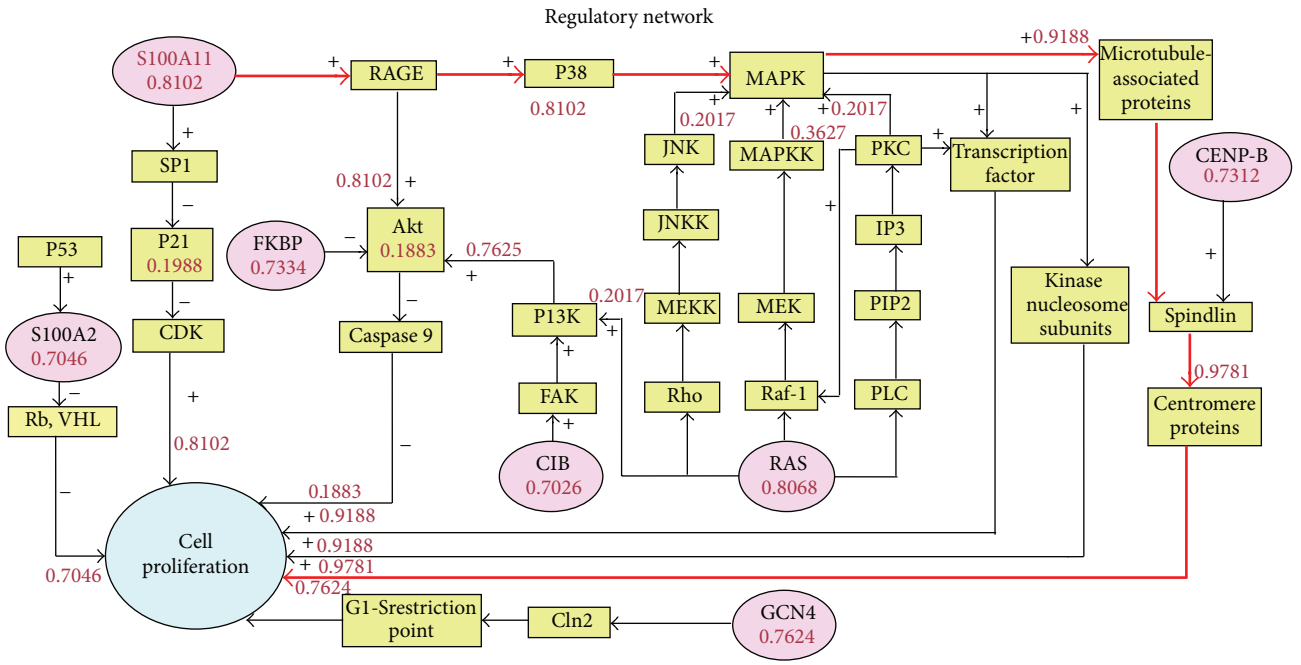


Figure 3: The most possible acting pathway of p42.3 protein by optimization of Bayes theorem.

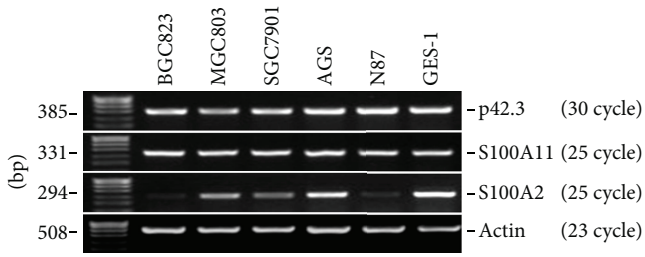


Figure 4: Expressions of p42.3, S100A11, and S100A2 in the cell lines of gastric carcinoma.

[10, 11]. It was found that an EF-hand structural domain existed in the N-terminal amino acid sequence of p42.3 protein, which also presented in the S100 family of proteins [36]. The EF-hand structure consists of a typical helix-loop-helix structural unit; that is, two alpha helixes linked by a Ca^{2+} chelate ring [37]. In all the reports about EF-hand structures, the majority of EF-hand structural domains are even number and form structural domain pairs, separated by connexin, or homologous or heterologous dimmers, such as S100 family proteins with two EF-hand structural domains [38, 39]. Proteins with odd number of structural domains usually need to form homologous or heterologous dimers and their activity is presented in the form of dimer. CC-domain is a kind of super-secondary structure of protein, intertwined by two to seven α helices (most commonly two or four) to form a braided structure [40]. Many proteins with coiled helical structures have significant biological functions, such as the transcription factor in the regulation of gene expression [40]. The most well-known proteins containing coiled helical structures are oncoprotein and tropomyosin. To study the action mechanism of p42.3, a similarity algorithm with multiparameter calculation was adopted to find proteins with high structural similarity to p42.3. As a result, proteins that might be related to the occurrence of gastric carcinoma were screened and treated as gene regulatory path nodes [11]. By a series of probability calculation, it was found that the possible action mechanism of p42.3 in the pathogenesis of gastric carcinoma was S100A11 → RAGE → P38 → MAPK → Microtubule-associated protein → Spindle protein → Centromere protein → Cell proliferation (Figure 3). And the initial molecule experiments also confirmed the consistency of p42.3 and S100A11 gene expression in gastric carcinoma cell (Figure 4). The study of gene regulatory networks can be used to quantitatively mine information regarding gene expression regulation from one side. Through extracting and analyzing this information, gene function and genetic networks can be understood, and the pathogenesis of the disease will be clear. The study of gene regulatory networks aids in the exploration of gene function in the overall framework [11]. Genes' functions should be studied not only from a structural level but also from a network level. Genes affect each other and work together in intricate networks, which consequently contain new functions that cannot be fully revealed by the DNA sequence.

The S100 proteins are a group of calcium-binding proteins with low molecular weight (10–12 kDa). Its amino acid sequence is highly conserved in vertebrates [41]. S100 proteins share a high degree of homology with calmodulin and other EF-Hand calcium binding proteins [41]. From the biological function, specific expression and chromosomal localization in tumor of S100 protein family and the intimate relation between S100 protein and tumor can be found. Recently, studies have indicated that S100A11 (S100C) can serve as

a tumor suppressor protein in some tumors and a tumor promoter in other tumors [42]. S100A11 is upregulated in breast cancer, prostate cancer, and nonsmall cell lung cancer, where it promotes tumor metastasis and invasion [43, 44]. On the contrary, S100A11 acts as a tumor suppressor in urinary bladder and renal carcinoma [45]. Our experimental results present an upregulated expression of S100A11 in gastric cancer. As a candidate tumor suppressor protein, the expression of S100A2 is significantly lower in a variety of malignancies, such as breast, liver, prostatic, and esophageal cancer [46–49]. Studies have indicated that S100A2 can inhibit cell proliferation and invasion and act as a tumor suppressor involved in the occurrence and metastasis of gastric carcinoma [50], which is in agreement with our findings. Through analysis of expression of S100A11 and S100A2 in gastric cancer, both of which contained EF-Hand structure, it was verified that p42.3 could participate in the occurrence and development of gastric cancer from both consistent and opposite to the p42.3 effect direction.

Currently, there are various ways to compare protein structures, each with their own advantages and disadvantages [51]. By analyzing the structure of the proteins, most of them calculate the similarity value of a pair of proteins by applying a mathematical algorithm. That is, from the spatial conformation of protein, they all have only analyzed the characteristics of spatial structure of proteins. The similarity of proteins in other aspects was not taken into account. For example, element P and element S are crucial to the functions of proteins. Using the multiparameter comprehensive comparison method, this study not only compared the differences between the two proteins in the spatial atomic density but also considered the similarity of many other aspects. When conducting the weighted summation of each parameter, the weights used all came from training of diverse data not from subjective weighting. It guarantees the accuracy of weight of each parameter and avoids the mistakes that some parameter is of little importance to the overall similarity but with high weight. Consequently, the similarity of two proteins was figured out more accurately. All the process of calculation was carried out by the M file compiled by MATLAB. Batch comparison of any amount of proteins could be carried out easily and quickly.

5. Conclusions

Here, the ligand-binding model of the EF-hand structure of p42.3 was successfully predicted. Meanwhile, a Bayesian network using the corresponding mathematical algorithm was constructed and optimized to predict the most likely pathway. On the other hand, molecular biology experiments indicated that p42.3 and S100A11 may be with the commonplace in character, and this provided a hypothesis for us to conduct further research. In a word, our findings provide important research directions for exploring the mechanism of action of p42.3 in gastric cancer.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Thanks should be given to workmates of the authors within the same department for their general support. In addition, this work was supported by the National Natural Science Foundation of China (60971110), Science and Technology Corporation Project of Henan Province (122106000042), and Open Science and Technology Corporation Project of Henan Province in 2013 (132106000064).

References

[1] B. G. Zhang, J. F. Li, B. Q. Yu, Z. G. Zhu, B. Y. Liu, and M. Yan, "microRNA-21 promotes tumor proliferation and invasion in gastric cancer by targeting PTEN," *Oncology Reports*, vol. 27, no. 4, pp. 1019–1026, 2012.

[2] G.-H. Zhao, T.-C. Li, L.-H. Shi et al., "Relationship between inactivation of p16 gene and gastric carcinoma," *World Journal of Gastroenterology*, vol. 9, no. 5, pp. 905–909, 2003.

[3] Y.-H. Seo, Y.-E. Joo, S.-K. Choi, J.-S. Rew, C.-S. Park, and S.-J. Kim, "Prognostic significance of p21 and p53 expression in gastric cancer," *The Korean Journal of Internal Medicine*, vol. 18, no. 2, pp. 98–103, 2003.

[4] B. Xiao, E.-D. Zhu, N. Li et al., "Increased miR-146a in gastric cancer directly targets SMAD4 and is involved in modulating cell proliferation and apoptosis," *Oncology Reports*, vol. 27, no. 2, pp. 559–566, 2012.

[5] E. Rippa, G. la Monica, R. Allocca, M. F. Romano, M. de Palma, and P. Arcari, "Overexpression of gastrokine 1 in gastric cancer cells induces Fas-mediated apoptosis," *Journal of Cellular Physiology*, vol. 226, no. 10, pp. 2571–2578, 2011.

[6] Y.-Y. Du, D.-Q. Dai, and Z. Yang, "Role of RECK methylation in gastric cancer and its clinical significance," *World Journal of Gastroenterology*, vol. 16, no. 7, pp. 904–908, 2010.

[7] L. Zhang, Y. Hou, H. Ashktorab et al., "The impact of C-MYC gene expression on gastric cancer cell," *Molecular and Cellular Biochemistry*, vol. 344, no. 1-2, pp. 125–135, 2010.

[8] C. Holmberg, B. Ghesquière, F. Impens et al., "Mapping proteolytic processing in the secretome of gastric cancer-associated myofibroblasts reveals activation of MMP-1, MMP-2, and MMP-3," *Journal of Proteome Research*, vol. 12, no. 7, pp. 3413–3422, 2013.

[9] X. Xu, W. Li, X. Fan et al., "Identification and characterization of a novel p42.3 gene as tumor-specific and mitosis phase-dependent expression in gastric cancer," *Oncogene*, vol. 26, no. 52, pp. 7371–7379, 2007.

[10] W. Sun, *Expression Change and Biological Significance of P42.3 in Gastric Mucosal Lesion*, Beijing Cancer Hospital, Beijing, China, 2012.

[11] J. Zhang, C. Lu, Z. Shang, R. Xing, L. Shi, and Y. Lv, "p42.3 gene expression in gastric cancer cell and its protein regulatory network analysis," *Theoretical Biology and Medical Modelling*, vol. 9, no. 1, article 53, 2012.

[12] M. F. Shao, *Threading Method Research on Protein Structure Prediction*, Graduate School of Chinese Academy of Sciences, 2011.

[13] C. Sandeep, V. Ravindra, J. Basuthkar et al., "Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms," *F1000 Research*, vol. 2, p. 211, 2013.

[14] Y. Li, A. Roy, and Y. Zhang, "HAAD: a quick algorithm for accurate prediction of hydrogen atoms in protein structures," *PLoS ONE*, vol. 4, no. 8, Article ID e6701, 2009.

[15] I. Kato, K. Narita, and I. Fuke, "Topological considerations in protein structure. I. Disulfide linkage," *Biopolymers*, vol. 4, no. 7, pp. 737–746, 1966.

[16] L. M. Gregoret, S. D. Rader, R. J. Fletterick, and F. E. Cohen, "Hydrogen bonds involving sulfur atoms in proteins," *Proteins: Structure, Function and Genetics*, vol. 9, no. 2, pp. 99–107, 1991.

[17] M. N. Wass, L. A. Kelley, and M. J. E. Sternberg, "3DLigandSite: predicting ligand-binding sites using similar structures," *Nucleic Acids Research*, vol. 38, no. 2, pp. W469–W473, 2010.

[18] E. Santonico, S. Panni, M. Falconi, L. Castagnoli, and G. Cesareni, "Binding to DPF-motif by the POB1 EH domain is responsible for POB1-Eps15 interaction," *BMC Biochemistry*, vol. 8, article 29, 2007.

[19] S. Koshiba, T. Kigawa, J. Iwahara, A. Kikuchi, and S. Yokoyama, "Solution structure of the Eps15 homology domain of a human POB1 (partner of RalBP1)," *FEBS Letters*, vol. 442, no. 2-3, pp. 138–142, 1999.

[20] J. F. Harper, M. R. Sussman, G. E. Schaller, C. Putnam-Evans, H. Charbonneau, and A. C. Harmon, "A calcium-dependent protein kinase with a regulatory domain similar to calmodulin," *Science*, vol. 252, no. 5008, pp. 951–954, 1991.

[21] A. M. Weljie and H. J. Vogel, "Unexpected structure of the Ca^{2+}-regulatory region from soybean calcium-dependent protein kinase-α," *The Journal of Biological Chemistry*, vol. 279, no. 34, pp. 35494–35502, 2004.

[22] S. Nakamura, T. Takemura, L. Tan et al., "Small GTPase RAB45-mediated p38 activation in apoptosis of chronic myeloid leukemia progenitor cells," *Carcinogenesis*, vol. 32, no. 12, pp. 1758–1772, 2011.

[23] M. Shintani, M. Tada, T. Kobayashi, H. Kajiho, K. Kontani, and T. Katada, "Characterization of Rab45/RASEF containing EF-hand domain and a coiled-coil motif as a self-associating GTPase," *Biochemical and Biophysical Research Communications*, vol. 357, no. 3, pp. 661–667, 2007.

[24] Y. Sudhakar Babu, C. E. Bugg, and W. J. Cook, "Structure of calmodulin refined at 2.2 Å resolution," *Journal of Molecular Biology*, vol. 204, no. 1, pp. 191–204, 1988.

[25] M. Koller, B. Schnyder, and E. E. Strehler, "Structural organization of the human CaMIII calmodulin gene," *Biochimica et Biophysica Acta—Gene Structure and Expression*, vol. 1087, no. 2, pp. 180–189, 1990.

[26] M. Koller and E. E. Strehler, "Functional analysis of the promoters of the human CaMIII calmodulin gene and of the intronless gene coding for a calmodulin-like protein," *Biochimica et Biophysica Acta*, vol. 1163, no. 1, pp. 1–9, 1993.

[27] H. J. Yuasa, J. A. Cox, and T. Takagi, "Genomic structure of the amphioxus calcium vector protein," *Journal of Biochemistry*, vol. 126, no. 3, pp. 572–577, 1999.

[28] I. Théret, S. Baladi, J. A. Cox, H. Sakamoto, and C. T. Craescu, "Sequential calcium binding to the regulatory domain of calcium vector protein reveals functional asymmetry and a novel mode of structural rearrangement," *Biochemistry*, vol. 39, no. 27, pp. 7920–7926, 2000.

[29] J. Jia, S. Tarabykina, C. Hansen, M. Berchtold, and M. Cygler, "Structure of apoptosis-linked protein ALG-2: insights into Ca^{2+}-induced changes in penta-EF-hand proteins," *Structure*, vol. 9, no. 4, pp. 267–275, 2001.

[30] J. Song, Q. Zhao, S. Thao, R. O. Frederick, and J. L. Markley, "Letter to the editor: solution structure of a calmodulin-like calcium-binding domain from *Arabidopsis thaliana*," *Journal of Biomolecular NMR*, vol. 30, no. 4, pp. 451–456, 2004.

[31] E. McCormack and J. Braam, "Calmodulins and related potential calcium sensors of Arabidopsis," *New Phytologist*, vol. 159, no. 3, pp. 585–598, 2003.

[32] M. W. Berchtold, C. W. Heizmann, and K. J. Wilson, "Primary structure of parvalbumin from rat skeletal muscle," *European Journal of Biochemistry*, vol. 127, no. 2, pp. 381–389, 1982.

[33] T. C. Williams, D. C. Corson, K. Oikawa, W. D. McCubbin, C. M. Kay, and B. D. Sykes, "1H NMR spectroscopic studies of calcium-binding proteins. 3. Solution conformations of rat apo-α-parvalbumin and metal-bound rat α-parvalbumin," *Biochemistry*, vol. 25, no. 7, pp. 1835–1846, 1986.

[34] D. M. Watterson, F. Sharief, and T. C. Vanaman, "The complete amino acid sequence of the Ca^{2+}-dependent modulator protein (calmodulin) of bovine brain," *Journal of Biological Chemistry*, vol. 255, no. 3, pp. 962–975, 1980.

[35] K. Ogura, H. Kumeta, K. Takahasi et al., "Solution structures of yeast *Saccharomyces cerevisiae* calmodulin in calcium- and target peptide-bound states reveal similarities and differences to vertebrate calmodulin," *Genes to Cells*, vol. 17, no. 3, pp. 159–172, 2012.

[36] J. L. Gifford, M. P. Walsh, and H. J. Vogel, "Structures and metal-ion-binding properties of the Ca^{2+}-binding helix-loop-helix EF-hand motifs," *Biochemical Journal*, vol. 405, no. 2, pp. 199–221, 2007.

[37] A. C. Drohat, D. M. Baldisseri, R. R. Rustandi, and D. J. Weber, "Solution structure of calcium-bound rat S100B($\beta\beta$) as determined by nuclear magnetic resonance spectroscopy," *Biochemistry*, vol. 37, no. 9, pp. 2729–2740, 1998.

[38] P. A. Hessian and L. Fisher, "The heterodimeric complex of MRP-8 (S100A8) and MRP-14 (S100A9) antibody recognition, epitope definition and the implications for structure," *European Journal of Biochemistry*, vol. 268, no. 2, pp. 353–363, 2001.

[39] J. Liu, Q. Zheng, Y. Deng, C.-S. Cheng, N. R. Kallenbach, and M. Lu, "A seven-helix coiled coil," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 42, pp. 15457–15462, 2006.

[40] O. V. Moroz, A. A. Antson, G. N. Murshudov et al., "The three-dimensional structure of human S100A12," *Acta Crystallographica Section D: Biological Crystallography*, vol. 57, no. 1, pp. 20–29, 2001.

[41] M. Sakaguchi, H. Sonegawa, H. Murata et al., "S100A11, an dual mediator for growth regulation of human keratinocytes," *Molecular Biology of the Cell*, vol. 19, no. 1, pp. 78–85, 2008.

[42] S. Hiratsuka, A. Watanabe, H. Aburatani, and Y. Maru, "Tumour-mediated upregulation of chemoattractants and recruitment of myeloid cells predetermines lung metastasis," *Nature Cell Biology*, vol. 8, no. 12, pp. 1369–1375, 2006.

[43] X. G. Liu, X. P. Wang, W. F. Li et al., "Ca^{2+}-binding protein S100A11: a novel diagnostic marker for breast carcinoma," *Oncology Reports*, vol. 23, no. 5, pp. 1301–1308, 2010.

[44] R. Yao, D. D. Davidson, A. Lopez-Beltran, G. T. MacLennan, R. Montironi, and L. Cheng, "The S100 proteins for screening and prognostic grading of bladder cancer," *Histology and Histopathology*, vol. 22, no. 7–9, pp. 1025–1032, 2007.

[45] S. W. Lee, C. Tomasetto, K. Swisshelm, K. Keyomarsi, and R. Sager, "Down-regulation of a member of the S100 gene family in mammary carcinoma cells and reexpression by azadeoxycytidine treatment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 6, pp. 2504–2508, 1992.

[46] G. Feng, X. Xu, E. M. Youssef, and R. Lotan, "Diminished expression of S100A2, a putative tumor suppressor, at early stage

of human lung carcinogenesis," *Cancer Research*, vol. 61, no. 21, pp. 7999–8004, 2001.

[47] I. D. Kyriazanos, M. Tachibana, D. K. Dhar et al., "Expression and prognostic significance of S100A2 protein in squamous cell carcinoma of the esophagus," *Oncology Reports*, vol. 9, no. 3, pp. 503–510, 2002.

[48] S. Gupta, T. Hussain, G. T. MacLennan, P. Fu, J. Patel, and H. Mukhtar, "Differential expression of S100A2 and S100A4 during progression of human prostate adenocarcinoma," *Journal of Clinical Oncology*, vol. 21, no. 1, pp. 106–112, 2003.

[49] X. Su, Z. You, and Z. Zheng, "Effects of gene silencing in S100A2 and E-cadherin by RNA interference on gastric cancer cell proliferation and invasiveness," *Journal of China Medical University*, vol. 41, no. 4, pp. 297–306, 2012.

[50] L. Ming, S. Xianzhong, and Y. Min, "Progress in Protein stuction prediction," *Biotechnology*, vol. 19, no. 3, pp. 87–90, 2009.

[51] M. N. Lang, *The Method and Application of Protein Stucture Alignment*, Jilin University, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 893507, 18 pages
http://dx.doi.org/10.1155/2015/893507

Research Article

Unified Modeling of Familial Mediterranean Fever and Cryopyrin Associated Periodic Syndromes

Yasemin Bozkurt,[1] Alper Demir,[1] Burak Erman,[1] and Ahmet Gül[2]

[1] *Computational and Quantitative Biology Lab, Koc University, 34450 Istanbul, Turkey*
[2] *Division of Rheumatology, Department of Internal Medicine, Istanbul Faculty of Medicine, Istanbul University, 34093 Istanbul, Turkey*

Correspondence should be addressed to Yasemin Bozkurt; yabozkurt@ku.edu.tr

Received 18 September 2014; Accepted 24 November 2014

Academic Editor: Francesco Camastra

Familial mediterranean fever (FMF) and Cryopyrin associated periodic syndromes (CAPS) are two prototypical hereditary autoinflammatory diseases, characterized by recurrent episodes of fever and inflammation as a result of mutations in *MEFV* and *NLRP3* genes encoding Pyrin and Cryopyrin proteins, respectively. Pyrin and Cryopyrin play key roles in the multiprotein inflammasome complex assembly, which regulates activity of an enzyme, Caspase 1, and its target cytokine, IL-1β. Overproduction of IL-1β by Caspase 1 is the main cause of episodic fever and inflammatory findings in FMF and CAPS. We present a unifying dynamical model for FMF and CAPS in the form of coupled nonlinear ordinary differential equations. The model is composed of two subsystems, which capture the interactions and dynamics of the key molecular players and the insults on the immune system. One of the subsystems, which contains a coupled positive-negative feedback motif, captures the dynamics of inflammation formation and regulation. We perform a comprehensive bifurcation analysis of the model and show that it exhibits three modes, capturing the Healthy, FMF, and CAPS cases. The mutations in Pyrin and Cryopyrin are reflected in the values of three parameters in the model. We present extensive simulation results for the model that match clinical observations.

1. Introduction

Inflammation is the organisms' protective response to remove the insult and initiate the healing. Inflammatory reaction triggered by the recognition of pathogen or damage associated molecular patterns (PAMP or DAMP) usually results in the elimination of the stimuli and protection of the body integrity. Inflammasomes, multiprotein oligomers, form a link between the sensing of microbial products and developing an immune response to the danger signals via the regulation of the secretion of proinflammatory cytokines [1]. Inflammasome activation is highly regulated and the level of activation is critical to generate a proper immune response [2]. However, when the insult level is higher than the immune system can handle or due to problems in the control mechanisms regulating the reaction against these harmful stimuli, inflammatory disorders arise leading to various forms of discomfort, tissue or organ damage, or even death, depending on the magnitude of the insult or the nature of the insufficiency in the control mechanisms.

Quantitative approaches have recently been used in order to attain a better understanding of the immune system and immunity related diseases. Quantitative models have been used in order to understand how the immune cells and key molecular players interact with each other to constitute the total activity of the immune system [3]. Mathematical models are, of course, not yet detailed enough to describe the quantitative behavior of all immune cells and molecular species involved. However, the overall observed clinical behaviors can be modeled using several well established tools of chemical kinetics and dynamical systems theory. These models can then help explain observed clinical behavior and may further be used in designing custom drug therapies for immune system related diseases.

Autoinflammatory syndromes, also called hereditary periodic fever syndromes, are a class of disorders characterized by recurrent episodes of inflammation in tissues such as joints, skin, gut, and eyes, generally accompanied by fever, in the absence of any adaptive immune response by

cytotoxic T cells or pathogenic autoantibodies [4]. Familial mediterranean fever (FMF) and Cryopyrin associated periodic syndromes (CAPS), which we consider in this paper, are among the prototypical members of this autoinflammatory disease class. FMF is caused by inherited loss-of-function mutations in Pyrin, and CAPS by gain-of-function mutations in Cryopyrin, two proteins that play key roles in the control and regulation of inflammation along with an enzyme, Caspase 1, and its target cytokine, IL-1β. Normally, triggered by a microbial or a sterile insult, Cryopyrin forms an inflammasome that converts precursor procaspase 1 into active Caspase 1. Once activated, Caspase 1 proteolytically cleaves proIL-1β into an active IL-1β with a proinflammatory effect [5]. Pyrin, on the other hand, acts as an anti-inflammatory mediator that controls and prevents Cryopyrin inflammasome formation by interacting with both Caspase 1 and Cryopyrin protein (Nlrp3). Mutated Pyrin can no longer act as an effective suppressor of inflammasome formation [6]. Mutations in Cryopyrin, on the other hand, result in elevated inflammasome formation even in the absence of a trigger or insult [7]. Overproduction of IL-1β by Caspase 1 as a result is the main cause of fever and inflammatory episodes in FMF and CAPS [8].

In this paper, we present a unifying dynamical model for FMF and CAPS in the form of coupled nonlinear ordinary differential equations. The model is composed of two subsystems. The first subsystem captures the interactions and dynamics of Pyrin, Cryopyrin, procaspase 1, Caspase 1, and inflammasome formation, including the effect of triggers. The second subsystem, which contains a coupled positive-negative feedback motif, captures the dynamics of IL-1β, its receptor, and receptor antagonist, including Caspase 1 independent cleavage of IL-1β. The two subsystems are coupled via the key player Caspase 1. We perform a comprehensive bifurcation analysis of the model and show that it exhibits three modes, capturing the Healthy, FMF, and CAPS cases. The mutations in Pyrin and Cryopyrin are reflected in the values of three parameters in the model. We then present extensive simulation results for the model, which match clinical observations. In the presence of a trigger, there is a normal increase in inflammation initiators in the Healthy mode. In FMF, the response to trigger introduction is more intense and a severe inflammatory cascade is activated. In CAPS, on the other hand, even in the absence of a trigger, periodically recurring, severe inflammatory episodes are observed. The proposed model also explains why a procaspase 1 inhibitor can be effective in treating FMF, but not CAPS, for which drugs that directly inhibit IL-1β are used.

The paper is organized as follows. First, background information on FMF and CAPS pathogeneses is given based on an extensive literature review. Next, the unified model that captures key aspects of FMF and CAPS is introduced. Detailed bifurcation analyses are performed on the model. Simulation results obtained from the proposed model are presented and discussed. Finally, conclusions are drawn.

2. Biological Background on FMF

FMF is a hereditary autoinflammatory disease associated with mutations in the *MEFV* (mediterranean fever) gene [9].

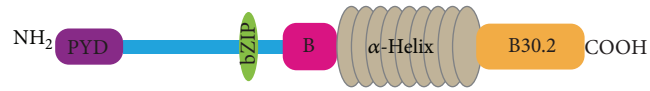


Figure 1: Structure of Pyrin [14].

MEFV gene is located in the short arm of chromosome 16 and encodes a 781-amino acid protein called Pyrin [10]. FMF is inherited in an autosomal recessive fashion (although there are some reported heterozygotes manifesting FMF) and mostly affects people originating from around the Mediterranean Sea, for example, mostly Armenians, Jews, Turks, and Arabs [11]. The main symptoms are irregular attacks of fever (38–41°C), abdominal pain, and chest and joint pain. The duration of the flare (1–3 days) and remission (weeks or months) period varies considerably. The acute phase response includes an increase in the white blood cell count, erythrocyte sedimentation rate, fibrinogen, C-reactive protein, serum amyloid A, and phospholipase A_2 [12]. Acute phase response increases during the attack and decreases in the attack-free period.

Although Pyrin's structure and function have not been completely elucidated yet, it is clear that it has a role in the control of the production of IL-1β and other proinflammatory cytokines. Mutations in exon 10 of the *MEFV* gene leads to ineffective Pyrin function. Pyrin is expressed mainly in neutrophils, eosinophils, monocytes, dendritic cells, and synovial and peritoneal fibroblasts (mostly in the innate immune system cells) and regulates Caspase 1 activation [13]. Caspase 1 is responsible for cleavage of proIL-1β and the subsequent release of bioactive IL-1β [5]. IL-1β is a very important pyrogenic cytokine in the inflammatory process. In FMF, overstimulation of IL-1β production causes increased inflammatory and febrile response.

Pyrin is composed of 5 domains: Pyrin (PYD), bZIP transcription factor basic, B-box zinc finger, α-helical (coiled coil), and B30.2 (PRY-SPRY) domains [14]. The structure of Pyrin is shown in Figure 1.

Cryopyrin inflammasome, inflammasome of interest in FMF, is a multiprotein oligomer containing ASC, procaspase 1, Cryopyrin, and Cardinal proteins [15]. Inflammasome formation takes place, depending on the type of the inflammasome, in response to the specific pathogen associated molecular patterns (PAMP) such as lipopolysaccharide (LPS), dsRNA, and peptidoglycan or damage associated molecular patterns (DAMP) such as ATP and uric acid. This protein complex regulates and proteolytically activates important cytokines such as IL-1β and IL-18 through the activation of Caspase 1 [16].

Cryopyrin protein is encoded by the *NLRP3* gene and belongs to the nucleotide-binding oligomerization domain receptors (NLR) protein family. Cryopyrin can detect a variety of danger signals such as dsRNA, uric acid, bacterial ligands, and imidazoquinolines [17]. PYD is a common domain found in proteins such as Pyrin, Cryopyrin, and some other NLR proteins. This domain allows the homotypic interaction with other proteins containing PYD. For example, since ASC contains a PYD, Pyrin and Cryopyrin interact

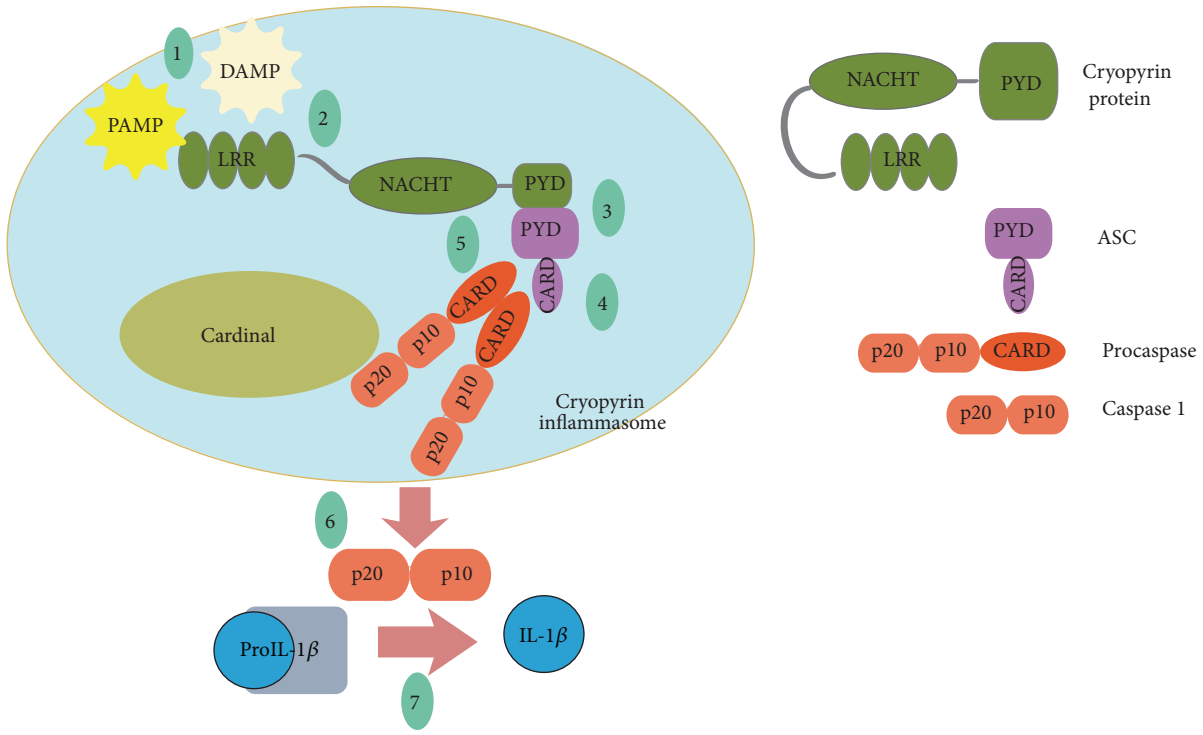


FIGURE 2: Cryopyrin inflammasome formation steps. (1) With DAMP or PAMP, inflammasome formation process is triggered. (2) A conformational change occurs in Cryopyrin protein due to the trigger. (3) Interaction between PYDs of ASC and Cryopyrin protein becomes possible after the conformational change. (4) CARD of ASC and procaspase 1 interact. (5) Cardinal brings another procaspase 1 to the system. (6) Induced proximity mediated autocatalysis results in the activation of Caspase 1. (7) Caspase 1 bioactivates IL-1β by cleavage.

with ASC through this domain. Cryopyrin is an essential component of the Cryopyrin inflammasome.

Without any trigger (DAMP or PAMP), interaction between PYD of Cryopyrin and ASC is not possible due to the nominal structure of Cryopyrin. With the presence of a stimulus, the interaction of the PYD domain of Cryopyrin and ASC becomes possible. ASC interacts with the PYD of Cryopyrin through its N terminal PYD. C terminal CARD domain of ASC is in interaction with the CARD domain of procaspase 1. The Cardinal brings another procaspase 1 to the system. Procaspase 1 molecules become closer to each other (proximity-induced-mediated-autocatalysis) which results in the proteolytic activation and release of active Caspase 1 protein with two chains of p20 and p10 [14]. ProIL-1β is a precursor which is proteolytically cleaved to a shorter, active form, IL-1β, by Caspase 1. Figure 2 summarizes the inflammasome formation steps. Mature IL-1β molecules induce expression of other cytokines. Secondary cytokines such as IL-6 recruit immune cells to the site of inflammation. There is a balance between the levels of activation and inhibition of inflammasome formation. Although IL-1β molecules help combat the infection, spontaneous or triggered overproduction of IL-1β causes adverse effects that are associated with inflammatory diseases such as FMF and CAPS.

There is evidence for both the anti-inflammatory [6, 14, 18] and proinflammatory roles of Pyrin [19, 20]. Anti-inflammatory role includes the inhibition of Caspase 1 action and IL-1β processing. Proinflammatory role portrays Pyrin as a constituent of an unknown inflammasome resulting in more IL-1β activation. Since there is more literature for the suppressive effect of Pyrin on inflammation, throughout this paper, we will take Pyrin as an anti-inflammatory mediator.

Some of the mechanisms on the contribution of Pyrin in inhibiting Caspase 1 activation are listed here.

(i) ASC binds to Pyrin through its PYD. Thus, ASC cannot participate in the formation of the Cryopyrin inflammasome. Since Cryopyrin inflammasome assembly does not take place, Caspase 1 is not activated, and hence IL-1β is not produced (step 1 in Figure 3).

(ii) Due to the interactions between B30.2 domain of WT Pyrin and the p20 and p10 domains of Caspase 1, Pyrin can directly bind to both procaspase 1 and Caspase 1 and hence indirectly prevent IL-1β activation. In FMF, the most common mutations (M608I, M694V, and V726A) are generally in the C terminal B30.2 domain of Pyrin. If there is a mutation in B30.2 domain of Pyrin, the interaction between this domain and p20 and p10 subunits of Caspase 1 is diminished [6]. p20 and p10 form a heterodimer; that is, Caspase 1 becomes activated (step 2 in Figure 3).

(iii) Pyrin also interacts directly with proIL-1β which is an additional inhibition factor on IL-1β secretion.

Maturation and secretion of IL-1β require two distinct signals, priming and activation [21]. With the priming signal,

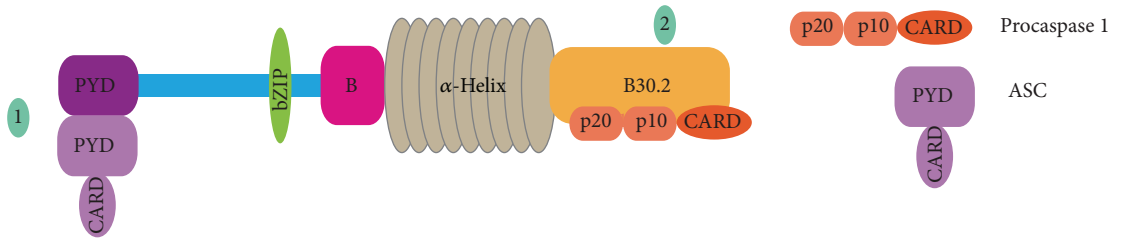


Figure 3: Anti-inflammatory role of Pyrin.

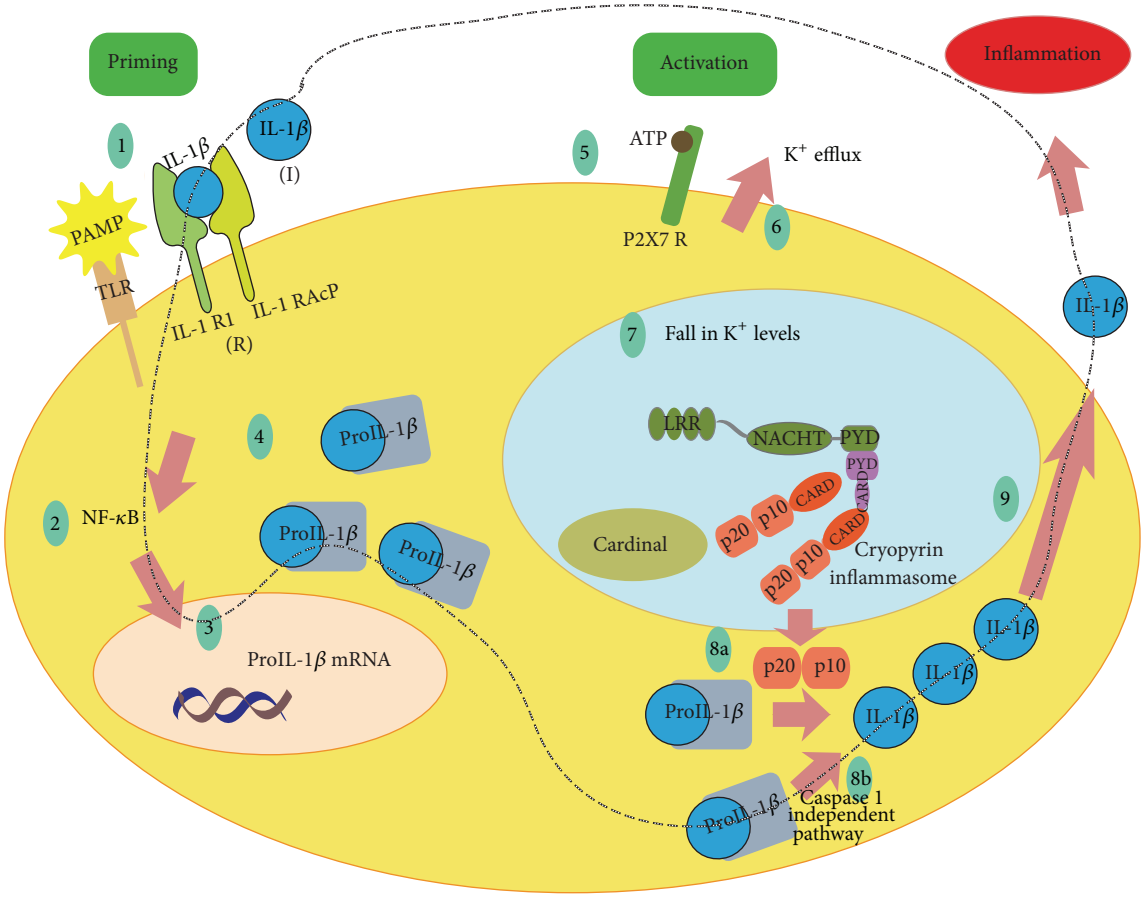


Figure 4: Maturation and secretion of IL-1β requires two signals. Binding of ligand (PAMP or IL-1β) to TLR (1) activates NF-κB pathway (2). NF-κB signaling results in proIL-1β transcription and secretion (3 and 4). Binding of ATP to the P2X7 receptor is considered as the activation signal (5) which causes K^+ efflux (6) and subsequently a decrease in intracellular K^+ concentration (7). Together with PAMP or DAMP, fall in K^+ acts as the initiator of the inflammasome formation (7). ProIL-1β is cleaved by the product of the inflammasome, Caspase 1 (8a). However, Caspase 1 is not the only protease that can cleave proIL-1β (8b). Mature IL-1β is secreted and inflammation process is then started (9). Dashed path shows the positive feedback relationship between the free IL-1β (I) and bound IL-1β (R).

transcription and translation of proIL-1β take place. Activation signal results in the assembly of Cryopyrin inflammasome, Caspase 1 activation, and subsequently IL-1β activation and release.

IL-1R1 is a member of TLR superfamily [22]. Binding of mature IL-1β to IL-1R1 results in the downstream signaling of NF-κB pathway leading to the production of proIL-1β [23]. Also, recognition of the microbial ligands by TLRs initiates proIL-1β transcription. Microbial ligands and endogenous cytokines are the priming signals for proIL-1β production (step 1 in Figure 4).

Activation signal (such as ion/membrane perturbations, reactive oxygen species (ROS), pore-forming toxins, crystals, and ATP) promotes the indirect activation of IL-1β secretion [24]. In Figure 4, ATP activating Cryopyrin inflammasome is shown as a representative example. Activation of P2X7

receptor by extracellular ATP results in the K^+ efflux. Fall in the intracellular K^+ levels triggers Cryopyrin inflammasome formation [25]. Inflammasome oligomerization leads to activation of Caspase 1, followed by the maturation and secretion of IL-1β. We should also note here the possible effect of other proteases. Caspase 1 specifically cleaves 31-kd precursor proIL-1β to 17-kd biologically active IL-1β. Caspase 1 belongs to the caspase family of cysteine proteases. Some of the other members of protease family, such as neutrophil serine proteases, proteinase 3, and mast cell derived serine proteases, can also cleave proIL-1β [26]. This suggests a redundancy in the mechanisms of IL-1β processing. The role of other proteases is much more apparent in some models of inflammatory diseases such as arthritis animal models [27]. Even though in FMF and CAPS overproduction of IL-1β is mostly related to the high levels of Caspase 1 concentrations, we will also include the possible effect of other proteases in the model.

In Figure 4, the dashed path shows the positive feedback relationship between the bound IL-1β (R) and free IL-1β (I). Binding of IL-1β to its receptor IL-1R1 results in more proIL-1β transcription and thus more IL-1β. In FMF patients, inflammasome activation is higher when compared to the healthy patients [6]. Overactivation of inflammasome causes overproduction of IL-1β. This process is partially self-sustaining; that is, IL-1β causes more IL-1β production (either via Caspase 1 or through Caspase 1 independent pathway) since IL-1β is one of the primary signals for IL-1β maturation and secretion.

IL-1β and other proinflammatory innate cytokines are the main mediators of autoinflammatory diseases. The mechanism for the cytokine induction pathway is not fully clear. FMF symptoms develop as a result of mutations disrupting functions of Pyrin, which result in overproduction of IL-1β. However, not only IL-1β but also a downstream cytokine IL-6 is essential in attack and fever development in FMF. High levels of IL-1β cause an increase in IL-6 [28].

IL-6 has both proinflammatory and anti-inflammatory roles. IL-6 changes the fever set point in the hypothalamus, responsible for the high fever in the attack period of FMF. Additionally, it mediates acute phase proteins. Levels of leukocyte count, ESR, CRP, sIL-2R, IL-6, and IL-10 increase considerably during an attack [29]. High levels of IL-6 may have suppressive effect in the production of other cytokines such as IL-1β and TNF-α while activating the antagonist of IL-1β, IL-1Ra [30].

IL-1β is the most critical endogenous pyrogen in autoinflammatory diseases and its control is extremely important. Its receptor binding gives some clues regarding the possible mechanisms that may lead to ending the active phase in FMF. There are natural inhibitors of IL-1 such as IL-1Ra, decoy receptor of IL-1R2, and other soluble receptors. In knockout mice lacking IL-1Ra, excessive inflammation has been observed. These mice develop spontaneous joint inflammation, vasculitis, and skin inflammation [31]. The biological activities of IL-1 are initiated by binding of IL-1α and/or IL-1β to the same receptor, namely, IL-1R1. IL-1R1 exists on the surface of a wide variety of cells. Binding of IL-1α and/or IL-1β causes a conformational change in IL-1R1 and recruits an accessory protein, IL-1RAcP [32]. Once IL-1/IL-1R1/IL-1RAcP complex is formed, signaling through other cascades such as NF-κB is activated. This is the only active form of this complex. Other scenarios fail to generate an active signal. Some of the cases that may cause no signal are listed here.

(i) *IL-1Ra.* It competes with IL-1β for binding to its receptor IL-1R1 and prevents binding of IL-1β and the subsequent downstream signaling [33].

(ii) *IL-1R2.* It is a decoy receptor similar to IL-1R1. IL-1β binds to IL-1R2 instead of IL-1R1. Cascades of other cytokines are not activated when IL-1β does not bind to IL-1R1 and transmit downstream signals [34].

(iii) *SIGIRR.* It prevents IL-1R1/IL-1RAcP heterodimerization [35].

(iv) *Soluble IL-1R1 or R2.* They are soluble receptors that can bind to IL-1 and IL-1RAcP but are incapable of propagating a signal [36].

3. Biological Background on CAPS

CAPS comprises a spectrum of rare autoinflammatory syndromes, ranging from FCAS (familial cold autoinflammatory syndrome) and MWS (Muckle-Wells syndrome) to NOMID (also called as CINCA) (neonatal-onset multisystem inflammatory disease). These diseases are caused by autosomal dominantly inherited gain-of-function or *de novo* mutations in various domains of *NLRP3* gene (also known as *CIAS1* gene), located on chromosome 1 (1q44), which encodes a Pyrin like protein, Cryopyrin (or Nlrp3 protein). Approximately 100 different mutations (mostly missense mutations) in exon 3 of this gene have been identified. Cryopyrin is expressed in monocytes, neutrophils, and chondrocytes [37]. Cryopyrin is one of the NLR family proteins with a critical role in the regulation of the inflammatory response. CAPS, or Cryopyrinopathies, lead to increased and spontaneous activity of Nlrp3-associated Caspase 1 activating inflammasome, that is, Cryopyrin inflammasome. The more inflammasome formation takes place, the more conversion of proIL-1β into IL-1β occurs. IL-1β not only activates the fever pathway, but also causes pain sensitization and bone and cartilage destruction and activates acute phase response [38]. Although the initial steps of the pathogenesis of FMF and CAPS are different, they become quite similar in terms of the main causative pathway and some of the design strategies in the treatment. IL-1 blocking therapies are applied to CAPS patients as well as colchicine-resistant FMF patients successfully. Similarities between FMF and CAPS pathogeneses are not repeated in this section; only the differences are indicated instead.

In Table 1, CAPS types and their symptoms are listed. The severity of the disease is also indicated.

Although CAPS attacks might occur following the initiation of a stimulus such as cold exposure in the mildest form of CAPS spectrum, that is, FCAS, no trigger is identified or associated in most disease attacks in more severe forms of MWS and NOMID [39]. This suggests that CAPS arise as recurrent episodes of fever even in the absence of any insult

Table 1: CAPS types and their symptoms.

Disease	Major symptoms	Severity
FCAS	Cold-induced urticarial rash and arthralgia Fever after exposure to cold	The mildest
MWS	Shared symptoms with FCAS and NOMID Inflammation is seen without provocation Cold, stress, and exercise also trigger the inflammation	Intermediate
NOMID	Neonatal-onset high fever, persistent rash, aseptic meningitis, mental retardation, sensory deafness, papilledema, arthritis with bone overgrowth, and secondary amyloidosis	The most severe

due to spontaneous activation of Cryopyrin inflammasome, and, in severe cases such as NOMID, an episode-free continuous inflammation can be observed. How exposure to cold in patients with FCAS induces the inflammatory disease flares remains unknown [39].

Similar symptoms and characteristics with different severity among FCAS, MWS, and NOMID suggest that it is possible to capture all of the spectrum in the same model.

4. FMF and CAPS Pathogeneses with a Modeling Perspective

The external causes triggering FMF attacks were studied and tested thoroughly [16, 40]. CAPS flares, on the other hand, seem to be self-sustaining. At least, triggers for CAPS have not been identified in the same extent as for FMF. This suggests that FMF attacks occur as a result of yet unknown external or endogenous triggers. The triggers (PAMP or DAMP) activate inflammasome formation and subsequently overproduction of IL-1β due to malfunctioning Pyrin in FMF patients, as described in Section 2. The immune response is then activated and the disease symptoms are observed in the attack period. When the trigger is deactivated by the immune action, disease enters a quiescent phase, where most of the disease symptoms disappear. CAPS attacks, on the other hand, seem to be related to the autonomous dynamic properties of the key players' relative concentrations. Positive and negative feedback mechanisms cannot keep these concentrations in normal ranges as a result of the mutations in Cryopyrin, which causes periodic flare/remission cycles.

In FMF and CAPS, mutations in the wildtype proteins (Pyrin and Cryopyrin) result in overproduction of IL-1β, stimulating an inflammatory response. As pointed out in Sections 2 and 3, interactions between positive and negative feedback mechanisms determine the overall behavior. Coupled positive and negative feedback loops are a widely seen motif that can exhibit various types of dynamics [41]. Here, we summarize important positive and negative feedback interactions that take part in both FMF and CAPS.

Positive Feedback Mechanisms

(i) Signaling by Receptor-IL-1β complex increases IL-1β levels, through further transcription of proIL-1β.

(ii) Additionally, Caspase 1 independent processing of IL-1β increases the active Receptor-IL-1β complex level.

Negative Feedback Mechanisms

(i) Receptor-IL-1β complex also induces the expression of the antagonist.

(ii) Binding of the antagonist to the receptor does not result in IL-1β signaling and hence decreases the active Receptor-IL-1β complex levels.

The key variables in FMF and CAPS pathogeneses are as follows.

(i) Trigger (T): it represents PAMP and DAMP which activates Cryopyrin inflammasome formation.

(ii) IL-1β (I): it indicates the serum levels of IL-1β.

(iii) Antagonist (A): it represents the total activity that decreases the binding of IL-1β to its receptor (e.g., IL-1Ra, sIL-1R2).

(iv) Receptor (R): it is used to indicate the amount of bound IL-1β, that is, receptor-IL-1β complex.

(v) Caspase 1 (C): it is Caspase 1 level.

(vi) Procaspase 1 (PC): it represents the free (not-Pyrin bound) procaspase 1 levels.

(vii) Pyrin (P): it represents the amount of Pyrin (assumed to be constant).

We consider overproduction of IL-1β as the driving cause of inflammation. PAMP and DAMP are taken as triggers which activate inflammasome formation and the subsequent processes in Pyrin mutants. When trigger is not present, FMF is in its quiescent phase. The action of the regulators of IL-1β levels (IL-1Ra, sIL-1R2, etc.) is lumped and considered as one variable, the antagonist A. Figures 5 and 6 illustrate the most crucial steps we have identified in FMF and CAPS pathogeneses which form the foundation of the mathematical model that will be described in Section 5.

5. Modeling of FMF and CAPS

Motivated by the considerations in Sections 2, 3, and 4, we construct a unifying disease model which captures both FMF and CAPS. Parameters used in the model are listed in Table 2.

The kinetic equations are given in (1a), (1b), (1c), (1d), and (1e) in the form of ordinary differential equations (ODEs) that describe the interactions of the key variables involved.

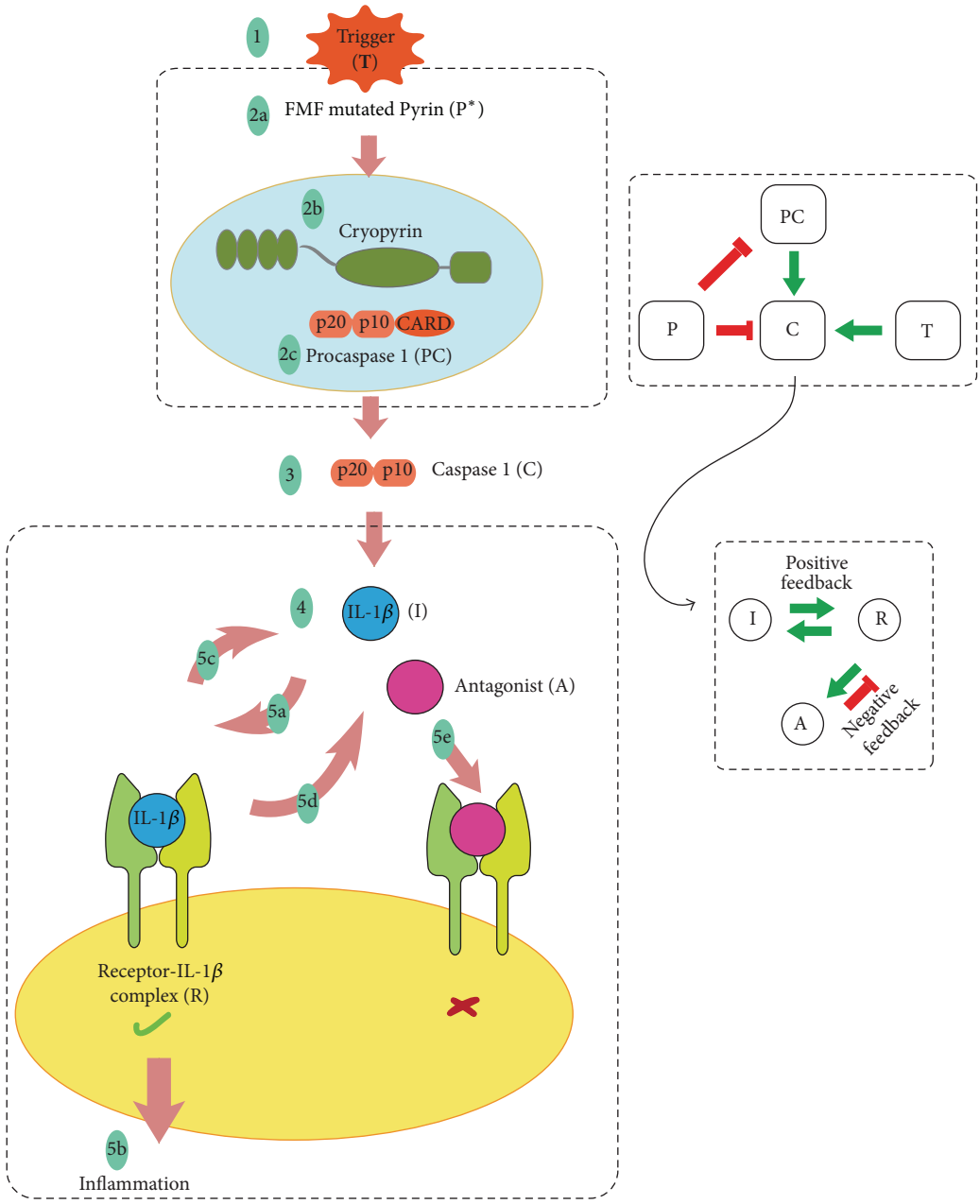


Figure 5: Summary of the pathogenesis of FMF. With trigger, inflammasome formation process is triggered in Pyrin mutants (1 and 2a). Cryopyrin protein and procaspase 1 complex into an inflammasome due to nonfunctioning Pyrin (2b and 2c). Procaspase 1 turns into Caspase 1 by the inflammasome action (3). IL-1β is maturated by Caspase 1 (4). IL-1β binds to the receptor (5a). IL-1β signaling cascade followed by inflammation is activated by the binding of IL-1β (5b). Signaling by the Receptor-IL-1β complex leads to further transcription of proIL-1β, resulting in an increase in IL-1β levels. Here, we had only considered the Caspase 1 independent processing of IL-1β (5c). Signaling by the Receptor-IL-1β complex also stimulates the antagonist production (5d). Binding of antagonist to the receptor does not result in active IL-1β signaling (5e).

The kinetic system is formed as a composition of competitive and noncompetitive inhibition models used in biochemistry [42]. We follow a similar approach in model development as in [41] for the coupled positive and negative feedback processes. Here, to represent the interactions between R, I, and A, we utilize competitive inhibition, due to the fact that the antagonist also binds to the same receptor and is structurally similar to the substrate, IL-1β. The inflammasome formation process is modeled based on noncompetitive inhibition reactions. Procaspase 1 is considered as a substrate. Pyrin, on the other hand, inhibits inflammasome formation by binding to PC and C. The product of the process is C.

Hill effect is also taken into consideration in both inhibition reactions [43]. Consider

$$\frac{d[\mathrm{R}]}{dt} = V_r \cdot \frac{\left([\mathrm{I}]/K_{ir}\right)^n}{1 + \left([\mathrm{I}]/K_{ir}\right)^n + \left([\mathrm{A}]/K_{ar}\right)^n} - k_{dr} \cdot [\mathrm{R}] + k_{br}, \tag{1a}$$

$$\frac{d[\mathrm{I}]}{dt} = V_i \cdot \frac{\left([\mathrm{R}]/K_{ri}\right)^n}{1 + \left([\mathrm{R}]/K_{ri}\right)^n} - k_{di} \cdot [\mathrm{I}] + \alpha \cdot [\mathrm{C}], \tag{1b}$$

$$\frac{d[\mathrm{A}]}{dt} = V_a \cdot \frac{\left([\mathrm{R}]/K_{ra}\right)^n}{1 + \left([\mathrm{R}]/K_{ra}\right)^n} - k_{da} \cdot [\mathrm{A}] + k_{ba}, \tag{1c}$$

$$\frac{d[\mathrm{C}]}{dt} = V_c \cdot \frac{\left([\mathrm{PC}]/K_{pcc}\right)^n}{1 + \left([\mathrm{PC}]/K_{pcc}\right)^n} \cdot \frac{\left([\mathrm{T}]/K_{tc}\right)^n}{1 + \left([\mathrm{T}]/K_{tc}\right)^n} \cdot \frac{1}{1 + \left([\mathrm{P}]/K_{pc}\right)^n} - k_{dc} \cdot [\mathrm{C}] + k_{bc}, \tag{1d}$$

$$\frac{d[\mathrm{PC}]}{dt} = V_{pc} \cdot \frac{1}{1 + \left([\mathrm{P}]/K_{ppc}\right)^n} - k_c \cdot [\mathrm{C}] - k_{dpc} \cdot [\mathrm{PC}] + k_{bpc}. \tag{1e}$$

For all the variables (R, I, A, C, and PC) we have assumed a basal synthesis rate. Since the production rate of I is directly related to the C concentration according to our model, basal synthesis rate of I is not included explicitly; that is, k_{bi} does not appear in (1b). The rate of degradation for a certain variable is assumed to be proportional to the concentration of that variable.

Equation (1a) captures competitive inhibition reactions where I is the substrate and A is the inhibitor. Equation (1b) represents the effect of Caspase 1 independent positive feedback mechanism between R and I. Activation of A by R is captured in (1c).

C, on the other hand, is generated as a result of a noncompetitive binding reaction as represented by (1d). Here, binding is not used in a strict sense, instead it captures the possible interactions. With the initiation of T, PC forms the inflammasome at a maximal rate of V_c. P inhibits inflammasome formation since it binds to PC and C. In inflammasome formation, other proteins such as Cryopyrin, ASC, and Cardinal concentrations are taken as constant and not represented explicitly. Conversion of PC into C indicates the formation of the inflammasome. Decrease in PC levels due to conversion to C is captured in (1e) with a rate constant k_c. C sets the basal synthesis rate for I proportional to a constant α.

5.1. Selection of Parameter Values and Capturing Mutations. The proposed model captures Healthy, FMF, and CAPS cases in a unified manner, via changes in the values of only three model parameters, without the addition or removal of equations to/from the model. In Healthy mode, nominal values of the parameters that are listed in Table 3 are used. Some of these nominal parameter values were simply normalized to 1; others were chosen based on the model parameter values given in [41] and by running extensive bifurcation analyses (described in Section 6) and parameter sweeps. The parameter values were chosen in such a way so that the characteristic and clinical features of both diseased FMF and CAPS systems, along with the healthy system, can be observed in the same model, by modifying only a small subset of the parameters. FMF is a result of a mutation in Pyrin. This mutation is attributed to an increase in the threshold for P to suppress C and PC, captured by two parameters K_{pc} and K_{ppc} in the model. The values of these two parameters are increased to 10 from their nominal value at 1 in order to reflect the FMF mutation. CAPS is associated with a mutation in Cryopyrin and subsequently increased inflammasome formation. This mutation is captured by the parameter V_c, which controls the rate of inflammasome formation in the model. The value of this parameter was increased to 450 from its nominal value of 1.15 in order to reflect the CAPS mutation and observe the clinical features of CAPS in the model. Various members of the CAPS disease family (FCAS, MWS, and NOMID) can also be captured by the model via changes in the value of V_c and by setting the trigger level T to an appropriate level.

Table 2: The parameters used in the model.

	Explanation
V_r	Maximum rate of I and A binding to the receptor
V_i	Strength of positive feedback
V_a	Strength of negative feedback
V_c	Maximum rate to form Cryopyrin inflammasome
V_{pc}	Maximum rate of PC production
K_{ir}	Threshold for I to induce R
K_{ar}	Threshold for A to suppress R
K_{ri}	Threshold for R to induce I
K_{ra}	Threshold for R to induce A
K_{pcc}	Threshold for PC to induce C
K_{tc}	Threshold for T to induce C
K_{pc}	Threshold for P to suppress C
K_{ppc}	Threshold for P to suppress PC
k_{dr}	Degradation rate of R
k_{di}	Degradation rate of I
k_{da}	Degradation rate of A
k_{dc}	Degradation rate of C
k_{dpc}	Degradation rate of PC
k_{br}	Basal synthesis rate of R
k_{ba}	Basal synthesis rate of A
k_{bc}	Basal synthesis rate of C
k_{bpc}	Basal synthesis rate of PC
k_c	Conversion rate of PC into C
n	Hill function cooperativity exponent
α	Proportionality constant of C and I

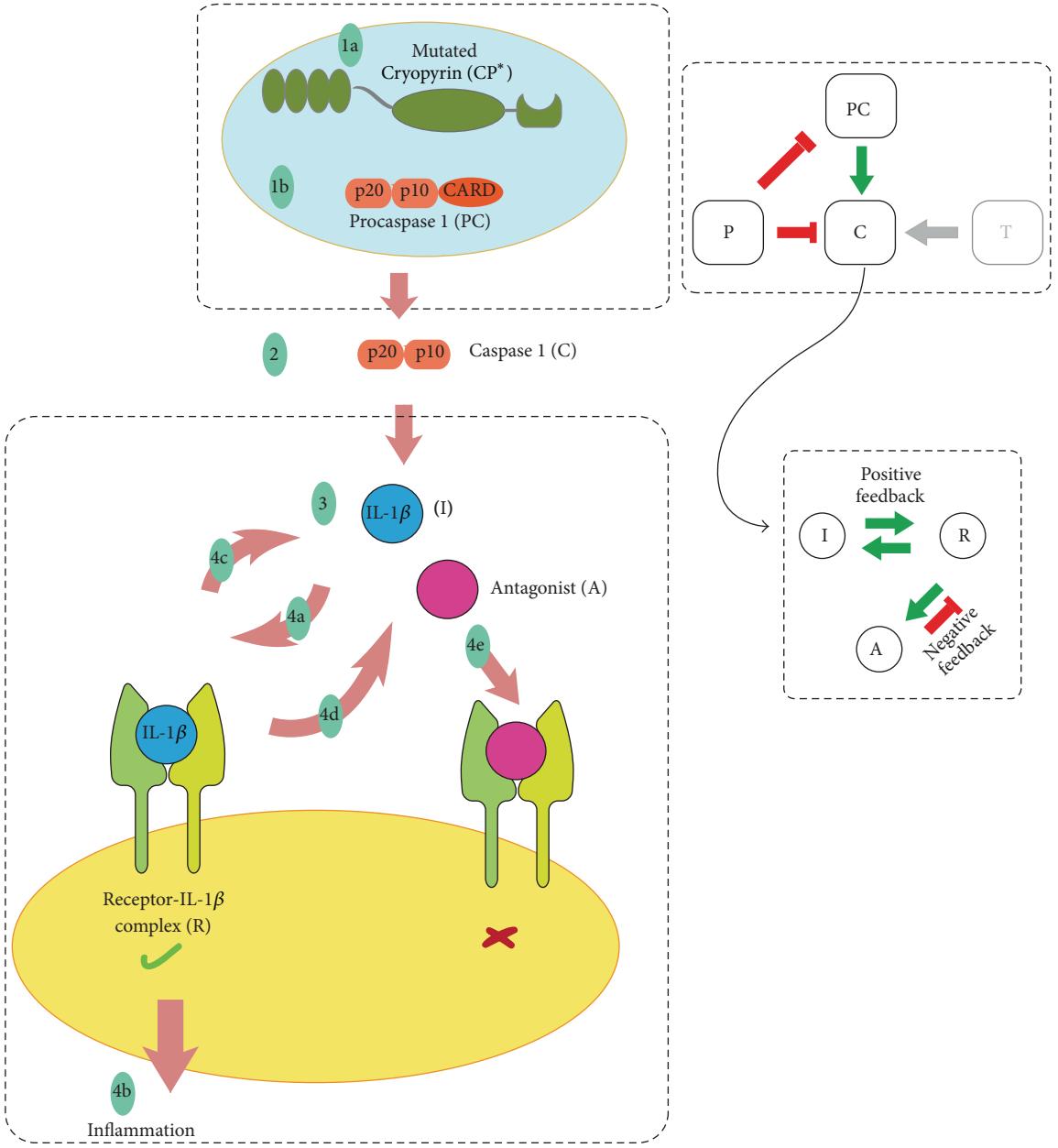


FIGURE 6: Summary of the pathogenesis of CAPS. Even without the trigger, inflammasome formation process takes place in Cryopyrin mutants (1a and 1b). Procaspase 1 turns into Caspase 1 by the inflammasome action (2). IL-1β is maturated by Caspase 1 (3). IL-1β binds to the receptor (4a). IL-1β signaling cascade followed by inflammation is activated by the binding of IL-1β (4b). Signaling by the Receptor-IL-1β complex leads to further transcription of proIL-1β and thus increase in Caspase 1 independent IL-1β levels (4c). Signaling by the Receptor-IL-1β complex also stimulates the antagonist production (4d). Binding of antagonist to the receptor does not result in active IL-1β signaling (4e).

5.2. The Model as a Composition of Two Subsystems. The model can be separated into two distinct subsystems as shown in Figure 7. *Subsystem 1* is composed of noncompetitive inhibition reactions among PC, T, P, and C and essentially determines the amount of Caspase 1. These interactions are captured by (1d) and (1e). Proportional to a constant α, C (Caspase 1) concentration then contributes to production of I. The only link between the competitive binding reactions among R, I, and A that make up *Subsystem 2* and *Subsystem 1* is via the C effect in I. *Subsystem 2* is represented by (1a), (1b), and (1c). Emerging as an input to *Subsystem 2* and produced by *Subsystem 1*, this model composed of two subsystems suggests that Caspase 1 level is the most critical parameter of the total system. Depending on the amount of Caspase 1 level, *Subsystem 2* exhibits varying characteristics, that is, monostability, excitability, and oscillatory behavior, which

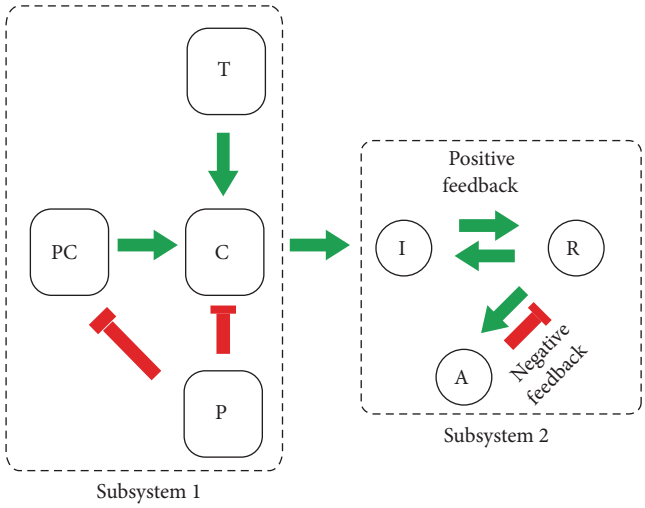


Figure 7: The model as a composition of two subsystems. The direction of each green (red) arrow represents a stimulation (inhibition) effect.

we analyze next in detail. These behaviors correspond to three distinct modes of the model, Healthy, FMF, and CAPS, respectively.

6. Bifurcation Analyses

The three distinct behaviors emerging from the model, Healthy, FMF, and CAPS, are revealed here with bifurcation analyses. Since Caspase 1 level is considered as the most critical component of the total system, it is selected as the bifurcation parameter. Caspase 1, being the output of *Subsystem 1*, has a constant, steady-state value that depends on the values of the model parameters; that is, it can be shown that *Subsystem 1* always has one stable fixed-point. This observation enables uncoupled analysis of *Subsytem 1* and *Subsystem 2*. We perform bifurcation analyses on *Subsystem 2* by considering the Caspase 1 concentration C as a bifurcation parameter that is swept in a certain range. One can think of each C value in this sweep to correspond to a stable fixed-point of *Subsytem 1* for a certain assignment to the model parameters of *Subsytem 1* and a certain trigger level as represented by T. The mutations in Pyrin and Cryopyrin are modeled as changes in the values of certain parameters in *Subsytem 1* as described before and result in an increase in the C as the output of *Subsytem 1*. In addition, higher values of trigger T also result in higher values for C.

In the bifurcation analysis we perform, as C is swept in a certain range, we determine the characteristics of the steady-state solutions of *Subsystem 2*. The results of this analysis are presented in Figure 8, where stable steady-state solutions are shown as a solid line and the unstable ones are represented by dashed lines. Periodic stable (unstable) solutions are shown by green solid (blue dashed) lines. The bifurcation diagrams for all *Subsystem 2* variables, that is, for R, I, and A, follow the same stability pattern.

Table 3: The parameter values used in the model simulations.

Parameter	Healthy	FMF	CAPS
V_r	1		
V_i	1.4		
V_a	0.17		
V_c	1.15		450
V_{pc}	1		
K_{ir}	1		
K_{ar}	1		
K_{ri}	1		
K_{ra}	1		
K_{pcc}	1		
K_{tc}	1		
K_{pc}	1	10	
K_{ppc}	1	10	
k_{dr}	0.2		
k_{di}	0.2		
k_{da}	0.02		
k_{dc}	10		
k_{dpc}	0.5		
k_{br}	0.01		
k_{ba}	0.1		
k_{bc}	1		
k_{bpc}	1		
k_c	0.3		
n	2		
α	0.1		
P	2		

In the three modes, the increase in C levels due to a pulse of trigger T with amplitude varying from 0.1 to 0.94 is shown in Figure 9. Nonzero low level of trigger T reflects the base trigger level which is always present in the environment. Classification of the three modes is as follows.

(i) *Healthy (Monostable-Low) Mode.* For low values of C (C = 1 to C = 1.44), *Subsystem 2* has one stable fixed-point. A trigger pulse causes a slight increase in the C level in this range but does not result in a qualitative change in the characteristics of the solution of *Subsystem 2*, as seen in Figure 9. Increasing values of C in this region does not cause drastic changes in R, I, and A.

(ii) *FMF (Excitable).* In the region between C = 1.44 and C = 1.48, *Subsystem 2* has two unstable fixed-points in addition to a stable fixed-point. When a trigger pulse pushes the system into this region, immune response becomes stronger compared with the ones observed for lower values of C. This becomes possible due to the effect of the Pyrin mutation which elevates the C level.

(iii) *CAPS (Oscillatory and Monostable-High).* When C is increased further (C = 1.48 to C = 1.9) due to the type of Cryopyrin mutations, *Subsystem 2* has one stable

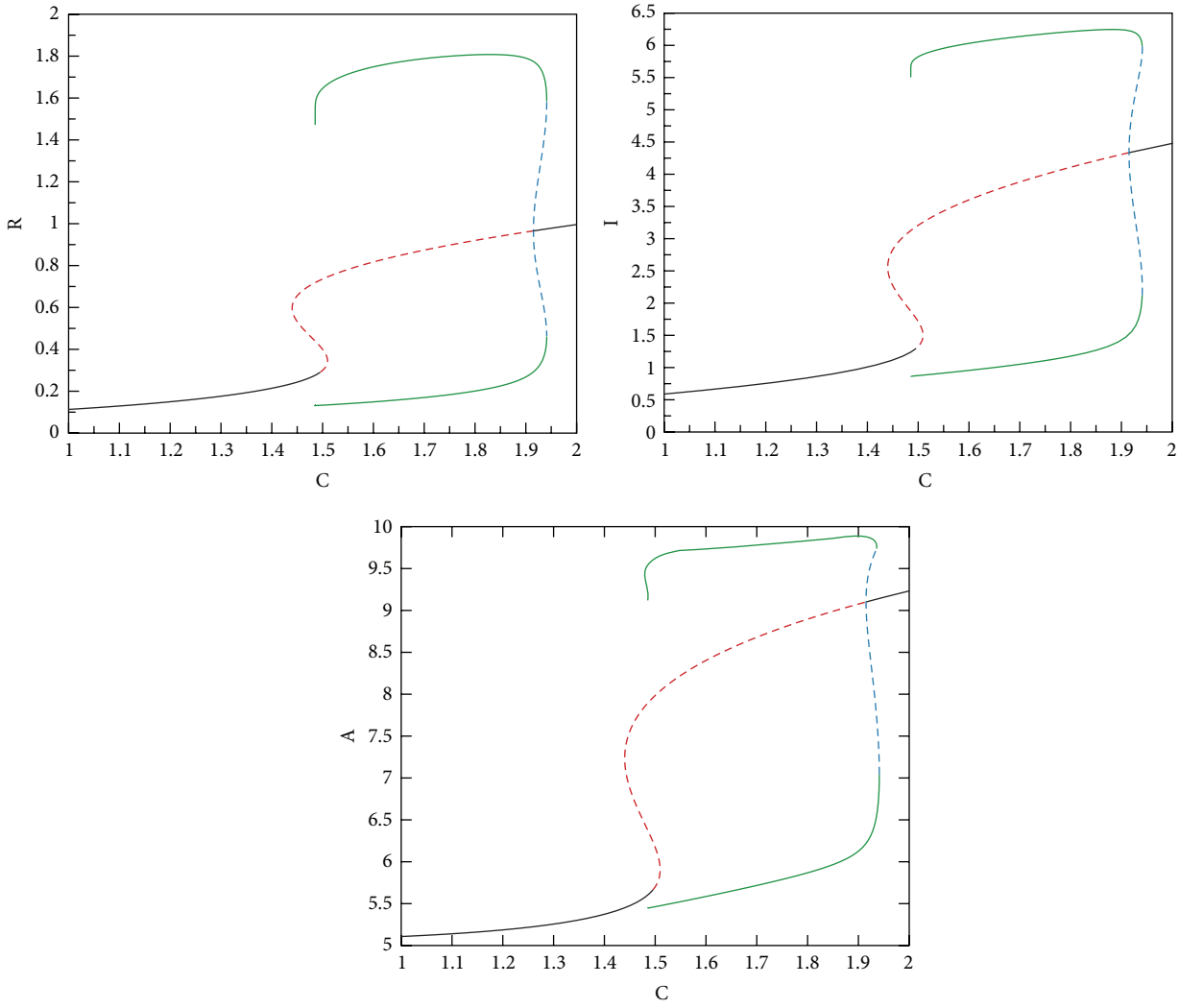


Figure 8: Bifurcation analyses. Stable (unstable) solutions are shown as a solid black (red dashed) line. Periodic stable (unstable) solutions are represented by green solid (blue dashed) lines.

periodic solution, that is, a limit cycle, and an unstable fixed-point. In this mode, limit cycle oscillations in R, I, and A are observed. In CAPS, C levels are always high due to the autosomal dominantly inherited mutations in Cryopyrin, resulting in oscillatory behavior even without a trigger pulse. The maximum levels observed for R, I, and A are the highest in this mode. Going beyond C = 1.9 corresponds to a more severe member of the CAPS disease family (NOMID), which is shown in Figure 10. In this range, oscillations (flare-remission cycles) disappear; there is one stable fixed-point. However, the values of R, I, and A stay at elevated levels indicating a much stronger immune response and hence continuous inflammation and fever.

Numerical bifurcation analyses presented in Figure 8 were performed with the XPPAUT software package [44]. Stability of the steady-state solutions was further verified by time-domain simulations using MATLAB [45].

7. Results and Discussion

We now present time-domain simulations for R, I, A, PC, and C in response to a pulse trigger input in Healthy and FMF modes and a constant trigger in CAPS, as shown in Figures 11, 12, 13, 14, and 15. In Healthy and FMF modes, magnitude of T is stepped from 0.1 to 0.94 and back to 0.1. The trigger is applied at time = 100 and duration of the trigger pulse is set to 150. In CAPS, a constant base trigger level of T = 0.1 is applied. The dashed lines represent T in the plots.

Due to a pulse of trigger, there is a slight increase in R, in the Healthy mode. The increase in R is not large enough to activate the cytokine cascade for inflammation. In FMF, larger R values are observed, initiating the inflammation. When the trigger disappears, R values return to normal ranges in a while. High R values correspond to the flare (attack) period in FMF. When the trigger level is subdued, remission period continues. Depending on the magnitude and duration of the trigger pulse, attacks occur or do not

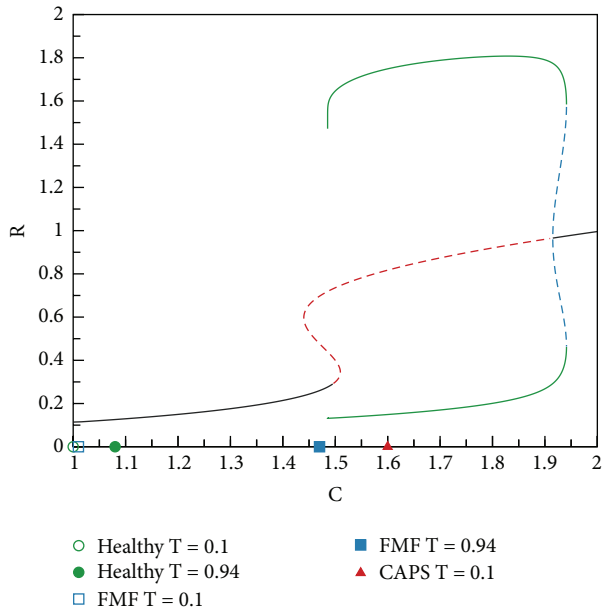


Figure 9: Classification of the three modes according to the bifurcation analysis.

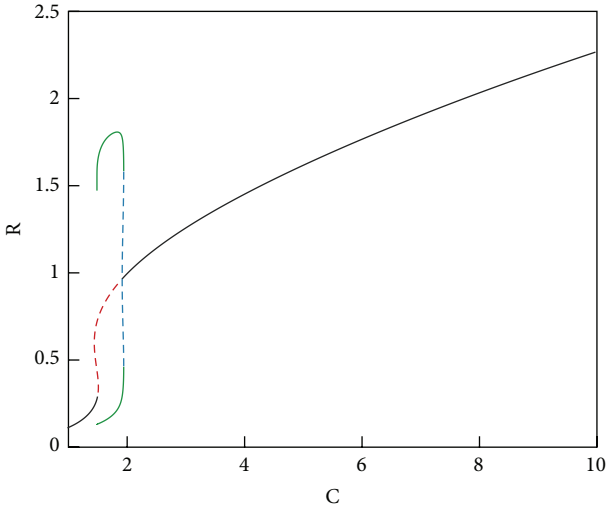


Figure 10: Extended bifurcation analysis.

occur, which fit well with the clinical experiences reflecting irregular periodicity of attacks ranging from weekly to ones that occur once in several months. In CAPS, even though T is kept at a low nominal value (T = 0.1), R levels oscillate due to the dynamics of the interaction between R, I, and A. The maximum R level is higher than that of FMF. As captured by our model, CAPS patients follow this oscillatory flare/remission pattern. I and A exhibit the same behavior as R.

Due to the initiation of triggers, procaspase 1 is converted to Caspase 1 through Cryopyrin inflammasome formation, as captured by *Subsystem 1*. Due to the mutation in Pyrin in FMF, Pyrin cannot suppress inflammasome formation by binding to PC. As shown in Figure 14, in FMF, PC level is the highest due to the mutation in Pyrin. In CAPS, as a result of the Cryopyrin mutation, conversion of PC to C is more favored. Therefore, PC is the lowest in the CAPS mode. Although FMF and CAPS are similar diseases in terms of inflammasome-associated overproduction of IL-1β, targets for drugs used in the treatment are different. IL-1β blocking strategies are the general approach in CAPS, while *colchicine* is the gold standard to reduce the number of FMF flares. Ineffectiveness of colchicine in CAPS suggests that colchicine may have a prophylactic effect only in a specific PC/C range. Low levels of PC and higher levels of activated form of C may correspond to the ineffective range of colchicine, due to inability to use it in higher doses because of potentially fatal adverse effects. Although the mechanism of colchicine is not known yet, colchicine is speculated to suppress Cryopyrin inflammasome activation [46].

Both time-domain simulations and bifurcation analyses justify the classification of model/disease modes we have identified. In Healthy mode, C levels are low, independent of T. In FMF, when T is low, so is C. Higher T levels cause an increase in C values into the excitable mode. In CAPS, C levels are always high, where other key players are in the oscillatory region. Disease severity in the CAPS mode can be correlated with the level of C. Furthermore, Figure 16 shows that the recurrence frequency of the attacks in CAPS increases with increasing values of C.

We have so far observed that CAPS attacks occur even in the absence of a trigger pulse. When there is also a trigger pulse, the attacks become more severe and frequent. Furthermore, C levels may go beyond the end of the oscillatory region, where there is again a monotonically increasing stable solution as shown in Figure 10. Inflammatory response in the CAPS mode with trigger initiation is more severe and this may correspond to the more severe forms of CAPS such as NOMID, in which continuous inflammatory activity is rule. R levels for the CAPS mode in response to a trigger pulse are shown in Figure 17. Base trigger level case is provided for comparison.

8. Conclusion

We have presented a unified mathematical model for two of the well characterized autoinflammatory syndromes, associated with increased innate immunity related inflammation, FMF and CAPS. These diseases are caused by the mutations in the regulatory inflammasome proteins, Pyrin and Cryopyrin, respectively. Although the mutations and their inheritance patterns are different, FMF and CAPS are closely related in terms of their pathogenic pathways. Overproduction of IL-1β due to increase in Caspase 1 activity is the main cause that triggers the inflammation process in both diseases. That is why we have endeavoured to develop a unified model in order to capture both FMF and CAPS, in addition to the healthy immune system behavior, by adjusting only three parameters in the model. By introducing the effect of relevant mutations, we were able to mimic, in the model, the observed clinical behaviors in terms of recurrence rate, triggers of inflammation, and disease severity which may elucidate

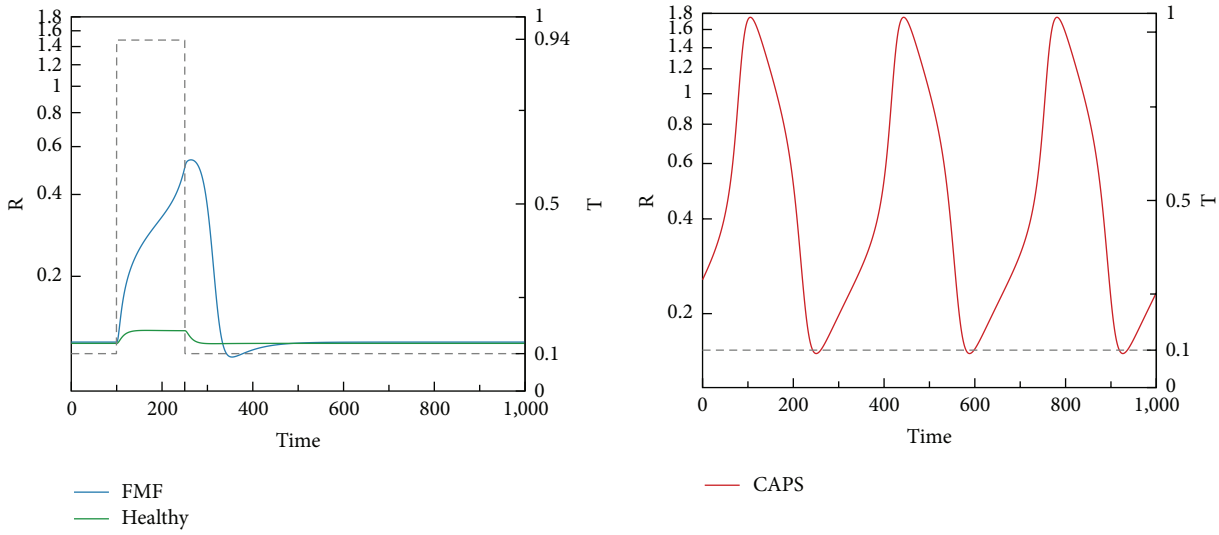


Figure 11: R levels in Healthy, FMF, and CAPS cases.

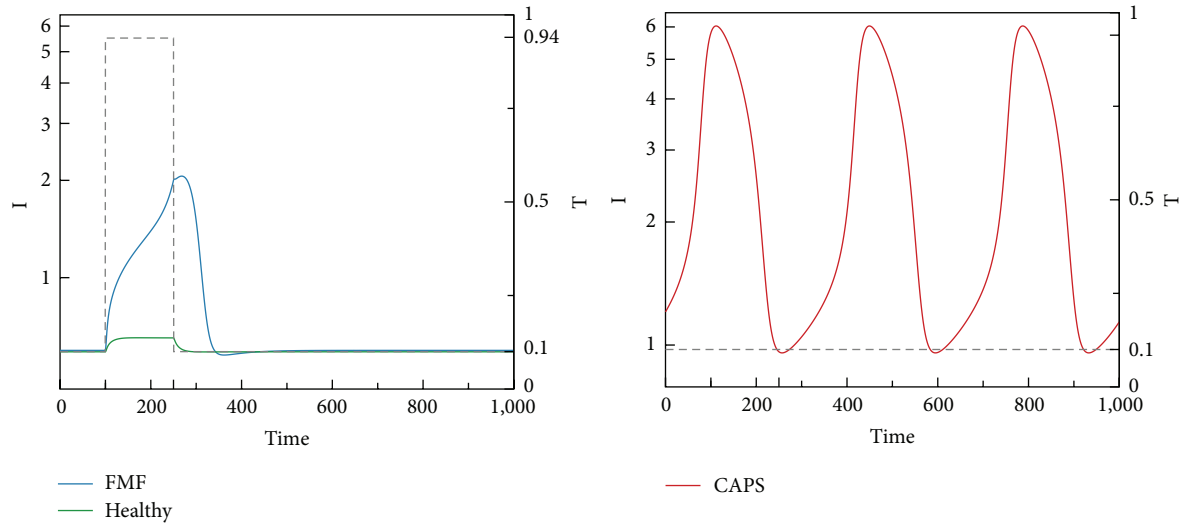


Figure 12: I levels in Healthy, FMF, and CAPS cases.

the differences in the disease mechanisms. According to the bifurcation analyses performed and simulation results obtained from the model, Caspase 1 level is the most critical parameter in determining the three modes that the model exhibits, Healthy, FMF, and CAPS. In accordance with the clinical literature, FMF comes out as trigger-dependent while CAPS is mostly due to autonomous self-dynamics of the immunity related protein concentrations. As a result, FMF attacks occur only when a trigger or insult is introduced to the system. CAPS, on the other hand, has an autonomous periodic nature, and even when there is no trigger present, attacks occur, if not treated. The model proposed in the paper matches and explains such clinical observations, as distilled from the results presented in the paper and summarized below.

(i) Clinical Observation 1. FMF attacks occur irregularly, possibly as a result of external or endogenous triggers such as stress, heavy exercise, and infections.

Corresponding Model Behavior and Outcome. FMF attacks are stimulated by a pulse trigger (T) in the model. In Pyrin mutants (i.e., FMF patients), the introduction of a trigger with a large enough magnitude and duration results in higher Caspase 1 levels and a transition from monostable behavior to excitability. This causes increases in the amount

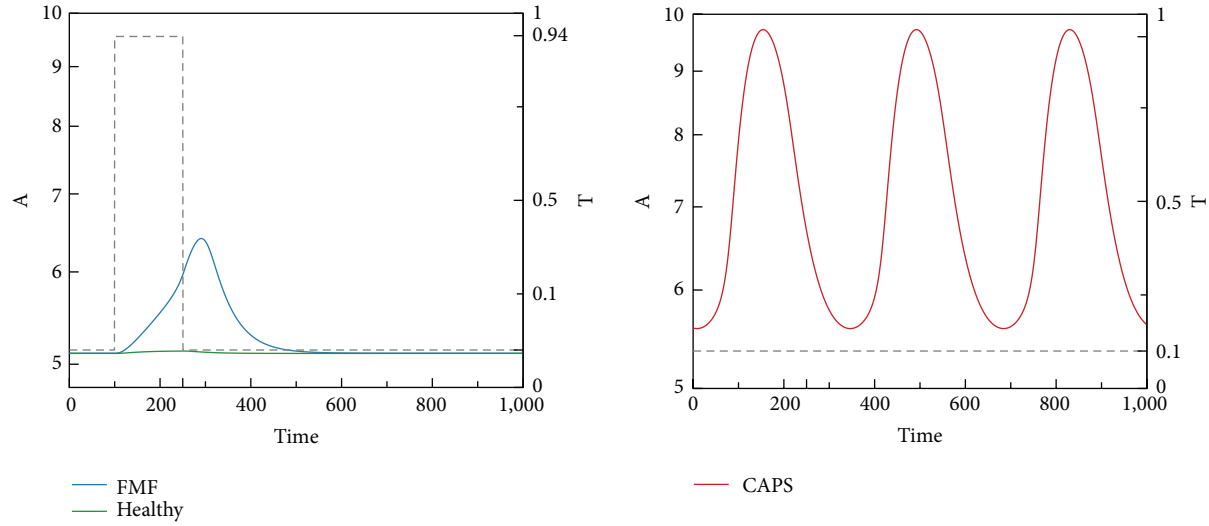


Figure 13: A levels in Healthy, FMF, and CAPS cases.

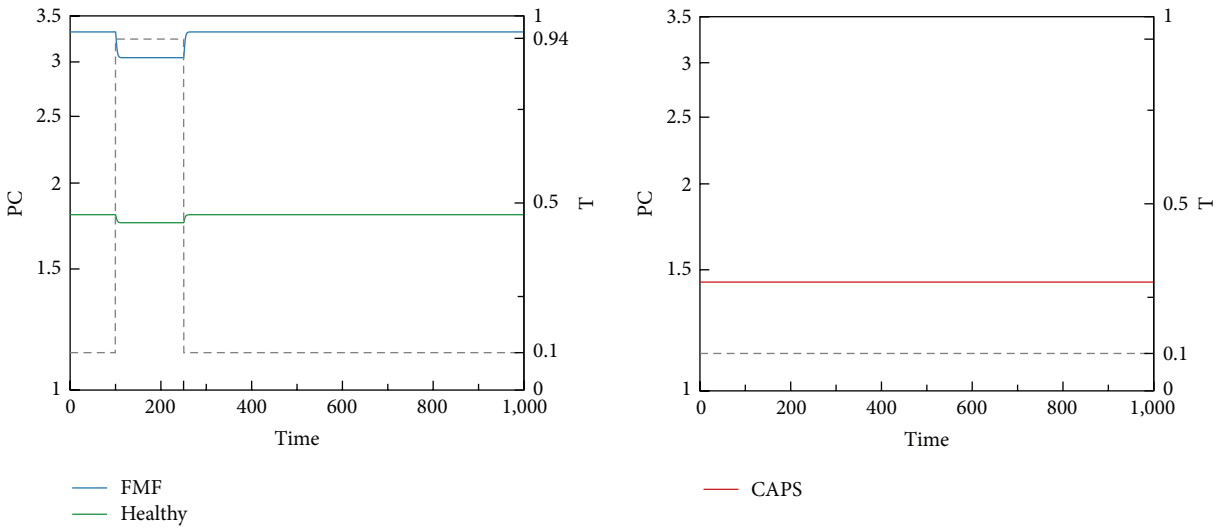


Figure 14: PC levels in Healthy, FMF, and CAPS cases.

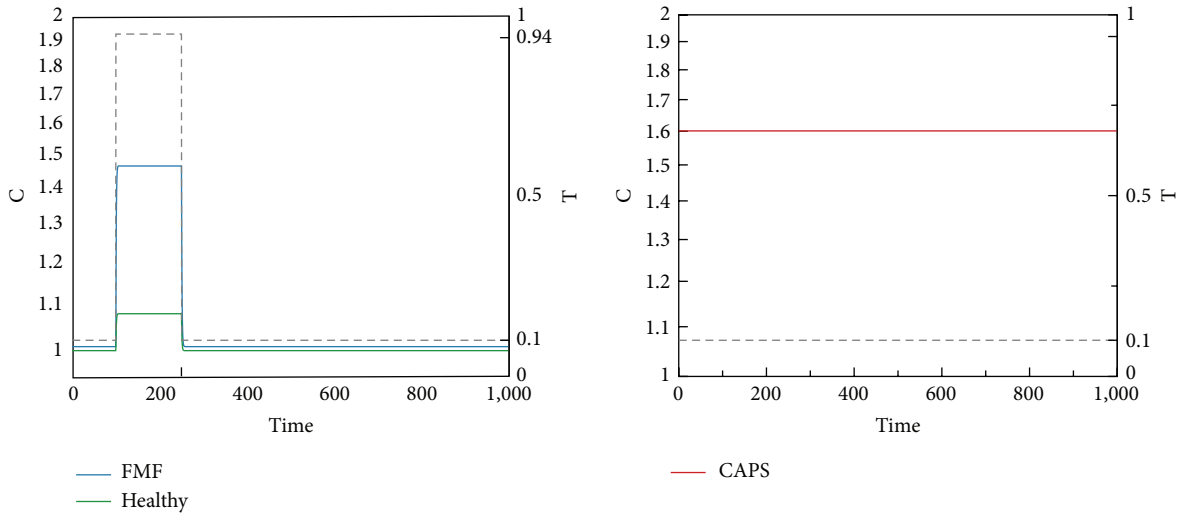


Figure 15: C levels in Healthy, FMF, and CAPS cases.

of the key variables (i.e., free (I) and bound IL-1β (R) and the antagonist (A)), indicating an attack period. When the trigger is removed from the system, the amounts of these key players settle back to their normal values, indicating a remission period.

(ii) Clinical Observation 2. FMF attacks have irregular periodicity, ranging from weekly attacks to ones that occur once in several months.

Corresponding Model Behavior and Outcome. In the FMF mode of the model, the attack characteristics are determined by the magnitude and the duration of the introduced trigger. Attacks in this mode usually do not have a regular pattern.

(iii) Clinical Observation 3. Disease and attack severity exhibits considerable variability among FMF patients. Furthermore, the different attacks experienced by a particular patient may be of varying severity.

Corresponding Model Behavior and Outcome. Person to person variations in the attack durations and severity and diverse attack patterns in FMF patients can be explained by irregular trigger actions based on environmental or seasonal variations. This can also be explained based on the difference in the sensitivity to external triggers among individuals along with the penetrance of the mutation, which can be reflected in the model by adjusting the values of the three key model parameters.

(iv) Clinical Observation 4. CAPS arise as recurrent episodes of fever, even in the absence of any insult; that is, CAPS flares seem to be self-sustaining. CAPS is a more severe disease than FMF with more frequent and relatively regular attacks.

Corresponding Model Behavior and Outcome. When the parameter (that corresponds to the mutation in Cryopyrin) is increased, oscillatory behavior in the key players of the immune system is observed even without any trigger. In other words, CAPS attacks become inevitable.

(v) Clinical Observation 5. CAPS is a spectrum of diseases with varying severities. In more severe forms of CAPS, such as NOMID, there is continuous inflammatory activity.

Corresponding Model Behavior and Outcome. Our model captures the disease severity in CAPS in two ways. First, even when there is no trigger, the system variables are in oscillatory region due to constant high Caspase 1 levels. The period of the attacks decreases when the degree of penetrance of the mutations in Cryopyrin is increased, as shown in Figure 16. When also a trigger is introduced into the system, the model exhibits a constant, that is, nonoscillatory, but much stronger response, which corresponds to a continuous inflammation and fever as observed in NOMID.

(vi) Clinical Observation 6. Although FMF and CAPS are similar diseases in terms of inflammasome-associated overproduction of IL-1β, targets for drugs used in the treatment are different. IL-1β blocking strategies are the general approach

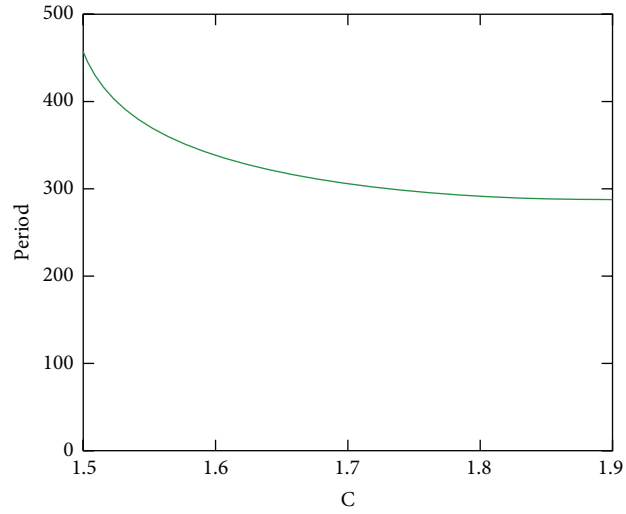


Figure 16: Period of the attacks in CAPS as a function of Caspase 1 level.

in CAPS, while *colchicine* is the gold standard in FMF in reducing the flare frequency and the symptoms.

Corresponding Model Behavior and Outcome. Due to the mutation in Pyrin in FMF, Pyrin cannot suppress inflammasome formation by binding to procaspase 1 and Caspase 1. In FMF, procaspase 1 level and procaspase 1/Caspase 1 ratio are higher due to the effects of mutations in Pyrin on inflammasome regulation. Because of trigger-dependent irregular attack development in FMF, procaspase 1 is the dominant form found in the cytoplasm during attack-free periods, and it is overprocessed into Caspase 1 with trigger, which exceeds physiologic control mechanisms and results in inflammatory attacks. Colchicine is effective as a prophylactic treatment by possibly showing its efficacy when procaspase 1/Caspase 1 ratio is high, and it has no effect during an attack. In CAPS, as a result of the dominantly inherited gain-of-function Cryopyrin mutations, which result in spontaneous activation of the inflammasome, conversion of procaspase 1 to Caspase 1 is highly favored. Therefore, procaspase 1 level and procaspase 1/Caspase 1 ratio are lower in the CAPS mode. Colchicine has no efficacy in CAPS patients, and this clinical observation may also suggest that colchicine's prophylactic efficacy can be observed only in a limited procaspase 1/Caspase 1 range. Low levels of procaspase 1 and higher levels of activated Caspase 1 may correspond to the ineffective range of colchicine, due to its inability to control higher Caspase 1 activity and its limited use at higher doses because of potentially fatal adverse effects. Although the mechanism of colchicine is not known yet, colchicine is speculated to suppress overall Cryopyrin associated inflammasome activation [46] and hence suppress conversion of procaspase 1 into active Caspase 1. This is effective in FMF due to higher procaspase 1 levels, but not in CAPS, because procaspase 1 is already depleted in CAPS according to the model (i.e., converted into Caspase 1). A possible inflammasome-related mechanism for colchicine as a procaspase 1 inhibitor also emerges from the model.

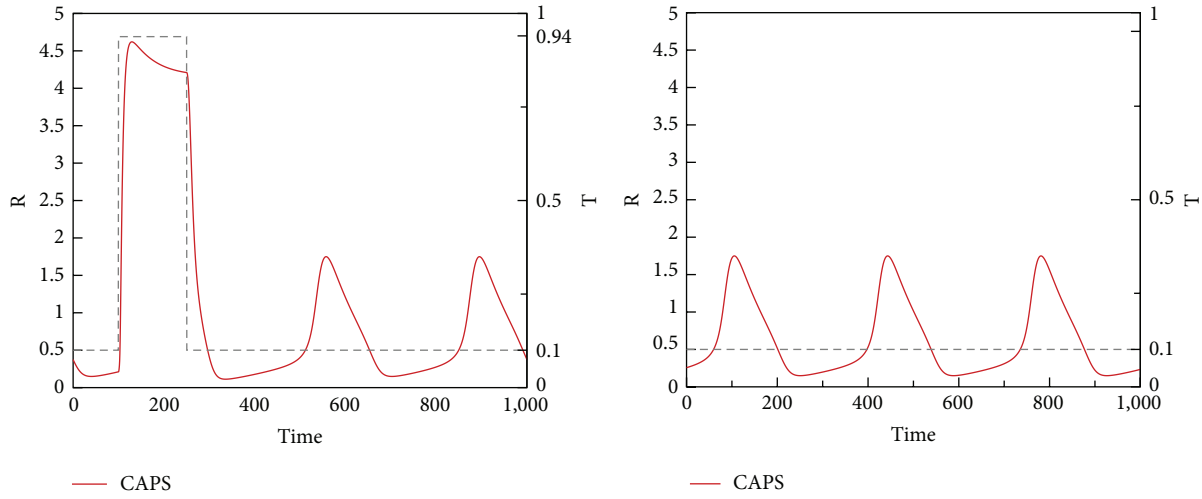


Figure 17: R levels in CAPS with and without trigger.

The proposed model explains the clinical observations, as described above, very well in a qualitative sense, and the model variables do have a direct correspondence to key molecular players in the immune system. Elevations in the amounts of the selected key variables (i.e., free (I) and bound IL-1β (R) and the antagonist (A)) in the attack period of both FMF and CAPS are captured by our model. The model also shows the recurrence patterns observed clinically in both diseases: irregular attacks in FMF following the stimuli and more frequent and relatively regular attacks in CAPS even without any stimuli depending on mutation-dependent disease severity. On the other hand, the specific values of the parameters used in our model are not yet based on experimental findings. As such, it is not yet possible to correlate the detailed quantitative outcomes of the model with actual quantitative clinical measurements. At this stage, however, the elevations in the concentrations of the aforementioned key variables act as the initiators of the subsequent cascades, which then result in the elevation of easily measurable inflammation parameters such as CRP and ESR in the attacks. Therefore, the values of the key variables that exceed chosen thresholds are considered as the markers of the attack period. Although the parameters are not physiologically meaningful when considered alone, following the adjustment of the parameters, relative changes in the key variables follow a clinically similar scenario.

In our future work, we will strive to link all model variables and parameter values to *in vivo* clinical observations, measurements from *in vitro* experiments based on cell lines, and also *in silico* experiments. In these *in silico* studies based on data from clinical and *in vitro* studies, we will formulate and run detailed molecular dynamics simulations in order to quantify the interactions of various molecular species involved.

Conflict of Interests

The authors declare that there are no conflict of interests regarding the publication of this paper.

Acknowledgment

Yasemin Bozkurt was supported by a Koc University Graduate Fellowship.

References

[1] L. Franchi, R. Muñoz-Planillo, T. Reimer, T. Eigenbrod, and G. Núñez, "Inflammasomes as microbial sensors," *European Journal of Immunology*, vol. 40, no. 3, pp. 611–615, 2010.

[2] E. Latz, T. S. Xiao, and A. Stutz, "Activation and regulation of the inflammasomes," *Nature Reviews Immunology*, vol. 13, no. 6, pp. 397–411, 2013.

[3] A. S. Perelson, "Modelling viral and immune system dynamics," *Nature Reviews Immunology*, vol. 2, no. 1, pp. 28–36, 2002.

[4] M. Galeazzi, G. Gasbarrini, A. Ghirardello et al., "Autoinflammatory syndromes," *Clinical and Experimental Rheumatology*, vol. 24, no. 40, pp. S79–S85, 2006.

[5] T. A. Doherty, S. D. Brydges, and H. M. Hoffman, "Autoinflammation: translating mechanism to therapy," *Journal of Leukocyte Biology*, vol. 90, no. 1, pp. 37–47, 2011.

[6] J. J. Chae, G. Wood, S. L. Masters et al., "The B30.2 domain of pyrin, the familial mediterranean fever protein, interacts directly with caspase-1 to modulate IL-1β production," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9982–9987, 2006.

[7] S. D. Brydges, J. L. Mueller, M. D. McGeough et al., "Inflammasome-mediated disease animal models reveal roles for innate but not adaptive immunity," *Immunity*, vol. 30, no. 6, pp. 875–887, 2009.

[8] F. Martinon, L. Agostini, E. Meylan, and J. Tschopp, "Identification of bacterial muramyl dipeptide as activator of the NALP3/Cryopyrin inflammasome," *Current Biology*, vol. 14, no. 21, pp. 1929–1934, 2004.

[9] French FMF Consortium, "A candidate gene for familial mediterranean fever," *Nature Genetics*, vol. 17, no. 1, pp. 25–31, 1997.

[10] E. Pras, I. Aksentijevich, L. Gruberg et al., "Mapping of a gene causing familial mediterranean fever to the short arm of chromosome 16," *The New England Journal of Medicine*, vol. 326, no. 23, pp. 1509–1513, 1992.

[11] C. Cazeneuve, Z. Hovannesyan, D. Geneviève et al., "Familial Mediterranean fever among patients from Karabakh and the diagnostic value of *MEFV* gene analysis in all classically affected populations," *Arthritis & Rheumatism*, vol. 48, no. 8, pp. 2324–2331, 2003.

[12] A. Livneh and P. Langevitz, "Diagnostic and treatment concerns in familial Mediterranean fever," *Bailliere's Best Practice and Research in Clinical Rheumatology*, vol. 14, no. 3, pp. 477–498, 2000.

[13] M. Centola, G. Wood, D. M. Frucht et al., "The gene for familial Mediterranean fever, MEFV, is expressed in early leukocyte development and is regulated in response to inflammatory mediators," *Blood*, vol. 95, no. 10, pp. 3223–3231, 2000.

[14] J. J. Chae, I. Aksentijevich, and D. L. Kastner, "Advances in the understanding of familial Mediterranean fever and possibilities for targeted therapy," *British Journal of Haematology*, vol. 146, no. 5, pp. 467–478, 2009.

[15] M. Inoue and M. L. Shinohara, "NLRP3 inflammasome and MS/EAE," *Autoimmune Diseases*, vol. 2013, Article ID 859145, 8 pages, 2013.

[16] K. Schroder and J. Tschopp, "The Inflammasomes," *Cell*, vol. 140, no. 6, pp. 821–832, 2010.

[17] M. C. Hochberg, A. J. Silman, J. S. Smolen, M. E. Weinblatt, and M. H. Weisman, *Rheumatology*, Elsevier, New York, NY, USA, 2011.

[18] J. J. Chae, H. D. Komarow, J. Cheng et al., "Targeted disruption of pyrin, the FMF protein, causes heightened sensitivity to endotoxin and a defect in macrophage apoptosis," *Molecular Cell*, vol. 11, no. 3, pp. 591–604, 2003.

[19] J.-W. Yu, J. Wu, Z. Zhang et al., "Cryopyrin and pyrin activate caspase-1, but not NF-κB, via ASC oligomerization," *Cell Death and Differentiation*, vol. 13, no. 2, pp. 236–249, 2006.

[20] S. Papin, S. Cuenin, L. Agostini et al., "The SPRY domain of pyrin, mutated in familial mediterranean fever patients, interacts with inflammasome components and inhibits proIL-1β processing," *Cell Death and Differentiation*, vol. 14, no. 8, pp. 1457–1466, 2007.

[21] E. Latz, "The inflammasomes: mechanisms of activation and function," *Current Opinion in Immunology*, vol. 22, no. 1, pp. 28–33, 2010.

[22] M. Yamamoto, S. Sato, K. Mori et al., "Cutting edge: a novel Toll/IL-1 receptor domain-containing adapter that preferentially activates the IFN-β promoter in the Toll-like receptor signaling," *Journal of Immunology*, vol. 169, no. 12, pp. 6668–6672, 2002.

[23] E. Thomassen, T. A. Bird, B. R. Renshaw, M. K. Kennedy, and J. E. Sims, "Binding of interleukin-18 to the interleukin-1 receptor homologous receptor IL-1Rrp1 leads to activation of signaling pathways similar to those used by interleukin-1," *Journal of Interferon and Cytokine Research*, vol. 18, no. 12, pp. 1077–1088, 1998.

[24] P. Pelegrín, "Inflammasome activation by danger signals," in *The Inflammasomes*, pp. 101–121, Springer, 2011.

[25] A. Rubartelli, M. Gattorno, M. G. Netea, and C. A. Dinarello, "Interplay between redox status and inflammasome activation," *Trends in Immunology*, vol. 32, no. 12, pp. 559–566, 2011.

[26] M. Guma, L. Ronacher, R. Liu-Bryan, S. Takai, M. Karin, and M. Corr, "Caspase 1-independent activation of interleukin-1β in neutrophil-predominant inflammation," *Arthritis & Rheumatism*, vol. 60, no. 12, pp. 3642–3650, 2009.

[27] L. A. B. Joosten, M. G. Netea, G. Fantuzzi et al., "Inflammatory arthritis in caspase 1 gene-deficient mice: contribution of proteinase 3 to caspase 1-independent production of bioactive interleukin-1β," *Arthritis & Rheumatism*, vol. 60, no. 12, pp. 3651–3662, 2009.

[28] G. P. Manukyan, K. A. Ghazaryan, Z. A. Ktsoyan et al., "Cytokine profile of Armenian patients with familial Mediterranean fever," *Clinical Biochemistry*, vol. 41, no. 10-11, pp. 920–922, 2008.

[29] S. Bagci, B. Toy, A. Tuzun et al., "Continuity of cytokine activation in patients with familial Mediterranean fever," *Clinical Rheumatology*, vol. 23, no. 4, pp. 333–337, 2004.

[30] N. Gang, J. P. H. Drenth, P. Langevitz et al., "Activation of the cytokine network in familial Mediterranean fever," *The Journal of Rheumatology*, vol. 26, no. 4, pp. 890–897, 1999.

[31] C. Gabay, C. Lamacchia, and G. Palmer, "IL-1 pathways in inflammation and human diseases," *Nature Reviews Rheumatology*, vol. 6, no. 4, pp. 232–241, 2010.

[32] E. B. Cullinan, L. Kwee, P. Nunes et al., "IL-1 receptor accessory protein is an essential component of the IL-1 receptor," *The Journal of Immunology*, vol. 161, no. 10, pp. 5614–5620, 1998.

[33] M. H. Schiff, "Role of interleukin 1 and interleukin 1 receptor antagonist in the mediation of rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 59, no. 1, pp. i103–i108, 2000.

[34] F. Colotta, S. K. Dower, J. E. Sims, and A. Mantovani, "The type II "decoy" receptor: a novel regulatory pathway for interleukin 1," *Immunology Today*, vol. 15, no. 12, pp. 562–566, 1994.

[35] N. Polentarutti, G. P. Rol, M. Muzio et al., "Unique pattern of expression and inhibition of IL-1 signaling by the IL-1 receptor family member TIR8/SIGIRR," *European Cytokine Network*, vol. 14, no. 4, pp. 211–218, 2003.

[36] J. A. Symons, P. R. Young, and G. W. Duff, "Soluble type II interleukin 1 (IL-1) receptor binds and blocks processing of IL-1β precursor and loses affinity for IL-1 receptor antagonist," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 5, pp. 1714–1718, 1995.

[37] J. Feldmann, A.-M. Prieur, P. Quartier et al., "Chronic infantile neurological cutaneous and articular syndrome is caused by mutations in CIAS1, a gene highly expressed in polymorphonuclear cells and chondrocytes," *The American Journal of Human Genetics*, vol. 71, no. 1, pp. 198–203, 2002.

[38] H. J. Lachmann, I. Kone-Paut, J. B. Kuemmerle-Deschner et al., "Use of canakinumab in the cryopyrin-associated periodic syndrome," *The New England Journal of Medicine*, vol. 360, no. 23, pp. 2416–2425, 2009.

[39] B. Neven, A.-M. Prieur, and P. Q. dit Maire, "Cryopyrinopathies: update on pathogenesis and treatment," *Nature Clinical Practice Rheumatology*, vol. 4, no. 9, pp. 481–489, 2008.

[40] G. Yenokyan and H. K. Armenian, "Triggers for attacks in familial mediterranean fever: application of the case-crossover design," *American Journal of Epidemiology*, vol. 175, no. 10, pp. 1054–1061, 2012.

[41] X.-J. Tian, X.-P. Zhang, F. Liu, and W. Wang, "Interlinking positive and negative feedback loops creates a tunable motif in gene regulatory networks," *Physical Review E*, vol. 80, no. 1, Article ID 011926, 2009.

[42] M. H. Sauro, *Enzyme Kinetics for Systems Biology*, Ambrosius, 2013.

[43] A. Cortés, M. Cascante, M. L. Cárdenas, and A. Cornish-Bowden, "Relationships between inhibition constants, inhibitor concentrations for 50% inhibition and types of inhibition: new ways of analysing data," *Biochemical Journal*, vol. 357, no. 1, pp. 263–268, 2001.

[44] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, vol. 14, SIAM, 2002.

[45] *MATLAB. R2011b*, The MathWorks, Natick, Mass, USA, 2011.

[46] G. Nuki, "Colchicine: its mechanism of action and efficacy in crystal-induced inflammation," *Current Rheumatology Reports*, vol. 10, no. 3, pp. 218–227, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 641393, 6 pages
http://dx.doi.org/10.1155/2015/641393

Research Article

Evolutionary Influenced Interaction Pattern as Indicator for the Investigation of Natural Variants Causing Nephrogenic Diabetes Insipidus

Steffen Grunert and Dirk Labudde

Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, 09648 Mittweida, Germany

Correspondence should be addressed to Steffen Grunert; sgrunert@hs-mittweida.de

Received 19 September 2014; Accepted 3 December 2014

Academic Editor: Maria D. Taranto

The importance of short membrane sequence motifs has been shown in many works and emphasizes the related sequence motif analysis. Together with specific transmembrane helix-helix interactions, the analysis of interacting sequence parts is helpful for understanding the process during membrane protein folding and in retaining the three-dimensional fold. Here we present a simple high-throughput analysis method for deriving mutational information of interacting sequence parts. Applied on aquaporin water channel proteins, our approach supports the analysis of mutational variants within different interacting subsequences and finally the investigation of natural variants which cause diseases like, for example, nephrogenic diabetes insipidus. In this work we demonstrate a simple method for massive membrane protein data analysis. As shown, the presented in silico analyses provide information about interacting sequence parts which are constrained by protein evolution. We present a simple graphical visualization medium for the representation of evolutionary influenced interaction pattern pairs (EIPPs) adapted to mutagen investigations of aquaporin-2, a protein whose mutants are involved in the rare endocrine disorder known as nephrogenic diabetes insipidus, and membrane proteins in general. Furthermore, we present a new method to derive new evolutionary variations within EIPPs which can be used for further mutagen laboratory investigations.

1. Introduction

Integral membrane proteins are coded by 20–30% of all open reading frames of known genomes [1–3]. As elements in accomplishing numerous molecular processes, that is, signal transduction and passive and active transport of an extensive number of chemical compounds and ions, mutations in genes coding for membrane proteins are often linked to diseases [4]. Despite their biological importance, relatively little is known about folding, functional mechanics, and synthesis of membrane proteins [1]. This is due to experimentally costly and complex procedures, since membrane proteins are difficult to handle in lab experiments [5]. To understand correspondence between genetic mutations and the effects on protein mechanics, the development of novel theoretical approaches is highly demanded. In our work we demonstrate a high-throughput in silico approach for the investigation of the influences of genetic variations within interacting sequence part in membrane proteins, which are directly linked to nephrogenic diabetes insipidus. Nephrogenic diabetes insipidus (NDI) is a disorder which can be acquired as a side effect of surpassing drug taking or which is caused by inherited genetic mutations. Autosomal recessive and dominant inherited NDI are linked to mutations in genes encoding the integral membrane aquaporin-2 water channel [6, 7]. X-linked inheritable NDI is caused by mutations in the gene encoding the AVP type-2 receptor membrane protein (V2R) [8, 9]. In the general population, inherited NDI shows a low prevalence of one case per 20,000–30,000 people [10–12]. Aquaporin-2 water channels and V2R are essential elements in the water reabsorption through the apical cell membrane. This water composes the main part of preurine, a product that results from ultrafiltration in the kidney. The process of water reabsorption from the preurine

is essential to ensure the body's fluid balance and is realised by membrane-integrated aquaporin-2 water channels. The insertion of aquaporin-2 into the human kidney cell membrane is triggered by the antidiuretic hormone, which is also referred to as arginine vasopressin (AVP). The AVP blood concentration is regulated by the controlled release of AVP in the pituitary gland which is adapted according to the body's fluid balance. In the process, the binding of AVP to V2R leads to the activation of the receptor. In this state, V2R is able to interact with the guanine nucleotide-binding G(s) subunit alpha [13, 14]. Subsequently, the activation of adenylcyclase 6 takes place, leading to cAMP synthesis and increase of cAMP concentration in the cell plasma [15, 16]. By means of protein kinase A, cAMP triggers the phosphorylation of aquaporin-2 molecules which are stored in cytoplasmic vesicles that have bound to the endoplasmic reticulum. The phosphorylation induces the translocation and fusion of the cytoplasmic vesicles into the plasma membrane and finally leads to the insertion of aquaporin-2 molecules into the apical membrane [17]. Inactive mutants of V2R and aquaporin-2 cause a reduced water reabsorption in the kidneys [18]. Consequences are the typical symptoms of NDI, for example, sensorineural deafness, urinary tract anatomy, ataxia, peripheral neuropathy, mental retardation, psychiatric illness, a daily output of 15–20 L highly dilute (<100 mOsmol/kg) urine (polyuria), and compensatory excessive liquid intake [18–20]. In newborn infants, NDI is characterized by dehydration symptoms, irritability, and poor feeding as well as poor weight gain. A schematic illustration of these molecular coherences is given in Figure 1. The direct inspection of the aquaporin-2 gene as well as the V2 receptor gene (AVPR2) has become accomplishable in clinical practice [21] for differential NDI diagnosis and has been substituting dehydration testing over the last years [18].

2. Materials and Methods

As the first step, we want to realise a task which is involved in the prediction of homologue sequence parts within transmembrane α-helices. This means that aquaporin specific evolutionary interaction pattern pairs (EIPPs) were generated as described in current work of [22]. In this work, Grunert and Labudde show that the combination of interaction information and sequence motifs with evolutionary variation can be used for 3D structure prediction. They obtained key information from homologue sequences to separate and predict membrane protein structures in the context of interacting pattern and their evolutionary variation. Patterns as motif representatives are investigated for evolutionary covariation. Here, a motif has been described in previous work of [23] and can be written in a generalized, regular expression-like form of XYn, where X and Y correspond to amino acids separated by $n - 1$ highly variable positions. Interaction information contributes to detecting interacting pattern with evolutionary background. This means that evolutionary variation at pattern positions was marked as X. Here, different mutation types like that described in [22] may apply at specific X-position. Subsequently, in this work recently published proteins with PDB-Ids: 4nef, 4oj2 were used to transfer family specific EIPPs to these aquaporin-2 representative proteins. For mention, the protein structure (PDB-Id: 4nef, 4oj2) was published by Frick et al. [24] and Vahedi-Faridi et al. [25] Beyond, both protein structures were considered as unknown structures at time of EIPP generation caused by missing Pfam entries. This led to no consideration of both proteins by EIPPs generation. Aquaporin specific EIPPs were derived from known structures of the corresponding PF00230 family. After obtaining of EIPPs, they were employed to generate interaction block schemes (IBSs). Here we try to illustrate that IBSs are useful graphical visualisation media to represent different interacting patterns which distinguish evolutionary. More specifically, we are able to show if a mutation within a pattern has influence on the evolutionary variability of the interacting counterpart. Eventually, IBSs can be used to support the understanding of the three-dimensional fold for the respective interaction partners and the whole protein structure. Moreover, transmembrane helical information was derived from PDBTM [26] for the proteins to be investigated (PDB-Ids: 4nef, 4oj2). Afterwards, EIPPs were applied on helices and sequence similarity of the incurred interacting ranges compared to known structures of the other family members was calculated. For further investigation, mutations occurring in nephrogenic diabetes insipidus patients were aggregated from recently published works [27–34] and registered on sequences of proteins to be investigated. These natural variants of NDI can also be obtained from UniProt (http://www.uniprot.org/uniprot/P41181). Finally, IBSs were applied to similar sequence parts which include NDI mutational effects.

2.1. Evolutionary Variations within EIPPs. In this section we describe a method to derive variation at X-positions from evolutionary sequence record. To realise this task, the full unknown seed structure dataset (9641 proteins) of the representative family (PF00230) was derived from Pfam database [35]. Transmembrane helical information was obtained using TMHMM Server v.2.0 [36]. A variety of methods have been developed to predict structural features from sequence, such as α-helical membrane-spanning helices and extra/intracellular domains. Basically, TMHMM performs a prediction of intra/extracellular regions and integral membrane helices based on sequence. Beside per-residue predictions TMHMM also lists underlying per-residue assignment probabilities as an indicator of prediction uncertainty. Consequently, helical information was used to apply our derived EIPPs at unknown structures. Here, X-positions were investigated in a closer way when both existing EIPP counterparts were registered in different helices. For the detecting of new evolutionary variations, the amino acid occupancy from unknown structure information was used to compare with amino acids of known structures at specific X-positions. At last, new amino acids at variable X-position were registered and added. Finally, with this method we are able to extend evolutionary information within interacting sequence parts which can be used for further mutational or

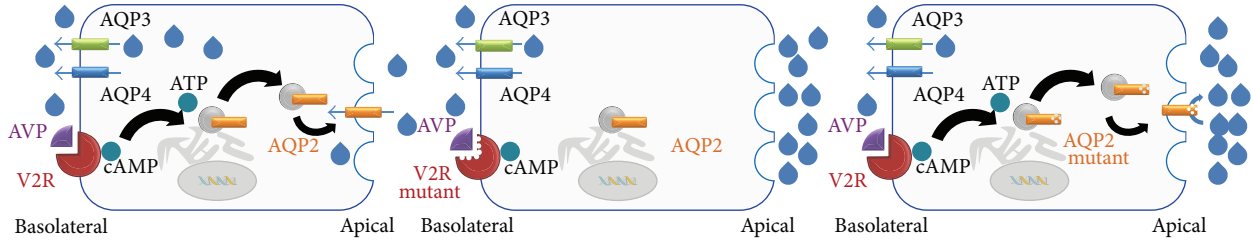


FIGURE 1: (a) In normally regulated water absorption in kidney cells, the antidiuretic hormone arginine vasopressin (AVP) is released in the pituitary gland, binds to the V2 receptor (V2R), and subsequently induces a series of phosphorylation reactions which lead to the insertion of aquaporin-2 water channels in the apical membrane that allow water molecules to pass the membrane. (b) Genetic mutations in the gene encoding V2R lead to reduced binding affinity and protein stability in V2R. Dysfunctional V2R mutants cause a significantly reduced amount of inserted aquaporin-2 proteins and thus decrease the water flux through the apical membrane. On the other hand, dysfunctional aquaporin-2 mutants decrease the water reabsorption as well (see (c)). Reduced water reabsorption is directly linked to an increased output of highly diluted urine (polyuria) and excessive drinking (polydipsia) which are the most severe symptoms observable in nephrogenic diabetes insipidus patients [12, 18, 20].

TABLE 1: Structural similar helical ranges of two given aquaporin-2 representatives in relation to aquaporin family members (PF00230). Similarity values describe what percentages are consistent helical ranges which are covered by EIPPs.

PDB-Id: 4nef		PDB-Id: 4oj2	
Helix	Similarity	Helix	Similarity
1	100%	1	100%
2	100%	2	100%
3	79.1%	3	79.3%
4	100%	4	100%
5	94.4%	5	95.2%
6	100%	6	100%
7	100%	7	100%

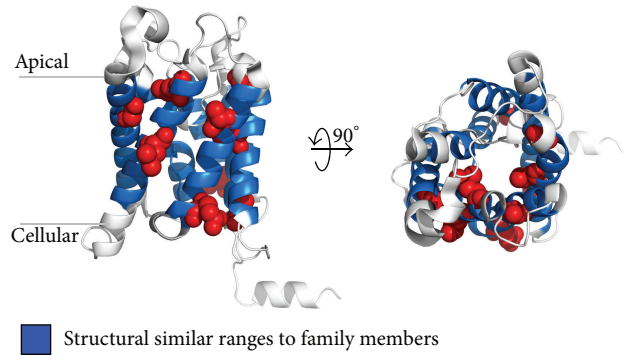


FIGURE 2: Structural colouring of helical ranges with high similarity to aquaporin family members (PF00230). Side and top-down view of the aquaporin-2 representative protein (PDB-Id: 4nef). Blue coloured cartoon residue ranges are present. These consist of family specific EIPPs which were found in known structures of the given family (PF00230). Red coloured spheres illustrate natural variants derived from UniProt (http://www.uniprot.org/uniprot/P41181) of the protein sequence caused by nephrogenic diabetes insipidus. All figures were rendered with PyMOL.

general investigations of membrane proteins or in this case specifically aquaporin water channel proteins.

3. Results and Discussion

Our structure prediction shows, if an unknown structure tends to a family affiliation, family specific EIPPs have to resurface on the protein sequence. Here, EIPPs were derived from known crystal structures of the aquaporin family (PF00230) and marked to α-helical structure of recently published aquaporin-2 representative proteins (PDB-Ids: 4nef, 4oj2). As mentioned before, aquaporin-2 representatives have not been considered in the EIPP generation process and make them to transparent unknown structures. Similarity results are shown in Figures 2 and 3 and confirm the already enlightened family affiliation and predicted structures. This means in all TM-helices EIPPs could be found and cover the helical range with up to 100% as listed in Table 1.

Here our prediction results explain the mightiness of EIPPs. On the one hand they provide useful and powerful information to predict α-helical structures within the transmembrane environment of homologue membrane proteins. On the other hand, we are able to describe selected interacting areas which are constrained by evolution. To evaluate this assumption, different IBSs were generated and applied to highly conserved interacting sequence parts which were derived from Pfam HMM-logos [35]. One IBS example is shown in Figure 4 which illustrates two interacting patterns. These patterns are part of the aquaporin-2 representative protein with PDB-Id: 4nef. Within an X-positioned pattern, we are able to register evolutionarily designed variable positions. Our IBSs additionally show that an interaction with another pattern takes place. In this work, the goal was not to show which pattern position is involved in spatial interaction but rather to show that two patterns build an interacting block. Figure 4 shows examples of variable positions, which can be occupied by different natural variants. With our IBSs, we can show that an interaction between two blocks is given, when the respective positions are occupied by the

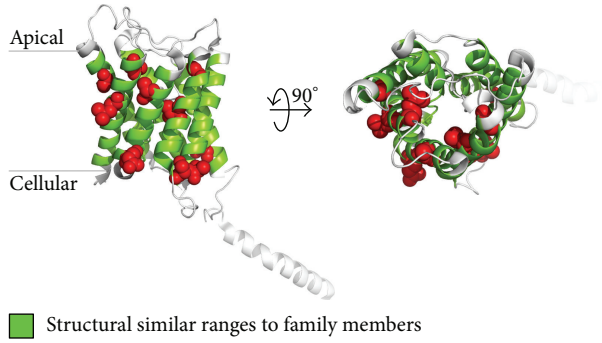


Figure 3: Structural colouring of helical ranges with high similarity to aquaporins (PF00230). Side and top-down view of the aquaporin-2 representative protein (PDB-Id: 4oj2). Green coloured cartoon residue ranges are present. These consist of family specific EIPPs which were found in known structures of the given family (PF00230). Red coloured spheres illustrate natural variants derived from UniProt (http://www.uniprot.org/uniprot/P41181) of the protein sequence caused by nephrogenic diabetes insipidus. All figures were rendered with PyMOL.

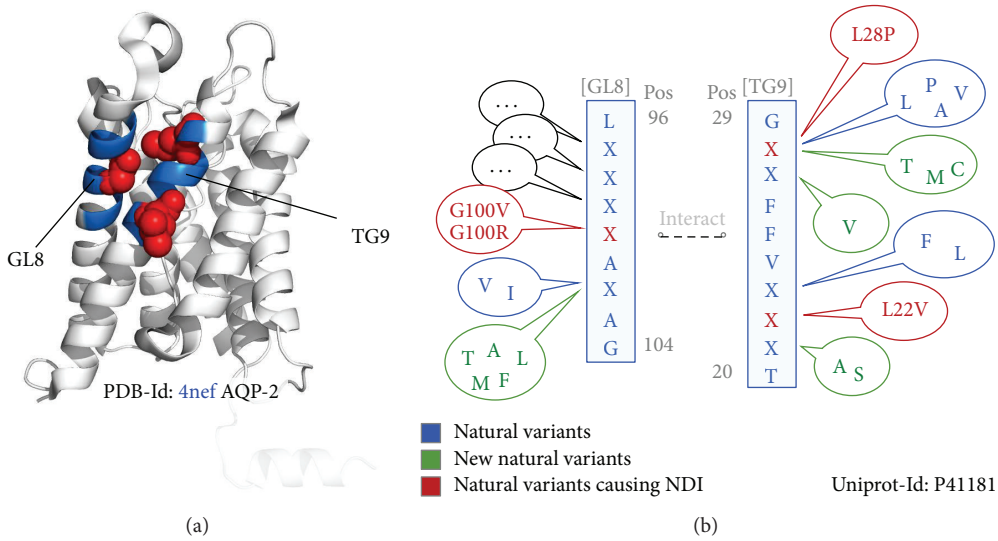


Figure 4: Examples of two spatially interacting subsequences. (a) More specifically, GL8 motif representative pattern (left) interacts with TG9 (right). Both patterns are coloured in blue. Red spheres representing natural variants causing nephrogenic diabetes insipidus (NDI). If we look at these patterns, it is not important to know which pattern position is involved in a spatial interaction. It is more important, that these two pattern build an interacting block. (b) The corresponding interaction block scheme (IBS), with the interacting pattern in blue letters. Red letters illustrating positions causing NDI. All patterns have variable positions. These positions are marked with X. Different coloured bubbles are present and address an X-position with possible natural variants. These are examples of possible occupations, which can be realized by natural variants derived from known structures (blue), natural variants causing NDI (red), or new natural variants derived from unknown structures (green).

amino acids. This implies that a TG9-GL8 interaction is given with Phe23-Ala101 or Phe26 and Ala101 or in the case of NDI caused by Leu28Pro [28], a TG9-GL8 interaction is given with Leu28 and Val102 (red coloured amino acid occupation) at specific positions. Here, IBSs give us a quick overview, that within or across a block evolutionary changes take place. That a destabilizing amino acid substitution is compensated by another position over the evolutionary time scale has been explained in detail in previous work of Morcos et al. [37]. Mutational variation information at specific X-positions can close this gap. This leads to further results in our work, the detection of new X-positioned variations caused by evolution. Here, many variations of new possible amino acid substitutions within different sequence pattern could be obtained. As one example, the TG9 motif representative pattern TLIXVFFXXG is given. For this, X-positions can be occupied with the following amino acids: X3F, X3L, X7G, X8V, X7A, X8P, and X8L, which lead to a final regular expression similar to PROSITE [38–41] pattern syntax TLI[F,L]VFF[G,A][V,A,P,L]G. The evaluation of the VG5 submotif representative pattern VFFGXG shows the flexibility of evolution. Referring to natural variant causing NDI L28P, the fourth position (starting from zero) can be occupied with the following amino acids: X4L, X4P. Here, our

method has spawned new amino acids for this position. X4T, X4M, and X4C are able to complete this block which leads to final regular expression VFFG[V,A,P,L,T,M,C]G. Ultimately, this shows a greater variability outside of VG5. For further tasks in genomics or proteomics like protein modelling or mutational investigations, new amino acid replacements can be included. This expands the view on what structural mechanisms could also be possible to realise the three-dimensional fold within the respective sequence part and finally to ensure the protein function.

4. Conclusion

In the present work, we have applied a new approach for extracting short, spatially interacting amino acid sequence parts, so-called evolutionary interaction pattern pairs (EEIPs), from known structures of membrane proteins, more specifically aquaporin water channel proteins. Based on EIPPs, structure similarity of recently published aquaporin-2 representative proteins was determined and this in silico analysis confirms the aquaporin family affiliation. EIPPs were obtained and employed to generate interaction block schemes of highly conserved sequence parts annotated with natural variants caused by diabetes insipidus. Newly amino acid variations have been discovered. In our further works we will prove the relevance. In conclusion, it is a fact that disease patterns play an important role in membrane proteins but currently few involved structures are available. Different works have shown mutations on a membrane protein sequence influencing disease patterns. Besides, mutations are used in the diagnosis of biomarkers. However, the application of interaction block schemes can lead to better indicators and this in silico analysis can support laboratory mutagen investigations.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Steffen Grunert and Dirk Labudde participated in the design of this study. Steffen Grunert designed all methods and performed the implementation. Steffen Grunert evaluated the results. Dirk Labudde provided valuable consultation on structural biology and procedural steps. Steffen Grunert and Dirk Labudde wrote the paper. All authors read and approved the final paper.

Acknowledgments

The authors would like to thank the Free State of Saxony and the European Social Fond (ESF) for financial support.

References

[1] A. Marsico, D. Labudde, T. Sapra, D. J. Muller, and M. Schroeder, "A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy," *Bioinformatics*, vol. 23, no. 2, pp. e231–e236, 2007.

[2] G. C. Brito and D. W. Andrews, "Removing bias against membrane proteins in interaction networks," *BMC Systems Biology*, vol. 5, article 169, 2011.

[3] S. Tan, T. T. Hwee, and M. C. M. Chung, "Membrane proteins and membrane proteomics," *Proteomics*, vol. 8, no. 19, pp. 3924–3932, 2008.

[4] M. Luckey, *Membrane Structural Biology: With Biochemical and Biophysical Foundations*, Cambridge University Press, 2008.

[5] P. G. Sadowski, A. J. Groen, P. Dupree, and K. S. Lilley, "Subcellular localization of membrane proteins," *Proteomics*, vol. 8, no. 19, pp. 3991–4011, 2008.

[6] P. M. T. Deen, M. A. J. Verdijk, N. V. A. M. Knoers et al., "Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine," *Science*, vol. 264, no. 5155, pp. 92–95, 1994.

[7] S. M. Mulders, D. G. Bichet, J. P. L. Rijss et al., "An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the golgi complex," *The Journal of Clinical Investigation*, vol. 102, no. 1, pp. 57–66, 1998.

[8] A. M. W. van den Ouweland, J. C. F. M. Dreesen, M. Verdijk et al., "Mutations in the vasopressin type 2 receptor gene (AVPR2) associated with nephrogenic diabetes insipidus," *Nature Genetics*, vol. 2, no. 2, pp. 99–102, 1992.

[9] W. Rosenthal, A. Seibold, A. Antaramian et al., "Molecular identification of the gene responsible for congenital nephrogenic diabetes insipidus," *Nature*, vol. 359, no. 6392, pp. 233–235, 1992.

[10] S. Ananthakrishnan, "Diabetes insipidus in pregnancy: etiology, eva luation, and management," *Endocrine Practice*, vol. 15, no. 4, pp. 377–382, 2009.

[11] R. Krysiak, I. Kobielusz-Gembala, and B. Okopien, "Recurrent pregnancy-induced diabetes insipidus in a woman with hemochromatosis," *Endocrine Journal*, vol. 57, no. 12, pp. 1023–1028, 2010.

[12] G. L. Robertson, "Diabetes insipidus," *Endocrinology & Metabolism Clinics of North America*, vol. 24, no. 3, pp. 549–572, 1995.

[13] N. Wettschureck and S. Offermanns, "Mammalian G proteins and their cell type specific functions," *Physiological Reviews*, vol. 85, no. 4, pp. 1159–1204, 2005.

[14] G. Milligan and E. Kostenis, "Heterotrimeric G-proteins: a short history," *British Journal of Pharmacology*, vol. 147, no. 1, pp. S46–S55, 2006.

[15] N. Defer, M. Best-Belpomme, and J. Hanoune, "Tissue specificity and physiological relevance of various isoforms of adenylyl cyclase," *American Journal of Physiology: Renal Physiology*, vol. 279, no. 3, pp. F400–F416, 2000.

[16] J. Hanoune, Y. Pouille, E. Tzavara et al., "Adenylyl cyclases: structure, regulation and function in an enzyme superfamily," *Molecular and Cellular Endocrinology*, vol. 128, no. 1-2, pp. 179–194, 1997.

[17] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[18] E. L. Los, P. M. T. Deen, and J. H. Robben, "Potential of nonpeptide (ant)agonists to rescue vasopressin v2 receptor mutants for the treatment of X-linked nephrogenic diabetes insipidus," *Journal of Neuroendocrinology*, vol. 22, no. 5, pp. 393–399, 2010.

[19] T. M. Strom, K. Hörtnagel, S. Hofmann et al., "Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein," *Human Molecular Genetics*, vol. 7, no. 13, pp. 2021–2028, 1998.

[20] M. Birnbaumer, "V2R structure and diabetes insipidus," *Receptors and Channels*, vol. 8, no. 1, pp. 51–56, 2002.

[21] T. M. Fujiwara and D. G. Bichet, "Molecular biology of hereditary diabetes insipidus," *Journal of the American Society of Nephrology*, vol. 16, no. 10, pp. 2836–2846, 2005.

[22] S. Grunert and D. Labudde, "The observation of evolutionary interaction pattern pairs in membrane proteins," Submitted to *BMC Structural Biology*.

[23] Y. Liu, D. M. Engelman, and M. Gerstein, "Genomic analysis of membrane protein families: abundance and conserved motifs," *Genome biology*, vol. 3, no. 10, 2002.

[24] A. Frick, U. K. Eriksson, F. de Mattia et al., "X-ray structure of human aquaporin 2 and its implications for nephrogenic diabetes insipidus and trafficking," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 17, pp. 6305–6310, 2014.

[25] A. Vahedi-Faridi, D. Lodowski, A. Schenk et al., The structure of aquaporin, Journal, To be Published.

[26] D. Kozma, I. Simon, and G. E. Tusnády, "PDBTM: protein data bank of transmembrane proteins after 8 years," *Nucleic Acids Research*, vol. 41, no. 1, pp. D524–D529, 2013.

[27] M. C. Canfield, B. K. Tamarappoo, A. M. Moses, A. S. Verkman, and E. J. Holtzman, "Identification and characterization of aquaporin-2 water channel mutations causing nephrogenic diabetes insipidus with partial vasopressin response," *Human Molecular Genetics*, vol. 6, no. 11, pp. 1865–1871, 1997.

[28] N. Marr, D. G. Bichet, S. Hoefs et al., "Cell-biologic and functional analyses of five new Aquaporin-2 missense mutations that cause recessive nephrogenic diabetes insipidus," *Journal of the American Society of Nephrology*, vol. 13, no. 9, pp. 2267–2277, 2002.

[29] S.-H. Lin, D. G. Bichet, S. Sasaki et al., "Two novel aquaporin-2 mutations responsible for congenital nephrogenic diabetes insipidus in Chinese families," *The Journal of Clinical Endocrinology & Metabolism*, vol. 87, no. 6, pp. 2694–2700, 2002.

[30] P. Carroll, H. Al-Mojalli, A. Al-Abbad et al., "Novel mutations underlying nephrogenic diabetes insipidus in Arab families," *Genetics in Medicine*, vol. 8, no. 7, pp. 443–447, 2006.

[31] S. M. Mulders, N. V. A. M. Knoers, A. F. van Lieburg et al., "New mutations in the aqp2 gene in nephrogenic diabetes insipidus resulting in functional but misrouted water channels," *Journal of the American Society of Nephrology*, vol. 8, no. 2, pp. 242–248, 1997.

[32] R. Vargas-Poussou, L. Forestier, M. D. Dautzenberg, P. Niaudet, M. Déchaux, and C. Antignac, "Mutations in the vasopressin v2 receptor and aquaporin-2 genes in 12 families with congenital nephrogenic diabetes insipidus," *Journal of the American Society of Nephrology*, vol. 8, no. 12, pp. 1855–1862, 1997.

[33] K. Goji, M. Kuwahara, Y. Gu, M. Matsuo, F. Marumo, and S. Sasaki, "Novel mutations in aquaporin-2 gene in female siblings with nephrogenic diabetes insipidus: evidence of disrupted water channel function," *The Journal of Clinical Endocrinology & Metabolism*, vol. 83, no. 9, pp. 3205–3209, 1998.

[34] M. Kuwahara, "Aquaporin-2, a vasopressin-sensitive water channel, and nephrogenic diabetes insipidus," *Internal Medicine*, vol. 37, no. 2, pp. 215–217, 1998.

[35] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[36] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[37] F. Morcos, A. Pagnani, B. Lunt et al., "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[38] C. J. A. Sigrist, E. De Castro, L. Cerutti et al., "New and continuing developments at PROSITE," *Nucleic Acids Research*, vol. 41, no. 1, pp. D344–D347, 2013.

[39] C. J. A. Sigrist, L. Cerutti, N. Hulo et al., "Prosite: a documented database using patterns and profiles as motif descriptors," *Briefings in bioinformatics*, vol. 3, no. 3, pp. 265–274, 2002.

[40] E. de Castro, C. J. A. Sigrist, A. Gattiker et al., "Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins," *Nucleic Acids Research*, vol. 34, supplement 2, pp. W362–W365, 2006.

[41] C. J. A. Sigrist, E. de Castro, P. S. Langendijk-Genevaux, V. Le Saux, A. Bairoch, and N. Hulo, "ProRule: a new database containing functional and structural information on PROSITE profiles," *Bioinformatics*, vol. 21, no. 21, pp. 4060–4066, 2005.