BBA - Molecular Basis of Disease xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - Molecular Basis of Disease

journal homepage: www.elsevier.com/locate/bbadis

Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine

Yu-Hang Zhang, Yu Hu, Yuchao Zhang, Lan-Dian Hu*, Xiangyin Kong**

Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

ARTICLE INFO

Keywords:
Hematopoiesis
Hematopoietic stem cells
Support vector machine
Sequential minimum optimization
Minimal redundancy maximal relevance

ABSTRACT

Hematopoiesis is a complicated process involving a series of biological sub-processes that lead to the formation of various blood components. A widely accepted model of early hematopoiesis proceeds from long-term hematopoietic stem cells (LT-HSCs) to multipotent progenitors (MPPs) and then to lineage-committed progenitors. However, the molecular mechanisms of early hematopoiesis have not been fully characterized. In this study, we applied a computational strategy to identify the gene expression signatures distinguishing three types of closely related hematopoietic cells collected in recent studies: (1) hematopoietic stem cell/multipotent progenitor cells; (2) LT-HSCs; and (3) hematopoietic progenitor cells. Each cell in these cell types was represented by its gene expression profile among a total number of 20,475 genes. The expression features were analyzed by a Monte-Carlo Feature Selection (MCFS) method, resulting in a feature list. Then, the incremental feature selection (IFS) and a support vector machine (SVM) optimized with a sequential minimum optimization (SMO) algorithm were employed to access the optimal classifier with the highest Matthews correlation coefficient (*MCC*) value of 0.889, in which 6698 features were used to represent cells. In addition, through an updated program of MCFS method, seventeen decision rules can be obtained, which can classify the three cell types with an overall accuracy of 0.812. Using a literature review, both the rules and the top features used for building the optimal classifier were confirmed to be commonly used or potential biological markers for distinguishing the three cell types of HSPCs. This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

1. Introduction

Hematopoiesis (also spelled hemopoiesis) is the formation of various complex blood components, including hematopoietic cells and related non-cellular substrates like platelets [1,2]. The hematopoietic stem cells (HSCs) give rise to both the two major cellular component of blood components: myeloid and lymphoid lineages, reflecting its unique regulatory role in the hematopoietic system [3].

For centuries, many studies have concentrated on the circulatory system, especially the blood circulation system [4–6]. A detailed atlas for the cellular and non-cellular components of the peripheral blood in various species has been fully elucidated [4]. Even the subgroups of functional components in peripheral blood like lymphocytes and monocytes have been analyzed in detail based on morphological characteristics and molecular markers [7]. For example, lymphocytes are functional white blood cells in the vertebrate immune system that include NK cells (Natural Killer cells), T cells and B cells. Cells in this group can be further classified into three subgroups based on their specific phenotypic markers [7]. For instance, NK cells have specific biological markers such as CD16 and CD56, whereas T helper cells express TCRαβ, CD3 and CD4, implying that morphological characteristics and molecular markers may be a crucial foundation for the identification of various hematopoiesis outcomes [8–10]. However, such methods are not suitable for HSCs and related progenitor cells because it is difficult to further classify HSCs and progenitor cells only based on morphological features and characteristic molecular markers [11–13].

With the development of high-throughput sequencing, including whole transcriptome sequencing, the differential gene expression profiles of blood cells at different developmental stages provide new insight into the biological processes occurring during hematopoiesis and further reveal the specific transcriptome markers for various cell types in the hematopoietic system [14]. Whole transcriptome sequencing can precisely reveal what components are actually expressed at the

* Corresponding author at: Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Shanghai, People's Republic of China.
** Corresponding author at: Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China.
E-mail addresses: zhangyuchao@sibs.ac.cn (Y. Zhang), ldhu@sibs.ac.cn (L.-D. Hu), xykong@sibs.ac.cn (X. Kong).

https://doi.org/10.1016/j.bbadis.2017.12.003
Received 11 October 2017; Received in revised form 20 November 2017; Accepted 1 December 2017

transcriptome level, which is close to the actual protein level [15,16]. However, it is not sufficient to only focus on the transcriptome. Previous studies have mainly investigated the gene expression profiles of various cell subgroups at the macro level [15,16]. Considering that even cells of the same type may also have gene expression profiles with slight differences, sequencing at the macro level may not only differentiate cells of different subtypes but might also obscure specific gene expression markers in each subtype. Advances in genetic sequencing technology have allowed sequencing with single-cell resolution, providing us with an ideal tool to study the complicated hematopoietic system [17–19].

According to recent publications, hematopoietic stem and progenitor cells (HSPCs) can be differentiated through three sorting gates corresponding to three cell types: hematopoietic stem cell/multipotent progenitor (HSPC gate), long-term hematopoietic stem cell (LT-HSC gate) and hematopoietic progenitor cell (Prog gate) [20]. HSPC-gate cells include short-term HSCs (ST-HSCs), MPPs and lymphoid multipotent progenitors (LMPPs), whereas Prog gate cells include megakaryocyte-erythrocyte progenitors (MEPs), common myeloid progenitors (CMPs), and granulocyte-monocyte progenitors (GMPs) [20]. The self-renewable LT-HSCs have been confirmed to initially give rise to MPPs and Prog cells, which lose self-renewal capacity [21,22]. Further downstream, MPPs give rise to common lymphoid progenitor (CLP) and CMP cells [23,24], which advance to MEP and GMP [25,26]. These Prog gate cells have partially lost pluripotency for further differentiation and have specific lineage restriction [27]. According to recent publications, these three subtypes of hematopoietic cells have similar morphological features and molecular markers. However, they may play different roles in the hematopoietic system [28–30]. Although various research efforts have tried to determine expression profiles that can be used to distinguish single hematopoietic cell types in mice, the core expression characteristics of HSCs, including the three major subtypes, and the origin of the hematopoietic cell differentiation are not fully understood.

In this study, based on the single-cell sequencing data of a recent study on the mouse hematopoietic system [20], we computationally analyzed the data to extract the core, characteristic expressed genes that may contribute to the development and biological functions of hematopoietic cells at their specific developmental stages. As mentioned above, hematopoietic cells are grouped into three subtypes: (1) HSPC gate cells; (2) LT-HSC gate cells; and (3) Prog gate cells [20]. Each cell was encoded based on its gene expression profiles among a total number of 20,475 genes. We applied an MCFS method [31,32] to identify the decision rules classifying the three cell types and ranked the aforementioned 20,475 features. Then, an incremental feature selection (IFS) method, together with a support vector machine (SVM) [33,34], were adopted to extract optimal gene expression features. Through these features, we were able to build an optimal classifier with the best performance in distinguishing three subtypes of hematopoietic cells, as evidenced by the highest Matthews correlation coefficient value of 0.889. Further analysis using a literature review of the top ranked 80 features and the seventeen decision rules confirmed their relevance for HSCP cell types. This study sheds light on the molecular mechanisms of early hematopoiesis.

2. Materials and methods

2.1. Gene expression profiles of hematopoietic cells

We downloaded the RNA-sequencing gene expression profiles of 1920 hematopoietic cell samples from Gene Expression Omnibus (GEO) under the accession number GSE81682 [20]. These cells were classified into 852 HSPCs, 216 LT-HSCs and 852 Prog. Genes with non-zero expression in more than 10% of cells were selected, resulting in 20475 genes. Each cell sample was represented based on its expression levels of these 20475 genes.

The gene expression profiles for the three cell sample subtypes were further processed with quantile normalization and $\log_2$ transformed. Finally, each cell sample can be represented by 20,457 features, each feature representing the expression level of a gene in the cell samples. Because three subtypes of cell samples were considered in this study, the problem can be transformed to a three-class classification problem.

2.2. Feature ranking and decision rule discovery

In data mining, several feature selection methods have been proposed to extract important features, such as minimal redundancy maximal relevance (mRMR) [35], ReliefF [36], maximum relevance maximum distance (MRMD) [37], etc. These methods have been successfully applied to tackle various biological problems [38–51].

In this study, we investigated 20475 features in 1920 samples, a high-dimensional dataset. The feature selection method, MCFS method [31,32], which is quite different from the aforementioned methods, was adopted to analyze the importance of the 20475 features because this method has been validated to be suitable for dealing with high-dimensional datasets. In addition, it has been successfully used to tackle biological problems [31,32,52–56]. A brief description of this method was as follows.

To score all features and select informative ones, the MCFS method always constructs tree classifiers from an original training dataset. If a feature participates more in the classification among the constructed tree classifiers, it is deemed to be more important. In detail, it constructs s feature subsets, each built by randomly selecting m features from all features, where m should be much smaller than the total number of features, denoted as d. Then, for each feature set, t tree classifiers are built and trained by a random selection of training and testing datasets using the original training set. Thus, a total of $s \cdot t$ classifier trees are constructed. Based on these trees, the relative importance (RI) of a feature g is determined by

$$\mathrm{RI}_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{no.\ \ in\ n_g(\tau)}{no.\ \ in\ \tau} \right)^v \tag{1}$$

where $wAcc$ represents the weighted accuracy of the $\tau - th$ tree, $\mathrm{IG}(n_g(\tau))$ represents the information gain of node $n_g(\tau)$, (no. in $n_g(\tau)$) represents the number of samples in node $n_g(\tau)$, (no. in τ) represents the number of samples in the $\tau - th$ tree, and u and v are fixed real numbers. According to the RI value assigned to each feature, all features can be ranked by decreasing order of their RI values, resulting in a feature list. Here, we formulated this feature list as

$$\mathrm{F} = [f_1, f_2, \cdots, f_d] \tag{2}$$

where d is the total number of features (d = 20,457 in this study).

In this study, we downloaded the program of MCFS at http://www.ipipan.eu/staff/m.draminski/mcfs.html. Through this program, we can assess not only the feature list as formulated in Eq. (2) but also several decision rules. The procedures of yielding these decision rules were as follows.

Based on the feature list yielded by the MCFS method, user can select the most informative features by setting the percentage of features users want to select. According to these informative features, a rule network [52] can be built separately. In detail, n subsets are generated by subsampling the original dataset harboring these informative features n times. Rule-based classifiers (e.g., the rough sets in [57]) can be built on each subset, thereby producing a number of IF-THEN rules. Each rule contains one or more than one condition determining a decision attribute. For each subsampled dataset, a Johnson Reducer algorithm implemented in ROSETTA software is used to generate reducts, the minimal sets of features for a classification task, and the associated rules. In this study, the obtained rules yielded by the program of MCFS on 1920 hematopoietic cell samples was investigated, which can provide direct insight into the differences of the three types of