International Journal of
Molecular Sciences

MDPI

Article

Tissue Expression Difference between mRNAs and lncRNAs

Lei Chen [1,2,3], Yu-Hang Zhang [4], Xiaoyong Pan [5], Min Liu [2], Shaopeng Wang [1], Tao Huang [4,*] and Yu-Dong Cai [1,*]

1 School of Life Sciences, Shanghai University, Shanghai 200444, China; chen_lei1@163.com (L.C.); wsptfb@163.com (S.W.)
2 College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; liumin@shmtu.edu.cn
3 Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China
4 Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; zhangyh825@163.com
5 Department of Medical Informatics, Erasmus MC, 3000 CA Rotterdam, The Netherlands; xypan172436@gmail.com
* Correspondence: tohuangtao@126.com (T.H.); cai_yud@126.com (Y.-D.C.); Tel.: +86-21-5492-3269 (T.H.); +86-21-6613-6132 (Y.-D.C.)

Received: 12 September 2018; Accepted: 28 October 2018; Published: 31 October 2018

Abstract: Messenger RNA (mRNA) and long noncoding RNA (lncRNA) are two main subgroups of RNAs participating in transcription regulation. With the development of next generation sequencing, increasing lncRNAs are identified. Many hidden functions of lncRNAs are also revealed. However, the differences in lncRNAs and mRNAs are still unclear. For example, we need to determine whether lncRNAs have stronger tissue specificity than mRNAs and which tissues have more lncRNAs expressed. To investigate such tissue expression difference between mRNAs and lncRNAs, we encoded 9339 lncRNAs and 14,294 mRNAs with 71 expression features, including 69 maximum expression features for 69 types of cells, one feature for the maximum expression in all cells, and one expression specificity feature that was measured as Chao-Shen-corrected Shannon's entropy. With advanced feature selection methods, such as maximum relevance minimum redundancy, incremental feature selection methods, and random forest algorithm, 13 features presented the dissimilarity of lncRNAs and mRNAs. The 11 cell subtype features indicated which cell types of the lncRNAs and mRNAs had the largest expression difference. Such cell subtypes may be the potential cell models for lncRNA identification and function investigation. The expression specificity feature suggested that the cell types to express mRNAs and lncRNAs were different. The maximum expression feature suggested that the maximum expression levels of mRNAs and lncRNAs were different. In addition, the rule learning algorithm, repeated incremental pruning to produce error reduction algorithm, was also employed to produce effective classification rules for classifying lncRNAs and mRNAs, which gave competitive results compared with random forest and could give a clearer picture of different expression patterns between lncRNAs and mRNAs. Results not only revealed the heterogeneous expression pattern of lncRNA and mRNA, but also gave rise to the development of a new tool to identify the potential biological functions of such RNA subgroups.

Keywords: mRNA; lncRNA; expression specificity; cell type; feature selection

1. Introduction

Ribonucleic acid (RNA) is one of the most significant nucleic acid components in all living creatures except for some viruses [1,2]. In general, nucleic acid is one of the three major macromolecules

that maintain the fundamental biological processes of all known forms of life [2]. Among the different subtypes of nucleic acid, RNA is a specific functional component that has specific structures and functions. As for the structures, generally, DNA, as another major component of nucleic acid in our living cells, exists as desoxy-ribonucleic acid of double-strand status with bases as A, G, C, and T in the cell nucleus as the transporter of genetic materials [3]. However, RNA that plays different biological roles generally acts as single-strand ribonucleic acid, or in other words, a chain of ribonucleic acid with specific base U but not T [4–6]. Aside from the structure differences, the biological functions of RNA are definitely specific [7]. Usually, RNA contributes to the maintenance and regulation of gene expression and function, regulating the coding, decoding, transcription, and translation of a specific protein, the final functional component of most biological processes; these factors are quite different from the biological functions of other nucleic acid subtypes, such as DNA, as the genetic passenger in the cell nucleus [7].

Similar to other material categories, RNA is also not a consistent group of subtype members with similar structure and functions. Messenger RNA (mRNA) and long noncoding RNA (lncRNA) are the two main subgroups of RNAs that participate in transcription regulation [8,9]. In general, RNA can be divided into two main subgroups, namely, coding RNA (mRNA) and noncoding RNA (ncRNA), according to their basic biological function, whether they can be further translated into proteins. For a long time, investigators mainly focus on the mRNA with specific protein-coding potentials and have been considered to be the only functional component of RNAs in living cells for a long time [9]. However, with the development of biotechnologies and the deepening of understanding on mRNA transcription, translation, and protein coding, ncRNAs are obtaining considerable attention in multiple biological research fields due to their irreplaceable regulatory roles of such biological processes [10–12]. Among these ncRNAs, lncRNAs turn out to be one of the particular subgroups with specific biological structures and functions.

In general, lncRNAs refer to nonprotein coding RNAs that are longer than 200 nucleotides [11,13]. Considering the distinctive structure of lncRNAs, which is different from other regulatory ncRNAs with quite short sequences similar to microRNAs (miRNAs) and short-interfering RNAs (siRNAs), lncRNAs have specific biological functions and, therefore, may be quite irreplaceable for the physical regulation [14–16]. Gene transcription regulation, post transcriptional regulation, and epigenetic regulation are the three main biological functions of lncRNAs according to existing literature [17–20]. The major biological functions of lncRNAs are gene-specific transcription regulation, which can be concluded into the gene transcription regulation subgroup, considering the specific gene *TP53*, a famous tumor suppressor, as an example. Recent publications confirmed that natural antisense transcripts of *TP53*, named *Wrap53*, which is also a functional lncRNA, directly participates in the regulation of *TP53* transcription and translation by targeting the 5′ untranslated region of *TP53*, enhancing the expression of such gene; this condition implies the gene-specific regulatory function of lncRNAs [21–23]. Considering that lncRNAs may have gene specific regulatory functions in living cells, as a regulator, lncRNAs may also have cell-type specific expression pattern, corresponding to their mRNA targets and reflecting biological functional features of such cell type similar to the cell-type specific expression of functional genes. Although lncRNAs have been widely reported to be multifunctional and play an irreplaceable role in various biological processes, identifying the distributive features and expression pattern of lncRNAs, especially the cell-type specific expression pattern, is still quite difficult.

In this study, we tried to identify specific cell types, in which lncRNAs may play the most irreplaceable roles. Considering that lncRNAs tend to have lower expression level comparing to coding RNAs, the direct comparison of lncRNA expression level with detailed fragments per kilobase million (FPKM) may not be suitable for further cell typing in this study, introducing systematic errors. Therefore, we introduced a specific benchmark for further screening. This method turns relies on the dissimilarity of the lncRNA and mRNA expression pattern. A recent study on the expression pattern of human lncRNAs confirmed the cell-type specific lncRNA expression pattern [24]. For such dataset,

we proposed a computational scheme incorporating several advanced computational methods, such as maximum relevance minimum redundancy (mRMR) method [25], incremental feature selection (IFS) method [26], random forest (RF) algorithm [27], and repeated incremental pruning to produce error reduction (RIPPER) algorithm [28], to analyze this dataset. For the first time, we identified a group of cell types as well as expression pattern features and rules that may be associated with distinctive expression pattern of lncRNAs and mRNAs, contributing to the identification of lncRNA specific expression pattern and further revealing the potential biological functions of lncRNAs.

2. Results

In this study, we built a computation scheme to analyze the expression features of lncRNAs and mRNAs, trying to extract essential differences between these two types of RNAs. The entire procedures are illustrated in Figure 1.

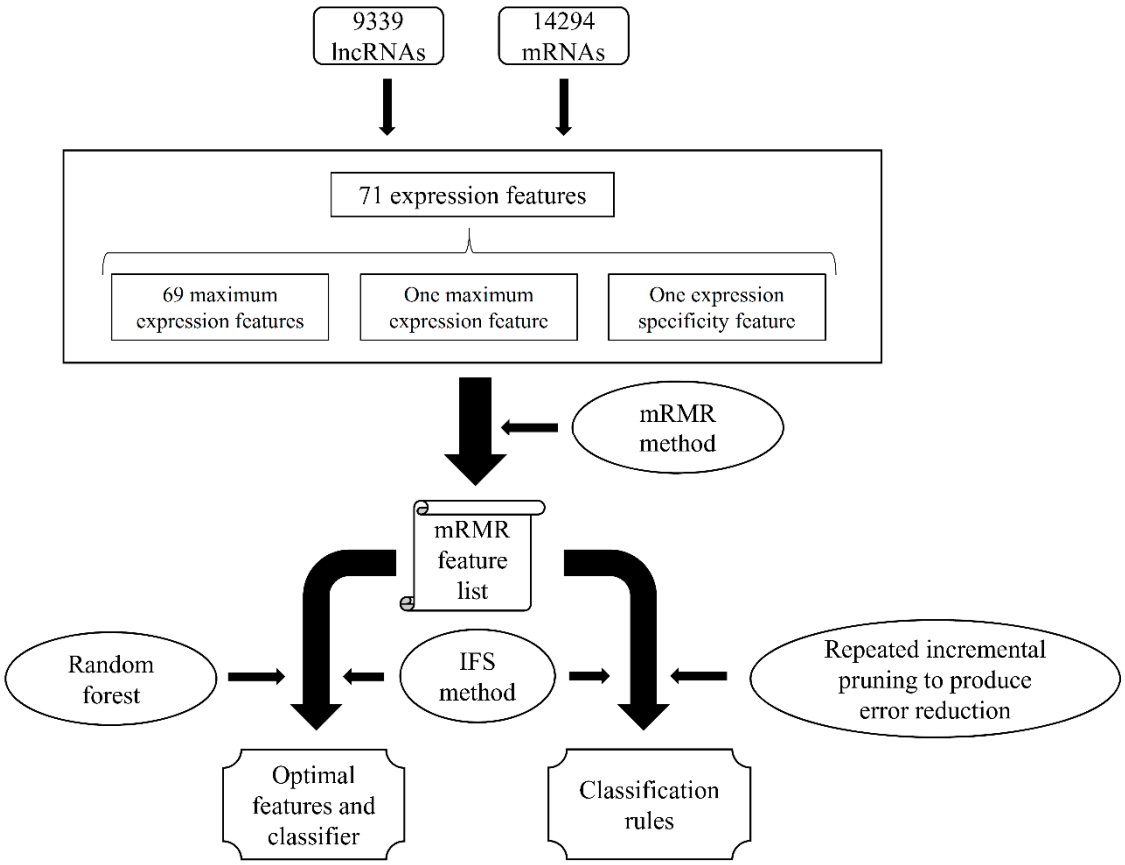

Figure 1. Entire procedures of the computational scheme for investigating lncRNA and mRNA with expression features. The 71 expression features were analyzed by the maximum relevance minimum redundancy (mRMR) method, resulting in an mRMR feature list. Then, the incremental feature selection (IFS) method with the random forest (RF) algorithm was used to extract the optimal features and build the optimal RF classifier. At the same time, the IFS method with the repeated incremental pruning to produce error reduction (RIPPER) algorithm was adopted to learn classification rules.

2.1. Results of the Maximum Relevance Minimum Redundancy (mRMR) Method

Each lncRNA or mRNA was encoded into 71 expression features that are described in detail in Section 4.1. Then, a popular feature selection procedure, mRMR method [25], was applied to analyze these features, producing the mRMR feature list, wherein the 71 features were sorted according to their relevance to target variable and redundancies to other features. The obtained mRMR feature list is presented in Table S1.

2.2. Results of the Incremental Feature Selection (IFS) Method with Random Forest (RF)

In the mRMR feature list, features were sorted in descending order according to their importance on discriminating lncRNAs from mRNAs. Based on this view, the combination of some top features in the list can yield accurate prediction. Thus, we applied the IFS method, together with RF as classification algorithm, to extract the optimal combination of features, thereby building an optimal RF classifier. In detail, we constructed 71 feature subsets according to the mRMR feature list, each of which contained some top features in the list. Then, for each feature subset, all RNAs were represented by features in this set and a RF classifier was built on them. A 10-fold cross-validation was adopted to evaluate the performance of each classifier. The prediction performances of classifiers represented by the four measurements (sensitivity (SN), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC)) are listed in Table S2.

For ease of observation, a curve, named IFS curve, was plotted by setting the *MCC* values as the Y-axis and the number of features participating in building classifiers as the X-axis. As shown in Figure 2, *MCC* reaches the optimal value of 0.895 when the first 53 features were used in classification. Accordingly, these 53 features were picked up as optimal features for RF. In addition, the corresponding RF classifier, using these features to represent lncRNAs and mRNAs, was denoted as the optimal RF classifier. Moreover, the optimal classifier yielded the SN of 0.963, SP of 0.940, and ACC of 0.949, listed in Table 1. This result indicated that the constructed optimal classifier performed well.

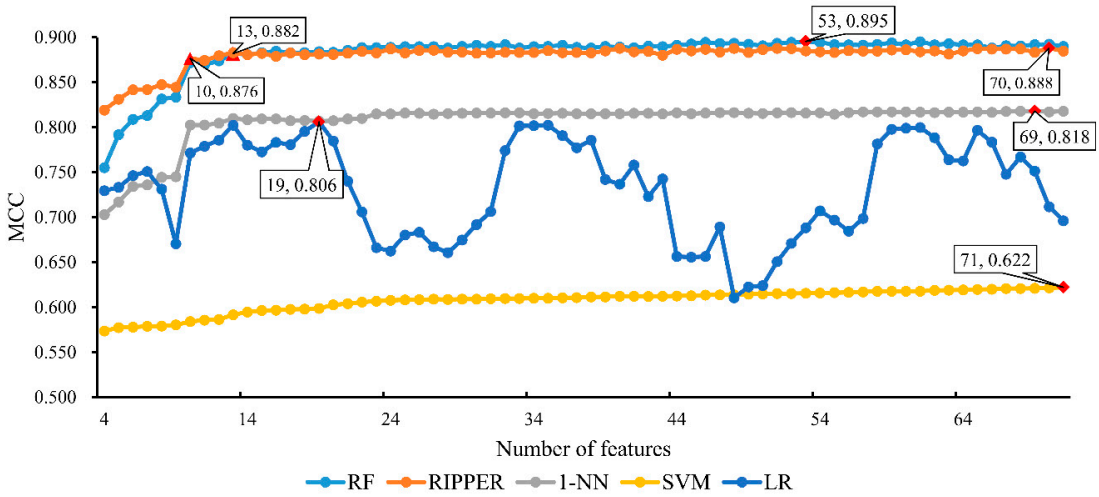

Figure 2. IFS curves based on the predicted results yielded by the IFS method with five different classification algorithms. The X-axis represents the number of features participating in the classification, and the Y-axis represents the Matthew's correlation coefficients (*MCC*s). The optimal *MCC* (marked with red diamonds) for random forest (RF), repeated incremental pruning to produce error reduction (RIPPER), nearest neighbor algorithm (1-NN), support vector machine (SVM) and logistic regression (LR) is 0.895, 0.888, 0.818, 0.622 and 0.806, respectively. For RF, when top 13 features were used, the *MCC* value first overcomes 0.880. While for RIPPER, the MCC value first achieves 0.870 when top 10 features were employed.

Table 1. Performance of the optimal random forest (RF), repeated incremental pruning to produce error reduction (RIPPER), nearest neighbor algorithm (1-NN), support vector machine (SVM), and logistic regression (LR) classifier.

Classification Algorithm	Number of Used Features	SN	SP	ACC	MCC
RF	53	**0.963**	0.940	**0.949**	**0.895**
RIPPER	70	0.952	**0.942**	0.946	0.888
1-NN	69	0.932	0.896	0.911	0.818
SVM	71	0.758	0.861	0.820	0.622
LR	19	0.944	0.876	0.903	0.806

2.3. Results of the IFS Method with Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

Based on the IFS method and RF, we built an optimal RF classifier with good performance (MCC = 0.895). However, this classifier was a black box, which cannot clearly indicate the differences between lncRNAs and mRNAs. In view of this, the rule learning algorithm, RIPPER algorithm, was adopted to do the same procedures that were done for RF. The purpose was to extract informative rules that can give a clear picture on different expression patterns between lncRNAs and mRNAs.

Similar to the IFS method with RF, for each of constructed feature subsets, we constructed an RIPPER classifier, which contained several classification rules, and evaluated its performance with 10-fold cross-validation. The performance of these classifiers is listed in Table S3. For easy observation, we also plotted an IFS curve in Figure 2. It can be observed that when the first 70 features were used to construct the RIPPER classifier, it yielded the best MCC of 0.888, which was slightly lower than that yielded by the optimal RF classifier. Accordingly, the optimal RIPPER classifier was built using the first 70 features in the mRMR feature list. The detailed performance, including SN, SP and ACC, is listed in Table 1. Compared with the performance of the optimal RF classifier, RIPPER classifier yielded higher SP and lower SN, ACC and MCC. However, they were only slightly lower than those of the optimal RF classifier. Thus, the performance of the optimal RIPPER classifier was competitive compared with optimal RF classifier.

2.4. Comparison of the IFS Method with Other Classification Algorithms

As mentioned above, we selected the RF as the classification algorithm for constructing the best classifier. From the results in Section 2.3, we can see that RF was slightly superior to RIPPER. Here, we further employed other three classification algorithms, namely, nearest neighbor algorithm (1-NN) [29], support vector machine (SVM) [30], and logistic regression (LR) [31], to clarify that RF is a proper choice. To this end, we employed three tools, called "IBk", "SMO" and "Logistic", in Weka, which implement the aforementioned algorithms. For convenience, these three tools were executed with their default parameters. The procedures for these three algorithms were almost the same as those for RF and RIPPER. The performance of these three algorithms on the constructed feature subsets is listed in Table S4. In addition, an IFS curve was plotted for the results obtained by each algorithm, as shown in Figure 2. The highest MCC for 1-NN was 0.818 when top 69 features in the mRMR feature list were used, whereas for SVM, the highest MCC was 0.622 when all features (71 features) were used, and the highest MCC was 0.806 for LR, which was obtained based on 19 features in the mRMR feature list. Obviously, MCCs yielded by these three algorithms were all much lower than that yielded by the optimal RF classifier. The detailed performance of the optimal classifiers based on five classification algorithms is listed in Table 1, from which we can see that all four measurements produced by the optimal RF classifier were higher than those obtained by the optimal 1-NN, SVM and LR classifier. This finding suggested that the selection of the RF as the classification algorithm is a good choice. Furthermore, Figure 2 shows that the performance of the RF classifier based on each constructed feature subset is better than that of other three classifiers because the IFS curve of RF is always above the IFS curves of 1-NN, SVM and LR. This outcome further proved the abovementioned conclusions.

3. Discussion

3.1. Analysis of Important Features for Constructing Optimal RF Classifier

As mentioned in Section 2.2, we extracted 53 optimal features for RF. Among them, the most were regarding cell types, wherein two types of RNAs might express differentially. However, analyzing all these features one by one and finding the central tendency of cell types are quite difficult. Meanwhile, these processes are helpful to discriminate different expression patterns of the two RNA types. By observing Figure 2, when top 13 features (listed in Table 2) were used to construct the RF classifier, the *MCC* value first overcomes 0.880 that was almost equal to that of the optimal RF classifier. Among these 13 features, one feature is expression specificity, describing the general gene-based expression

levels and specificity in primary cell facets. To indicate the statistical significance of these 13 features, we randomly constructed 1000 feature subsets, each of which contained the feature of expression specificity and 12 other features randomly selected from rest 70 features. Based on each of these feature subsets, we built a RF classifier and evaluated its performance with 10-fold cross-validation. Accordingly, we accessed 1000 *MCC*s and draw their box plot in Figure 3, from which we can see that 0.880, yielded by the first 13 features in the mRMR feature list, lies at the top of the box plot. Furthermore, the mean and standard deviation of 1000 *MCC*s were 0.860 and 0.015. 0.880 was larger than $0.860 + 1.3 \times 0.015$, suggesting it was high with statistical significance. Therefore, these 13 features contained essential information about cell types that can be used to distinguish lncRNAs from mRNAs and are reviewed in this section.

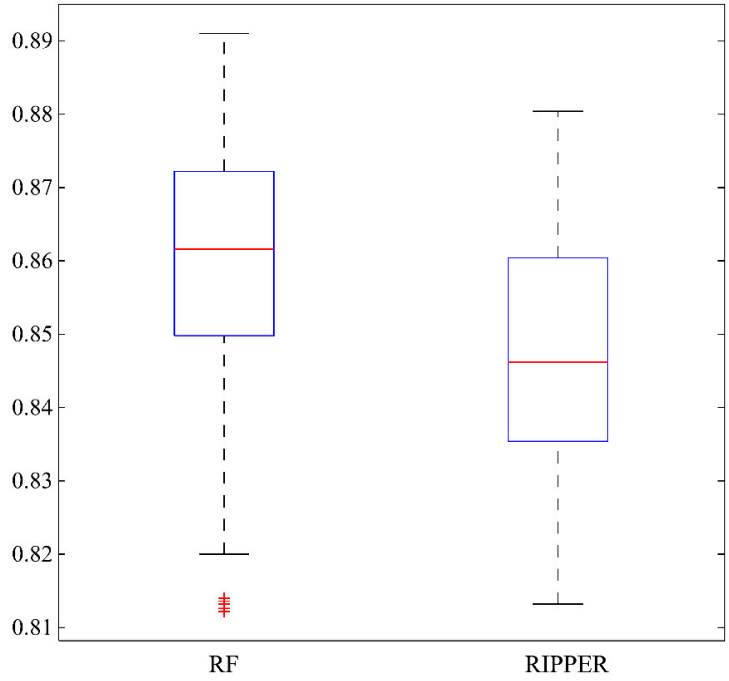

Figure 3. A box plot to illustrate the performance, measured by Matthew's correlation coefficient, of random forest (RF) and repeated incremental pruning to produce error reduction (RIPPER) algorithms on 1000 randomly produced feature subsets. For RF, each feature subset contained the feature of expression specificity and 12 other features randomly selected from rest 70 features, while for RIPPER, each feature subset contained the feature of expression specificity and 9 other features randomly selected from rest 70 features.

Table 2. Top 13 features in maximum relevance minimum redundancy (mRMR) feature list.

No.	Feature Name
1	Expression specificity
2	Intestinal epithelial cell
3	Neutrophil
4	Hepatocyte
5	Mast cell
6	Fibroblast of the conjuctiva
7	Reticulocyte
8	Mesenchymal cell
9	Lymphocyte of b lineage
10	Neuronal stem cell
11	Macrophage
12	Pericyte cell
13	Max cpm in all facet

Among such top 13 features, 11 features are about detailed cell subtypes, in which lncRNAs and mRNAs may have quite different expression patterns. This condition implied that such cell subtype may be a potential benchmark for further identification of specific lncRNA functions. As for the two remaining features, one of them, namely, expression specificity, describes the general gene-based expression levels and specificity in primary cell facets. This aspect is definitely associated with lncRNA regulation. Another feature, named max cpm (count per million) in all facets, describes the detailed expression quantity of either lncRNAs or mRNAs. Therefore, such regulatory factor may also be quite essential for the distinction of lncRNA and mRNA, considering the distinctive expression pattern of such two RNA subtypes according to recent publications [32–34]. Based on recent publications, all top 13 features, or in other words, cell subtypes can be confirmed to have specific lncRNA and mRNA expression pattern. The detailed analysis can be seen below.

Among the top 13 features, 11 were involved in cell subtypes that were determined to have specific distinctive lncRNA expression pattern compared with respective mRNAs. This condition indicated that in such cell subtypes, lncRNA may play specific biological roles, which can be validated by recent publications. Among them, intestinal epithelial cell (with rank 2 in the mRMR feature list) was deemed to be the optimal cell type with distinctive lncRNA expression pattern. Recent publications confirmed that various lncRNAs are upregulated in the intestinal epithelial cells, contributing to various gene expression regulations in response to environmental stimulus [35,36]. In general, intestinal epithelial cell directly interacts with the exogenous diet. In response to the stimulus of exogenous food or microbes, some certain biological functions of intestinal epithelial cells need to be modulated in time [37]. However, considering that the pre-transcriptional regulation of gene expression is quite hard and time consuming, lncRNAs as a functional component of post-transcription regulation may definitely act as a major regulator against such stimulus, resulting in the distinctive and flexible expression pattern of lncRNAs and mRNAs [38,39]. Aside from intestinal epithelial cell, which is a functional component of the digestive system, another specific cell subtype (hepatocyte) may have its respective lncRNA expression pattern, as compared with the mRNA expression of itself. In 2014, a specific study on the expression pattern of lncRNAs in the liver tissue revealed that during the developmental processes, the expression pattern of lncRNAs may change accordingly, participating in various regulatory processes for liver maturation [40]. Similar to intestinal epithelial cells, hepatocytes also have distinctive lncRNA expression pattern. However, this pattern can attribute to two major inducements, namely, exogenous and endogenous. As for the endogenous factors, in such literature, during the liver development, lncRNA has to be expressed highly flexibly for the accurate expression regulation in time, resulting in the distinctive expression pattern [40]. Exogenous factors, on the contrary, affect the liver maturation. Recent publications also confirmed that as the major organs/cell subtypes for biological transformation, hepatocyte may also have different lncRNA expression pattern after the impaired transformative function induced by partial hepatectomy [41]. Therefore, similar to intestinal epithelial cells, hepatocyte may also have its specific lncRNA expression pattern due to the fickle exogenous stimulus. Similar mechanisms may also be applied to conjunctiva fibroblast, mesenchymal cells, and pericyte cells that mainly interact with the stimulus factors either internal or external [42–45].

Aside from such epithelial or mesenchymal cells that usually have specific lncRNA expression pattern compared with mRNA due to the exogenous or endogenous stimulus factors, five specific stimulus-responsive cell subtypes are available that can be clustered into another group due to their specific immune-associated functions. Different from the aforementioned cell subtypes that interact with stimulus factors, immune cells, including five cell subtypes we extracted, may actively identify and respond to the exogenous or endogenous stimulus usually as antigens. Neutrophil, as one of the major types of granulocytes in the white blood cell subgroups, has been widely reported to have specific reflective pattern against endogenous or exogenous stimulus [46–48]. During the regulatory processes, especially the modulation against exogenous stimulus like pathogens, lncRNA has also been reported to be differentially regulated, inducing a specific relative expression pattern of

lncRNA compared with mRNAs in this cell type [49–51]. Actually, lncRNAs have been reported to be quite functional and regulatory in multiple immune-associated cells, such as macrophage, dendritic cells, neutrophils, and lymphocytes (both B and T cells), confirming the prediction of B cells and macrophage [52,53]. As for mast cells, another functional component of immune-associated cells, recent publications on the lncRNA expression pattern of such cell subtype directly confirmed that the expression of lncRNAs in mast cells may modulate the inflammatory and immune recognition characteristics of such cells [54]. During the inflammatory processes induced by either tumorigenesis or exogenous infections, the unique expression pattern of certain functional lncRNAs compared with mRNAs may contribute to the accurate and controllable immune response against exogenous or endogenous stimulus [54].

Two unique non-terminally differentiated cells, namely, reticulocytes and neuronal stem cells, relating to two optimal features exist. Reticulocytes, as the immature status of erythrocytes, have been widely reported to have a specific lncRNA expression pattern, contributing to the maturation and enucleation processes [55,56]. Considering that during the maturation of reticulocytes, the mRNA composition of such cell types changed from multicomponent to single component and the hemoglobin transcription [57]. As for the lncRNAs, recent publications also confirmed that during the maturation processes, lncRNAs maintained in the erythrocytes can consistently regulate their respective biological functions, preventing the mature red blood cells from death [58]. Therefore, the expression pattern of lncRNAs and mRNAs of reticulocytes may both be different from other cell types. As for the remaining cell subtype, namely, neuron stem cells, these cells are quite a heterogeneous population, including quiescent neural stem cell, active neural stem cell, niche astrocytes, and neural progenitor cells [59,60]. Recent publications also confirmed that the mRNA and lncRNA expression pattern of neural stem cells may be quite heterogeneous. This finding implied the unique RNA expression pattern of such two cell subtypes.

3.2. Analysis of Classification Rules Yielded by RIPPER

In Section 2.3, we built an optimal RIPPER classifier with top 70 features in the mRMR feature list. Based on these features, RIPPER can produce 31 rules for classifying lncRNAs and mRNAs. These rules were too many for us to analyze them one by one. In view of this, we carefully checked the IFS curve of RIPPER in Figure 2 and found that when the top ten features were selected, the corresponding RIPPER classifier can yield an MCC of 0.876, which was quite close to the MCC of the optimal RIPPER classifier. Similar to RF, we also tested the statistical significance of these ten features. 1000 feature subsets with 10 features were randomly produced, each of which contained the feature of expression specificity and nine features randomly selected from rest 70 features. The 1000 RIPPER classifiers were built on these feature subsets, which were further evaluated via 10-fold cross-validation. Based on the obtained 1000 MCCs, a box plot was drawn in Figure 3. It can be observed that 0.876, obtained by the first ten features in the mRMR feature list, also lies at the top of the box plot. In addition, we also counted the mean and standard deviation of these 1000 MCCs, obtaining the mean of 0.847 and the standard deviation of 0.015. 0.876 was higher than $0.847 + 1.8 \times 0.015$, implying the MCC yielded by the first 10 features in the list was high with statistical significance. Thus, we selected these ten features to produce classification rules via RIPPER, obtaining 18 rules listed in Table 3. This section gave some analyses on these rules. According to recent publications, several rules can be confirmed. Here, we screened five rules for detailed analysis and the remainder of the rules were left to readers.

The first rule (Rule-1) involves the RNA expression pattern of four tissues and the expression specificity described by Chao-Shen-corrected Shannon's entropy, reflecting the randomized expression pattern of certain RNA subtypes. The higher such parameter that describing expression specificity is, the more exclusively such screened out RNA may be expressed and the higher expression specificity such RNA may have [61]. Lower expression of RNA in four tissues (neuronal stem cell, mesenchymal cell, intestinal epithelial cell and mesenchymal cell) with respective threshold has been predicted to indicate such expressed RNAs turns out to be lncRNAs, corresponding with the general lower lncRNA

expression pattern comparing to mRNAs [62]. As for the predicted exclusive expression pattern of lncRNAs, although various publications inferred that the expression specificity of lncRNAs are lower than that of mRNAs [63–65], an independent study [66] reported in 2016, implied that lncRNAs should have higher expression specificity than mRNAs and the identified lack of tissue specificity and exclusivity of lncRNAs can be attribute to the high detection limit of RNA sequencing/detection approaches and the low expression pattern of such cluster of RNAs. As for the second rule (Rule-2), similarly with the first rule, the expression level of our target RNA, expression level lower than respective threshold in neuronal stem cell (2.58), hepatocyte (0) and mast cell (0.151) simultaneously turns out to be lncRNA. According to recent publications, in 2013, a systematic research on the expression level of lncRNAs in multiple neural stem cells confirmed that the average general expression of lncRNAs in such cell subtypes are lower than our parameter, thereby validating this rule [67].

Table 3. 18 Classification rules yielded by repeated incremental pruning to produce error reduction (RIPPER) on top ten features.

Rule Number	Condition	Outcome
Rule-1	(Neuronal stem cell $\leq$ 2.58) and (Mesenchymal cell $\leq$ 1.44) and (Intestinal epithelial cell $\leq$ 0) and (Mesenchymal cell $\leq$ 0.197) and (Expression specificity $\geq$ 0.552495)	lncRNA
Rule-2	(Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 0) and (Mast cell $\leq$ 0.151)	lncRNA
Rule-3	(Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 0) and (Intestinal epithelial cell $\leq$ 0.246) and (Mesenchymal cell $\leq$ 2.04) and (Neuronal stem cell $\leq$ 0)	lncRNA
Rule-4	(Neuronal stem cell $\leq$ 5.17) and (Mast cell $\leq$ 0.842) and (Intestinal epithelial cell $\leq$ 0.737) and (Neuronal stem cell $\leq$ 0.542) and (Mesenchymal cell $\leq$ 2.75)	lncRNA
Rule-5	(Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 1.73) and (Intestinal epithelial cell $\leq$ 0.246) and (Mesenchymal cell $\leq$ 2.63) and (Neuronal stem cell $\leq$ 0)	lncRNA
Rule-6	(Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 1.73) and (Mast cell $\leq$ 0.352) and (Expression specificity $\leq$ 0.299388) and (Mast cell $\leq$ 0)	lncRNA
Rule-7	(Neuronal stem cell $\leq$ 5.17) and (Hepatocyte $\leq$ 1.73) and (Intestinal epithelial cell $\leq$ 0.246) and (Expression specificity $\geq$ 0.497214) and (Lymphocyte of b lineage $\geq$ 1.23)	lncRNA
Rule-8	(Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 1.73) and (Mesenchymal cell $\leq$ 1.44) and (Intestinal epithelial cell $\leq$ 0)	lncRNA
Rule-9	(Neuronal stem cell $\leq$ 5.17) and (Mesenchymal cell $\leq$ 4.52) and (Reticulocyte $\leq$ 0) and (Fibroblast of the conjuctiva $\leq$ 0) and (Hepatocyte $\leq$ 3.46) and (Expression specificity $\geq$ 0.311106) and (Intestinal epithelial cell $\leq$ 4.18) and (Neuronal stem cell $\leq$ 2.71)	lncRNA
Rule-10	(Neuronal stem cell $\leq$ 5.17) and (Mast cell $\leq$ 0.842) and (Intestinal epithelial cell $\leq$ 1.72) and (Hepatocyte $\leq$ 0) and (Fibroblast of the conjuctiva $\leq$ 0)	lncRNA
Rule-11	(Neuronal stem cell $\leq$ 5.17) and (Hepatocyte $\leq$ 5.19) and (Intestinal epithelial cell $\leq$ 1.23) and (Neuronal stem cell $\leq$ 0.542) and (Mesenchymal cell $\leq$ 10.1) and (Mast cell $\leq$ 0.907)	lncRNA
Rule-12	(Neuronal stem cell $\leq$ 2.58) and (Mesenchymal cell $\leq$ 4.5) and (Intestinal epithelial cell $\leq$ 0.983) and (Hepatocyte $\leq$ 0) and (Fibroblast of the conjuctiva $\leq$ 0)	lncRNA
Rule-13	(Neuronal stem cell $\leq$ 5.17) and (Mesenchymal cell $\leq$ 5.18) and (Mesenchymal cell $\leq$ 1.45) and (Neutrophil $\geq$ 0.848) and (Expression specificity $\leq$ 0.333673) and (Mesenchymal cell $\leq$ 0.904) and (Fibroblast of the conjuctiva $\leq$ 0)	lncRNA
Rule-14	(Neuronal stem cell $\leq$ 5.17) and (Mesenchymal cell $\leq$ 5.18) and (Neuronal stem cell $\leq$ 2.58) and (Intestinal epithelial cell $\leq$ 0.737) and (Neuronal stem cell $\leq$ 0) and (Expression specificity $\leq$ 0.517863) and (Mesenchymal cell $\geq$ 2.87)	lncRNA
Rule-15	(Neuronal stem cell $\leq$ 5.17) and (Mast cell $\leq$ 1.06) and (Hepatocyte $\leq$ 5.19) and (Neuronal stem cell $\leq$ 0.542) and (Intestinal epithelial cell $\leq$ 6.88) and (Mesenchymal cell $\leq$ 28.5)	lncRNA
Rule-16	(Neuronal stem cell $\leq$ 5.17) and (Mesenchymal cell $\leq$ 5.28) and (Fibroblast of the conjuctiva $\leq$ 0) and (Mast cell $\leq$ 0.842) and (Intestinal epithelial cell $\leq$ 11.8) and (Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 20.1)	lncRNA
Rule-17	(Neuronal stem cell $\leq$ 7.75) and (Intestinal epithelial cell $\leq$ 1.72) and (Mesenchymal cell $\leq$ 4.52) and (Neuronal stem cell $\leq$ 2.58) and (Lymphocyte of b lineage $\leq$ 4.07) and (Neutrophil $\geq$ 5.72) and (Mast cell $\leq$ 19) and (Expression specificity $\geq$ 0.115115)	lncRNA
Rule-18	Otherwise	mRNA

The third and fourth, fifth rules (Rule-3, Rule-4 and Rule-5) all involve neuronal stem cell, hepatocyte, intestinal epithelial cell, mesenchymal cell. According to such rules, quite lower expression level of our target RNAs in such cells may indicate such RNAs turn out to be lncRNAs, corresponding with general expression tendency. Take a specific parameter of intestinal epithelial cell expression pattern of rule three as an example. Two independent studies [68,69] on the expression pattern of lncRNAs in intestinal epithelial cells confirmed that generally, the expression level of such RNAs are lower than mRNAs and RNAs expression level lower than 0.5 tend to be lncRNAs, corresponding with these rules. Further, RNAs with expression level lower than 2 in mesenchymal cells have also been indirectly validated by a specific study on mRNA expression analysis of mesenchymal cells [70]. Therefore, as we have discussed above, the first five rules for distinguishing lncRNAs and mRNAs can be validated by recent publications, reflecting the reasonability of screened rules and indicating that lncRNAs and mRNAs definitely have distinctive expression level inclination.

4. Materials and Methods

4.1. Dataset and Feature Representation

To construct a reliable dataset, we selected 9339 lncRNAs and 14,294 mRNAs from Hon et al.'s study [24], wherein all RNAs were validated and evaluated by experimental bioinformatics methods. Each RNA sample in the dataset was encoded by 71 expression features, including 69 maximum expression features for 69 types of cells, one feature for the maximum expression in all cells, and one expression specificity feature [24]. Based on Hon et al. [24], the expression values of Counts Per Million (CPM) were from FANTOM5 cap analysis of gene expression (CAGE) data. The expression levels were normalized across samples with edgeR. The 69 maximum expression features for 69 types of cells and the one feature for the maximum expression in all cells were the corresponding CPM expression levels. The expression specificity feature was measured as Chao-Shen-corrected Shannon's entropy of the CPM expression levels. Based on the encoded RNA samples, we can explore the different expression patterns of lncRNAs and mRNAs in various cell types. To perform this method, we built a binary classification problem, treating lncRNAs as positive samples and mRNAs as negative samples.

4.2. Feature Selection Method

Each lncRNA or mRNA in our dataset was encoded by 71 expression features as described in Section 4.1. Clearly, not all features contributed equally to discriminating these two types of RNAs, i.e., some of them may play more important roles than others. Thus, we performed a two-stage feature selection method to investigate the association between features and RNA types.

4.2.1. mRMR Method

In the first stage, all features were ranked by a powerful feature selection method, i.e., the mRMR method [25]. This method has been widely used to rank features in several studies [71–80]. In this method, two criteria, namely, (1) Max-Relevance and (2) Min-Redundancy, were adopted to rank features singly or simultaneously. If only the first one was used, features are sorted according to their relevance to target variable, resulting in a feature list, named MaxRel feature list. Meanwhile, if both of two criteria were used, i.e., each feature is evaluated according to the relevance between it and target variable and also the redundancies to other features, another feature list, called mRMR feature list, can be produced. Therefore, users can obtain two output feature lists: (1) MaxRel feature list and (2) mRMR feature list, via the mRMR method. In this study, we not only extracted important features that can mark the differences between lncRNAs and mRNAs but also built a good classifier for discriminating these two types of RNAs. Thus, we only used the second list, i.e., the mRMR feature list. In this list, features with high ranks must have high relevance to target variable and low redundancies to features that are listed before it.

After the mRMR method had been applied to analyze the data of lncRNAs and mRNAs, we obtained the mRMR feature list, denoted as the following:

$$F = [F_1, F_2, \ldots, F_N] \quad (1)$$

where N is the total number of features ($N = 71$ in this study). A detailed description on the mRMR method and how to construct the mRMR feature list can be found in Peng et al.'s study [25]. In this study, we used the mRMR program downloaded at http://home.penglab.com/proj/mRMR/index.htm and its default parameters were used.

4.2.2. IFS Method

In the second stage, we further used the mRMR feature list derived from the mRMR method. However, the mRMR feature list only sorts features according to their importance, which features are more important than others is still a problem. Here, the IFS method [26] was adopted to determine which features were important. This method first produced several feature subsets, denoted as:

$$FS_i = \{F_1, F_2, \ldots, F_i\} \ (1 \leq i \leq \mathrm{N}), \quad (2)$$

Notably, FS_i consists the first i features in F. Then, for each FS_i, all RNAs mentioned in Section 4.1 were encoded by features in FS_i, and a classification algorithm was performed on this dataset, evaluated by a 10-fold cross-validation [81–85]. After all feature subsets had been tested, the feature subset that can yield the best performance was selected. Optimal features were defined as features in this set, and the classifier with these optimal features was termed as optimal classifier.

4.3. Classification Algorithm

In this study, we employed two classification algorithms: RF [27] and RIPPER [28], where RF was adopted to build a classifier with good performance, while RIPPER was used to extract informative rules that can clearly indicate the difference between lncRNAs and mRNAs. Their brief descriptions were as follows.

RF is an integrated classifier. RF always comprises several decision trees [27]. In the training stage, two statistical techniques, namely, bootstrap method [86], and random selection of features [87] were adopted to create decision trees for a training dataset with N samples. In the bootstrap procedure, N samples were randomly selected from the training dataset, with replacement, to constitute a new dataset. This way, B (B is a parameter indicating the number of decision trees) datasets can be constructed. For each constructed dataset with N samples, a decision tree can be built. Random selection of features is adopted in the construction procedure of decision trees. We randomly selected some features, which are much lesser than total features, and determined the optimal splitting way to extend a tree at one node. Finally, B decision trees can be built. For a query sample, each decision tree provides a predicted result. RF integrates these results by using majority voting. RF is deemed to be an effective classification algorithm and has been adopted to tackle different biological problems to date [85,88–92].

RIPPER [28], proposed by Cohen in 1995, is a classic rule learning algorithm based on rough set theory. It is an optimized version of Incremental Reduced Error Pruning (IREP) [93]. Compared with C4.5, RIPPER is much more efficient on large datasets and provides competitive results. Figure 4 illustrates the detailed procedures of rule learning using RIPPER. The output rules of RIPPER always contain the conditions, listed in the left hand, and the outcome, listed in the right hand. For example, (Neuronal stem cell $\leq$ 2.58) and (Hepatocyte $\leq$ 0) and (Mast cell $\leq$ 0.151) $\geq$ class = lncRNA.

Initialize a set E to be the training set
Choose a class C that contains least instances
    Initialize a rule R to have an empty left-hand side that predicts C
    Split E into growing and pruning sets
    While there are positive samples (instances of C) in the growing set, or the description length (DL) is 64 bits greater than the smallest DL found so far, or the error rate is greater than 50%
        Until R is perfect (or no more attributes to add)
            For each attribute a not included in R, and for each value v,
                Consider $a = v$ to add to the left-hand side of R
                Choose the a and v that have the highest Foil's information gain
                Add $a = v$ to R
                Prune R using reduced error pruning
        Remove the instances covered by R from the growing set
Global optimization strategy is applied to further prune the rule.

Figure 4. The rule learning procedures of repeated incremental pruning to produce error reduction (RIPPER) algorithm [85].

Weka [94] (version 3.8.0) is a powerful suit of software, collecting several state-of-the-art machine learning algorithms and tools, which can be downloaded at https://www.cs.waikato.ac.nz/ml/weka/. One classifier in Weka, called "RandomForest", implements the abovementioned RF algorithm. And the other classifier "JRip" implements the RIPPER algorithm. Here, these two tools were directly employed in this study and were used with their default setting.

4.4. Performance Measurements

Several classifiers with different classification algorithms were constructed on features in different feature subsets. The 10-fold cross-validation [95] was adopted to evaluate the prediction performance of each classifier. As a binary classification problem, four values can be counted according to the predicted results: (1) True positive (*TP*); (2) True negative (*TN*); (3) False positive (*FP*); and (4) False negative (*FN*).

Based on the abovementioned values, four measurements, named *SN*, *SP*, *ACC* [96,97], and *MCC* [98], can be computed to evaluate the prediction ability of each classifier. Their definitions are as follows:

$$SN = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TN + FP} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Among the four measurements, *MCC* is a comprehensive measurement ranging from -1 to 1. *MCC* can more accurately evaluate the performance of each classifier than *ACC* because *MCC* is a balanced measure even if the sizes of classes are very different. Thus, *MCC* was selected as a major one in this study, whereas other measurements were provided as references.

5. Conclusions

Based on the expression data of lncRNA and mRNA retrieved from a recent publication, we proposed a computational scheme to identify a group of functional features and rules, corresponding to cell subtypes, which have distinctive expression patterns of such two RNA subtypes. According to recent publications, several top features can be confirmed to have such function and obtained rules can really indicate the different expression patterns on two RNA subtypes. The new findings of this study may not only reveal the heterogeneous expression patterns of lncRNA and mRNA but also provide us a new tool to identify the potential biological functions of different RNA types. In addition, we provided the whole dataset investigated in this study in Supplementary Data 1, with which readers can easily replicate our work.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/19/11/3416/s1.

Author Contributions: Conceptualization, T.H. and Y.-D.C.; methodology, L.C., X.P., M.L. and Y.-D.C.; validation, T.H.; formal analysis, Y.-H.Z. and S.W.; writing—original draft preparation, L.C. and Y.-H.Z.; writing—review and editing, T.H. supervision, Y.-D.C.; funding acquisition, L.C., T.H. and Y.-D.C.

Funding: This research was funded by the National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program, the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maurel, M.C. Rna in evolution—A review. *J. Evol. Biol.* **1992**, *5*, 173–188. [CrossRef]
2. De Oliveira, D.E. DNA viruses in human cancer: An integrated overview on fundamental mechanisms of viral carcinogenesis. *Cancer Lett.* **2007**, *247*, 182–196. [CrossRef] [PubMed]
3. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **1993**, *362*, 709–715. [CrossRef] [PubMed]
4. Dorman, C.J.; Colgan, A.; Dorman, M.J. Bacterial pathogen gene regulation: A DNA-structure-centred view of a protein-dominated domain. *Clin. Sci.* **2016**, *130*, 1165–1177. [CrossRef] [PubMed]
5. Ibragimova, M.; Akberova, N.I.; Tarasov, D.S.; Izotova, E.D.; Alimova, F.K.; Zhdanov, R.I. Fatty acid regulation of gene expression: Bioinformatics view to structure and dynamics of DNA-fatty acid complexation. *FEBS J.* **2013**, *280*, 550.
6. Gagna, C.E.; Chan, N.J.; Spinelli, D.; Lambert, W.C. Regulation of gene expression by novel antisense technology, based on structures of DNA and RNA: Structural transitional genomics (and proteomics). *Biophys. J.* **2005**, *88*, 571a.
7. Kapranov, P.; Cheng, J.; Dike, S.; Nix, D.A.; Duttagupta, R.; Willingham, A.T.; Stadler, P.F.; Hertel, J.; Hackermuller, J.; Hofacker, I.L.; et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **2007**, *316*, 1484–1488. [CrossRef] [PubMed]
8. Sigova, A.A.; Mullen, A.C.; Molinie, B.; Gupta, S.; Orlando, D.A.; Guenther, M.G.; Almada, A.E.; Lin, C.; Sharp, P.A.; Giallourakis, C.C.; et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 2876–2881. [CrossRef] [PubMed]
9. Bernard, D.; Prasanth, K.V.; Tripathi, V.; Colasse, S.; Nakamura, T.; Xuan, Z.Y.; Zhang, M.Q.; Sedel, F.; Jourdren, L.; Coulpier, F.; et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J.* **2010**, *29*, 3082–3093. [CrossRef] [PubMed]
10. Tomaru, Y.; Hayashizaki, Y. Cancer research with non-coding RNA. *Cancer Sci.* **2006**, *97*, 1285–1290. [CrossRef] [PubMed]
11. Mattick, J.S.; Makunin, I.V. Non-coding RNA. *Hum. Mol. Genet.* **2006**, *15*, R17–R29. [CrossRef] [PubMed]
12. Presutti, C.; Rosati, J.; Vincenti, S.; Nasi, S. Non coding RNA and brain. *BMC Neurosci.* **2006**, *7*, S5. [CrossRef] [PubMed]
13. O'Gorman, W.; Kwek, K.Y.; Thomas, B.; Akoulitchev, A. Non-coding RNA in transcription initiation. *Biochem. Soc. Symp.* **2006**, *73*, 131–140. [CrossRef]

14. Vendramin, R.; Marine, J.C.; Leucci, E. Non-coding RNAs: The dark side of nuclear-mitochondrial communication. *EMBO J.* **2017**, *36*, 1123–1133. [CrossRef] [PubMed]
15. Boland, C.R. Non-coding RNA: It's not junk. *Dig. Dis. Sci.* **2017**, *62*, 1107–1109. [CrossRef] [PubMed]
16. Gupta, S.C.; Tripathi, Y.N. Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets. *Int. J. Cancer* **2017**, *140*, 1955–1967. [CrossRef] [PubMed]
17. Mathiyalagan, P.; Okabe, J.; Kaipananickal, H.; Du, X.J.; El-Oata, A. Epigenetic regulation of gene expression by long intergenic non-coding RNA encoded by cardiac myosin heavy chain genes. *J. Card. Fail.* **2014**, *20*, S81–S82. [CrossRef]
18. Maass, P.G. Long non-coding RNA (lncRNA). Gene and genome regulation. *Med. Genet.* **2014**, *26*, 5–10.
19. Rozovski, U.; Calin, G.; Tetsirp, S.; D'Abundo, L.; Harris, D.; Li, P.; Liu, Z.M.; Grgurevic, S.; Ferrajoli, A.; Faderl, S.; et al. Signal transducer and activator of transcription (STAT)-3-dependent regulation of non-coding RNA genes in chronic lymphocytic leukemia (CLL). *Blood* **2012**, *120*, 1886.
20. Pan, Y.F.; Feng, L.; Zhang, X.Q.; Song, L.J.; Liang, H.X.; Li, Z.Q.; Tao, F.B. Role of long non-coding RNAs in gene regulation and oncogenesis. *Chin. Med. J.* **2011**, *124*, 2378–2383. [PubMed]
21. Jain, A.K.; Xi, Y.X.; McCarthy, R.; Allton, K.; Akdemir, K.C.; Patel, L.R.; Aronow, B.; Lin, C.R.; Li, W.; Yang, L.Q.; et al. Lncpress1 is a p53-regulated lncrna that safeguards pluripotency by disrupting SIRT6-mediated de-acetylation of histone H3K56. *Mol. Cell* **2016**, *64*, 967–981. [CrossRef] [PubMed]
22. Hunten, S.; Kaller, M.; Drepper, F.; Oeljeklaus, S.; Bonfert, T.; Erhard, F.; Dueck, A.; Eichner, N.; Friedel, C.C.; Meister, G.; et al. P53-regulated networks of protein, mRNA, miRNA, and lncRNA expression revealed by integrated pulsed stable isotope labeling with amino acids in cell culture (pSILAC) and next generation sequencing (NGS) analyses. *Mol. Cell. Proteom.* **2015**, *14*, 2609–2629. [CrossRef] [PubMed]
23. Rassoolzadeh, H.; Bohm, S.; Hedstrom, E.; Gad, H.; Helleday, T.; Henriksson, S.; Farnebo, M. Overexpression of the scaffold WD40 protein wrap53β enhances the repair of and cell survival from DNA double-strand breaks. *Cell Death Dis.* **2016**, *7*, e2267. [CrossRef] [PubMed]
24. Hon, C.C.; Ramilowski, J.A.; Harshbarger, J.; Bertin, N.; Rackham, O.J.L.; Gough, J.; Denisenko, E.; Schmeier, S.; Poulsen, T.M.; Severin, J.; et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **2017**, *543*, 199–204. [CrossRef] [PubMed]
25. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
26. Huang, T.; Cui, W.R.; Hu, L.L.; Feng, K.Y.; Li, Y.X.; Cai, Y.D. Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. *PLoS ONE* **2009**, *4*, e8126. [CrossRef] [PubMed]
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Cohen, W.W. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 115–123.
29. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
30. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
31. Le Cessie, S.; Van Houwelingen, J.C. Ridge estimators in logistic regression. *Appl. Stat.* **1992**, *41*, 191–201. [CrossRef]
32. Xu, J.; Zhang, F.; Gao, C.; Ma, X.F.; Peng, X.L.; Kong, D.X.; Hao, J.W. Microarray analysis of lncRNA and mRNA expression profiles in patients with neuromyelitis optica. *Mol. Neurobiol.* **2017**, *54*, 2201–2208. [CrossRef] [PubMed]
33. Zhang, Z.L.; Zhao, L.J.; Chai, L.; Zhou, S.H.; Wang, F.; Wei, Y.; Xu, Y.P.; Zhao, P. Seven lncRNA-mRNA based risk score predicts the survival of head and neck squamous cell carcinoma. *Sci. Rep.* **2017**, *7*, 309. [CrossRef] [PubMed]
34. Zhang, T.; Zhang, X.Q.; Han, K.P.; Zhang, G.X.; Wang, J.Y.; Xie, K.Z.; Xue, Q. Genome-wide analysis of lncRNA and mRNA expression during differentiation of abdominal preadipocytes in the chicken. *G3 Genes Genomes Genet.* **2017**, *7*, 953–966. [CrossRef] [PubMed]
35. Mirza, A.H.; Berthelsen, C.H.B.; Seemann, S.E.; Pan, X.Y.; Frederiksen, K.S.; Vilien, M.; Gorodkin, J.; Pociot, F. Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome Med.* **2015**, *7*, 39. [CrossRef] [PubMed]

36. Liang, L.X.; Ai, L.Y.; Qian, J.; Fang, J.Y.; Xu, J. Long noncoding RNA expression profiles in gut tissues constitute molecular signatures that reflect the types of microbes. *Sci. Rep.* **2015**, *5*, 11763. [CrossRef] [PubMed]
37. Dempsey, J.L.; Cui, J.Y. Long non-coding RNAs: A novel paradigm for toxicology. *Toxicol. Sci.* **2017**, *155*, 3–21. [CrossRef] [PubMed]
38. Granneman, S.; Baserga, S.J. Crosstalk in gene expression: Coupling and co-regulation of rDNA transcription, pre-ribosome assembly and pre-rRNA processing. *Curr. Opin. Cell Biol.* **2005**, *17*, 281–286. [CrossRef] [PubMed]
39. Parra, M.; Gee, S.; Ramez, M.; Gascard, P.; Mohandas, N.; Conboy, J. Regulation of protein 4.1r gene expression: Complex 5′ exon structure and apparent coupling between transcription and alternative pre-mRNA splicing events. *Blood* **2001**, *98*, 8A–9A.
40. Peng, L.; Paulson, A.; Li, H.; Piekos, S.; He, X.; Li, L.H.; Zhong, X.B. Developmental programming of long non-coding RNAs during postnatal liver maturation in mice. *PLoS ONE* **2014**, *9*, e114917. [CrossRef] [PubMed]
41. Huang, L.L.; Damle, S.S.; Booten, S.; Singh, P.; Sabripour, M.; Hsu, J.; Jo, M.J.; Katz, M.; Watt, A.; Hart, C.E.; et al. Partial hepatectomy induced long noncoding RNA inhibits hepatocyte proliferation during liver regeneration. *PLoS ONE* **2015**, *10*, e0132798. [CrossRef] [PubMed]
42. Buchanan, R.A.; Wagner, R.C. Morphometric changes in pericyte-capillary endothelial-cell associations correlated with vasoactive stimulus. *Microcirc. Endothel. Lymphat.* **1990**, *6*, 159–181.
43. Miao, Y.Y.; Liu, J.; Zhu, J.; Tao, Y.L.; Zhang, J.A.; Luo, D.; Zhou, B.R. The effect of botulinum toxin type a on expression profiling of long noncoding RNAs in human dermal fibroblasts. *BioMed Res. Int.* **2017**, *2017*, 2957941. [CrossRef] [PubMed]
44. Li, X.Y.; Zhang, L.; Liang, J.J. Unraveling the expression profiles of long noncoding RNAs in rat cardiac hypertrophy and functions of lncRNA bc088254 in cardiac hypertrophy induced by transverse aortic constriction. *Cardiology* **2016**, *134*, 84–98. [CrossRef] [PubMed]
45. Jiang, X.Y.; Ning, Q.L. Expression profiling of long noncoding RNAs and the dynamic changes of lncRNA-NR024118 and CDKN1C in angiotensin II-treated cardiac fibroblasts. *Int. J. Clin. Exp. Pathol.* **2014**, *7*, 1325–1336. [PubMed]
46. Honda, T.; Uehara, T.; Matsumoto, G.; Arai, S.; Sugano, M. Neutrophil left shift and white blood cell count as markers of bacterial infection. *Clin. Chim. Acta* **2016**, *457*, 46–53. [CrossRef] [PubMed]
47. Yilmaz, M.; Tenekecioglu, E.; Arslan, B.; Bekler, A.; Ozluk, O.A.; Karaagac, K.; Agca, F.V.; Peker, T.; Akgumus, A. White blood cell subtypes and neutrophil-lymphocyte ratio in prediction of coronary thrombus formation in non-ST-segment elevated acute coronary syndrome. *Clin. Appl. Thromb. Hemost.* **2015**, *21*, 446–452. [CrossRef] [PubMed]
48. Sakai, A.; Ohira, T.; Hosoya, M.; Ohtsuru, A.; Satoh, H.; Kawasaki, Y.; Suzuki, H.; Takahashi, A.; Kobashi, G.; Ozasa, K.; et al. White blood cell, neutrophil, and lymphocyte counts in individuals in the evacuation zone designated by the government after the fukushima daiichi nuclear power plant accident: The fukushima health management survey. *J. Epidemiol.* **2015**, *25*, 80–87. [CrossRef] [PubMed]
49. Guo, Q.Y.; Cheng, Y.; Liang, T.; He, Y.A.; Ren, C.C.; Sun, L.Y.; Zhang, G.M. Comprehensive analysis of lncRNA-mRNA co-expression patterns identifies immune-associated lncRNA biomarkers in ovarian cancer malignant progression. *Sci. Rep.* **2015**, *5*, 17683. [CrossRef] [PubMed]
50. Li, Z.H.; Rana, T.M. Decoding the noncoding: Prospective of lncRNA-mediated innate immune regulation. *RNA Biol.* **2014**, *11*, 979–985. [CrossRef] [PubMed]
51. Wei, N.; Pang, W.J.; Wang, Y.; Xiong, Y.; Xu, R.X.; Wu, W.J.; Zhao, C.Z.; Yang, G.S. Knockdown of PU. 1 mRNA and AS lncRNA regulates expression of immune-related genes in zebrafish danio rerio. *Dev. Comp. Immunol.* **2014**, *44*, 315–319. [CrossRef] [PubMed]
52. Kumar, U.; Shahid, M.; Tripathi, R.; Mohanty, S.; Kumar, A.; Bhattacharyya, P.; Lal, B.; Gautam, P.; Raja, R.; Panda, B.B.; et al. Variation of functional diversity of soil microbial community in sub-humid tropical rice-rice cropping system under long-term organic and inorganic fertilization. *Ecol. Indic.* **2017**, *73*, 536–543. [CrossRef]
53. Hu, J.L.; Lin, X.G.; Wang, J.H.; Dai, J.; Chen, R.R.; Zhang, J.B.; Wong, M.H. Microbial functional diversity, metabolic quotient, and invertase activity of a sandy loam soil as affected by long-term application of organic amendment and mineral fertilizer. *J. Soils Sediments* **2011**, *11*, 271–280. [CrossRef]

54. Li, L.; Dang, Q.; Xie, H.J.; Yang, Z.; He, D.L.; Liang, L.; Song, W.B.; Yeh, S.Y.; Chang, C.S. Correction: Infiltrating mast cells enhance prostate cancer invasion via altering lncRNA-HOTAIR/PRC2-androgen receptor (AR)-MMP9 signals and increased stem/progenitor cell population. *Oncotarget* **2016**, *7*, 83828. [CrossRef] [PubMed]
55. Alvarez-Dominguez, J.R.; Hu, W.Q.; Yuan, B.B.; Shi, J.H.; Park, S.S.; Gromatzky, A.A.; van Oudenaarden, A.; Lodish, H.F. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **2014**, *123*, 570–581. [CrossRef] [PubMed]
56. Li, S.L. The functional role of long non-coding RNAs and epigenetics. *Biol. Proced. Online* **2014**. [CrossRef]
57. Kung, J.T.Y.; Colognori, D.; Lee, J.T. Long noncoding RNAs: Past, present, and future. *Genetics* **2013**, *193*, 651–669. [CrossRef] [PubMed]
58. Kulczynska, K.; Siatecka, M. A regulatory function of long non-coding RNAs in red blood cell development. *Acta Biochim. Pol.* **2016**, *63*, 675–680. [CrossRef] [PubMed]
59. Marvin, M.; McKay, R. Multipotential stem cells in the vertebrate cns. *Semin. Cell Biol.* **1992**, *3*, 401–411. [CrossRef]
60. Gonzalez-Perez, O.; Quinones-Hinojosa, A. Astrocytes as neural stem cells in the adult brain. *J. Stem Cells* **2012**, *7*, 181–188. [PubMed]
61. Andersson, R.; Gebhard, C.; Miguel-Escalada, I.; Hoof, I.; Bornholdt, J.; Boyd, M.; Chen, Y.; Zhao, X.; Schmidl, C.; Suzuki, T.; et al. An atlas of active enhancers across human cell types and tissues. *Nature* **2014**, *507*, 455–461. [CrossRef] [PubMed]
62. Ulitsky, I. Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **2016**, *17*, 601–614. [CrossRef] [PubMed]
63. Kern, C.; Wang, Y.; Chitwood, J.; Korf, I.; Delany, M.; Cheng, H.; Medrano, J.F.; Van Eenennaam, A.L.; Ernst, C.; Ross, P.; et al. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genom.* **2018**, *19*, 684. [CrossRef] [PubMed]
64. An, N.; Fan, S.; Wang, Y.; Zhang, L.; Gao, C.; Zhang, D.; Han, M. Genome-wide identification, characterization and expression analysis of long non-coding RNAs in different tissues of apple. *Gene* **2018**, *666*, 44–57. [CrossRef] [PubMed]
65. Washietl, S.; Kellis, M.; Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* **2014**, *24*, 616–628. [CrossRef] [PubMed]
66. Gloss, B.S.; Dinger, M.E. The specificity of long noncoding RNA expression. *Biochim. Biophys. Acta* **2016**, *1859*, 16–22. [CrossRef] [PubMed]
67. Ramos, A.D.; Diaz, A.; Nellore, A.; Delgado, R.N.; Park, K.Y.; Gonzales-Roybal, G.; Oldham, M.C.; Song, J.S.; Lim, D.A. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell* **2013**, *12*, 616–628. [CrossRef] [PubMed]
68. Wang, J.Y.; Xiao, L.; Wang, J.Y. Posttranscriptional regulation of intestinal epithelial integrity by noncoding RNAs. *Wiley Interdiscip. Rev.* **2017**, *8*, e1399. [CrossRef] [PubMed]
69. Chen, J.; Wan, J.; Ye, J.; Xia, L.; Lu, N. Emerging role of lncRNAs in the normal and diseased intestinal barrier. *Inflamm. Res.* **2018**, *67*, 757–764. [CrossRef] [PubMed]
70. Ryan, J.M.; Pettit, A.R.; Guillot, P.V.; Chan, J.K.; Fisk, N.M. Unravelling the pluripotency paradox in fetal and placental mesenchymal stem cells: Oct-4 expression and the case of the emperor's new clothes. *Stem Cell Rev. Rep.* **2013**, *9*, 408–421. [CrossRef] [PubMed]
71. Niu, S.; Hu, L.L.; Zheng, L.L.; Huang, T.; Feng, K.Y.; Cai, Y.D.; Li, H.P.; Li, Y.X.; Chou, K.C. Predicting protein oxidation sites with feature selection and analysis approach. *J. Biomol. Struct. Dyn.* **2012**, *29*, 650–658. [CrossRef] [PubMed]
72. Cai, Y.; He, J.; Lu, L. Predicting sumoylation site by feature selection method. *J. Biomol. Struct. Dyn.* **2011**, *28*, 797–804. [CrossRef] [PubMed]
73. Cai, Y.; Huang, T.; Hu, L.; Shi, X.; Xie, L.; Li, Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* **2012**, *42*, 1387–1395. [CrossRef] [PubMed]
74. Chen, L.; Chu, C.; Feng, K. Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization. *Comb. Chem. High Throughput Screen.* **2016**, *19*, 136–143. [CrossRef] [PubMed]

75. Chen, L.; Pan, X.; Hu, X.; Zhang, Y.-H.; Wang, S.; Huang, T.; Cai, Y.-D. Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* **2018**, *143*, 1731–1740. [CrossRef] [PubMed]
76. Zhao, X.; Chen, L.; Lu, J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* **2018**. [CrossRef] [PubMed]
77. He, L.; Cao, Z.; Wang, Y.; Du, W.; Liang, Y. An ensemble feature selection method based on mRMR for paired microarray data. *J. Comput. Inf. Syst.* **2014**, *10*, 4875–4882.
78. Zhang, Y.; Ding, C.; Li, T. Gene selection algorithm by combining relieff and mRMR. *BMC Genom.* **2008**, *9*, S27. [CrossRef] [PubMed]
79. Chen, L.; Wang, S.; Zhang, Y.-H.; Wei, L.; Xu, X.; Huang, T.; Cai, Y.-D. Prediction of nitrated tyrosine residues in protein sequences by extreme learning machine and feature selection methods. *Comb. Chem. High Throughput Screen.* **2018**, *21*, 393–402. [CrossRef] [PubMed]
80. Li, J.; Lu, L.; Zhang, Y.; Liu, M.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* **2018**. [CrossRef] [PubMed]
81. Chou, K.C.; Shen, H.B. Cell-ploc: A package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153–162. [CrossRef] [PubMed]
82. Chen, L.; Zhang, Y.-H.; Huang, G.; Pan, X.; Wang, S.; Huang, T.; Cai, Y.-D. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genom.* **2018**, *293*, 137–149. [CrossRef] [PubMed]
83. Chen, L.; Wang, S.; Zhang, Y.-H.; Li, J.; Xing, Z.-H.; Yang, J.; Huang, T.; Cai, Y.-D. Identify key sequence features to improve crispr sgRNA efficacy. *IEEE Access* **2017**, *5*, 26582–26590. [CrossRef]
84. Pan, X.; Hu, X.; Zhang, Y.-H.; Feng, K.; Wang, S.P.; Chen, L.; Huang, T.; Cai, Y.-D. Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* **2018**, *9*, 208. [CrossRef] [PubMed]
85. Wang, D.; Li, J.-R.; Zhang, Y.-H.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* **2018**, *9*, 155. [CrossRef] [PubMed]
86. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
87. Ting, K.M.; Witten, I.H. Stacking bagged and dagged models. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 1997.
88. Chen, L.; Zhang, Y.-H.; Zheng, M.; Huang, T.; Cai, Y.-D. Identification of compound–protein interactions through the analysis of gene ontology, kegg enrichment for proteins and molecular fragments of compounds. *Mol. Genet. Genom.* **2016**, *291*, 2065–2079. [CrossRef] [PubMed]
89. Casanova, R.; Saldana, S.; Chew, E.Y.; Danis, R.P.; Greven, C.M.; Ambrosius, W.T. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS ONE* **2014**, *9*, e98587. [CrossRef] [PubMed]
90. Pan, X.Y.; Zhu, L.; Fan, Y.X.; Yan, J.C. Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection. *Comput. Biol. Chem.* **2014**, *53*, 324–330. [CrossRef] [PubMed]
91. Liu, L.; Chen, L.; Zhang, Y.H.; Wei, L.; Cheng, S.; Kong, X.; Zheng, M.; Huang, T.; Cai, Y.D. Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection. *J. Biomol. Struct. Dyn.* **2017**, *35*, 312–329. [CrossRef] [PubMed]
92. Chen, L.; Chu, C.; Huang, T.; Kong, X.; Cai, Y.-D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino Acids* **2015**, *47*, 1485–1493. [CrossRef] [PubMed]
93. Johannes, F.; Widmer, G. Incremental reduced error pruning. In Proceedings of the Eleventh Annual Conference on Machine Learning, New Brunswick, NJ, USA, 10–13 July 1994.
94. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: San Francisco, CA, USA, 2005.
95. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Lawrence Erlbaum Associates Ltd.: Mahwah, NJ, USA, 1995; pp. 1137–1145.

96. Chen, L.; Chu, C.; Zhang, Y.-H.; Zheng, M.-Y.; Zhu, L.; Kong, X.; Huang, T. Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* **2017**, *12*, 526–534. [CrossRef]
97. Chen, L.; Feng, K.Y.; Cai, Y.D.; Chou, K.C.; Li, H.P. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. *BMC Bioinform.* **2010**, *11*, 293. [CrossRef] [PubMed]
98. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451. [CrossRef]