Maryam Pashaias, Khodadad Khodadadi, Amir Hossein Kayvanjoo, Roghiyeh Pashaei-asl, Esmaeil Ebrahimie, Mansour Ebrahimi

Unravelling evolution of Nanog, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics

Gene, 2016; 578(2):194-204

Unravelling evolution of *Nanog*, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics

Maryam Pashaiasl, Khodadad Khodadadi, Amir Hossein Kayvanjoo, Roghiyeh Pashaei-asl, Esmaeil Ebrahimie, Mansour Ebrahimi

PII: S0378-1119(15)01509-7
DOI: doi: 10.1016/j.gene.2015.12.023
Reference: GENE 41055

To appear in: *Gene*

Received date: 26 June 2015
Revised date: 10 December 2015
Accepted date: 10 December 2015

Please cite this article as: Pashaiasl, Maryam, Khodadadi, Khodadad, Kayvanjoo, Amir Hossein, Pashaei-asl, Roghiyeh, Ebrahimie, Esmaeil, Ebrahimi, Mansour, Unravelling evolution of *Nanog*, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics, *Gene* (2015), doi: 10.1016/j.gene.2015.12.023

Unravelling evolution of *Nanog*, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics

Maryam Pashaiasl[1,2,3,4]*, Khodadad Khodadadi[5], Amir Hossein Kayvanjoo[6], Roghiyeh Pashaei-asl[2], Esmaeil Ebrahimie[7,8,9,10], Mansour Ebrahimi[6*]

1. Drug Applied Research Center, Tabriz University of Medical Sciences, Tabriz, Iran.
2. Department of Anatomical Sciences, Faculty of Medicine, Tabriz University of Medical Sciences, Iran
3. Department of Molecular Medicine, School of Advanced Medical Sciences, Tabriz University of Medical Sciences, Tabriz, Iran
4. Women's Reproductive Health Research Centre, Tabriz University of Medical Sciences, Tabriz, Iran
5. Genetic Research Theme, Murdoch Children's Research Institute, Royal Children's Hospital, The University of Melbourne, Australia
6. Department of Biology, University of Qom, Qom, Iran
7. Institute of Biotechnology, Shiraz University, Shiraz, Iran
8. School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, University of South Australia, Adelaide, Australia
9. Department of Genetics and Evolution, School of Biological Sciences, The University of Adelaide, Adelaide, Australia
10. School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia

Corresponding authors, both authors contributed equally

pashaim@tbzmed.ac.ir; mansour@future.org

Abstract

Nanog, an important transcription factor in embryonic stem cells (ESC), is the key factor in maintaining pluripotency to establish ESC identity and has the ability to induce embryonic germ layers. *Nanog* is responsible for self-renewal and pluripotency of stem cells as well as cancer invasiveness, tumor cell proliferation, motility and drug-resistance. Understanding the underlying mechanisms of *Nanog* evolution and regulation can lead to future advances in treatment of cancers. Recent integration of machine learning models with genetics has provided a powerful tool for knowledge discovery and uncovering evolutionary pathways. Herein, sequences of 47 *Nanog* genes from various species were extracted and two datasets of features were computationally extracted from these sequences. At the first dataset, 76 nucleotide acid attributes were calculated for each *Nanog* sequence. The second dataset was prepared based on the 10480 repeated nucleotide sequences (from 5 to 50 bp lengths). Then, various data mining algorithms such as decision tree models were applied on these datasets to find the evolutionary pathways of *Nanog* diversion. Attribute weighting models were highlighted features such as the frequencies of AA and GC as the most important genomic features in *Nanog* gene classification and differentiation. Similar findings were obtained by tree induction algorithms. Results from the second database showed that some short sequence strings, such as ACTACT, TCCTGA, CCTGA, GAAGAC, and TATCCC can be effectively used to identify *Nanog* genes in various species. The outcomes of this study, for the first time, unravels the importance of particular genomic features in *Nanog* gene evolution paving roads toward better understanding of stem cell development and human targeted disorder therapy.

Keywords: Gene Evolution; Bioinformatics; Machine Learning; *Nanog*; Tandem Repeats

Introduction

Embryonic stem cells (ESCs) are derived from inner cell mass (ICM) of the mammalian pre-implantation embryo (Evans and Kaufman, 1981). ESCs have two important characteristics: an unlimited ability of self-renewal and differentiation capacity. They are pluripotent with the capacity to differentiate into three embryonic germ layers (endoderm, mesoderm and ectoderm). In addition, these cells can integrate into embryos and contribute to functional tissue generation. The pluripotency of ESCs seems to be authorized by multiple transcriptional factors. Three core pluripotency factors Oct3/4, Sox2 and Nanog have been introduced as essential factors in maintenance of pluoripotency and self-renewal of ESCs (Pashaiasl et al., 2013; Ebrahimie et al., 2014; Mansouri et al., 2014). In developmental process of embryo, Oct4 activates by Nanog. It has been suggested that the expression of Oct4 and Sox2 is regulated by Nanog, and by the loss of Nanog function, cells will enter to differentiation status (Boyer et al., 2005; Loh et al., 2006; Ebrahimie et al., 2014; Mansouri et al., 2014). It has been also shown that over-expression of *Nanog* increases reprograming efficiency in cell fusion through stimulation and activation of gene in somatic cell genome yielding 200-fold more colonies than others and resetting pluripotency (Silva et al., 2006). In a recent genome-wide analysis, we demonstrated that the common organization of transcription factor binding sites on the non-coding promoter regions of *Nanog*,*Oct4* and *Sox2* can be used for discovery of novel genes involved in stem cell proliferation (irrespective of coding gene sequence) (Hosseinpour et al., 2013).

The *Nanog* is a family of homeobox genes which encodes homeodomain proteins and is part of the key set of transcription factors with a vital role in the second embryonic cell fate arrangement event (Cavaleri and Scholer, 2003; Mitsui et al., 2003). *Nanog* transcription factor is transcribed in the pluripotent cells of human, mouse, monkey, bovine and embryonic germ cells (Mitsui et al., 2009; Pashaiasl et al., 2010; Yang et al., 2012). Its over-expression robustly maintains ESC identity and proliferation as well as invasiveness of the cancer cells (Yang et al., 2012; Wang et al., 2013). *Nanog* expression increases in the ICM stage of embryo and is lost around the time of implantation. Persistence of *Nanog* expression will delay blastocyst implantation (Chambers et al., 2003; Mitsui et

al., 2003). Interestingly, disruption of the *Nanog* causes losing pluripotency in both ICM and ESCs. Its expression level is crucial for pluripotency and cells expressing high level of *Nanog* fulfil pluripotency. In contrast, low level *Nanog* expressing cells will move to differentiation procedure (Luo et al., 2012; Ebrahimie et al., 2014).

Abnormal expression of *Nanog* has been observed in several types of tumour and tumorigenic tissues. The level of *Nanog* is linked to the prognoses of tumorogenecity of the cells and its expression could support the proliferation and invasiveness of the cancer cells, inhibiting the apoptosis. Knockdown of *Nanog* in liver cancer cells (Zhou et al., 2014) and pancreatic cancer cells (Bluteau et al., 2013) significantly decreased pluripotent ability and increased chemosensitivity of cancer cells.

Despite its importance in both stem cell proliferation and cancer pathology, mechanisms underpinning *Nanog* function and regulation in genomic level have been poorly understood. Increasing the number of available gene and protein sequences in different species in line with the recent development of advanced mathematical formula such as feature selection (attribute weighting models), decision trees, support vector machine (SVM), association rule mining, and neural networks has opened a new avenue in genetics for understanding gene function and evolution (Tahrokh et al., 2011; Zinati et al., 2014; Ebrahimi et al., 2015). For application of the mentioned data mining models in genetics, it is essential to convert the gene/protein sequence to a series of attributes (features). Computationally calculated nucleotide attributes such as frequency of different nucleotides and di-neucleotides have been widely used in this context (KayvanJoo et al., 2014). String of tandem repeats, in view of existence and numbers, are another type of employed features. As example, combination of viral nucleotide attributes with machine learning successfully predicted the therapy outcome of interferon/ribavirin in hepatitis C (KayvanJoo et al., 2014). In another example, decision tree models unravelled the evolutionary pathway of ammonium transporters in different organisms based on di-peptide attributes (Tahrokh et al., 2011). Combination of data mining algorithms with tandem repeat features, compared to the common multivariate based models, has resulted in more precise genotype

discrimination and oped a new avenue for classification and prediction of different genotypes (Beiki et al., 2012; Nasiri et al., 2015; Torkzaban et al., 2015).

In this study, a range of nucleotide features and tandem repeats were calculated for *Nanog* sequences in different organisms, categorised in 2 different datasets. Various machine learning models were applied to (1) find the key discriminating genomic attributes governing the differentiation of *Nanog* transcription factors in different organisms, and (2) to find the best combination of nucleotide or tandem repeat features for unravelling evolutionary pathways of *Nanog* in different organisms.

Materials & Methods

Preparation of two datasets of nucleotide and tandem repeat attributes

Forty-seven *Nanong* genes from different species (Fish, Mouse, Primates, Cat, Birds and Domestic Mammals) were extracted from NCBI database. List of GIs of *Nanong* sequences is presented in Supplementary 1. Two distinct datasets created as follows:

Gene Attribute Dataset (GAD dataset)

Seventy six gene attributes – e.g. the count and frequency of each nucleotide, di-nucleotides, and molarities of salt contents (the concentration of monovalent cations in units of molar) were extracted using CLC bio main workbench (Qiagen). All features were classified as continuous variables, except the *Nanog* species which were classified as polynomial variable. A dataset of these genes features was imported into RapidMiner software [RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany], null data for type variable was discarded, and this feature was set as the output (target) variable and the other variables were set as input variables.

Repeated Sequences Database (RSD dataset)

Repeated sequences of nucleotides (from 5 to 50 repeats) were calculated based on in-house software developed in our laboratory. The generated dataset contained 10480 attributes (or features) of any possible repeated sequences for all 47 extracted gene sequences. Then, the steps detailed below were applied on both datasets.

Data cleaning

In data cleaning step, at first, all features were checked by comparing all examples with each other on the basis of the specified selection of attributes; if there was any duplication, we removed that (two examples were assumed equal if all values of all selected attributes were equal). Next, useless attributes (with low standard deviations) were removed from the dataset. Nominal attributes were regarded as useless when the most frequent values were contained in more or less than nominal useless above or below per cent of all examples. Numerical attributes which possessed standard

deviations less than or equal to a given deviation threshold (0.1) were assumed to be useless and removed. Finally, correlated features (with Pearson correlation greater than 0.95) were omitted. These databases were named as Final Cleaned database (FCdb).

Attribute Weighting

In order to identify the most important attributes and to find the possible patterns in features that contribute to divergence of *Nanog* genes in different species, 10 different algorithms of weighting models were applied to the final cleaned datasets (FCdb) as described below:

Weight by Information gain: this model calculates the relevance of an attribute by computing the information gain in class distribution.

Weight by Information Gain ratio: this operator calculates the relevance of an attribute by computing the information gain ratio for the class distribution.

Weight by Rule: method of this operator is calculating the relevance of an attribute by computing the error rate of a OneR Model on the example set without this feature.

Weight Deviation: this operator created weights from the standard deviations of all attributes. The values were normalised by the average, the minimum, or the maximum of the attribute.

Weight by Chi squared statistic: This weighting operator calculates the relevance of an attribute by computing, for each attribute of the input example set, the value of the chi-squared statistic with respect to the class attribute.

Weight by Gini index: This operator calculated the relevance of an attribute by computing the Gini index of the class distribution, if the given example set would have been split according to the feature.

Weight by Uncertainty: This operator calculated the relevance of an attribute by measuring the symmetrical uncertainty with respect to the class.

Weight by Relief: This operator measured the relevance of features by sampling examples and comparing the value of the current feature for the nearest example of the same and of a different class. This version also worked for multiple classes and regression data sets.

Weight by PCA: This operator is based on principle component analysis and takes the coefficients of the first component of PCA as feature weights.

Weight by SVM: This operator is based on linear Support Vector Machine (SVM) and takes the coefficients of the normal vector as weight of features.

The resulting weights were normalised into the interval between 0 and 1 to allow the comparison between different methods.

Comparing the results of different "Attribute Weighting" algorithms and "Attribute selection"

After running attribute weighting models on the datasets, each gene attribute (feature) gained a value between 0 and 1, which revealed the importance of that attribute with regards to a target attribute (Role of proteins in depression). We selected variables with weights higher than 0.5as important features according to the employed weighting model.

Trimming the original datasets according to "Attribute Weighting algorithms" and generating new datasets

For each of GAD and RSD datasets, 10 new datasets were created containing the features which were announced important in attribute weighting algorithms. These newly formed datasets were named according to their attribute weighting models (*Information gain, Information gain ratio, Rule, Deviation, Chi Squared, Gini index, Uncertainty, Relief, SVM and PCA*) and were used to join with subsequent predictive trees induction models. In total for each of GAD or RSD dataset, 11 datasets were used for "trees induction models": 1 original dataset, and 10 datasets with trimmed features according to attribute weighting models.

Trees Induction Models

Four tree induction models, including *Decision Tree, Decision Tree Parallel, Decision Stump and Random Forest,* were run on all 11 main datasets. Each tree induction model ran with the following four different criteria: *Gain Ratio, Information Gain, Gini Index and Accuracy*. As a result, 16 combinational machine learning models were applied including *decision tree Accuracy, decision tree Gain Ratio, decision tree Gini Index, decision tree Info Gain, decision tree Parallel Accuracy, decision tree Parallel Gain Ratio, decision tree Parallel Gini Index, decision tree Parallel Info Gain,*

decision tree Stump Accuracy, *decision tree Stump Gain Ratio*, *decision tree Stump Gini Index*, *decision tree Stump Info Gain*, *decision tree Random Forest Accuracy*, d*ecision tree Random Forest Gain Ratio*, *decision tree Random Forest Gini Index*, and *decision tree Random Forest Info Gain*. As *Random Forest* model generates 10 different trees for each criterion, 572 trees were induced by tree induction models.

To calculate the accuracy of each model, 10-fold cross validation (Habashy et al., 2010) was used to train and test models on all patterns. To perform cross validation, all the records (47) were randomly divided into 10 parts; 9 sets were used for training and the 10th one for testing. The process was repeated 10 times and the accuracy for true, false and total accuracy calculated. The final accuracy was reported as the average of the accuracy in all ten tests.

Application of multivariate and univariate methods on selected important methods by attribute weighting models

Selected attribute by machine learning models (such as attribute weighting models) were further used for application of common multivariate and univariate methods as previously described (Zinati et al., 2014). To this end, the important features selected by intersection of the above mentioned 10 attribute weighting algorithms were used for clustering, MANOVA (multiple ANOVA), ANOVA, clustering, and PCA. Clustering carried out based on Average Linkage and Euclidean Distance. Features were standardised before clustering. Univariate and multivariate analysis carried out by MINITAB 17.

Results

Generated datasets

The generated Gene Attribute Dataset (GAD dataset) is presented in Supplementary 2. This dataset provides a comprehensive view on underlying nucleotide attributes of *Nanog* genes in different organisms for running attribute weighting and pattern recognition models.

The dataset of Repeated Sequences Dataset (RSD dataset) is presented at Supplementary 3. RSD dataset comprehensively monitored the existence and the number of repeated sequences (from 5 to 50 repeats) in *Nanog* nucleotide sequences and provided a big dataset for pattern recognition.

Data cleaning

The initial datasets contained 47 records with 76 and 10486 attributes, in the first (GAD) and second (RSD) datasets, respectively. The 47 records belonged to 6 groups of Domestic Mammals (18), Mouse (11), Primates (8), Fish (5), Birds (3), and Felis Catus (2). Following the removal of duplicates, useless attributes, and correlated features (data cleaning) 31 attributes were remained in GAD (genomic features dataset) and 10479 in RSD (repeated sequences dataset).

Attribute weighting

Data were normalized before running the models, and 10 different attribute weighting models (as described in material and methods) were run on GAD and RSD datasets. Each attribute was weighted between 0 and 1. These weights determined the importance of attributes in *Nanog* differentiation and evolution. Attributes which gained weight equal to 0.5 or higher by at least five weighting models were selected.

Table 1 shows the important genomic attributes of *Nanog* genes in different organisms selected by weighting algorithms. The frequencies of AA and GC were the two top features which were selected

by 70% of attribute weighting models. The detailed weights and the results of attribute weighting models on genomic attributes are presented at supplementary 4.

Interestingly, as presented in Table 2, 13 strings of tandem repeats were selected as the most important features by all attribute weighting algorithms in RSD dataset. These repeats were: TATCCC, AGCTATA, CCAGAC, GACCTG, AGATGC, GCAGCC, ACTACT, AGACCT, ACTTGG, GAAGAC, TCCTGA, GCAGC, and CCTGA. Within these repeats, AGCTATA was the longest selected string. Supplementary 5 shows the detailed weights and results of attribute weighting models on the repeated sequence features.

Secondary generated datasets via trimming the original datasets by attribute weighting models

In addition to finding the important features, we used the attribute weighting models for generating the secondary datasets. These datasets were only contained the features which were announced important by the corresponding weighting model. These datasets and their features are presented at Supplementary 6. Also, Table 3 presents the number of remaining features in each of GAD and RSD datasets after selection of important features with attribute weighting models. Attribute weighting models were remarkably different in selection of attributes. The size of new generated gene attribute datasets was remarkably different from only 1 feature in weighting by PCA to 26 features in weighting by Info Gain Ratio and weighting by Uncertainty (Table 3, Supplementary 6). Also, the size of new repeated sequence datasets was distinguishly different from only 1 feature in weighting by PCA to 1420 repeated sequence features in weighting by Info Gain Ratio. The goal of generating new branched datasets was to evaluate the effect of feature selection on increasing the accuracy of tree induction models to find the best combination of tree induction model and attribute weighting model. Also, attribute weighting models are important techniques to prevent overfitting.

Trees Induction and pattern recognition in genomic features (GAD dataset)

From 572 decisions tree induced by tree induction models for GAD dataset, none of them was able to completely distinguish between labels (organisms). Within the applied models, Decision Tree

Random Forest when run on dataset either filtered by Gini Index or SVM criterion was the best model in differentiation of *Nanog* genes (Figures 1 and Figure 2).

Figure 1 shows the *Nanog* evolutionary pattern *for Decision Tree Random Forest* with *Gain Ratio* criterion when ran on GAD dataset pre-filtered by *Gini Index*. This model highlights the frequency of CG as the main attribute in *Nanog* differentiation. If frequency of this feature was more than 0.03 in the example, the record fell into Fish. However, if frequency of CG di-nucleotide was less or equal to 0.03, then it depends on the level of next attribute which is Salt 0.1M. If Salt 0.1M was more than 86.480, the record fell into the Birds group and if it's less or equal to 86.480, the class group depends on next attribute and so on.

Figure 2 presents the model for *Decision Tree Random Forest* ran based on *Gini Index* criterion on *SVM* dataset of genomic features. Similar to previous model, frequency of CG is the key attribute where in combination with frequency of CT discriminate *Nanog* sequences between different organisms.

Accuracy of different tree induction models in combination with GAD datasets trimmed with attribute weighting models (as well as original non-trimmed dataset) in prediction the origin of *Nanog* sequences are presented in Table 4 (based on 10-fold cross validation). Within the decision tree models, Gini Index, Random Forest Gini Index, and Random Forest Info Gain had the highest average of accuracy with 66.91%, 64.95%, 63.41%, respectively (Table 4). Within the attribute weighting models, employed for trimming GAD dataset, Ginin Index and SVM showed performance with 63.72% and 62.28% average of accuracy which was higher than Fcdb (without feature selection with 60.6%). This shows that pre- feature selection of genomic attributes with the Ginin Index and SVM models and importing the important features are able to increase the prediction efficiency of tree induction models. Ginin Index and SVM attribute weighting models reduced the number of genomic attributes from 30 attributes to 10 and 11 attributes, respectively (Table 3). As presented in Table 4, the highest

accuracies in prediction of origin organism of *Nanog* sequences based on genomic attributes were obtained in the following combinations: (1) Decision tree Random Forest Gini Index+ GAD dataset trimmed by SVM attribute weighting with 82% accuracy, and (2) Decision tree Random Forest Gain Ratio + GAD dataset trimmed by Gini Index with 81.5% accuracy (Table 4).

Trees Induction and pattern recognition in repeated Sequence features (RSD dataset)

Tree induction models generated 572 trees on RSD datasets. Noticeably, many trees were able to clearly distinguish between *Nanog* genes from various organisms. As example, Figure 3 shows that *Decision Tree algorithm* with *Gini Index* criterion is able to fully distinguish between *Nanog* genes. GCCCAG was the root feature and the most important feature to induce the tree and determine *Nanog's* organism.

Table 5 shows the accuracy of different decision tree models in combination of pre-feature selection of repeated sequence features with attribute weighting models based on 10-fold cross validation. The models with higher accuracy have the higher performance in discrimination and classification of *Nanog*'s organism based on the discovery of the pattern of repeated sequences. Decision tree Random Forest Info Gain in combination of RSD dataset trimmed by Info Gain attribute weighting was successful in 87% of cases to predict the organism origin of *Nanog* sequences based on its repeated sequence attributes (Table 5). Overall, decision tree Gini Index (with average accuracy of 67.27%), decision tree Random Forest with Gini Index criterion (with average accuracy of 66.91%), and decision tree Random Forest with Info Gain criterion (with average accuracy of 65.73%) were the best tree induction models (Table 5). Interestingly, compared to the original dataset of repeated sequence (FCdb, without feature selection), in average, feature selection by Info Gain

attribute weighting increased the accuracy of prediction by 4% (from 68.5% in FCdb to 72.5% in Info Gain dataset) (Table 5). This demonstrates the importance of pre-feature selection of RSD dataset before running of predictive tree induction models. Trimming RSD dataset with Info Gain feature selection dataset reduced the number of attributes from 10479 to 2101 attributes (Table 3).

Further investigations of selected genomic features by attribute weighting and decision tree models

Statistics of 10 important features selected by attribute weighting models in different organisms is presented in Table 6. Mean and variance of features in different organisms are variable which show importance of these features in evolution of *Nanog*. Coefficient of Variance (CV) showed that Frequency of AA and Frequency of CG were highly variable in Birds, Domestic Mammals, Fish, Mouse and Primates. High variation of most of features in Primates is noticeable. Frequency of TG and Frequency of GG, Frequency of CT and salt 0.1M had high variation in fish (Table 6).

Clustering of *Nanog* sequences based on important genomic features is presented in Figure 4. As it can be inferred from Figure 4, *Nanog* sequences in birds and fish make a separate cluster compared to the other organisms.

Analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) of 10 important genomic features, selected by attribute weighting methods (Table 1), among different organisms is presented in Supplementary 7. MANOVA carried out using Wilks', Lawley-Hotelling, and Pillai's criteria. ANOVA of all features as well as MANOVA analysis showed that the selected features were all significant between organisms at $p = 0.05$. The attributes which were the main features in pattern recognition and rule discovery Pattern

recognition in differentiation of *Nanog* genes between different organisms based on *Decision Tree Random Forest* (Figure 2), including Frequency of CG (R-square = 70.76%), Frequency of CT (R-square = 56.42%), Frequency of GC (R-square = 76.33%),

These findings reconfirm the importance of selected attributes and highlight the efficiency of attribute weighting ad decision tree models in finding the important features within a large number of features.

Principle Component Analysis (PCA) of *Nanog* sequences based on important genomic features, selected by attribute weighting models (Table 1), is presented at Supplementary 8. First and second PCAs could explain 68.9% of variations. PCA plot of first and second components showed that bird and fish sequences group together and primates have highly diverse genomic attributes. Coefficients of PCA components shows that salt 0.1M and Frequency of GC in first component as well as Frequency of Adenine, Frequency of AA, and Frequency of CT in second component are important features which reconfirms the high efficiency of the Decision Tree Random Forest in finding evolutionary pathways of *Nanog* sequences.

A practical guide for running data mining models using Rapid Miner package is presented at Supplementary 9.

Discussion

Nanog is an important regulator both in stem cells and cancer research. *Nanog* acts differently in various species and could be targeted for therapeutic aims. *Nanog* is a valuable target for manipulation, drug discovery, and possible gene and cancer therapy.

To find simple and efficient way to investigate the structural differences at genomic levels of *Nanog* genes in different organisms, the current study has looked deeply into the matter by using various bioinformatics tools. Two different databases were created; one based on nucleotide attributes of genes (GAD) and the other one based on tandem repeats of gene sequences (RAD); these databases created for the first time and provided very useful base for algorithm application.

Many resources are needed to run an algorithm on a big database (such as RAD, which contained more than 10000 columns), so data reduction algorithms should be applied on huge datasets to prevent burden on processing facilities. Data cleaning algorithms such as remove useless or remove correlated attributes used here and the size of GAD reduced by 60% but RAD database freed by just 1% showing nearly all attributes were necessary and not-redundant.

Attribute weighting algorithms weighs the importance of each attribute in distinguishing between *Nanog* genes from various organisms. The frequencies of AA and GC in GAD and 12 strings in RAD selected as the most important features; AGCTATA was the longest sequence used by these methods to clearly distinguish between *Nanog* genes from various organisms. For the first time, herein, a small sequence based on genomic sequences of *Nanog* genes has been proposed to identify them in different organisms.

Tree induction algorithms also highlighted the importance of features weighed top in weighting models. Decision tree models selected GCCCGA as the most important feature to build the tree root and then two other strings (one with 27 nucleotides) employed to distinguish between Fish and other organisms. Decision tree also showed the importance of di-nucleotides feature of CG to build the tree based on this feature first. These findings for the first time highlight the importance of di-nucelotides frequencies in phylogenic structure of *Nanog* genes.

Nanog is an important regulator of functional genomics in different animals and understanding its regulation and expression could pave roads toward customized human targeted disorder therapy. In fact, identification of *Nanog* gene expression and regulation is a crucial step in understanding early embryogenesis, pluripotent cells development and cancer cells proliferations. Considering the importance of *Nanog* gene evolution and its diversity in different organisms, herein, for the first time we proposed new criteria based on bioinformatics tools to easily identify them based on genomic features. The importance of features based on genomic attributes of *Nanog* genes in various organisms is highlighted. The findings may pave the roads to understand the evolutionary path of genes contributing to new therapeutic aims for cancer and also stem cell therapy.

Acknowledgment

This study is supported by research grant of Tabriz University of Medical Sciences, Tabriz, Iran awarded to Dr. Maryam Pashaiasl.

References

Beiki, A.H., Saboor, S. and Ebrahimi, M. A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. *PloS one* **7** (2012), p. e44164.

Bluteau, O., Langlois, T., Rivera-Munoz, P., Favale, F., Rameau, P., Meurice, G., Dessen, P., Solary, E., Raslova, H., Mercher, T., Debili, N. and Vainchenker, W. Developmental changes in human megakaryopoiesis. *J Thromb Haemost* **11** (2013), pp. 1730-41.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R. and Young, R.A. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122** (2005), pp. 947-56.

Cavaleri, F. and Scholer, H.R.: Nanog: a new recruit to the embryonic stem cell orchestra. Cell Press (2003), pp. 551-557.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S. and Smith, A.: Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells, Cell Press (2003), pp. 643-655.

Ebrahimi, M., Ebrahimie, E. and Bull, C.M. Minimizing the cost of translocation failure with decision-tree models that predict species' behavioral response in translocation sites. *Conservation Biology* (2015).

Ebrahimie, M., Esmaeili, F., Cheraghi, S., Houshmand, F., Shabani, L. and Ebrahimie, E. Efficient and simple production of insulin-producing cells from embryonal carcinoma stem cells using mouse neonate pancreas extract, as a natural inducer. *PloS one* **9** (2014), p. e90885.

Evans, M.J. and Kaufman, M.H. Establishment in culture of pluripotential cells from mouse embryos. *nature* **292** (1981), pp. 154-156.

Habashy, H.O., Powe, D.G., Glaab, E., Ball, G., Spiteri, I., Krasnogor, N., Garibaldi, J.M., Rakha, E.A., Green, A.R., Caldas, C. and Ellis, I.O. RERG (Ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: a marker of ER-positive luminal-like subtype. *Breast Cancer Res Treat* (2010).

Hosseinpour, B., Bakhtiarizadeh, M.R., Khosravi, P. and Ebrahimie, E. Predicting distinct organization of transcription factor binding sites on the promoter regions: a new genome-based approach to expand human embryonic stem cell regulatory network. *Gene* **531** (2013), pp. 212-219.

KayvanJoo, A.H., Ebrahimi, M. and Haqshenas, G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. *BMC research notes* **7** (2014), p. 565.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.Y., Sung, K.W., Lee, C.W., Zhao, X.D., Chiu, K.P., Lipovich, L., Kuznetsov, V.A., Robson, P., Stanton, L.W., Wei, C.L., Ruan, Y., Lim, B. and Ng, H.H. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38** (2006), pp. 431-40.

Luo, Y., Lim, C.L., Nichols, J., Martinez-Arias, A. and Wernisch, L. Cell signalling regulates dynamics of Nanog distribution in embryonic stem cell populations. *Journal of the Royal Society, Interface / the Royal Society* (2012).

Mansouri, A., Esmaeili, F., Nejatpour, A., Houshmand, F., Shabani, L. and Ebrahimie, E. Differentiation of P19 embryonal carcinoma stem cells into insulin-producing cells promoted by pancreas-conditioned medium. *Journal of tissue engineering and regenerative medicine* (2014).

Mitsui, K., Suzuki, K., Aizawa, E., Kawase, E., Suemori, H., Nakatsuji, N. and Mitani, K. Gene targeting in human pluripotent stem cells with adeno-associated virus vectors. *Biochem Biophys Res Commun* **388** (2009), pp. 711-7.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M. and Yamanaka, S. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113** (2003), pp. 631-42.

Nasiri, J., Naghavi, M.R., Kayvanjoo, A.H., Nasiri, M. and Ebrahimi, M. Precision assessment of some supervised and unsupervised algorithms for genotype discrimination in the genus pisum using SSR molecular data. *Journal of theoretical biology* (2015).

Pashaiasl, M., Khodadadi, K., Holland, M.K. and Verma, P.J. The efficient generation of cell lines from bovine parthenotes. *Cell Reprogram* **12** (2010), pp. 571-9.

Pashaiasl, M., Khodadadi, K., Richings, N.M., Holland, M.K. and Verma, P.J. Cryopreservation and long-term maintenance of bovine embryo-derived cell lines. *Reprod Fertil Dev* **25** (2013), pp. 707-18.

Silva, J., Chambers, I., Pollard, S. and Smith, A. Nanog promotes transfer of pluripotency after cell fusion. *Nature* **441** (2006), pp. 997-1001.

Tahrokh, E., Ebrahimi, M., Ebrahimi, M., Zamansani, F., Sarvestani, N.R., Mohammadi-Dehcheshmeh, M., Ghaemi, M.R. and Ebrahimie, E. Comparative study of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms. *Genes & Genomics* **33** (2011), pp. 565-575.

Torkzaban, B., Kayvanjoo, A.H., Ardalan, A., Mousavi, S., Mariotti, R., Baldoni, L., Ebrahimie, E., Ebrahimi, M. and Hosseini-Mazinani, M. Machine Learning Based Classification of Microsatellite Variation: An Effective Approach for Phylogeographic Characterization of Olive Populations. *PloS one* **10** (2015).

Wang, M.-L., Chiou, S.-H. and Wu, C.-W. Targeting cancer stem cells: emerging role of Nanog transcription factor. *OncoTargets and therapy* **6** (2013), p. 1207.

Yang, L., Zhang, X., Zhang, M., Zhang, J., Sheng, Y., Sun, X., Chen, Q. and Wang, L.-X. Increased nanog expression promotes tumor development and Cisplatin resistance in human esophageal cancer cells. *Cellular Physiology and Biochemistry* **30** (2012), pp. 943-952.

Zhou, J.J., Deng, X.G., He, X.Y., Zhou, Y., Yu, M., Gao, W.C., Zeng, B., Zhou, Q.B., Li, Z.H. and Chen, R.F. Knockdown of NANOG enhances chemosensitivity of liver cancer cells to doxorubicin by reducing MDR1 expression. *Int J Oncol* **44** (2014), pp. 2034-40.

Zinati, Z., Zamansani, F., KayvanJoo, A.H., Ebrahimi, M., Ebrahimi, M., Ebrahimie, E. and Dehcheshmeh, M.M. New layers in understanding and predicting α-linolenic acid content in plants using amino acid characteristics of omega-3 fatty acid desaturase. *Computers in biology and medicine* **54** (2014), pp. 14-23.

Table 1. Important genomic attributes of *Nanog* sequences that gained weight higher or equal to 0.5 by at least 3 weighting model in Nanog's differentiation between different species based Gene Attribute Dataset (GAD dataset) [*]

Importance ranking of genomic attribute	Genomic attribute	Number of weighting models which announced attribute as important[**]	Importance ranking of genomic attribute	Genomic attribute	Number of weighting models which announced attribute as important[**]
1	Frequency of AA	7	15	Frequency of AT	5
2	Frequency of GC	7	16	Frequency of AG	5
3	Frequency of GG	6	17	Frequency of Cytosine	5
4	Frequency of CT	6	18	DS Frequency of carbon	4
5	Frequency of CG	6	19	Frequency of carbon	4
6	Frequency of TG	6	20	DS Frequency of hydrogen	4
7	Frequency of AC	6	21	Frequency of hydrogen	4
8	Frequency of Adenine	6	22	Frequency of GA	4
9	Frequency of CA	6	23	Frequency of Thymine	4
10	salt 0.1M	6	24	DS Frequency of nitrogen	3
11	Frequency of oxygen	5	25	Frequency of nitrogen	3
12	Frequency of Guanine	5	26	Frequency of TC	3
13	Length	5	27	Frequency of GT	3
14	Frequency of TA	5			

[*] In total, 76 genomic attributes were calculated for each *Nanog* sequence.

[**]Seven attribute weighting models were tested. Importance of di- nucleotides of AA and GC were confirmed by all (100%) of tested attribute weighting models.

Table 2. Tandem repeat sequences which can significantly distinguish *Nanog* in different organisms as they received higher weight (equal to or higher than 0.5) by all weighting models based on Repeated Sequences Dataset (RSD) [*]

Importance ranking of tandem repeat attribute	**Tandem repeat attribute**	**Number of weighting models which announced attribute as important** [**]
1	TATCCC	7
2	AGCTATA	7
3	CCAGAC	7
4	GACCTG	7
5	AGATGC	7
6	GCAGCC	7
7	ACTACT	7
8	AGACCT	7
9	ACTTGG	7
10	GAAGAC	7
11	TCCTGA	7
12	GCAGC	7
13	CCTGA	7

[*] In total, 10480 tandem repeat attributes were calculated for each *Nanog* sequence.

[**]Seven attribute weighting models were tested. The importance of above tandem repeates were confirmed by all (100%) of tested attribute weighting models.

Table 3. New generated datasets by filtering attributes with various attribute weighting models in Gene Attribute Dataset and Repeated Sequence Dataset. The detailed list of attributes in each new dataset is presented at Supplementary 6.

Attributes weighting model	**Number of remained attributes in Gene Attribute Dataset**	**Number of remained attributes in Repeated Sequence Dataset**
Weighting by PCA	1	1
Weighting by Deviation	1	4
Weighting by Relief	4	208
Weighting by SVM	10	1923
Weighting by Gini Index	11	969
Weighting by Rule	11	67
Weighting by Chi Squared	20	105
Weighting by Info Gain	25	2101
Weighting by Uncertainty	26	718
Weighting by Info Gain Ratio	26	1420
FCdb (Origibnal data set) after data cleaning	30	10479

Table 4. Comparison the accuracy of different tree induction model in combination of datasets trimmed with different attribute weighting algorithms to predict the origin of *Nanog* sequences based on the genomic features. Ten-fold cross validation was used for comparison of the models*.

Data Base trimmed by	DT Accuracy	DT Gain Ratio	DT Gini Index	DT Info Gain	DT Parallel Accuracy	DT Parallel Gain Ratio	DT Parallel Gini Index	DT Parallel Info Gain	DT Stump Accuracy	DT Stump Gain Ratio	DT Stump Gini Index	DT Stump Info Gain	DT Random Forest Accuracy	DT Random Forest Gain Ratio	DT Random Forest Gini Index	DT Random Forest Info Gain	**Maximum accuracy of attribute weighting**	**Minimum accuracy of attribute weighting**	**Average accuracy of attribute weighting**
Chi Squared	59.50 %	63.50 %	70.50 %	64.00 %	64.00 %	70.00 %	71.00 %	62.00 %	49.00 %	53.50 %	49.50 %	40.50 %	49.50 %	74.50 %	72.50 %	73.00 %	**74.50 %**	**40.50 %**	**61.66 %**
Info Gain	65.50 %	63.50 %	73.00 %	64.00 %	57.50 %	61.50 %	73.00 %	68.00 %	49.00 %	49.00 %	49.50 %	40.50 %	56.00 %	63.50 %	64.00 %	71.00 %	**73.00 %**	**40.50 %**	**60.53 %**
Deviation	38.50 %	28.50 %	44.50 %	42.50 %	38.50 %	28.50 %	44.50 %	42.50 %	34.00 %	30.00 %	39.00 %	32.00 %	38.50 %	40.50 %	45.50 %	41.00 %	**45.50 %**	**28.50 %**	**38.03 %**
Gini Index	65.50 %	66.00 %	77.00 %	64.00 %	70.00 %	66.00 %	68.50 %	70.50 %	49.00 %	49.00 %	49.50 %	40.50 %	55.50 %	81.50 %	72.50 %	74.50 %	**81.50 %**	**40.50 %**	**63.72 %**
Info Gain Ratio	65.50 %	63.50 %	70.50 %	64.00 %	65.50 %	65.00 %	68.50 %	64.50 %	49.00 %	53.50 %	49.50 %	40.50 %	43.50 %	64.00 %	78.00 %	69.00 %	**78.00 %**	**40.50 %**	**60.88 %**
PCA	38.50 %	28.50 %	44.50 %	42.50 %	38.50 %	28.50 %	44.50 %	42.50 %	34.00 %	30.00 %	39.00 %	32.00 %	38.50 %	40.50 %	45.50 %	41.00 %	**45.50 %**	**28.50 %**	**38.03 %**
Relief	61.50 %	57.50 %	63.50 %	57.50 %	61.50 %	54.00 %	63.50 %	55.50 %	38.50 %	38.50 %	38.50 %	38.50 %	46.50 %	48.00 %	43.50 %	49.50 %	**63.50 %**	**38.50 %**	**51.00 %**
Rule	66.50 %	59.50 %	77.00 %	68.00 %	68.50 %	65.00 %	69.00 %	74.50 %	49.00 %	49.00 %	53.50 %	40.50 %	49.00 %	55.00 %	65.00 %	68.50 %	**77.00 %**	**40.50 %**	**61.09 %**
Uncertainty	65.50 %	63.50 %	70.50 %	64.00 %	65.50 %	66.00 %	71.00 %	68.00 %	49.00 %	53.50 %	49.50 %	40.50 %	43.50 %	64.00 %	78.00 %	69.00 %	**78.00 %**	**40.50 %**	**61.31 %**
SVM	64.00 %	63.50 %	72.50 %	70.50 %	66.50 %	61.50 %	70.50 %	70.50 %	38.50 %	43.00 %	49.50 %	40.50 %	55.50 %	69.00 %	82.00 %	79.00 %	**82.00 %**	**38.50 %**	**62.28 %**
FCdb (Original	65.50 %	63.50 %	72.50 %	66.00 %	57.00 %	61.50 %	72.50 %	64.50 %	49.00 %	53.50 %	49.50 %	40.50 %	47.00 %	68.50 %	68.00 %	62.00 %	**72.50 %**	**40.50 %**	**60.06 %**

data set)																
Maximum accuracy of DT	**66.50%**	**66.00%**	**77.00%**	**70.50%**	**70.00%**	**70.00%**	**73.00%**	**74.50%**	**49.00%**	**53.50%**	**53.50%**	**40.50%**	**56.00%**	**81.50%**	**82.00%**	**79.00%**
Minimum accuracy of DT	**38.50%**	**28.50%**	**44.50%**	**42.50%**	**38.50%**	**28.50%**	**44.50%**	**42.50%**	**34.00%**	**30.00%**	**38.50%**	**32.00%**	**38.50%**	**40.50%**	**43.50%**	**41.00%**
Average accuracy of DT	**59.64%**	**56.45%**	**66.91%**	**60.64%**	**59.36%**	**57.05%**	**65.14%**	**62.09%**	**44.36%**	**45.68%**	**46.95%**	**38.77%**	**47.55%**	**60.82%**	**64.95%**	**63.41%**

*DT is the abbreviation of decision tree.

Table 5. Comparison the accuracy of different tree induction model in combination of datasets trimmed with different attribute weighting algorithms to predict the origin of *Nanog* sequences based on the number of repeated sequence features. Ten-fold cross validation was used for comparison of the models.

Data Base trimmed by	DT Accuracy	DT Gain Ratio	DT Gini Index	DT Info Gain	DT Parallel Accuracy	DT Parallel Gain Ratio	DT Parallel Gini Index	DT Parallel Info Gain	DT Stump Accuracy	DT Stump Gain Ratio	DT Stump Gini Index	DT Stump Info Gain	DT Random Forest Accuracy	DT Random Forest Gain Ratio	DT Random Forest Gini Index	**Maximum accuracy of attribute weighting**	**Minimum accuracy of attribute weighting**	**Average accuracy of attribute weighting**
Chi Squared	66.50%	73.50%	74.00%	67.00%	68.50%	71.50%	76.00%	67.00%	59.00%	44.00%	54.50%	52.00%	69.00%	86.50%	80.50%	**82.50%**	**86.50%**	**44.00%**
Info Gain	81.50%	75.50%	73.50%	71.50%	81.50%	77.00%	75.50%	79.50%	56.50%	42.00%	61.00%	57.00%	73.00%	72.50%	77.50%	**87.00%**	**87.00%**	**42.00%**
Deviation	38.50%	28.50%	44.50%	42.50%	38.50%	28.50%	44.50%	42.50%	34.00%	30.00%	39.00%	32.00%	38.50%	40.50%	45.50%	**41.00%**	**45.50%**	**28.50%**
Gini Index	65.50%	66.00%	77.00%	64.00%	70.00%	66.00%	68.50%	70.50%	49.00%	49.00%	49.50%	40.50%	55.50%	81.50%	72.50%	**74.50%**	**81.50%**	**40.50%**
Info Gain Ratio	65.50%	63.50%	70.50%	64.00%	65.50%	65.00%	68.50%	64.50%	49.00%	53.50%	49.50%	40.50%	43.50%	64.00%	78.00%	**69.00%**	**78.00%**	**40.50%**
PCA	38.50%	28.50%	44.50%	42.50%	38.50%	28.50%	44.50%	42.50%	34.00%	30.00%	39.00%	32.00%	38.50%	40.50%	45.50%	**41.00%**	**45.50%**	**28.50%**
Relief	61.50%	57.50%	63.50%	57.50%	61.50%	54.00%	63.50%	55.50%	38.50%	38.50%	38.50%	38.50%	46.50%	48.00%	43.50%	**49.50%**	**63.50%**	**38.50%**
Rule	66.50%	59.50%	77.00%	68.00%	68.50%	65.00%	69.00%	74.50%	49.00%	49.00%	53.50%	40.50%	49.00%	55.00%	65.00%	**68.50%**	**77.00%**	**40.50%**
Uncertainty	65.50%	63.50%	70.50%	64.00%	65.50%	66.00%	71.00%	68.00%	49.00%	53.50%	49.50%	40.50%	43.50%	64.00%	78.00%	**69.00%**	**78.00%**	**40.50%**
FCdb	65.50%	63.50%	72.50%	66.00%	57.00%	61.50%	72.50%	64.50%	49.00%	53.50%	49.50%	40.50%	47.00%	68.50%	68.00%	**62.00%**	**72.50%**	**40.50%**
SVM	64.00%	63.50%	72.50%	70.50%	66.50%	61.50%	70.50%	70.50%	38.50%	43.00%	49.50%	40.50%	55.50%	69.00%	82.00%	**79.00%**	**82.00%**	**38.50%**
Maximum accuracy of DT	**81.50%**	**75.50%**	**77.00%**	**71.50%**	**81.50%**	**77.00%**	**76.00%**	**79.50%**	**59.00%**	**53.50%**	**61.00%**	**57.00%**	**73.00%**	**86.50%**	**82.00%**	**87.00%**	**81.50%**	
Minimum	**38.50%**	**28.50%**	**44.50%**	**42.50%**	**38.50%**	**28.50%**	**44.50%**	**42.50%**	**34.00%**	**30.00%**	**38.50%**	**32.00%**	**38.50%**	**40.50%**	**43.50%**	**41.00%**	**38.50%**	

accuracy of DT																
Average accuracy of DT	**61.73%**	**58.45%**	**67.27%**	**61.59%**	**61.95%**	**58.59%**	**65.82%**	**63.59%**	**45.95%**	**44.18%**	**48.45%**	**41.32%**	**50.86%**	**62.73%**	**66.91%**	**65.73%**	**61.73%**

*DT is the abbreviation of decision tree.

Table 6. Comparative statistics of 10 important features selected by attribute weighting models in different organisms

Important genomic feature	Organism	Count	Mean	StDev	Variance	CoefVar
salt 0.1M	Birds	3	87.6	0.987	0.975	1.13
	Domestic Mammals	18	83.356	1.611	2.594	1.93
	Felis catus	2	85.93	0	0	0
	Fish	5	86.558	1.751	3.065	2.02
	Mouse	11	84.716	0.935	0.875	1.1
	Primates	8	83.7	1.221	1.491	1.46
Frequency of Adenine	Birds	3	0.246	0.00346	0.00001	1.41
	Domestic Mammals	18	0.28028	0.01786	0.00032	6.37
	Felis catus	2	0.297	0	0	0
	Fish	5	0.2606	0.00695	0.00005	2.67
	Mouse	11	0.24645	0.00792	0.00006	3.21
	Primates	8	0.2726	0.0457	0.0021	16.75
Frequency of AA	Birds	3	0.04933	0.00751	0.00006	15.21
	Domestic Mammals	18	0.09061	0.01039	0.00011	11.46
	Felis catus	2	0.086	0	0	0
	Fish	5	0.0624	0.00631	0.00004	10.11
	Mouse	11	0.069	0.00557	0.00003	8.07
	Primates	8	0.086	0.0304	0.0009	35.36
Frequency of AC	Birds	3	0.06467	0.00404	0.00002	6.25
	Domestic Mammals	18	0.05628	0.00627	0.00004	11.14
	Felis catus	2	0.065	0	0	0
	Fish	5	0.0722	0.00841	0.00007	11.65
	Mouse	11	0.054091	0.00164	0.000003	3.03
	Primates	8	0.058	0.00782	0.00006	13.48
Frequency of CA	Birds	3	0.10233	0.00924	0.00009	9.03
	Domestic Mammals	18	0.078	0.01231	0.00015	15.78
	Felis catus	2	0.096	0	0	0
	Fish	5	0.0956	0.01078	0.00012	11.28
	Mouse	11	0.07355	0.00559	0.00003	7.6
	Primates	8	0.07288	0.0123	0.00015	16.88
Frequency of CG	Birds	3	0.02467	0.00289	0.00001	11.7
	Domestic Mammals	18	0.014778	0.003828	0.000015	25.9
	Felis catus	2	0.025	0	0	0
	Fish	5	0.0376	0.00627	0.00004	16.67
	Mouse	11	0.01627	0.00388	0.00002	23.81
	Primates	8	0.014	0.00659	0.00004	47.07
Frequency of CT	Birds	3	0.07667	0.00577	0.00003	7.53
	Domestic Mammals	18	0.07556	0.00511	0.00003	6.77
	Felis catus	2	0.066	0	0	0
	Fish	5	0.0634	0.00677	0.00005	10.67
	Mouse	11	0.09018	0.00412	0.00002	4.57
	Primates	8	0.08375	0.01565	0.00024	18.68
Frequency of GC	Birds	3	0.08033	0.00751	0.00006	9.34
	Domestic Mammals	18	0.04533	0.00683	0.00005	15.08

	Felis catus	2	0.063	0	0	0
	Fish	5	0.0704	0.00844	0.00007	11.99
	Mouse	11	0.05927	0.0066	0.00004	11.14
	Primates	8	0.04525	0.00518	0.00003	11.44
Frequency of GG	Birds	3	0.07967	0.01155	0.00013	14.49
	Domestic mammals	18	0.055	0.00941	0.00009	17.1
	Felis catus	2	0.049	0	0	0
	Fish	5	0.0614	0.01519	0.00023	24.74
	Mouse	11	0.060091	0.003145	0.00001	5.23
	Primates	8	0.0535	0.01707	0.00029	31.91
Frequency of TG	Birds	3	0.065	0.00346	0.00001	5.33
	Domestic mammals	18	0.068778	0.003606	0.000013	5.24
	Felis catus	2	0.061	0	0	0
	Fish	5	0.0676	0.00643	0.00004	9.51
	Mouse	11	0.076909	0.002386	0.000006	3.1
	Primates	8	0.07487	0.00631	0.00004	8.43

Figures

Figure 1. Pattern recognition in differentiation of *Nanog* genes between different organisms via *Decision Tree Random Forest* ran with *Gain Ratio* criterion on dataset pre-filtered *with Gini Index* attribute weighting models.

Figure 2. Pattern recognition in differentiation of *Nanog* genes between different organisms via *Decision Tree Random Forest* ran when ran with *Gini Index* criterion on dataset pre-filtered with *SVM* attribute weighting model.

Figure 3. Complete differentiation/prediction of *Nanog* genes between different organisms via *Decision Tree algorithm* ran on repeated sequence features with *Gini Index* criterion

Figure 4. Clustering of *Nanog* sequences based on the 10 fist important attributes of attribute weighting models.

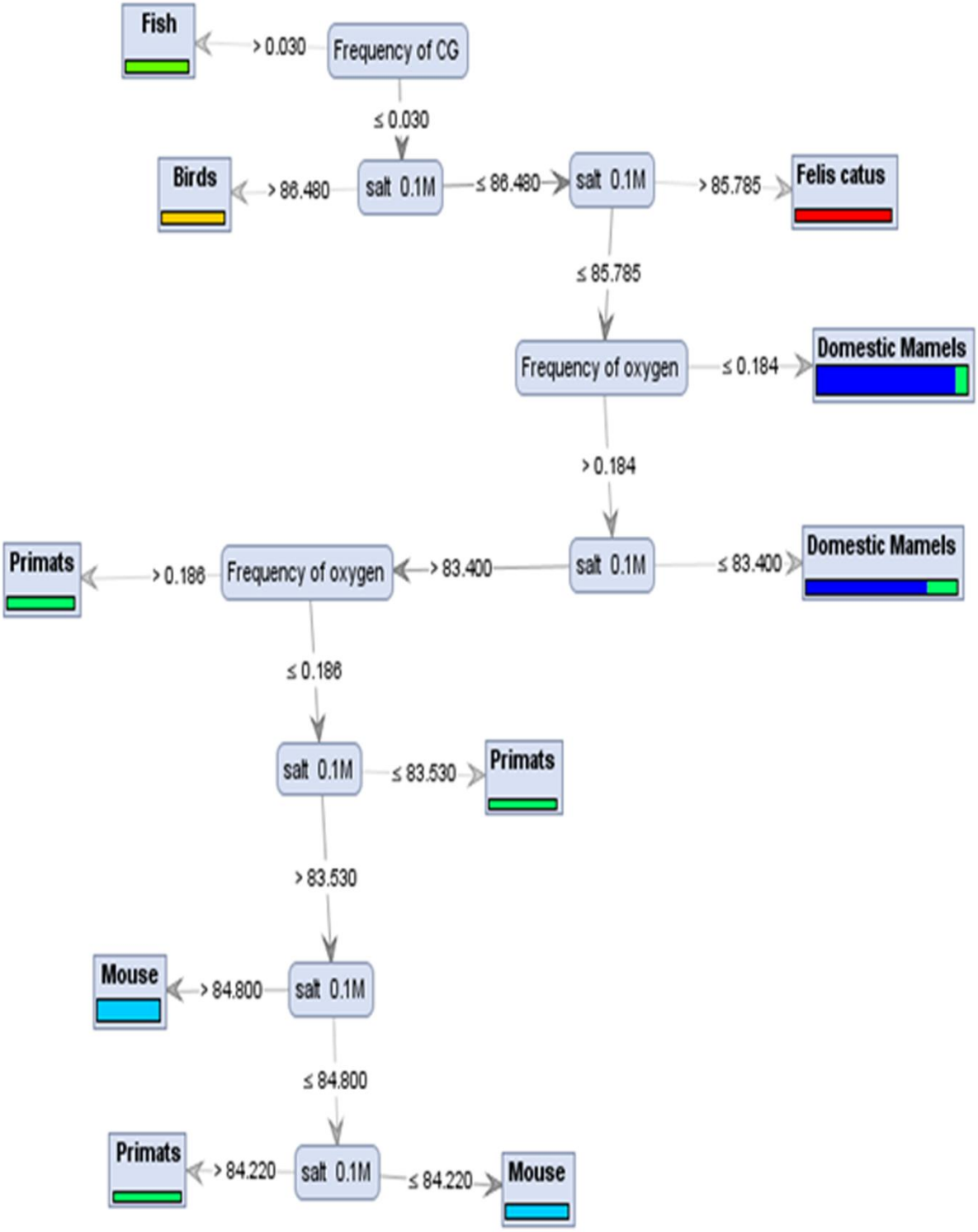

Figure 1

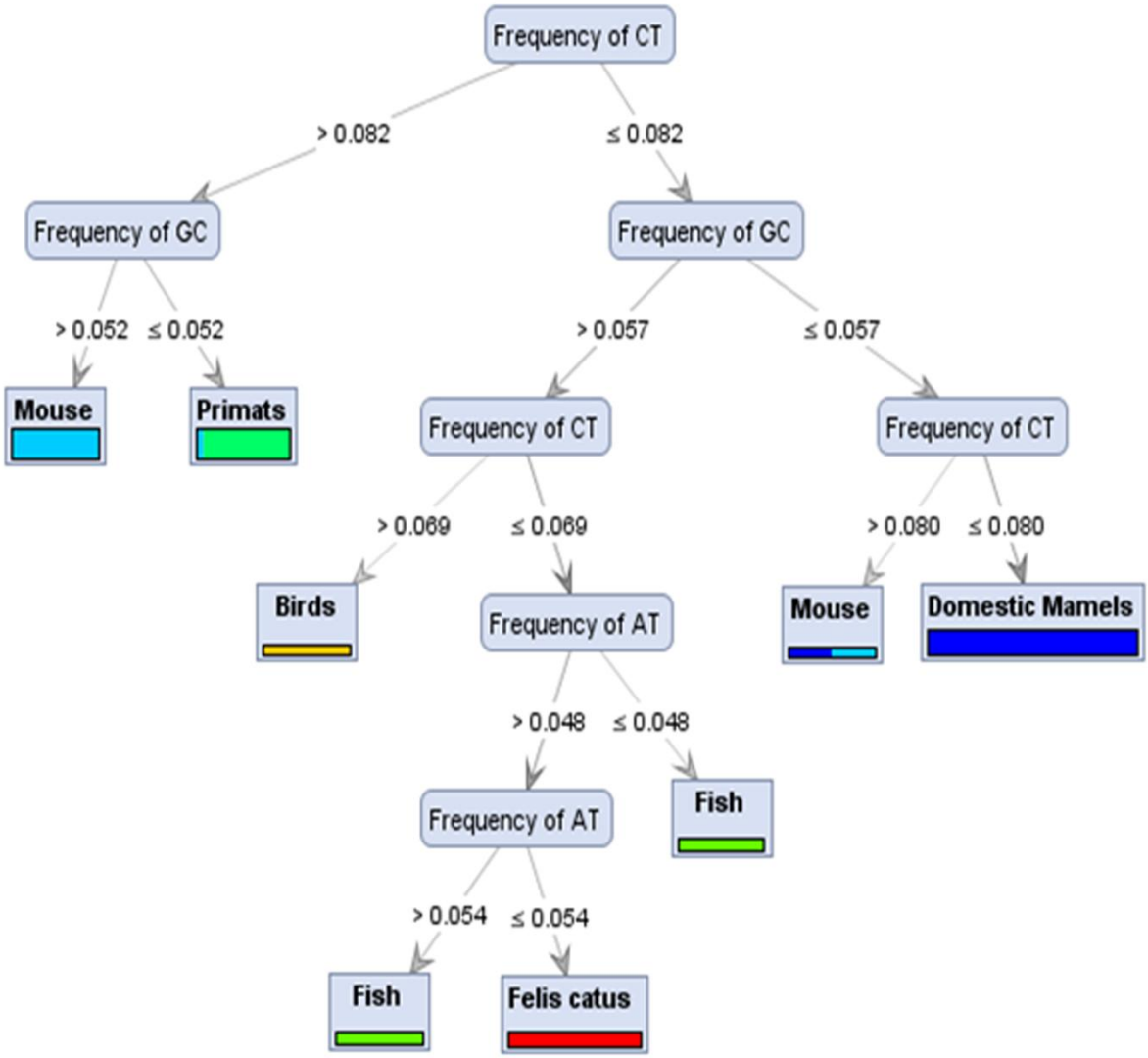

Figure 2

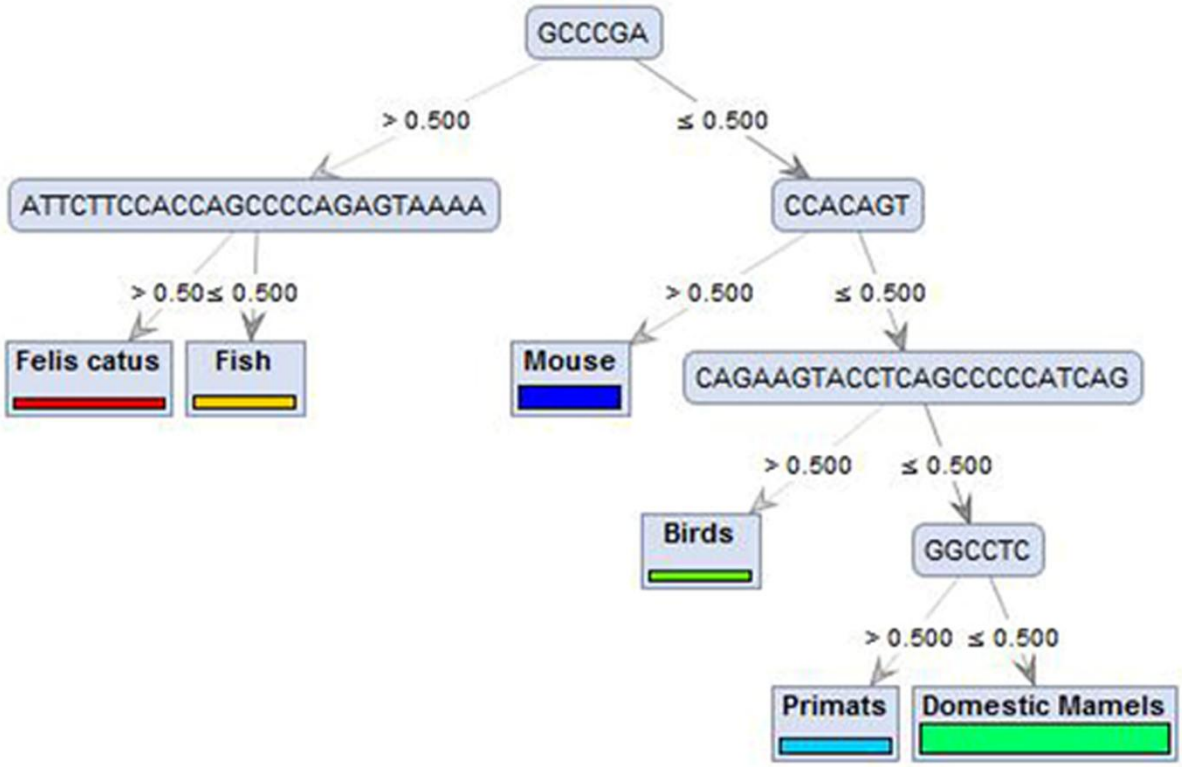

Figure 3

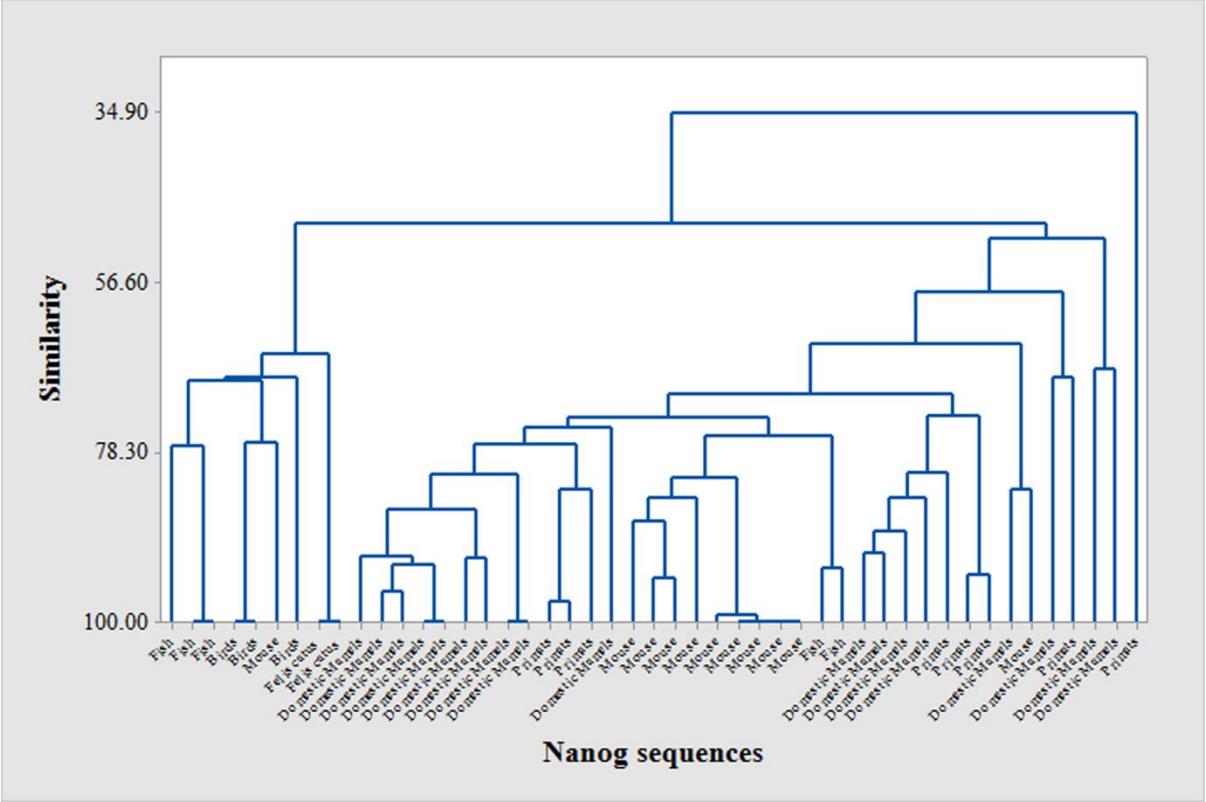

Figure 4

Abbreviation

DT	decision tree
ESC	embryonic stem cells
GAD	Gene Attribute Dataset
ICM	inner cell mass
PCA	Principle component analysis
RSD	Repeated Sequences Database (dataset)
SVM	support vector machine

Highlights

- Evolution study of key transcription factor in stem cell and tumor progression
- Pattern recognition of Nanog evolution by application of machine learning
- Finding the key genomic features governing Nanog evolution in different organisms
- Discovery of organism specific repeated sequences in Nanog gene sequences
- Documenting the high efficiency of Decision Tree Random Forest in Nanog evolution