Journal of
Proteomics & Bioinformatics

Wang and Xiao, J Proteomics Bioinform 2014, S9
DOI: 10.4172/0974-276X.S9-002

Research Article Open Access

NRPred-FS: A Feature Selection based Two-level Predictor for Nuclear Receptors

Pu Wang[1] and Xuan Xiao[1,2,3]*

[1]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China*
[2]*Information School, ZheJiang Textile & Fashion College, NingBo, 315211, China*
[3]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA*

Abstract

Motivation: Nuclear receptors (NRs) play a role in all developmental and physiological processes and are important drug targets in a wide variety of disease and healthy states. In the past years, to identify NRs and their subfamilies with high throughput and low-cost, many machine learning methods have been introduced. However, these predictors are all developed based on old dataset in the NucleaRDB, what's more, no feature selection technique is employed, so that the performances are very limited.

Result: In this study, a feature selection based two-level predictor, called NRPred-FS, is developed that can be used to identify a query protein as a nuclear receptor or not based on its sequence information alone, if it is, the prediction will be automatically continued to further identify it among the following eight subfamilies: (1) Thyroid hormone like (NR1), (2) HNF4-like (NR2), (3) Estrogen like, (4) Nerve growth factor IB-like (NR4), (5) Fushi tarazu-F1 like (NR5), (6) Germ cell nuclear factor like (NR6), (**7**) knirps like (NR0A), and (8) DAX like (NR0B). The nuclear receptor sequences are encoded as sequence-derived feature vectors formed by incorporating various physicochemical and statistical features. Furthermore, the features set are optimized by forward feature selection algorithm for reducing the feature dimensions and for getting higher classifying accuracy. As a demonstration, this method gone through rigorous testing on a benchmark datasets derived from the latest version of NucleaRDB and UniProt. The overall prediction accuracies of leave-one-out cross-validation were about 97% and 93% in the first and second level respectively. As a convenience to the users, the powerful predictor, NRPred-FS, is freely accessible at http://www.jci-bioinfo.cn/NRPred-FS. Hopefully it will be a useful vehicle for identifying NRs and their subfamilies.

Keywords: Nuclear receptor; Two-level predictor; Sequence-derived feature; Feature selection; Cross-validation

Introduction

Functioning as ligand-activated transcription factors, nuclear receptors have the ability to regulate gene expression by interacting with specific DNA sequences adjacent their target genes [1,2]. Because NRs can regulate a myriad of human developmental and physiological functions (reproduction, development, metabolism), they have been implicated in a wide range of diseases, such as cancer, diabetes, inflammatory diseases or osteoporosis [3,4].

The importance of NRs has prompted a rapid accumulation of the relevant data from a great diversity of fields of research. If searching in the comprehensive protein database UniProt (Release 2013_05) with the query words "nuclear hormone receptor family", you will obtain 7,752 results, from which you can access the information of protein attributes, comments, ontologies and so on. Specific databases about a single protein family can bring researchers great convenience in using all data needed for their research, while relieving them of the onerous tasks to retrieve many data from different sources [5]. As a professional database for NRs, NucleaRDB holds many different data types in a well-organized form, what's more, the data are validated, internally consistent and updated regularly [6,7].

These accumulated data are very helpful for data mining and knowledge discovery. There is a strong link between the function of a protein and the family or subfamily it belongs to, so it is very useful to develop bioinformatics tools for identifying NRs and their types rapidly and effectively. In recent years, researchers have made some studies and attempts for this problem.

Initial effort was made by Bhasin and Raghava [8] with amino acid and dipeptide compositions as input, a SVM-based model was developed for predicting four sub-families of NRs. Later, to identify a NR sequence among the same four sub-families as Bhasin and Raghava had worked upon, Gao et al. [9] reconstructed the dataset, and introduced the pseudo amino acid composition (PseAAC) [10] to represent the protein samples in hope to improve the prediction quality. However, the biggest weakness of the above predictors was that all the input sequences would be assumed to be NRs, obviously this might generate meaningless outcome. Recently two novel predictors were proposed, in which the prediction was carried through two steps. Firstly, the input protein sequences are screened, and secondly, if the input proteins are recognized as NRs, they will be further identified among seven sub-families [11,12].

All the aforementioned methods each have their own merits and did play a role in stimulating the development of this area, but they all have the following main shortcomings. (1) The datasets constructed to train the predictors were derived from the old version of NucleaRDB, which has been much updated recently. (2) Various feature extraction

***Corresponding author:** Xuan Xiao, Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China, Tel : +86-138-7980-9729; E-mail: jdzxiaoxuan@163.com

Received January 21, 2014; **Accepted** February 24, 2014; **Published** February 28, 2014

Citation: Wang P, Xiao X (2014) NRPred-FS: A Feature Selection based Two-level Predictor for Nuclear Receptors. J Proteomics Bioinform S9: 002. doi:10.4172/0974-276X. S9-002

methods were proposed yet no feature selection approach is employed, thus there may be bad features which would increase calculation and decrease the classification performance.

To solve the problem mentioned above, the present study proposed a new model. In order to extend the coverage scope of NR subfamilies and reduce the homology bias, new benchmark datasets were constructed based on the latest version of NucleaRDB (version 11.7.1), yet the old datasets used in Wang et al. and Xiao et al. [11,12] were derived from the very old version of NucleaRDB (version 5.0). The new datasets contains more recently updated NRs and cover eight sub-families, however the old dataset in Wang et al. and Xiao et al. [11,12] consisted of seven sub-families. The prediction was carried by the Support Vector Machine (SVM) classifier based on feature extraction by incorporating various physicochemical and statistical information derived from the protein sequences. What's more, the features extracted will be optimized by forward feature selection algorithm for improving performances. This crucial step was not performed in previous studies. As a result, the new proposed method only need fewer features to get better prediction performance than the methods in Wang et al. and Xiao et al. [11,12]. For more detail, see below.

Materials and Methods

Benchmark datasets

Nuclear receptor sequences were collected from the NucleaRDB (version 11.7.1) at http://www.receptors.org/nucleardb/, which is a molecular class-specific information system for NRs [7]. The database has collected and harvested all the eight subfamilies of nuclear receptors marked with (1) NR1: Thyroid hormone like (TR, RAR, ROR, PPAR, VDR), (2) NR2: HNF4-like (HNF4, RXR, TLL, COUP, USP), (3) NR3: Estrogen like (ER, ERR, GR, MR, PR, AR), (4) NR4: Nerve Growth factor IB-like (NGFIB, NURR), (5) NR5: Fushi tarazu-F1 like (SF1, FTF, FTZ-F1), (6) NR6: Germ cell nuclear factor like (GCNF1), (7) NR0A: Knirps like (KNI, KNRL, EGON, ODR7), and (8) NR0B: DAX like (DAX, SHP). The initial data set had 3016 sequences belonging to eight subfamilies of nuclear receptors. A redundancy cutoff was imposed with the program CD-HIT [13] to set the redundancy degree to 40% for NR1~NR5 and 80% for NR6, NR0A and NR0B, because the later contain too few sequences. If the 40% redundancy degree was also set on these classes, the samples left would be too few to have any statistical significance. The final dataset $\mathbf{S}^{\mathbf{NR}}$ contains 267 NR sequences belonging to eight different subfamilies as shown in Table 1. To identify query proteins between NRs and non-NRs, 1000 protein sequences that are not NRs were also collected in $\mathbf{S}^{\mathbf{nNR}}$ for training the 1st level predictor. The protein sequences in $\mathbf{S}^{\mathbf{nNR}}$ were randomly collected from the UniProt (Release 2013_05) at http://www.uniprot.org/ according their annotations in the "Keyword" field. The redundancy reduction was also operated in $\mathbf{S}^{\mathbf{nNR}}$, so that none of the proteins in $\mathbf{S}^{\mathbf{nNR}}$ has 40% pairwise sequence identity to any other. The accession numbers and sequences for the dataset $\mathbf{S}^{\mathbf{NR}}$ and $\mathbf{S}^{\mathbf{nNR}}$ are given in Supporting Information S1.

Family	Subfamily	Number of sequences
NR	NR1	82
	NR2	68
	NR3	33
	NR4	11
	NR5	15
	NR6	10
	NR0A	29
	NR0B	19
Non-NR	N/A	1000

Table 1: Breakdown of the benchmark dataset.

Sequence-derived features

A protein sequence **P** with L amino acid residues can be expressed as:

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\cdots\mathrm{R}_L \quad (1)$$

As pointed out in Chou [10], to develop a classifier for protein sequences, how to formulate the protein samples with an effective mathematical expression is the critical question. To answer this question, many features will be extracted from three different sources so as to capture as much useful information as possible.

Amino acid composition (AAC): As a simple and effective method, AAC was widely used for feature extraction (see, e.g., [14-16]). The AAC of a protein is defined as the normalized occurrence frequencies of 20 amino acids in that protein, i.e.,

$$\mathrm{AAC} = [f_1, f_2, \cdots, f_{20}]^{\mathbf{T}} \quad (2)$$

where $f_i = n_i / L$ with each i (= 1,2,···,20) corresponding to one of the 20 native amino acid types, and n_i is the number of type i amino acids in the protein, while **T** is the transpose operator.

Dipeptide composition (DC): One of the main drawbacks of amino acid composition is that it only emphasizes on overall sequence information but ignores the sequence order information. Dipeptide (amino acid pair) composition can make up for with capturing the local-order information of a protein sequence, which gives a fixed pattern length of 400 (20×20) [17], and can be generally formulated as

$$\mathrm{DC} = [d_1, d_2, \cdots, d_{400}]^{\mathrm{T}} \quad (3)$$

where d_i denotes the occurrence frequency of the i-th dipeptide as

$$d_i = \frac{\text{Total number of dip}(i)}{\text{Total number of all possible dipeptides}} \quad (4)$$

Where dip(i) (i = 1,2,···,400) is the i-th dipeptide.

Correlation factor (CF): Given a protein sequence **P**, suppose $H(R_1)$ is the certain physicochemical property value of the 1st residue R_1, $H(R_2)$ that of the 2nd residue R_2, and so forth. In terms of these property values the protein sequence can be converted to a digit signal $[H(R_1), H(R_2), \cdots H(R_L)]$, from which we can get correlation factors [18] as follow,

$$\theta_i = \frac{1}{L-i}\sum_{j=1}^{L-i} H(\mathrm{R}_j)\cdot H(\mathrm{R}_{j+i}) \qquad (i < L) \quad (5)$$

where θ_1 is the 1st-tier correlation factor, θ_2 the 2nd-tier correlation factor, and so forth. Here we only choose the first 10-tier correlation factors to be candidate features because the high-tier correlation factors made very little difference in prediction but increased much calculation in the feature selection procedure.

In this study, the following eight physical-chemical properties were taken into account: (1) Hydrophobicity index, (2) Hydropathy index, (3) pK-N, (4) pK-C, (5) Mean polarity, (6) Isoelectric point, (7) Molecular weight, (8) Normalized van der Waals volume. The values of these properties can be obtained by entry searching from **AAindex** (http://www.genome.jp/aaindex/), which is a database of numerical indices for various physicochemical and biochemical properties of

amino acids and pairs of amino acids. All data in this database [19] are derived from published literatures. Thus we will extract 8×10 = 80 correlation factors.

Feature normalization: Finally, we obtained a total of 500 feature elements, of which 20 are from AAC, 400 from DC, 80 from CF. We can easily find that different features have different scales, so every feature will be normalized as follow

$$y = (y_{\max} - y_{\min}) * \frac{x - x_{\min}}{x_{\max} - x_{\min}} + y_{\min} \quad (6)$$

where x is the original feature value, while y is the normalized feature value, $x_{\max}$ and $x_{\min}$ is the maximum and minimum value of the original feature respectively, $y_{\max} = 1$ and $y_{\min} = -1$, thus every feature value will be normalized in the range of -1 to 1.

Thus, a protein sample can be formulated as a normalized 500-D vector given by

$$= [\psi_1, \psi_2, \cdots, \psi_{500}] \quad (7)$$

Support vector machine

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns. The original SVM algorithm was invented by Vapnik [20] and the current standard incarnation (soft margin) was proposed by Cortes and Vapnik [21]. An SVM model is a representation of the examples as points in space, mapped by a kernel function so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Different kernel functions define different SVMs. In principle, SVM is a two-class classifier, but it can directly cope with multi-class classification problem through the one-against-all or pairwise method.

SVM has been widely used for predicting protein attributes (see, e.g. [8,22-26]). In this study, the LIBSVM package [27] was used as an implementation of SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/, the popular radial basis function (RBF) was taken as the kernel function, and there were two unknown parameter: penalty parameter C and kernel parameter γ. The values of the two parameters are closely related to the quality of the model, how to determine them will be discussed later.

Cross-validation and performance measures

In statistical prediction, cross-validation [28] is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. In K-fold cross-validation, the original sample is randomly partitioned into K equal size subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining K-1 subsamples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. Leave-one-out cross-validation (LOOCV) involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sampling. As elucidated and demonstrated by Eqs.28-32 of Chou [29], the LOOCV test has the least arbitrary that can always yield a unique result for a given benchmark dataset, and hence has been increasingly and widely used by investigators to examine the accuracy of various predictors (see, e.g. [30-36]). Accordingly, the LOOCV test was also adopted here to examine the quality of the present predictor.

For performance measures we used accuracy (ACC) and Matthew's correlation coefficient (MCC). Accuracy measures the proportion of correct predictions. MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC can be calculated using the formula:

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}} \quad (8)$$

where TP represents the true positive, TN, the true negative, FP, the false positive, and FN, the false negative. Specifically for a hypothetical class X, all the other classes are marked $\overline{X}$, then TP is the number of correctly predicted sequences that belong to X, TN is the number of correctly predicted sequences that belong to $\overline{X}$, FP is the number of sequences wrongly predicted to belong to X while FN is the number of sequences wrongly predicted to belong to $\overline{X}$. Eq.8 returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

Feature selection

As we know, not all the extracted features would contribute to the classification, so the feature selection procedure is always indispensable in a classification problem. In essence, feature selection is a combinatorial optimization problem. Its goal is to seek the feature subset that maximizes the performance of the predictor. To find the optimal feature subset from the original feature set, all the combination of features should be tried from the point of view of the exhaustion principle, which is of computational intractability. Hence we usually rely on some heuristics to overcome the complexity of exhaustive search. Sequential Forward Selection (SFS for short) proposed by Whitney [37] is one of the commonly used heuristic methods for feature selection. It involves the following steps:

(1) Use one classifier (in this case, SVM), and the cross-validation test for prediction accuracy estimate.

(2) Select the first feature that has the highest accuracy among all features.

(3) Select the feature, among all unselected features, together with the selected features that gives the highest accuracy.

(4) Repeat the previous process until you have selected enough number of features, or until the accuracy is good enough.

Results and Discussion

Before the procedure of features selection, the two uncertain parameters (C and γ) in SVM must be determined firstly. However, it would need a lot of computational time to find their optimal values by LOOCV. Therefore, as a first step, the values of the two parameters were determined by pursuing the highest prediction accuracy of 10-fold cross-validation through 2-D grid search as shown in Figure 1. The values thus obtained for the two parameters were given by

$$\begin{cases} C = 2^{11}, & \gamma = 2^{-14} \quad \text{for the } 1^{\text{st}}\text{-level prediction} \\ C = 2^{1}, & \gamma = 2^{-5} \quad \text{for the } 2^{\text{nd}}\text{-level prediction} \end{cases} \quad (9)$$

where the 1st-level prediction was for identifying a query protein as NR or non-NR, while the 2nd-level prediction was for identifying a NR among its eight subfamilies.

Next, the Sequential Forward Selection algorithm was implemented on the original features set. Again, to reduce time consuming of computation, prediction accuracy by 10-fold cross-validation was taken as the measure of the feature subset. For each feature newly selected, there would be overall accuracy accordingly. With the number of features as x-axis and overall accuracy as the y-axis, SFS curve was plotted to reveal the relation between the performance of the predictor and the feature subset. From the SFS curve in Figure 2 we can see that, in the 1st level prediction, the SFS curve peaks at 0.9795 when the feature set is comprised of the first 421 features, while in the 2nd level prediction, the SFS curve peaks at 0.9663 when the feature set is comprised of the first 455 features. The optimal feature subset is considered with the highest prediction accuracy, and the predictor thus obtained was used to identify the NRs and their subfamilies. From this figure we can also find that, in the 1st-level prediction, if input all the original 500 features, the prediction accuracy is about 97%, however if input the first 152 optimal features, the results are equivalent, and if only input the first 17 optimal features, more than 90% accuracy could be obtained. Similarly in the 2nd-level prediction, 92% accuracy is obtained with all the 500 original features as input, while only input the first 91 optimal features, the same result can be achieved. So feature selection is very useful here, and the contribution of most of the features is quite limited. The sequentially selected features by Sequential Forward Selection algorithm in the 1st and 2nd level prediction are given in Supporting Information S2, from which we can find that, for the 1st level, the previous features selected are mainly AAC, it seems that the nuclear receptors are clearly different from the other proteins in amino acid composition, for the 2nd level, DC and CF are selected firstly and contribute more to the classification, it shows that NR sequences in different NR subfamilies are similar in amino acid composition, and it need more sequence order and physical-chemical information to distinguish.

Finally, using the parameters values of Eq.9 for the SVM operation engine and optimal feature subset in the features selection procedure, the LOOCV was performed on the benchmark dataset. The results thus obtained in identifying proteins as NRs or non-NRs are given in Table 2, while those in identifying NRs among their eight subfamilies are given in Table 3.

To verify the effectiveness of the proposed model, we also compare the prediction results among **NRPred-FS**, **NR-2L** [11], and **iNR-PhysChem** [12], the latter two are the latest predictors for nuclear receptors. However, because the datasets for training and testing are not the same and the latter two can only identify NR among seven subfamilies in the 2nd level prediction. In the cause of fairness, **NRPred-FS** was also tested on the old dataset as used in **NR-2L** and **iNR-PhysChem**. The old dataset was also derived from the NucleaRDB (version 5.0) and UniProt (Release 2010_10). After redundancy cutoff, there were 500 non-NRs and 159 NRs which were classified into seven subfamilies. For more detailed information about the old dataset [11]. It needs to be stressed that, for the old dataset the calculation and optimization are carried out again as described above. All the results are listed in Tables 4 and 5 for the 1st and 2nd level prediction respectively, from which we can see that, for the same dataset the prediction quality of **NRPred-FS** is improved a lot. What's more, **NRPred-FS** need only 435 features in the 1st level and 340 features in the 2nd level to get the

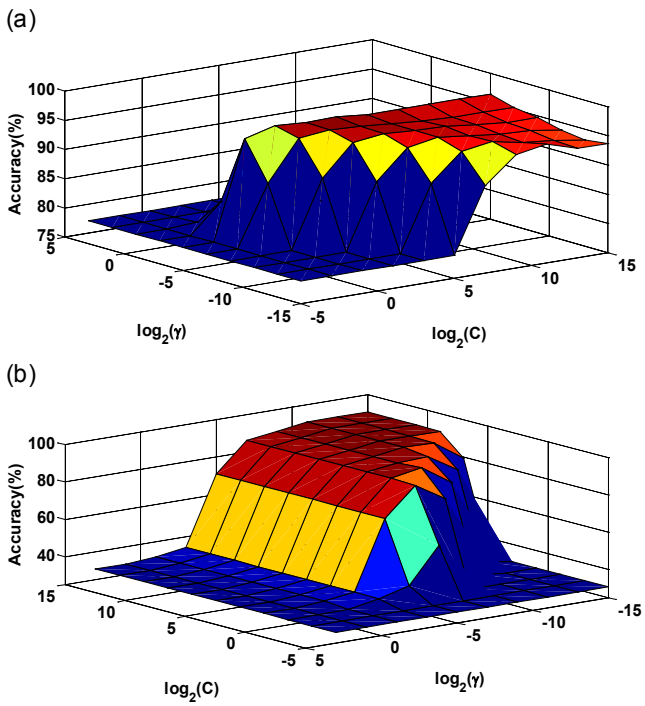

Figure 1: The 3D graph to show the prediction accuracies by the 10-fold cross-validation with different values of *C* and *γ* in the SVM engine. (a) The results obtained for the 1st-level prediction. (b) The results obtained for the 2nd-level prediction.

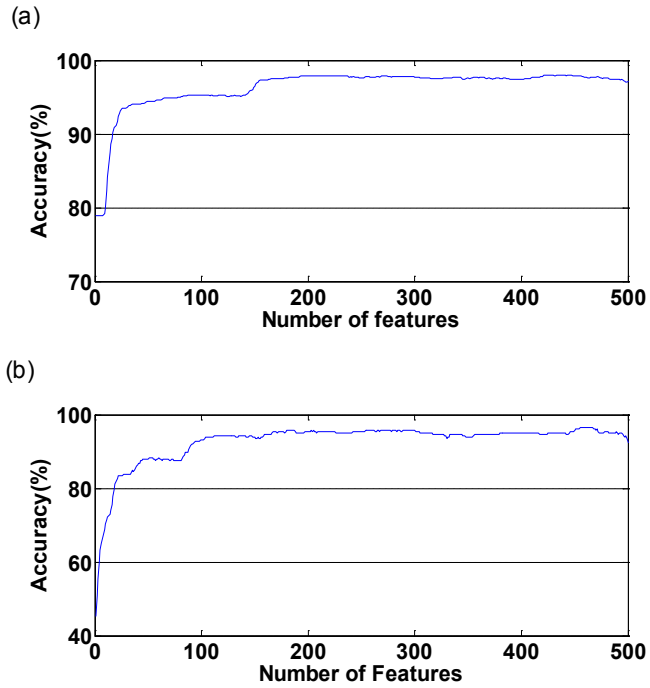

Figure 2: SFS curve reveal the relation between the performance of the predictor and the feature subset. (a) In the 1st level prediction, the SFS curve arrives at the apogee when the feature set is comprised of the first 421 features. (b) In the 2nd level prediction, the SFS curve arrives at the apogee when the feature set is comprised of the first 455 features.

Family	ACC	MCC
NR	$\frac{247}{267}=92.51\%$	0.91
Non-NR	$\frac{982}{1000}=98.20\%$	0.91
Overall	$\frac{1229}{1267}=97\%$	0.94

Table 2: ACC and MCC in the 1st-level prediction by LOOCV.

Subfamily	ACC	MCC
NR1	$\frac{79}{82}=96.34\%$	0.90
NR2	$\frac{63}{68}=92.65\%$	0.91
NR3	$\frac{28}{33}=84.85\%$	0.84
NR4	$\frac{10}{11}=90.91\%$	0.95
NR5	$\frac{12}{15}=80\%$	0.89
NR6	$\frac{8}{10}=80\%$	0.89
NR0A	$\frac{29}{29}=100\%$	0.96
NR0B	$\frac{19}{19}=100\%$	1
Overall	$\frac{248}{267}=92.88\%$	0.92

Table 3: ACC and MCC in the 2nd-level prediction by LOOCV.

Family	NRPred-FS		iNR-PhysChem		NR-2L	
	ACC	MCC	ACC	MCC	ACC	MCC
NR	$\frac{153}{159}=96.23\%$	0.97	$\frac{153}{159}=96.23\%$	0.95	$\frac{156}{159}=98.11\%$	0.83
Non-NR	$\frac{498}{500}=99.60\%$	0.97	$\frac{494}{500}=98.80\%$	0.95	$\frac{454}{500}=90.80\%$	0.83
Overall	$\frac{651}{659}=98.79\%$	0.98	$\frac{647}{659}=98.18\%$	0.96	$\frac{610}{659}=92.56\%$	0.85

Table 4: Comparison of the prediction results among **NRPred-FS, iNR-PhysChem** and **NR-2L** in identifying NRs and non-NRs by the LOOCV on the old dataset.

best results, while **iNR-PhysChem** used 1000 features and 500 features are adopt in **NR-2L**.

Conclusion

In this study, the feature selection method with various physicochemical and statistical features is implemented for improving prediction performance. The prediction accuracy on the newly constructed benchmark dataset is 97% and 93% in the 1st-level and 2nd-level classifier. It is anticipated that **NRPred-FS** may become a useful tool for identifying NRs and their functional types. Feature selection procedure is very important and necessary for protein attribute prediction based on machine learning, because some features contribute very little to, even interfere with the decision-making, especially when you don't know the intrinsic correlation between the features extracted and the attribute to be predicted. The feature subset optimizing is usually very time-consuming, however, it will make the final predictor faster and more efficient, this is worthwhile.

Subfamily	NRPred-FS		iNR-PhysChem		NR-2L	
	ACC	MCC	ACC	MCC	ACC	MCC
NR1	$\frac{50}{50}=100\%$	0.99	$\frac{47}{50}=94.00\%$	0.87	$\frac{43}{50}=86.00\%$	0.88
NR2	$\frac{36}{36}=100\%$	0.95	$\frac{35}{36}=97.22\%$	0.93	$\frac{31}{36}=86.11\%$	0.85
NR3	$\frac{37}{37}=100\%$	1.00	$\frac{37}{37}=100\%$	0.95	$\frac{37}{37}=100\%$	0.86
NR4	$\frac{7}{7}=100\%$	1.00	$\frac{5}{7}=71.43\%$	0.84	$\frac{6}{7}=85.71\%$	0.70
NR5	$\frac{10}{12}=83.33\%$	0.91	$\frac{10}{12}=83.33\%$	0.91	$\frac{10}{12}=83.33\%$	0.86
NR6	$\frac{5}{5}=100\%$	1.00	$\frac{5}{5}=100\%$	1.00	$\frac{5}{5}=100\%$	1.00
NR0	$\frac{10}{12}=83.33\%$	0.91	$\frac{8}{12}=66.67\%$	0.81	$\frac{9}{12}=75.00\%$	0.86
Overall	$\frac{155}{159}=97.48\%$	0.97	$\frac{147}{159}=92.45\%$	0.91	$\frac{141}{159}=88.68\%$	0.87

Table 5: Comparison of the prediction results among **NRPred-FS, iNR-PhysChem** and **NR-2L** in identifying the subfamilies of NRs by the LOOCV on the old dataset.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (No.60961003, No.6121027 and No.31260273), the Key Project of Chinese Ministry of Education (No. 210116), the Province National Natural Science Foundation of JiangXi (No.2010GZS0122, No.20114BAB211013 and No. 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008).

References

1. Robinson-Rechavi M1, Escriva Garcia H, Laudet V (2003) The nuclear receptor superfamily. J Cell Sci 116: 585-586.
2. Altucci L, Gronemeyer H (2001) Nuclear receptors in cell life and death. Trends Endocrinol Metab 12: 460-468.
3. Moore JT, Collins JL, Pearce KH (2006) The nuclear receptor superfamily and drug discovery. ChemMedChem 1: 504-523.
4. Huang P, Chandra V, Rastinejad F (2010) Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. Annu Rev Physiol 72: 247-272.
5. Folkertsma S, van Noort P, Van Durme J, Joosten HJ, Bettler E, et al. (2004) A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. J Mol Biol 341: 321-335.
6. Horn F, Vriend G, Cohen FE (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res 29: 346-349.
7. Vroling B, Thorne D, McDermott P, Joosten HJ, Attwood TK, et al. (2012) NucleaRDB: information system for nuclear receptors. Nucleic Acids Res 40: D377-380.
8. Bhasin M, Raghava GP (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279: 23262-23266.
9. Gao QB, Jin ZC, Ye XF, Wu C, He J (2009) Prediction of nuclear receptors with optimal pseudo amino acid composition. Anal Biochem 387: 54-59.
10. Chou KC (2009) Pseudo Amino Acid Composition and its Applications in Bioinformatics Proteomics and System Biology. Curr Proteomics 6: 262-274.

11. Wang P, Xiao X, Chou KC (2011) NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS One 6: e23505.
12. Xiao X, Wang P, Chou KC (2012) iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. PLoS One 7: e30869.
13. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
14. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99: 153-162.
15. Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729-738.
16. Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12: 107-118.
17. Reczko M, Bohr H (1994) The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res 22: 3616-3619.
18. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43: 246-255.
19. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36: D202-205.
20. Vapnik VN (1995) The Nature of statistical learning theory springer-verlag.
21. Cortes C, Vapnik V (1995) Support vector networks. Machine Learning 20: 273-293.
22. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765-45769.
23. Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257-3263.
24. Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14: 811-815.
25. Li D, Jiang Z, Yu W, Du L (2010) Predicting caspase substrate cleavage sites based on a hybrid SVM-PSSM method. Protein Pept Lett 17: 1566-1571.
26. Li YX, Shao YH, Deng NY (2011) Improved prediction of palmitoylation sites using PWMs and SVM. Protein Pept Lett 18: 186-193.
27. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines Software.
28. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2: 1137-1143.
29. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273: 236-247.
30. Liu T, Jia C (2010) A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J Theor Biol 267: 272-275.
31. Masso M, Vaisman II (2010) Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. J Theor Biol 266: 560-568.
32. Wang T, Xia T, Hu XM (2010) Geometry preserving projections algorithm for predicting membrane protein types. J Theor Biol 262: 208-213.
33. Joshi RR, Sekharan S (2010) Characteristic peptides of protein secondary structural motifs. Protein Pept Lett 17: 1198-1206.
34. Kandaswamy KK, Pugalenthi G, Möller S, Hartmann E, Kalies KU, et al. (2010) Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept Lett 17: 1473-1479.
35. Liu T, Zheng X, Wang C, Wang J (2010) Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept Lett 17: 1263-1269.
36. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17: 1207-1214.
37. Whitney AW (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput 20: 1100-1103.

This article was originally published in a special issue, **Computational Intelligence in Bioinformatics** handled by Editor(s). Dr. Jean-Christophe Nebel, Kingston University, London, UK