NRPreTo: A Machine Learning-Based Nuclear Receptor and Subfamily Prediction Tool

Sita Sirisha Madugula, Suman Pandey, Shreya Amalapurapu, and Serdar Bozdag*

ABSTRACT: The nuclear receptor (NR) superfamily includes phylogenetically related ligand-activated proteins, which play a key role in various cellular activities. NR proteins are subdivided into seven subfamilies based on their function, mechanism, and nature of the interacting ligand. Developing robust tools to identify NR could give insights into their functional relationships and involvement in disease pathways. Existing NR prediction tools only use a few types of sequence-based features and are tested on relatively similar independent datasets; thus, they may suffer from overfitting when extended to new genera of sequences. To address this problem, we developed Nuclear Receptor Prediction Tool (NRPreTo), a two-level NR prediction tool with a unique training approach where in addition to the sequence-based features used by existing NR prediction tools, six additional feature groups depicting various physiochemical, structural, and evolutionary features of proteins were utilized. The first level of NRPreTo allows for the successful prediction of a query protein as NR or non-NR and further subclasses the protein into one of the seven NR subfamilies in the second level. We developed Random Forest classifiers to test on benchmark datasets, as well as the entire human protein datasets from RefSeq and Human Protein Reference Database (HPRD). We observed that using additional feature groups improved the performance. We also observed that NRPreTo achieved high performance on the external datasets and predicted 59 novel NRs in the human proteome. The source code of NRPreTo is publicly available at https://github.com/bozdaglab/NRPreTo.

INTRODUCTION

The nuclear receptor (NR) superfamily are transcription factors that are involved in critical signaling pathways controlling physiological growth, differentiation, and cell maintenance. They respond to steroid-based signaling molecules like thyroid hormone, vitamin D3, and retinoids. Due to their involvement in critical cell regulating activities, NRs are also instrumental in the disease pathways of several physiological and reproductive diseases such as breast cancer, obesity, and diabetes. Structurally, all members of the NR superfamily have a common three-domain architecture. The N-terminal AF-1 region also known as the A/B region is a ligand-independent highly variable transcription activation region. They also contain two main domains, a moderately conserved ligand binding domain (LBD) involved in ligand recognition and a highly conserved DNA binding domain (DBD) involved in DNA binding at the C-terminal region. The DBD contains two conserved zinc finger motifs involved in hormone response element (HRE) recognition. Binding of the ligand at the LBD induces a conformational change in the DBD, leading to the activation of a specific HRE and subsequent downstream gene expression. Hence, both the conserved domains play important roles in NR-led gene regulation. The presence of both ligand and DNA binding domains in a protein sequence is therefore considered a signature hallmark of the NR superfamily. Since ligand binding is highly specific in the NR superfamily, these proteins are divided into seven subfamilies based on the type of ligand binding at the LBD. While NR1 to NR6 subfamilies are grouped clearly depending on their activating ligands, the NR0 subfamily are characterized by “orphan receptors” with no known ligands and may sometimes lack DBD. A detailed description of the subfamily names and their members can be found in Table S1.

NR is one of the most important superfamilies of druggable proteins. Some proteins of this superfamily are potential targets for developing therapeutic strategies for genetic diseases. Furthermore, NR proteins are regulated by small-molecule ligands, making them amicable for drug designing strategies. These properties make NR promising pharmacological targets. Moreover, with the enormous number of protein sequences emerging from next-generation sequencing efforts, there is a growing need to develop robust and accurate methods to...
identify NR and determine the subclass of the new incoming protein sequences.

In this regard, most existing tools for NR classification and subfamily prediction use only spectrum-kernel features, which are empirical in nature. One of the earliest NR prediction tools NRpred is a Support Vector Machine (SVM)-based tool that used only amino acid composition (AAC) and dipeptide composition (DPC) features to classify proteins into four NR subfamilies. Gao et al. then improved NRpred by using pseudo amino acid composition (PAAC) features and achieved higher accuracy. However, both tools directly predict the NR subfamily before predicting if a given protein is NR or not. Later, NR-2L and iNR-PhysChem took a two-level prediction approach to address this problem and broadened the prediction scope to seven classes. At the first level, the tools determine if an incoming protein is an NR, and at the second level, subfamily prediction is performed to classify the protein into one of the seven subclasses. This approach improved the overall accuracy but suffered from limitations like data redundancy and noisy features during model development. NRPreTo addressed this issue through feature selection and achieved improved accuracy. Later, NRfamPred used the same two-level approach and utilized AAC and DPC features for level 1 and level 2 predictions, respectively. NRfamPred was verified on an independent dataset as well. Additionally, they also carried out human proteome-wide NR subfamily prediction and identified 76 NR using the HPRD dataset. However, NRfamPred did not apply any feature selection and used only the spectrum-kernel-like features for model building, leading to lower overall performance. Recently, a Random Forest (RF)-based NR prediction tool RF-NR was developed using AAC, DPC, and tripeptide composition (TPC) features and achieved much higher performance. However, this tool has been tested only on independent datasets, which are relatively smaller; thus, RF-NR may not be generalized enough to make proteome-scale NR predictions. Besides, RF-NR is not a publicly available tool.

In the current study, we hypothesized that since NRs differ from each other at allosteric sites, merely using spectrum-like features to determine their subfamily cannot capture the structural differences among the proteins of different NR subfamilies. There is a need to consider additional features that account for evolutionary, co-evolutionary, and physiochemical properties of proteins. Existing prediction tools use only primary sequence-based features to classify NR subfamilies. Here, we investigated the role of various kinds of protein features that can govern NR subfamily classification. For this purpose, we developed Nuclear Receptor Prediction Tool (NRPreTo), which was trained on an extensive set of features to perform two-level NR prediction. Considering seven different feature groups including the three spectrum-kernel features (AAC, DPC, and TPC) used by previous NR prediction tools, we computed 13,494 features. After performing feature selection, we trained RF models to predict whether an incoming protein is a NR or not and further predict its subfamily if found to be NR. We applied NRPreTo on two different data setups: (a) building a model using “train” dataset of a previously published benchmark dataset (BD) and testing on its independent dataset and (b) combining the training and independent datasets of the respective BD to generate a “combined” dataset used for training, followed by testing on the human proteome collected from two external datasets: HPRD and NCBI RefSeq (humans). A schematic representation of the different components of NRPreTo is shown in Figure 1.

Figure 1. Overview of the NRPreTo workflow. Experiments 1 and 2 refer to testing the models on the independent datasets and the human proteome dataset, respectively. Level 1 is the prediction task to classify NR proteins, and level 2 is the prediction task to classify the subfamily of an NR protein.

Our proposed methodology has outperformed or been on par with all previous studies. Additionally, we also predicted 59 new NRs from the human proteome. Some of the identified proteins include pharmacologically important drug targets, which can be promising candidates for further studies. To the best of our knowledge, this is the first NR classification study to build classifiers using the entire spectrum of protein features in addition to the regularly used sequence-based features. Moreover, this is the first study to test its models not only on two independent datasets but also carry out a proteome scale prediction across two important public datasets: HPRD and RefSeq covering all sequences of the human proteome known so far. These results indicate that NR classification is influenced by other protein features in addition to sequence-based features.

## RESULTS

In the current study, we developed a two-level NR prediction tool called NRPreTo, built on a multitude of protein features (Table 1) including sequence-based features used by the previous studies. Using two experimental setups with different-sized training datasets (see Materials and Methods), we performed extensive evaluations of NRPreTo on two previously published BDs (i.e., BD1 and BD2) and the entire human proteome. We successfully identified 59 new NRs from the human proteome while improving the performance of the model.

**Level 1 Predictions.** We carried out level 1 prediction to screen-in query proteins that are NRs. Only the sequences
Table 1. List of Descriptors Used in the Study

| Descriptor group                      | Descriptor set                                           | Descriptors |
|--------------------------------------|----------------------------------------------------------|-------------|
| Amino acid composition               | Amino acid composition (AAC)                             | 20          |
|                                      | Composition of k-spaced amino acid pairs (CKSAAP)        | 2400        |
|                                      | Dipeptide composition (DPC)                              | 400         |
|                                      | Dipeptide deviation from expected mean (DDE)             | 400         |
|                                      | Tripeptide composition (TPC)                             | 8000        |
| Grouped amino acid composition       | Grouped amino acid composition (GAAC)                    | 5           |
|                                      | Composition of k-spaced amino acid group pairs (CKSAAGP) | 150         |
|                                      | Grouped dipeptide composition (GDPC)                     | 25          |
|                                      | Grouped tripeptide composition (GTPC)                    | 125         |
| Autocorrelation                      | Moran (Moran)                                           | 240         |
|                                      | Geary (Geary)                                           | 240         |
|                                      | Normalized Moreau-Broto (NMBroto)                       | 240         |
| C/T/D                                | Composition (CTDC)                                      | 39          |
|                                      | Transition (CTDT)                                       | 39          |
|                                      | Distribution (CTDD)                                     | 195         |
| Conjoint triad                       | Conjoint triad (CTriad)                                 | 343         |
|                                      | Conjoint k-spaced triad (KSCTriad)                      | 343         |
| Quasi-sequence-order                 | Sequence-order-coupling number (SOCNumber)              | 60          |
|                                      | Quasi-sequence-order descriptors (QSOrder)              | 100         |
| Pseudo-amino-acid composition        | Pseudo-amino-acid composition (PAAC)                    | 50          |
|                                      | Amphiphilic PAAC (APAAC)                                | 80          |

Table adopted from the iFeature package.

Figure 2. Feature set-wise distribution of the important descriptors of the level 1 prediction model. (A) Experiment-1 (train-only data) and (B) experiment-2 (combined dataset).
Table 2. Performance Comparison of NRPreTo Level 1 “Train-Only” Models with the Existing Tools on the Two Independent Datasets

| Studies        | BD1-Independent dataset | BD2-Independent dataset |
|----------------|-------------------------|-------------------------|
| NR-2L          | -1 0.980                | -1 0.96                 |
| NRfamPred       | -1 0.992                | -1 0.990                |
| RF-NR           | -1                      | -1                      |
| NRPreTo (best model) | 0.970 | 0.989         |
| NRPreTo (15 runs) | 0.967 ± 0.002 | 0.984 ± 0.002 |
|                | 0.960 ± 0.002 | 0.984 ± 0.002 |
|                | 0.924 ± 0.003 | 0.970 ± 0.004 |
|                | 0.997 ± 0.001 | 1.000 ± 0.000 |

For each cell, the score on BD1 is followed by the score on BD2 with a “|” symbol in between. “-” implies that the performance of the tool is not reported on the corresponding BD. Boldfaced values indicate the best scores for each metric. Acc: Accuracy. MCC: Matthew's correlation coefficient. AUC: Area under the curve.

Table 3. Summary of Level 1 Prediction Performance of Combined Models on the HPRD and RefSeq Datasets over 15 Runs

| External datasets | BD1 Combined | BD2 Combined |
|-------------------|--------------|--------------|
| HPRD              | 0.999 ± 0.001 | 0.999 ± 0.001 |
|                   | 0.986 ± 0.007 | 0.976 ± 0.008 |
|                   | 0.972 ± 0.014 | 0.952 ± 0.016 |
|                   | 0.999 ± 0.001 | 0.999 ± 0.001 |
| RefSeq            | 0.999 ± 0.001 | 0.999 ± 0.001 |
|                   | 0.985 ± 0.004 | 0.974 ± 0.006 |
|                   | 0.970 ± 0.009 | 0.947 ± 0.012 |
|                   | 0.999 ± 0.001 | 0.999 ± 0.001 |

For each cell, the score obtained by the model trained on BD1 is followed by the score obtained by the model trained on BD2 with a “|” symbol in between. Acc: accuracy. MCC: Matthew's correlation coefficient. AUC: area under the curve.

determined as NR using the level 1 models were taken ahead to the second level for subfamily classification. Since this level is a binary classification problem, both NR and non-NR sequences were used to train the models. We used the BorutaPy feature selection method to select separate important feature sets for level 1 models for the two experimental setups. Figure 2 shows a feature-set-wise summary of selected features of level 1 model for experiment-1 (“train” data) and experiment-2 (“combined” dataset). For experiment-1, 557 and 385 important features were selected for the two BDs, respectively. In both BDs, the maximum number of important descriptors were from the amino acid composition group, followed by C/T/D (Figure 2). The amino acid composition group in this study includes composition of k-spaced amino acid pairs (CKSAAP), TPC, and DDE descriptor sets, which constitute the maximum count of their group (see Table 1). These are sequence-based features, which are elemental to any protein and are widely employed in different protein identification and classification tasks. Additionally, other descriptor sets like PAAC (Pseudo amino acid composition), CKSAAP (Amino acid composition), and Moran (Autocorrelation) were also identified among the important features, which are also known to play a significant role in the classification of proteins, including NR. The PAAC, which represents the positional and compositional pattern of an amino acid in protein sequences, preserves the evolutionary patterns within the protein families. Therefore, it is widely used in problems like predicting various post-translational modification sites or identifying protein subcellular localization where evolutionary relationships of residues are important.

**Experiment-1 — Train Only.** Experimental setup-1 consists of models trained only on the “training” dataset of the two BDs and tested on their respective independent datasets. The BD1 training dataset contained 267 NR and 1000 non-NR sequences, while BD2 had 159 NR and 500 non-NR (Table S1). Due to the imbalance in the dataset, class weights were used to give equal importance to the two classes and negate the influence of the majority class during prediction. A comparison of the average performance of train-only models on independent datasets over 15 runs across the two BDs is shown in Table 2.

Further, we compared NRPreTo with four existing NR prediction tools, namely, NR-2L, NRfamPred, NRPred-FS, and RF-NR, to assess its performance. Since we used the same BDs that were used to develop the aforementioned tools, we did a one-to-one comparison of NRPreTo’s best model with the results reported by the existing tools. We selected the best model for NRPreTo based on F1 and AUC scores. We also reported the performance of our method over 15 runs to observe the variance of the model. On BD1, we compared our model performance with that of NRfamPred and RF-NR, which were trained on BD1, while for BD2, we compared our results with NR-2L, NRfamPred, and RF-NR, which were trained on BD2.

Due to the several missing values in the papers of the previous tools, a comprehensive comparison was not possible. The source codes of the tools were not available to test directly, either. Based on the available results, however, we observed that NRPreTo’s performance was on par with the performance of the other tools (Table 2).

**Experiment-2 (Combined Dataset).** To test the performance of NRPreTo on the entire human proteome, we trained a new RF model using a “combined dataset” made by combining the training and independent sets for each BD. The combined BD1 dataset contained 3016 NR and 2064 non-NR proteins, while combined BD2 consisted of 727 NR and 1000 non-NR proteins, respectively. A summary of the prediction performance of RF models on HPRD and RefSeq datasets over 15 runs using the combined models is given in Table 3.

Our “BD1-combined dataset” RF model successfully predicted 246 NRs out of 249 NRs in the RefSeq dataset and all 127 NRs from the HPRD dataset. On the other hand, the “BD2-combined dataset” RF model was able to correctly classify 234 and 122 NRs in the RefSeq and the HPRD datasets, respectively. Among the existing NR prediction tools, only NRfamPred carried out proteome-scale prediction. They
used only the HPRD dataset and utilized a limited number of features for model generation, which does not account for the different structural, functional profiles of the NR proteins. Besides, NRfamPred identified only 76 novel NRs across the HPRD dataset. It has been demonstrated that using a combination of different protein features along with appropriate feature selection methods can improve the prediction performance of classification models. In line with this, our tool was able to achieve higher prediction performance in terms of accurately identifying the maximum number of NRs across the human proteome reported so far by any previous studies or databases.

**Level 2 Predictions.** In level 2 prediction, the goal was to determine the subfamily of an NR protein. Only the proteins that were predicted as NR at level 1 were taken ahead for level 2 classification. We trained RF models using only the NR sequences belonging to the seven NR subfamilies considered in the study. We used a similar feature selection process in

**Figure 3.** Feature set-wise distribution of the important descriptors of the level 2 prediction models. (A) Experiment-1 (train only) and (B) experiment-2 (combined dataset).

**Table 4. Performance Comparison of NRPreTo Level 2 RF “Train-Only” Models with the Existing Tools on the Two Independent Datasets**

| Studies           | BD1-Independent dataset | BD2-Independent dataset |
|-------------------|-------------------------|-------------------------|
|                   | Acc                     | F1                      | MCC                     | AUC                     |
| NR-2L             | 0.996                   | -                       | -                       | -                       |
| NRPred-FS         | 0.974*                  | -                       | -                       | -                       |
| NRfamPred         | 0.995                   | -                       | -                       | -                       |
| RF-NR             | 0.996                   | 0.989                   | 0.980                   | 0.990                   |
| NRPreTo (Best     | 0.992 ± 0.004           | 0.998 ± 0.003           | 0.977 ± 0.005           | 0.991 ± 0.004           |
| model)            | 0.998 ± 0.004           | 0.999 ± 0.000           | 0.967 ± 0.015           | 0.987 ± 0.000           | 0.999 ± 0.000           |

For each cell, the score on BD1 is followed by the score on BD2 with a “|” symbol in between. * Indicates that the accuracy value reported for NRPred-FS is taken from the RF-NR publication. Since NRPred-FS was not evaluated on the BD1-independent dataset, this value is taken from the RF-NR publication who calculated NRPred-FS accuracy on the BD1-independent set. Boldfaced values indicate the best scores for each metric. Acc: accuracy. MCC: Matthew's correlation coefficient. AUC: area under the curve.
selecting important features for level 2 as in level 1. For experiment-1, 402 and 407 important features were selected for BD1 and BD2, respectively, whereas for experiment-2, 2870 and 835 important features were selected for BD1 and BD2, respectively (Figure 3). Amino acid composition group constituted the maximum number of important features in level 2 as in level 1.

Experiment-1 (Train Only). As in experimental setup 1 of level 1, this dataset also contains only the training set sequences (i.e., without including the independent set). However, since level 2 is a multiclass classification task, the classifiers were trained only on the positive set (i.e., NR sequences) to classify an incoming protein into one of the seven NR subfamilies. Thus, the training dataset in this experiment consisted of 267 and 159 NRs for BD1 and BD2, respectively. A breakup of the number of sequences in every subfamily class in BD1 and BD2 is given in Table S1. As is seen in Table S1, the number of training sequences in some classes were fewer compared to the others, leading to a highly imbalanced dataset. Therefore, class weights were used as in level 1 to give corresponding weights to all the seven classes during model training.

We compared the level 2 prediction performance of NRPreTo with existing tools. As can be seen in Table 4, most of the existing tools are built only on one of the BDs and tested on only one of the independent datasets. Also, these tools are reported using only few evaluation metrics on the independent dataset. Although RF-NR used both BDs in their pipeline, they reported the model performance using only MCC and recall for the BD2-independent dataset. NRPreTo is the first tool to build models using both BDs and evaluate them on their respective independent datasets. With the evaluation profiles of the existing tools, it is difficult to derive conclusions about the overall model performance. Based on the available values, the best NRPreTo model had almost comparable accuracy to RF-NR on BD1 but outperformed it on both BDs based on MCC and F1 scores, which are reliable metrics for imbalanced classes (Table 4). It also outperformed NRfamPred and NR-2L on BD2 based on accuracy and MCC, respectively.

Experiment-2 (Combined Dataset). To perform level 2 classification on the human proteome, we trained models on the combined datasets and tested these models on the human proteome. Here, the classifiers of BD1 and BD2 were trained on 3016 and 727 NR sequences, which makeup the number of sequences in the seven NR subfamilies. The class imbalance issue was addressed using class weights as in the previous experiments.

The results indicated that the models produced consistent results across both the external datasets over the 15 splits (Table 5). The combined BD2 model correctly classified the subfamilies of 120 out of 122 NRs identified at level 1 from

---

Table 5. Summary of Level 2 RF Model Prediction Performance Trained on Combined Model BD1 or BD2 Dataset and Tested on External Datasets over 15 Runs

| External datasets | Combined BD1 | Combined BD2 |
|-------------------|--------------|--------------|
| HPRD              | Acc ± 0.0084 | 0.9836 ± 0.0270 | 0.9423 ± 0.0018 | 0.9423 ± 0.0124 |
| RefSeq            | 0.9786 ± 0.0018 | 0.9829 ± 0.0068 | 0.9595 ± 0.0007 | 0.9620 ± 0.0245 |

For each cell, the score obtained by the model trained on BD1 is followed by the score obtained by the model trained on BD2 with a “|” symbol in between. Acc: accuracy. MCC: Matthew’s correlation coefficient. AUC: area under the curve.
HPRD. It also identified subfamilies of 230 RefSeq sequences correctly out of 234 NRs identified at level 1. The model trained on combined BD1 on the other end correctly classified 125 out of 127 HPRD NRs and 241 out of 246 RefSeq NRs identified at level 1. These results indicate that the descriptors chosen at this level were highly contributing and produced reliable results regardless of the number of splits.

**Analysis of the Importance of Protein Feature Categories.** We performed feature importance analysis for the models at both levels to assess the influence of different descriptor groups on the performance of our models. For this, we performed SHapley Additive exPlanations (SHAP) analysis, which utilizes a game theoretic approach to compute the contribution of each feature in predictions. Since our best performing models were on the combined datasets, we chose to do a feature importance analysis of RF models trained on the combined datasets. For this purpose, we used the best model of the combined dataset (Tables 3 and S) and computed the average SHAP value for each of the seven descriptor groups identified during feature selection.

The SHAP analysis results showed that for all models, the features from all descriptor groups had contribution (Figure 4). Particularly, features from the amino acid composition group, namely, AAC, DPC, TPC, CKSAAP, and dipeptide deviation from expected mean (DDE), had the highest mean SHAP value. This trend agrees with the fact that this group of descriptors consistently formed the highest fraction of the important features on both of the combined models. The previous studies utilized features from the amino acid composition group, too. However, they only considered AAC, DPC, and TPC feature sets. Our results indicate that features from CKSAAP and DDE were the most contributing for two models (Figure 4A,D) and the second most contributing in the other two models (Figure 4B,C). It is known that CKSAAP features can capture patterns of short linear motif information and thereby preserve the coevolutionary information of amino acid residues because of which they are employed in post-translational modification site identification.

**NRPreTo Predicts Known and Novel NRs.** NRPreTo classified a total of 253 sequences from both external sets. These include sequences reported by NRfamPred. We noticed that among the 76 sequences reported by NRfamPred, our tool classified 64 sequences as NR, and the remaining 12 sequences were predicted as non-NR. The 64 NRs were mostly from subfamilies 1–6 and included some widely recognized receptors such as thyroid, retinoic acid receptors (RARs), PPARs, vitamin D3-like (VDR), Pregnane-X receptors, retinoid X receptor (RXR), HNF-4α, progesterone, estrogen, and oxysterol receptors. These findings are reported in Table S2.

Upon careful analysis of the 12 sequences reported as non-NR by our tool in the NCBI-Protein database, we did not observe the signature ligand and DNA binding domains of NR. Moreover, UniProt and HPRD do not annotate these sequences as NRs (Table S3). Thus, these sequences are unlikely to be a NR. Apart from the 76 sequences reported by NRfamPred, we identified 61 sequences whose subfamily class is reported in HPRD and UniProt but not by NRfamPred. NRPreTo was able to predict the subfamily of all these sequences correctly (Table S4). NRPreTo also correctly predicted 60 sequences, which are not annotated as NR in HPRD, have not been reported in NRfamPred prediction, but UniProt annotated them as NR (Table S5). Additionally, we also predicted the subclasses of 59 RefSeq sequences, which are not annotated as NR in HPRD and UniProt (Table S6).

Upon further analysis of these sequences in the NCBI-protein database, we noticed that all the 59 sequences have the signature conserved ligand and DNA binding domains of a classic NR. Besides, we identified regions of similarity between our predicted sequences and known NR using the BLAST tool in the NCBI-Protein database. Every sequence had high similarity with known NR proteins in humans. Therefore, we propose these 59 sequences as the novel predictions of NRPreTo. We further wanted to investigate the intermolecular interactions of these receptors with approved drugs using in silico structure-based techniques like molecular docking or molecular dynamics simulations. However, these sequences do not have crystal structures yet limiting us from investigating the pharmacological associations of these sequences.

**DISCUSSION**

In the current study, we developed an NR prediction tool called NRPreTo, which performs two-level classification to identify NRs and their subfamilies. We evaluated NRPreTo on two BDs published in earlier studies and on the human proteome from HPRD and RefSeq datasets. Compared to the existing NR prediction tools, which are built only using some sequence-based features, NRPreTo was developed using a complete spectrum of protein features including the sequence-based features. We made NRPreTo publicly available at https://github.com/bozdaglab/NRPreTo under Creative Commons Attribution Non-Commercial 4.0 International Public License.

This study was carried out at under two experimental setups. In the first setup, we built a model using the training datasets of each BDs and test on the corresponding independent dataset. In the second setup, we trained a model on combined training and independent datasets of each BDs and tested directly on the human proteome from HPRD and RefSeq. We observed that the performance of the models (developed using different protein features) on independent datasets was on par with the existing tools. Moreover, the strategy of combining the datasets to train the classifiers further boosted the performance. While comparing Table 3 with Tables S7 and S8, it can be seen that the models trained on combined datasets achieved much higher performance than the models trained using the training data only. For level 2, the performance of combined models was higher than “train-only” models for all metrics except for F1 on the HPRD dataset (Table S8). Using this strategy, we not only predicted subfamilies of 60 sequences unannotated by HPRD but also proposed 59 novel NR sequences that are not annotated as NR in any databases. Using the feature importance analysis on the combined models, we observed features from CKSAAP and DDE descriptor groups contributed highly to the prediction tasks at both levels (Figure 4). Our comparisons with the existing tools were limited by the lack of scores for some evaluation metrics reported in their respective papers. All the existing tools had models trained and tested using only one of the BDs. They did not have the source code available for us to run their models. Moreover, except NRfamPred, no other tool has tested their model to identify new NRs from the human proteome. Thus, we could not compare the performance of NRPreTo with that of all tools on all BDs for all performance metrics.
CONCLUSIONS

In this study, we demonstrated the application of the complete protein feature spectrum for the classification of NR proteins. Our results indicated the contribution made by various features encoding the structural, special, physicochemical, and evolutionary information along with the regular sequence-based features toward the NR classification task. This study supports our hypothesis that a collective contribution of all the feature types can capture the patterns within the NR subfamilies compared to the use of only the sequence-based features. We developed ML-based pipelines involving feature generation and a robust feature selection algorithm to identify multiple important features to be used for model development. We used the same two-level approach as the previous studies to first identify the NR and thereafter assigning them into one of the seven NR subfamilies. We observed that NRPreTo has been on par or outperformed the existing NR prediction tools.

One of the main reasons for the improved performance of NRPreTo can be attributed to the use of a highly vivid feature set combined with the two-experimental setups used in our pipeline to train and validate our models. Specifically, we observed that at level 2, the strategy of combining training and independent datasets yielded a better performance compared to the models developed only on the training sets of the two BDs. This strategy makes NRPreTo unique compared to the existing tools. We also used our tool to identify proteome-wide predictions of NR across the human proteome deposited in two public databases NCBI-RefSeq and the HPRD. Our tool identified a total of 253 sequences from the two external datasets.

MATERIALS AND METHODS

Data Sources. Benchmark Datasets. We utilized two previously published benchmark datasets, namely, benchmark dataset 1 (BD1) of RF-NR and benchmark dataset 2 (BD2) of NR-2L. These BDs are composed of training and independent sets containing both NR and non-NR sequences (Table S1). Both BDs were prepared from the Nuclear Receptor Database (NucleaRDB Release5.0) containing 3016 NR sequences, which are phylogenetically classified into seven subfamilies with each subfamily containing NR sequences from human and different animal species. Before using the data for training, we also ensured that the FASTA sequences of the benchmark datasets are completed. The detailed procedure of generating the BDs is outlined in NR-2L and RF-NR studies.

External Datasets. To further evaluate the models trained on BD, we used two high-quality human proteome datasets, namely, HPRD and NCBI-RefSeq. The HPRD dataset contains 30,046 manually curated human proteins. Since the HPRD dataset could not be downloaded from the HPRD website, we used the data distributed by the OmniPath rescued data repository for this study. We verified that the data obtained from OmniPath is the latest version of HPRD (HPRD_Release 9.0) mentioned on the HPRD website. Since the version on the HPRD website is not updated after 2009, we also tested our models on another updated set of human protein sequences derived from NCBI-RefSeq. The NCBI-RefSeq database contains integrated, comprehensive, non-redundant, well-annotated sequences of genomic DNA, transcripts, and proteins for various organisms. For this study, we only considered human protein sequences reported in the RefSeq database, which had 116,585 records at the time of download. After eliminating the computationally predicted models (i.e., the records with accession numbers prefixed with “XP” or “YP”), we had 62,625 records, which were used in this study.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c00286.
Description of the subfamily names and the number of proteins in the two benchmark datasets; information of the sequences commonly identified by NRPreTo and NRfamPred and comparison with UniProt and HPRD; further information to support the novel sequences identified by NRPreTo from the human proteome; level 1 performance comparison of the “train-only” models of the two BDs; level 2 performance comparison of the “combined models” of the two BDs; and description of the hyperparameters tuned and their ranges while training the RF models (PDF)

**REFERENCES**

1. Mangelsdorf, D. J.; Thummel, C.; Beato, M.; Herrlich, P.; Schütz, G.; Umesono, K.; Blumberg, B.; Kastner, P.; Mark, M.; Chambon, P.; et al. The nuclear receptor superfamily: The second decade. *Cell* 1995, 83, 835–839.
2. Nagy, L.; Schwabe, J. W. R. Mechanism of the nuclear receptor molecular switch. *Trends Biochem. Sci.* 2004, 29, 317–324.
3. Gronemeyer, H.; Gustafsson, J-A.; Laudet, V. Principles for modulation of the nuclear receptor superfamily. *Nat. Rev. Drug Discov.* 2004, 3, 950–964.
4. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/ coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 1998, 95, 927–937.
5. Weikum, E. R.; Liu, X.; Ortlund, E. A. The nuclear receptor superfamily: A structural perspective. *Protein Sci.* 2018, 27, 1876–1892.
6. Evans, R. M. The steroid and thyroid hormone receptor superfamily. *Science* 1988, 240, 889–895.
7. Daniellian, P. S.; White, R.; Lees, J. A.; Parker, M. G. Identification of a conserved region required for hormone dependent transcriptional activation by steroid hormone receptors. *EMBO J.* 1992, 11, 1025–1033.
8. Zhang, C.; Zhang, B.; Zhang, X.; Sun, G.; Sun, X. Targeting orphan nuclear receptors NR4As for energy homeostasis and diabetes. *Front. Pharmacol.* 2020, 11, 587457.
9. Tata, J. R. Signalling through nuclear receptors. *Nat. Rev. Mol. Cell Biol.* 2002, 3, 702–710.
10. Aranda, A.; Pascual, A. Nuclear hormone receptors and gene expression. *Physiol. Rev.* 2001, 81, 1269–1304.
11. Mangelsdorf, D. A Unified Nomenclature System for the Nuclear Receptor Superfamily. *Cell* 1999, 97, 161–163.
12. Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* 2002, 1, 727–730.
13. Bhasin, M.; Raghava, G. P. S. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 2004, 279, 23262–23266.
14. Gao, Y.; Shao, S.; Xiao, X.; Ding, Y.; Huang, Y.; Huang, Z.; Chou, K. C. Using pseudo amino acid composition to predict protein subcellular location: Approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids* 2005, 28, 373–376.
15. Wang, P.; Xiao, X.; Chou, K.-C. NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features. *PLoS ONE* 2011, 6, e23505.
16. Xiao, X.; Wang, P.; Chou, K. C. INR-PhysChem: A Sequence-Based Predictor for Identifying Nuclear Receptors and Their Subfamilies via Physical-Chemical Property Matrix. *PLoS ONE* 2012, 7, No. e30869.
17. Wang, P.; Xiao, X. NRPre-FS A feature selection based two-level predictor for nuclear receptors. *J. Proteom. Bioinform.* 2014, 9. DOI: 10.4172/jpb.9-002
18. Kumar, R.; Kumari, B.; Srivastava, A.; Kumar, M. N. NRfamPred: A proteome-scale two level method for prediction of nuclear receptor proteins and their sub-families. *Sci. Rep.* 2014, 4, 6810.
19. Keshava Prasad, T. S.; Goel, R.; Kundasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafran, B.; Venugopal, A.; et al. Human protein reference database—2009 update. *Nucleic Acids Res.* 2009, 37, D767–D772.
20. Ismail, H. D.; Saigo, H.; Kc, D. B. RF-NR: Random Forest Based Approach for Improved Classification of Nuclear Receptors. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2018, 15, 1844–1852.
21. Fischer, A.; Smiesko, M. AllostERIC binding sites on nuclear receptors: focus on drug efficacy and selectivity. *Int. J. Mol. Sci.* 2020, 21, No. E534.
22. O’Leary, N. A.; Wright, M. W.; Brister, J. R.; Cuifo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI.
current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016, **44**, D733–D745.

(23) Kursa, M. B.; Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Soft.* 2010, **36**, DOI: 10.18637/jss.v036.i11.

(24) Ding, H.; Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 2015, **47**, 329–333.

(25) Jiao, S.; Zou, Q.; Guo, H.; Shi, L. iTTCA-RF: a random forest predictor for tumor T cell antigens. *J. Transl. Med.* 2021, **19**, 449.

(26) Bian, H.; Guo, M.; Wang, J. Recognition of mitochondrial proteins in Plasmodium based on the tripeptide composition. *Front. Cell Dev. Biol.* 2020, **8**, 578901.

(27) Ding, C.; Yuan, L.-F.; Guo, S.-H.; Lin, H.; Chen, W. Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteom.* 2012, **77**, 321–328.

(28) Yang, L.; Gao, H.; Wu, K.; Zhang, H.; Li, C.; Tang, L. Identification of cancerlectins using cascade linear discriminant analysis and optimal g-gap tripeptide composition. *CBIO* 2020, **15**, 528–537.

(29) Yang, L.; Gao, H.; Liu, Z.; Tang, L. Identification of phage virion proteins by using the g-gap tripeptide composition. *LOC* 2019, **16**, 332–339.

(30) Wang, K.; Li, S.; Wang, Q.; Hou, C. Identification of hormone-binding proteins using a novel ensemble classifier. *Computing* 2019, **101**, 693–703.

(31) You, Z.-H.; Yu, J.-Z.; Zhu, L.; Li, S.; Wen, Z.-K. A MapReduce based parallel SVM for large-scale predicting protein–protein interactions. *Neurocomputing* 2014, **145**, 37–43.

(32) Limongelli, I.; Marini, S.; Bellazzi, R. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinformatics* 2015, **16**, 123.

(33) Chen, Z.; Chen, Y. Z.; Wang, X. F.; Wang, C.; Yan, R. X.; Zhang, Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS ONE* 2011, **6**, e22930.

(34) Lv, Z.; Wang, D.; Ding, H.; Zhong, B.; Xu, L. Escherichia Coli DNA N-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology. *IEEE Access* 2020, **8**, 14851–14859.

(35) Li, Y.; Zhang, Z.; Teng, Z.; Liu, X. PredAmyl-MLP: Prediction of Amyloid Proteins Using Multilayer Perceptron. *Comput. Math. Methods Med.* 2020, **2020**, 1–12.

(36) Xu, R.; Zhou, J.; Liu, B.; He, Y.; Zou, Q.; Wang, X.; Chou, K. C. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. *Biomol. Struct. Dyn.* 2015, **33**, 1720–1730.

(37) Nanni, L.; Braham, S.; Luminii, A. Wavelet images and Chou’s pseudo amino acid composition for protein classification. *Amino Acids* 2012, **43**, 657–665.

(38) Lv, Z.; Zhang, J.; Ding, H.; Zou, Q. RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. *Front. Bioeng. Biotechnol.* 2020, **8**, DOI: 10.3389/fbioe.2020.00134.

(39) Sikander, R.; Ghulam, A.; Ali, F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. *Sci. Rep.* 2022, **12**, 5505.

(40) Sequeira, A. M.; Louza, D.; Rocha, M. P. P. ProPythia: A Python package for protein classification based on machine and deep learning. *Neurocomputing* 2022, **484**, 172–182.

(41) Xiao, X.; Shao, S.-H.; Huang, Z.-D.; Chou, K.-C. Using pseudo amino acid composition to predict protein structural classes: Approached with complexity measure factor. *J. Comput. Chem.* 2006, **27**, 478–482.

(42) Yu, H.; Yang, J.; Wang, W.; Han, J. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. *Computational Systems Bioinformatics. (CSB)2003. Proceedings of the 2003 {IEEE} Bioinformatics Conference. (CSB)2003 IEEE 2003, Vol. 2*, pp 220–228.

(43) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 2017, **30**, 20388 https://doi.org/10.1021/acsomega.3c02086

ACS Omega 2023, **8**, 20379−20388

ACS Omega is published by the American Chemical Society.

© 2023 American Chemical Society. See http://publishing.acs.org/acsomega/rights and permissions for rights and reprints.

20388