SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PROTEIN FUNCTION PREDICTIONS
SEQUENCE ANNOTATION

Received
26 June 2014

Accepted
9 October 2014

Published
29 October 2014

Correspondence and requests for materials should be addressed to M.K. (manish@south.du.ac.in)

* These authors contributed equally to this work.

NRfamPred: A proteome-scale two level method for prediction of nuclear receptor proteins and their sub-families

Ravindra Kumar*, Bandana Kumari*, Abhishikha Srivastava & Manish Kumar

Department of Biophysics, University of Delhi South Campus, Benito Juarez Road, New Delhi, India-110021.

Nuclear receptor proteins (NRP) are transcription factor that regulate many vital cellular processes in animal cells. NRPs form a super-family of phylogenetically related proteins and divided into different sub-families on the basis of ligand characteristics and their functions. In the post-genomic era, when new proteins are being added to the database in a high-throughput mode, it becomes imperative to identify new NRPs using information from amino acid sequence alone. In this study we report a SVM based two level prediction systems, NRfamPred, using dipeptide composition of proteins as input. At the 1st level, NRfamPred screens whether the query protein is NRP or non-NRP; if the query protein belongs to NRP class, prediction moves to 2nd level and predicts the sub-family. Using leave-one-out cross-validation, we were able to achieve an overall accuracy of 97.88% at the 1st level and an overall accuracy of 98.11% at the 2nd level with dipeptide composition. Benchmarking on independent datasets showed that NRfamPred had comparable accuracy to other existing methods, developed on the same dataset. Our method predicted the existence of 76 NRPs in the human proteome, out of which 14 are novel NRPs. NRfamPred also predicted the sub-families of these 14 NRPs.

Nuclear receptor proteins (NRP) are one of the most abundant type of transcription regulators, which are present exclusively in animals[1]. NRPs form an evolutionarily related super-family of proteins, which function as ligand-activated transcription factors, providing a direct link between signaling molecules that control these processes and transcriptional responses[1]. All NRPs share a common five-domain structure with a highly conserved DNA binding domain. Interaction of cognate ligands, which are mostly small hydrophobic compounds, such as steroids, retinoids, and thyroid hormones, trigger a conformational change in the receptor proteins. It enables interaction with specific cofactors and cis-regulatory DNA sequences called hormone response elements (HREs) thereby subsequently altering the gene expression.

Members of the NRP super-family has a conserved modular domain architecture: a domain of variable length, A/B domain having activation function (AF-1), the highly conserved C-region or DNA-binding domain (DBD), the hinge or the D- region, the E-region containing ligand-binding domain (LBD) and an F-domain that is present in few NRPs[2] (from N to C-terminus). The DBD contains two zinc finger motif in tandem (spanning nearly 80 amino acid residues in total) and are directly involved in recognition of the cognate HRE[3,4]. The LBD is responsible for both ligand recognition and regulation of protein–protein interactions[5]. The ligand binding in NRPs occurs through a receptor specific hydrophobic ligand-binding pocket, which is present deep in the core of LBD. Since DBD and LBD are the two most conserved domains of NRPs, they are regarded as dual signatures of this protein super-family. Nuclear receptors are ancient proteins that have been found in diverse clades like sponges, echinoderms, tunicates, arthropods and vertebrates, and are therefore believed to be present throughout the Metazoa[6]. Depending on the nature of the ligand, NRP super-family proteins have been sub-divided into six different sub-families while all unusual receptors that contain only one of the two conserved domains (C or E) were grouped into a separate sub-family NR0[7]. NRPs having no known ligand are classified as orphans[8]. Hence, the function of a NRP is closely related to the sub-family to which it belongs. Due to the vital importance of NRPs in many physiological and pathological aspects of metazoan life, they are considered as candidates of equal importance for drug development as are G-protein coupled receptors (GPCR), ion channels and kinases[9]. Another factor which makes NRPs a promising pharmacological target is the nature of ligands which are small lipophilic compounds such as steroids, thyroid hormone, vitamin D3, and retinoids[2,10], which regulate crucial biological functions like metabolism, homeostasis, development and disease[11]. Since ligands are small molecules, they can be easily modified by drug designing, making NRPs a promising pharmacological target.

Considering the pace with which new protein sequences are being generated in the post-genomic age, it is the need of the hour to develop automated methods for rapid and accurate identification of NRPs and their sub-families on the basis of amino acid sequence information. Bhasin and Raghava[12] made a pioneering effort in this direction by developing a support vector machine (SVM) based method for predicting four sub-families of NRPs (thyroid hormone-like, HNF4-like, estrogen-like, Fushi tarazu-F1-like) using amino acid and dipeptide compositions as the input. Later Gao et al.[13] used the pseudo amino acid composition[14] and a new dataset for the same four NRP sub-families that Bhasin and Raghava had worked upon and reported a higher prediction accuracy. Though Bhasin and Raghava's method is available to the scientific community via web-server (NRpred), no such provision was made by Gao et al.[13]. Besides the limited coverage for four NRP sub-families, one major limitation of NRpred is that it predicts sub-families without screening whether the query protein is actually a NRP or not. Thus even if the query protein doesn't belong to the NRP super-family, it would be classified in one of the four NRP sub-families. Recently two different predictors were proposed which extended the coverage scope of prediction to seven sub-families and carried the prediction cycle at two levels. At the 1st level they screen NRPs, while 2nd level identify the sub-family. The first method was named as NR-2L[15] while second was called iNR-PhysChem[16].

In the present study, we have described a method developed by us named NRfamPred, which identifies NRPs from primary amino acid sequence. NRfamPred is SVM based two level method for prediction of NRPs and their seven sub-families, which uses dipeptide compositions as input vector. We tested our method on an independent dataset and found that our method was better than other existing methods. The proposed method can also be used to annotate proteome. In this work we annotated human proteome and fetched 76 NRPs out of which 14 are novel.

Results

1st Level Prediction. We used the 1st level classifier to screen if the query protein belonged to the NRP super-family. Only those protein, which were predicted as NRPs using the 1st level classifier could proceed to the 2nd level for prediction of its sub-family. The SVM model used for 1st level prediction was generated using a non-redundant dataset of 500 non-NRPs and 159 NRPs (Supplementary Table S1) (described in methods). When the amino acid composition was used as the input, 93.32% accuracy with Matthew's correlation coefficient (MCC) 0.84 was achieved. The accuracy rose to 97.88% with MCC 0.94 when dipeptide composition was used as the input (Table 1). This showed that dipeptide composition encapsulated the sequence information better than that by amino acid composition.

2nd Level Prediction. To generate SVM model at the 2nd level we used only those NRPs (159 in total), which were labeled as positive class examples in 1st level prediction. At 2nd level, proteins belonging to a particular sub-family were considered as positive class while the remaining one as negative. For example, in order to predict proteins of NRP0 sub-family, all 12 NRP0 sub-family proteins (Supplementary Table S1) were used as positive class example, while remaining 147 proteins (belonging to NRP1-6 sub-families) were considered as negative class example.

It is evident from Table 2 that similar to the 1st level, dipeptide composition based SVM models performed better than amino acid composition based models. For each sub-family, we were able to achieve nearly $\geq$ 95% prediction accuracy. It is also pertinent to mention that except NRP0 and NRP2 sub-families, we achieved nearly 100% prediction specificity with dipeptide composition based models (Table 2). This showed that the SVM models developed for sub-family prediction not only predicted the proteins belonging to the same sub-family with high sensitivity but also with a very high specificity.

The developed dipeptide composition based SVM models (both 1st and 2nd levels) were collectively called as NRfamPred.

Receiver Operating Characteristics (ROC) Plot and Area Under ROC Curve (AUC) Analysis. When a classifier has to do the multi-class classification, especially on an imbalanced dataset like the present work, overall accuracy might be an unrealistic assessment of classifier's performance due to the correct classification of majority class. Hence, to avoid the influence of majority classes in performance estimation, the prediction capability of SVM model was assessed by both sensitivity and specificity, also taking into account that values of both were nearly equal. This also helped in eradicating an inequity in the accuracy value, which might have occurred due to incorporation of unequal number of positive and negative examples. Another way of unbiased estimation of classifier's accuracy is by using the receiver operating characteristic (ROC) plot[17,18], which is a very popular way of analyzing the overall performance of a classifier system. It shows the tradeoff between sensitivity and specificity at various thresholds and is created by plotting 'sensitivity' (True positive rate) vs. '100-specificity' (False positive rate). The area under the ROC curve (AUC)[19] is commonly used as a summary measure of diagnostic accuracy. The ROC plots (Supplementary Fig. S1) and corresponding AUC values (Table 1 and Table 2) also support the conclusion that dipeptide composition based SVM modules can predict NRPs at very high accuracy and were better than amino acid composition based SVM modules.

Performance of NRfamPred on Independent Dataset. The performance of NRfamPred was further evaluated on an independent dataset having 568 NRPs and 500 non-NRPs compiled by Wang et al.[15] (see methods for detail). As shown in Supplementary Table S2, the relative prediction rates for both positive and negative classes are consistent across both levels and all sub-families. This indicates that the NRfamPred not only identifies NRPs but it is also able to classify them very accurately.

Comparison with Existing Methods. A few methods have already been published for predicting NRP and/or their sub-families[12,15,16,20]. It is not practically possible to compare the performance with all existing methods due to difference in number of sub-families or training datasets. As NRfamPred was developed using the dataset

Table 1 | Performance of amino acid and dipeptide composition based SVM models during LOOCV at 1st level. All values except MCC and AUC are in percentage. Sens, Spec, Acc, MCC and AUC stand for sensitivity, specificity, accuracy, Matthew's correlation coefficient and area under ROC curve respectively

Amino Acid					Dipeptide				
Sens	Spec	Acc	MCC	AUC	Sens	Spec	Acc	MCC	AUC
96.86	92.20	93.32	0.84	0.98	96.23	98.40	97.88	0.94	1.00

Table 2 | Performance of amino acid and dipeptide composition based SVM models during LOOCV at 2nd level. All values except MCC and AUC are in percentage. Sens, Spec, Acc, MCC and AUC represent sensitivity, specificity, accuracy, Matthew's correlation coefficient and area under ROC curve respectively

	Amino Acid					Dipeptide				
Sub-family	Sens	Spec	Acc	MCC	AUC	Sens	Spec	Acc	MCC	AUC
NRP0	50.00	97.96	94.34	0.55	0.79	83.33	95.92	94.97	0.70	0.95
NRP1	80.00	82.57	81.76	0.60	0.86	98.00	99.08	98.74	0.97	0.99
NRP2	83.33	78.86	79.87	0.54	0.84	91.67	96.75	95.60	0.88	0.97
NRP3	78.38	98.36	93.71	0.82	0.98	100.00	99.18	99.37	0.98	1.00
NRP4	85.71	90.79	90.57	0.47	0.92	100.00	99.34	99.37	0.93	1.00
NRP5	58.33	97.96	94.97	0.61	0.84	83.33	100.00	98.74	0.91	0.98
NRP6	80.00	100.00	99.37	0.89	1.00	100.00	100.00	100.00	1.00	1.00
Overall	76.73	92.98	90.66	0.63	–	94.97	98.63	98.11	0.92	–

on which NR-2L and iNR-PhysChem was based, a direct one to one comparison with them will be more appropriate.

Leave-One-Out Cross-Validation (LOOCV) Performance. A comparative performance of NRfamPred vis-à-vis iNR-PhysChem and NR-2L is shown in Table 3 and Table 4. NRfamPred achieved 96.23% sensitivity and 98.40% specificity at the 1st level. With the same set of proteins, iNR-PhysChem sensitivity and specificity were 96.23% and 98.80% respectively. It clearly shows that NRfamPred has comparable prediction accuracies during LOOCV with iNR-PhysChem. For NR-2L the sensitivity and specificity was 98.11% and 90.80% respectively. At a preliminary glance it appeared that the sensitivity achieved by NRfamPred (96.23%) was lesser than NR-2L, thus we analyzed the different sensitivity and specificity values attained by NRfamPred during LOOCV in detail (Supplementary Table S3). NRfamPred was found to achieve a specificity of 96.40% at the corresponding sensitivity 98.11%, which was much higher than the specificity obtained by NR-2L (Table 3 and Supplementary Table S3). At 2nd level the sensitivity of NRfamPred was more than iNR-PhysChem for all sub-families (except for NRP2) while it had higher or equivalent prediction accuracy for all sub-families when compared to NR-2L (Table 4).

Performance on Independent Dataset. In their study Xiao et al.[16] had not benchmarked the performance of iNR-PhysChem on independent dataset, hence in this paper comparison on independent dataset (P^{IND}) was carried out only with NR-2L. As shown in Table 3 and Table 5, the overall performance of NRfamPred is better than NR-2L at both 1st and 2nd levels. It clearly shows that NRfamPred is better and more accurate than NR-2L for practical applications also.

NRfamPred Web-Server Performance. In real life, nature of proteins presented to NRfamPred for prediction will not be known in advance. Hence *one-vs-rest* approach of prediction will not work in the actual situation. In the NRfamPred web-server/standalone the first prediction would decide whether the query protein is a NRP or not? This is a threshold dependent prediction so that the user can select desired level of sensitivity and specificity. If the query protein is predicted to be a NRP, it will be forwarded for the sub-family prediction and will be classified into the sub-family corresponding to the highest SVM score. In order to compare the prediction capability of NRfamPred *vis-à-vis* iNR-PhysChem and NR-2L web-servers, all of the 1068 proteins of P^{IND} were submitted to the NRfamPred, iNR-PhysChem and NR-2L web-servers and prediction was done at default parameters. As shown in the Table 6, NRfamPred web-server predicted all 568 NRPs correctly as NRP and also their sub-families. On the other hand iNR-PhysChem and NR-2L were able to correctly predict and classify only 562 and 565 NRPs respectively. In case of non-NRP, out of 500 NRfamPred falsely predicted 12 proteins as NRPs. The performance of NRfamPred was better than NR-2L at this level also, which falsely predicted 19 proteins as NRPs. 11 proteins were wrongly predicted as NRPs by the iNR-PhysChem. It clearly shows that NRfamPred web-server can do prediction with higher accuracy in comparison of iNR-PhysChem and NR-2L even in blind condition.

Comparison with Other Prediction Approaches. Almost all nuclear receptors share a highly conserved zinc-finger DBD and a less conserved LBD[21]. The DBD and LBD are regarded as dual signatures of this protein super-family[22]. Hence, one of the most intuitive ways of filtering NRPs from non-NRPs is to build a profile Hidden Markov model (HMM) using NRP sequences and then perform searching against the query protein. Similarly sub-family prediction can be done using HMM build for individual sub-family. If an unknown sequence shows very high conservation with the NRP profile, it can be predicted to belong to NRP. Similarly a sequence showing high similarity to a particular sub-family profile can belong to the corresponding sub-family. In order to verify this approach, we built HMM profiles of total NRPs as well as each sub-

Table 3 | Comparative performance of NRfamPred vis-à-vis INR-PhysChem and NR-2L at 1st level. At LOOCV, comparison is made at the point where sensitivities of NR-2L and iNR-PhysChem were equal to NRfamPred. iNR-PhysChem was not evaluated using independent dataset in Ref. no. [16]. Hence, corresponding values of iNR-PhysChem is not shown. All values except MCC are in percentage. (#Ref. no. [16], ¶Ref. no. [15])

LOOCV				
Predictor	Sensitivity	Specificity	Accuracy	MCC
NRfamPred/iNR-PhysChem#	96.23/96.23	98.40/98.80	97.88/98.18	0.94/0.96
NRfamPred/NR-2L¶	98.11/98.11	96.40/90.80	96.81/92.56	0.92/0.83
$DATA^{IND}$				
NRfamPred	100.00	98.40	99.25	0.99
NR-2L¶	99.65	96.20	98.03	0.96

Table 4 | Comparison of LOOCV performance of NRfamPred, iNR-PhysChem and NR-2L at 2nd level of prediction. All values except MCC are in percentage. Sensitivities of iNR-PhysChem and NR-2L was reported as accuracy in #Ref. no. [16] and ¶Ref. no. [15] respectively

	NRfamPred		iNR-PhysChem		NR-2L	
Sub-family	Sensitivity	MCC	Sensitivity#	MCC#	Sensitivity¶	MCC¶
NRP0	83.33	0.70	66.67	0.81	75.00	0.86
NRP1	98.00	0.97	94.00	0.87	86.00	0.88
NRP2	91.67	0.88	97.22	0.93	86.11	0.85
NRP3	100.00	0.98	100.00	0.95	100.00	0.86
NRP4	100.00	0.93	71.43	0.84	85.71	0.70
NRP5	83.33	0.91	83.33	0.91	83.33	0.86
NRP6	100.00	1.00	100.00	1.00	100.00	1.00
Overall	94.97	0.92	92.45	0.91	88.68	0.87

family using the sequences of main dataset (P^{MAIN}) and searched proteins of P^{IND}. At 1st level total 564 NRPs were correctly predicted without any false positive (meaning no non-NRP was predicted as NRP). But at 2nd level we observed a long hit list with high score and very low E-values showing similarity to same as well as different sub-families. This made it difficult to differentiate the probable sub-family to which the query protein might belong, only on the basis of alignment score and E-values. In homology based clustering algorithms generally an E-values between 10^{-8} to 10^{-100} were considered as threshold to define homology between protein sequences[23]. At both levels the hits of HMM based search were within this range.

Other than HMM, a number of network-based unsupervised classification approaches are also available[24–26]. These methods use homology based clustering to group proteins of same class in same cluster. Similar to the HMM based prediction, this approach can also be very successful at 1st level but fail at 2nd level due to high sequence conservation among proteins of different sub-families. It shows that although HMM based searching and homology based clustering methods are intuitively a logical and most obvious step to find a novel NRP, but as shown in this work, inherent conservation of sequences across the super-family makes these approaches difficult to use in real life. In contrast to this, our method provides a clear-cut unambiguous answer of 'the sub-family to which the query protein may belongs'.

Proteome-scale Prediction of Nuclear Receptor Proteins. In order to show the efficiency of our method, we used two phylogenetically widely separated organisms *Arabidopsis thaliana* and *Homo sapiens*.

Table 5 | Comparison of performance of NRfamPred and NR-2L on P^{IND} at 2nd level of prediction using P^{IND}. iNR-PhysChem was not evaluated on P^{IND} in Ref. no. [16]. Hence, corresponding values of iNR-PhysChem is not shown. All values except MCC are in percentage. ¶Sensitivity of NR-2L was reported as accuracy in Ref. no. [15]

	NRfamPred		NR-2L	
Sub-family	Sensitivity	MCC	Sensitivity¶	MCC
NRP0	100.00	0.77	100.00	1.00
NRP1	100.00	1.00	99.13	0.99
NRP2	99.21	0.99	100.00	1.00
NRP3	100.00	0.97	100.00	1.00
NRP4	100.00	1.00	100.00	0.98
NRP5	100.00	1.00	100.00	0.98
NRP6	–	–	–	–
Overall	99.82	0.98	99.65	–

Table 6 | Comparative performance of NRfamPred, iNR-PhysChem and NR-2L web-servers on P^{IND}. NRP6 was not evaluated since P^{IND} doesn't have sub-family NR6 data

Sub-family	Number of proteins in P^{IND}	NRfamPred	iNR-PhysChem	NR-2L
NRP0	6	6	5	6
NRP1	231	231	229	228
NRP2	127	127	126	127
NRP3	148	148	147	148
NRP4	23	23	22	23
NRP5	33	33	33	33
NRP6	NA	NA	NA	NA
Total	568	568	562	565
Non-NRP	500	488	489	481

We opted to choose Human proteome due to its complex human hormone signaling pathways. Previous reports suggested 48 nuclear-receptor genes found in human[22,27]. The idea behind annotating Arabidopsis genome was to show the specificity of genome wide prediction of NRPs since nuclear hormone receptor homologs do not exist in plants[28].

NRfamPred predicted 76 NRPs out of total 30,046 human proteins (0.25% of total proteome). Our estimate is very near to the estimate of 75 nuclear receptors in the mammalian proteome[21]. We compared NRfamPred prediction with Human Protein Reference Database (HPRD)[29] and Uniprot annotations and observed that all three placed 27 proteins in the same sub-family. 29 proteins were predicted as the similar sub-families as annotated in Uniprot, however HPRD did not annotate them. Similarly, 6 proteins that are annotated in HPRD but not in Uniprot had similar results with NRfamPred. NRfamPred predicted 14 proteins as NRPs that were either annotated as non-NRP or had no information in Uniprot and HPRD (Supplementary Table S4).

Since plants doesn't have NRPs[28], we tested NRfamPred on negative control to see how many NRPs were predicted in plants proteome. For that we use *Arabidopsis thaliana* proteome, which contains 27,416 sequences. Our method predicts only three NRPs (AT1G12860.1, AT2G43945.1 and AT3G59870.1) in total proteome, which showed that NRfamPred could do proteome wide searching of NRPs with very high specificity.

Web-Server and Standalone Software. In order to make our prediction method available to the scientific community, a web-server has been established at http://14.139.227.92/mkumar/nrfampred, where user can submit up to 25 protein sequences at a time for prediction. The overall schema of prediction methodology used in the web-sever is described in Supplementary Fig. S2. We also developed standalone version of NRfamPred to automate the task of proteome wide prediction, which can be downloaded from http://14.139.227.92/mkumar/nrfampred/download.html.

Discussion

The aim of this study was to develop a reliable method, named as NRfamPred, to identify nuclear receptor proteins in proteome and group them into appropriate sub-families. The whole prediction approach was divided into two steps. First step discriminated between NRP and non-NRP while second step predicted the sub-family. We used earlier compiled datasets and two different forms of sequence information (amino acid and dipeptide compositions) as input to the SVM to develop the proposed method. Between the two input modes, dipeptide based SVM model performed better than the amino acid composition based model (Table 1 and Table 2). Performance on independent dataset (Supplementary Table S2) and the comparative study between NRfamPred and other available methods (Tables 3–6, Supplementary Table S2) also proved

NRfamPred as a better predictor. Performance of iNR-PhysChem (not available for independent dataset) prima facie seems comparable on the basis of LOOCV result but when we used its web-server for prediction of dataset P^{IND}, it failed to achieve the level of NRfamPred (Table 6).

We also used NRfamPred to predict NRPs, which are present in the human proteome and identified 14 novel NRPs, which have not been reported till date. NRfamPred also assigned sub-families to these novels NRPs. NRPs play a crucial role in diverse biological processes, including lipid and glucose homeostasis, detoxification, cellular differentiation and embryonic development, and mutations in nuclear receptors associated to many common and lethal disorders, including cancer, diabetes and heart disease[30,31]. Hence a proper investigation and experimental validation of these predicted NRPs might be useful for the scientific community. Another interesting finding was observed that NR0 sub-family proteins were not predicted in human proteome. There might be two possibilities behind this. Either NRfamPred failed to predict the members of sub-family NR0 or dataset on which annotation pipeline was executed, didn't had these proteins. The former probability was ruled out by submitting a compilation of NRPs of human, mouse and rat (compiled by Zhang et al.[22]) to the NRfamPred web-server. The result (Supplementary Table S4) showed that this assumption was not correct as NRfamPred rightly predicted the proteins of sub-family NR0, which were present in human, mouse and rate genomes (Supplementary Table S5).

Methods

Prediction Schema. In the present work we have tried to solve two different problems simultaneously. The 1st problem is to identify the proteins belonging to the family NRPs and the 2nd is to predict the sub-family to which a particular NRP belongs. It means 1st level is a binary classification problem, which can be addressed by a classifier that can classify the query protein into NRP or non-NRP. But the prediction of NRP sub-family was a typical example of multi-class classification. Here the objective was to identify the correct sub-family of a protein, predicted as NRP, in the previous step. A simple strategy to handle this type of problem is to divide multi-class classification into a series of binary classifications, popularly known as *one-vs-rest* approach. It involves development of a number of classifiers using one class as positive while remaining classes as negative. An SVM trained to predict proteins of a particular sub-family was trained on all samples of that sub-family with positive label and proteins of remaining sub-families with negative label. The same approach has been used in a number of earlier studies like prediction of sub-cellular localization[32–34], G-protein coupled receptors[35,36], NRP protein sub-family prediction[12,15,16,20].

In multi-class classification, it is an underlying assumption that the input/query sequence belongs to the family whose class we are going to predict. During training the assumption might be correct as it is being done on manually curated data. But in this post-genomic era when, annotation pipelines use prediction methods in assembly mode, therefore probability of getting a sequence, which doesn't belong to the same family, is fairly high. Hence in absence of a filter a non-family member might be predicted to belongs to the class to which it is unrelated. Further, one of the main aims of our work is to provide a tool that can be used to annotate uncharacterized proteins. If the input sequence doesn't belong to NRP super-family, the sub-family classification is actually meaningless. In order to reduce the likelihood of wrong classification, we have adopted the two level prediction approach. The 1st level is to screen NRPs, while the 2nd level identifies the sub-family to which it belongs (Supplementary Fig. S2). In summary the overall prediction works in following 3 steps: (a) The query protein is presented to the prediction algorithm. (b) If it is a non-NRP, the prediction stops after 1st level (c) If the query protein predicted as a NRP at the 1st level, it is forwarded to the 2nd level for sub-family prediction.

Main Dataset (P^{MAIN}). In the present work we have used the earlier published dataset, which was used for the development of NR-2L and iNR-PhysChem. It has 159 NRPs and 500 non-NRPs (Supplementary Table S1). NRPs were collected from nuclear receptor database (NucleaRDB release 5.0; http://www.receptors.org/NR/)[37]. The 500 non-NRP sequences were randomly collected from the UniProt (http://www.uniprot.org/) according to their annotations in the "Keyword" field. To remove redundancy, sequences having more than 60% pair-wise sequence identity to any other proteins in same sub-family were removed using CD-HIT[38].

Blind or Independent Dataset (P^{IND}). This data was used as an independent benchmarking dataset to evaluate real life performance of predictor developed using P^{MAIN}. Similar to P^{MAIN}, this was also originally compiled and used for benchmarking earlier method NR-2L. It has 568 NRPs and 500 non-NRPs (Supplementary Table S1).

Genome-scale Prediction of Nuclear Receptor Proteins. In order to show the real life usage and efficacy of our method, we annotated two proteomes namely, *Homo sapiens* as positive control, and *Arabidopsis thaliana* as negative control. The human proteome was downloaded from HPRD[29], which is a very high quality manually curated human protein database. It had 30,046 human protein sequences. The Arabidopsis proteome was downloaded from TAIR[39]. It contains 27,416 proteins of representative gene models.

Cross-Validation and Performance Evaluation. Cross-validation is a way to estimate the performance of a prediction model on a dataset, which is not used for generating it. It involves partitioning of data into complementary sub-sets, performing the analysis on one sub-set (called training set), and validating the analysis on other sub-set (called testing set). To reduce variability due to sample partition, multiple rounds of cross-validations are performed using different data partitions and results are averaged over all partitions.

In statistical prediction, the following three cross-validation methods are commonly used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling test and jackknife test. However, of the three test methods, the jackknife or leave-one-out test is considered least arbitrary and can yield a unique result for a given benchmark dataset whereas the other two test methods bear considerable arbitrariness, as elaborated by Chou[40]. For example a predictor achieving a higher success rate than other predictors for a given blind or independent testing dataset might fail to repeat the performance when tested by another blind or independent testing dataset. Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors[41,42].

In the present study, we have used *leave-one-out* cross-validation approach. It partitions entire data into N (=number of sequences in dataset) number of training and test set pairs. In each pair, training set contains all except one sequence, while testing set contains the sequence absent in training set. At a selected parameter, model is generated using the training set and prediction performance was evaluated on corresponding test set. During training for 1st level predictor, NRPs were considered as positive while non-NRPs were considered as negative. During 2nd level i.e. sub-family prediction, only NRPs were used during both training and testing. Proteins belonging to same sub-family were labeled as positive while remaining all proteins as negative.

For performance evaluation we used standard parameters regularly used in other similar classification and prediction works[34,43–47] viz. sensitivity, specificity, accuracy and MCC. These parameters defined by:

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \quad (1)$$

$$Specificity = \frac{TN}{TN+FN} \times 100 \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (3)$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

Where, TP represents true positive, TN represents true negative, FP represents false positive, FN represents false negative and MCC represents Matthew's correlation coefficient.

The schema of categorizing a prediction into different categories can be summarized as follows: At 1st level (Figure 1), TP represents the number of proteins, which are actually NRPs and also predicted as NRPs. TN represents the number of proteins which are actually non-NRPs and also predicted as non-NRPs. FP is number of non-NRPs, predicted as NRPs while FN is number of proteins which are actually NRPs predicted as non-NRPs. At 2nd level (Figure 1) since the classification was done to predict the sub-family, the meaning of TP, TN, FP and FN has also changed accordingly. For a hypothetical sub-family X, TP is the number of sequences correctly predicted to belong to sub-family X; TN is the number of non-family sequences predicted as non member of sub-family X; FP is the number of sequences wrongly predicted to belong to sub-family X while FN is the number of sequences which actually belong to sub-family X but predicted as non-family protein.

Support Vector Machine. In this study, we implemented SVM using SVM_light package[48], which allows us to choose a number of parameters and kernels (e.g. linear, polynomial, radial basis function, sigmoid or any user-defined kernel). SVM models were generated using different parameters and kernels. SVM model, which had best performance during LOOCV, was selected as the optimal model.

Amino Acid Composition. It is the fraction of each amino acid present in a protein, encapsulated in a vector of 20 dimensions. In the earlier studies also amino acid composition had been used for annotating different features of proteins[46,47]. It was calculated using the expression:

$$Comp(i) = \frac{Ri}{N} X100 \quad (5)$$

Where, Comp(i) is the amino acid composition of residue type Ri and N is the total number of amino acids.

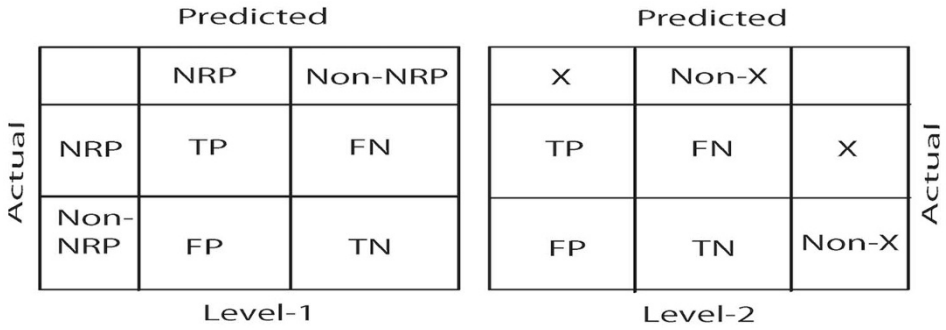

Figure 1 | Classification schema of prediction on the basis of actual and predicted state. At level-1, decision is made on the basis of whether the query protein is predicted as NRP or non-NRP. At level-2, the predicted NRP is categorized into same or different sub-family. At level-2 the schema is described for a hypothetical sub-family 'X'.

Dipeptide Composition. One of the main drawbacks of amino acid composition is that it only emphasizes on overall sequence information but ignores the local order information. In order to incorporate the local sequence order information along with amino acid composition, dipeptide composition was also used as input[34,49]. It was calculated using the expression:

$$Dipep(\mathrm{i}) = \frac{\text{Total number of Dipep}(i)}{\text{Total number of all possible dipeptides}} X100 \qquad (6)$$

Where, Dipep(i) = i-th dipeptide; i = 1 to 400.

1. Robinson-Rechavi, M., Escriva Garcia, H. & Laudet, V. The nuclear receptor superfamily. *J Cell Sci* **116**, 585–586 (2003).
2. Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: the second decade. *Cell* **83**, 835–839 (1995).
3. Evans, R. M. The steroid and thyroid hormone receptor superfamily. *Science* **240**, 889–895 (1988).
4. Danielian, 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.* **11**, 1025–1033 (1992).
5. Shiau, A. K. *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927–937 (1998).
6. Thornton, J. W. Nonmammalian nuclear receptors: Evolution and endocrine disruption. *Pure Appl. Chem.* **75**, 1827–1839 (2003).
7. Committee, N. R. N. A unified nomenclature system for the nuclear receptor superfamily. *Cell* **97**, 161–163 (1999).
8. Kliewer, S. A., Lehmann, J. M. & Willson, T. M. Orphan nuclear receptors: shifting endocrinology into reverse. *Science* **284**, 757–760 (1999).
9. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov.* **1**, 727–730 (2002).
10. Folkertsma, S. *et al.* A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol.* **341**, 321–335 (2004).
11. Aranda, A. & Pascual, A. Nuclear hormone receptors and gene expression. *Physiol Rev.* **81**, 1269–1304 (2001).
12. Bhasin, M. & Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem.* **279**, 23262–23266 (2004).
13. Gao, Y. *et al.* Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids* **28**, 373–376 (2005).
14. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**, 246–255 (2001).
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* **6**, e23505 (2011).
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* **7**, e30869 (2012).
17. Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **27**, 861–874 (2006).
18. Eng, J. Receiver operating characteristic analysis: a primer. *Acad Radiol.* **12**, 909–916 (2005).
19. Bradley, A. E. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
20. Gao, Q. B., Jin, Z. C., Ye, X. F., Wu, C. & He, J. Prediction of nuclear receptors with optimal pseudo amino acid composition. *Anal Biochem.* **387**, 54–59 (2009).
21. Robinson-Rechavi, M., Carpentier, A. S., Duffraisse, M. & Laudet, V. How many nuclear hormone receptors are there in the human genome? *Trends Genet.* **17**, 554–556 (2001).
22. Zhang, Z. *et al.* Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res.* **14**, 580–590 (2004).
23. Rottger, R. *et al.* Density parameter estimation for finding clusters of homologous proteins–tracing actinobacterial pathogenicity lifestyles. *Bioinformatics* **29**, 215–222 (2013).
24. Apeltsin, L., Morris, J. H., Babbitt, P. C. & Ferrin, T. E. Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. *Bioinformatics* **27**, 326–333 (2011).
25. Wittkop, T. *et al.* Partitioning biological data with transitivity clustering. *Nat Methods* **7**, 419–420 (2010).
26. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
27. Thomson, S. A., Baldwin, W. S., Wang, Y. H., Kwon, G. & Leblanc, G. A. Annotation, phylogenetics, and expression of the nuclear receptors in Daphnia pulex. *BMC Genomics* **10**, 500 (2009).
28. Lumba, S., Cutler, S. & McCourt, P. Plant nuclear hormone receptors: a role for small molecules in protein-protein interactions. *Annu Rev Cell Dev Biol.* **26**, 445–469 (2010).
29. Keshava Prasad, T. S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–772 (2009).
30. Francis, G. A., Fayard, E., Picard, F. & Auwerx, J. Nuclear receptors and the control of metabolism. *Annu Rev Physiol.* **65**, 261–311 (2003).
31. Chawla, A., Repa, J. J., Evans, R. M. & Mangelsdorf, D. J. Nuclear receptors and lipid physiology: opening the X-files. *Science* **294**, 1866–1870 (2001).
32. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**, 721–728 (2001).
33. Bhasin, M., Garg, A. & Raghava, G. P. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21**, 2522–2524 (2005).
34. Kumar, R., Jain, S., Kumari, B. & Kumar, M. Protein Sub-Nuclear Localization Prediction Using SVM and Pfam Domain Information. *PloS one* **9**, e98345 (2014).
35. Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. & Suwa, M. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Res.* **33**, W148–153 (2005).
36. Bhasin, M. & Raghava, G. P. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* **32**, W383–389 (2004).
37. Horn, F., Vriend, G. & Cohen, F. E. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.* **29**, 346–349 (2001).
38. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
39. Huala, E. *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **29**, 102–105 (2001).
40. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol.* **273**, 236–247 (2011).
41. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **41**, e68 (2013).
42. Mohabatkar, H., Mohammad Beigi, M. & Esmaeili, A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J Theor Biol.* **281**, 18–23 (2011).
43. Kumari, B., Kumar, R. & Kumar, M. PalmPred: An SVM Based Palmitoylation Prediction Method Using Sequence Profile Information. *PloS one* **9**, e89246 (2014).
44. Kumar, M., Gromiha, M. M. & Raghava, G. P. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit.* **24**, 303–313 (2010).
45. Kumar, M. & Raghava, G. P. Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics* **10**, 22 (2009).

46. Kumar, M., Gromiha, M. M. & Raghava, G. P. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* **8**, 463 (2007).
47. Kumar, M., Verma, R. & Raghava, G. P. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J Biol Chem.* **281**, 5357–5363 (2006).
48. Joachims. T. *Making Large Scale SVM Learning Practical.* (MIT Press Cambridge, 1999).
49. Bhasin, M. & Raghava, G. P. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* **32**, W414–419 (2004).

Acknowledgments

This work was funded by Science & Engineering Research Board (SERB), Department of Science & Technology, Government of India under Fast Track Scheme for Young Scientist (Grant no. SR/FT/LS-84/2010). Ravindra Kumar is supported as a Senior Research fellow grant number (20-12/2009(ii)EU-IV) from the University Grant Commission of India. Bandana Kumari is supported by SERB grant and Abhishikha Srivastava is supported by Indian Council of Medical Research grant (Grant no. AMR/17/2011-ECD-1). We gratefully acknowledge Dr. Neelja Singhal for critically reading the manuscript.

Author contributions

M.K. conceived and designed the experiment. R.K., B.K. and A.S. performed the experiment. The data was analyzed by M.K., R.K., B.K. and A.S. M.K., R.K. and B.K. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kumar, R., Kumari, B., Srivastava, A. & Kumar, M. NRfamPred: A proteome-scale two level method for prediction of nuclear receptor proteins and their sub-families. *Sci. Rep.* **4**, 6810; DOI:10.1038/srep06810 (2014).