Netw Model Anal Health Inform Bioinforma (2016)5:21
DOI 10.1007/s13721-016-0129-2

ORIGINAL ARTICLE

DeepLNC, a long non-coding RNA prediction tool using deep neural network

Rashmi Tripathi[1] · Sunil Patel[1] · Vandana Kumari[1] · Pavan Chakraborty[1] · Pritish Kumar Varadwaj[1]

Received: 22 April 2016 / Revised: 26 May 2016 / Accepted: 29 May 2016

Abstract The significant role of long non-coding RNAs (lncRNAs) in various cellular functions, such as gene imprinting, immune response, embryonic pluripotency, tumorogenesis, and genetic regulations, has been widely studied and reported in recent years. Several experimental and computational methods involving genome-wide search and screenings of ncRNAs are being proposed utilizing sequence features-length, occurrence, and composition of bases with various limitations. The proposed classifier, Deep Neural Network (DNN) is fast and an accurate alternative for the identification of lncRNAs as compared to other existing classifiers. The information content stored in *k-mer* pattern has been used as a sole feature for the DNN classifier using manually annotated training datasets from LNCipedia and RefSeq database, obtaining accuracy of 98.07 %, sensitivity of 98.98 %, and specificity of 97.19 %, respectively, on test dataset. The *k-mer* information content generated on the basis of Shannon entropy function has resulted in improved classifier accuracy. This classification framework was also tested on known human genome dataset, and the framework has successfully identified known lncRNAs with 99 % accuracy rate. The said algorithm has been implemented as a web prediction tool, which is available on server interface http://bioserver.iiita.ac.in/deeplnc.

Keywords Long non-coding RNAs (lncRNAs) · Machine learning · Deep neural network · *K-mer* features · Shannon entropy

1 Introduction

Ribonucleic acid (RNA) is a macromolecule which stores genetic information embedded in the form of sequence of nucleotide bases along a nucleic acid chain, thus passing it from one generation to another generation with high flexibility and high fidelity. Various types of RNA can be characterized by series of contrasting features like their activation, modification, transportation, and digression profiles which play crucial roles in central dogma (Harries 2012). On the basis of the function performed, RNA has been primarily classified into three types, viz. messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), acting as information-carrying intermediates in protein synthesis machinery (Berg et al. 2007). Moreover, the genetic material is also classified as coding RNAs and non-coding RNAs (ncRNAs), depending upon the fact that many of these transcripts do not encode for protein, instead, function as ncRNAs. Moreover, with the advancement in high-throughput RNA sequencing technologies (RNA-seq), ncRNAs can further roughly be classified as snRNAs (small-nuclear RNAs), snoRNAs

Electronic supplementary material The online version of this article (doi:10.1007/s13721-016-0129-2) contains supplementary material, which is available to authorized users.

✉ Pritish Kumar Varadwaj
pritish@iiita.ac.in

Rashmi Tripathi
rashmi.tripathi12@gmail.com

Sunil Patel
snlpatel001213@hotmail.com

Vandana Kumari
kvandana1992@gmail.com

Pavan Chakraborty
pavan@iiita.ac.in

[1] Department of Bioinformatics, Indian Institute of Information Technology-Allahabad, Allahabad, UP, India

Published online: 10 June 2016

(small-nucleolar RNAs), siRNAs (short-interfering RNAs), shRNAs (short-hairpin RNAs), miRNAs (micro-RNAs), piRNAs (piwi-interacting RNAs), circRNAs (circular-RNAs), and lncRNAs (long-non-coding RNAs) (Lasda and Roy 2014; Hüttenhofer et al. 2005). The exact number of ncRNAs encoded within the human genome is cryptic. Genome transcriptome analysis shows that more than 70 % of genome is likely to be transcribed into ncRNAs, whereas protein-coding transcripts account for only ~2–3 % of the genome (Lertampaiporn et al. 2014). On the basis of transcript size, ncRNAs are further grouped into two classes—(a) small ncRNAs (miRNA, piRNA, siRNA, spliRNAs, sdRNAs), and (b) long non-coding RNAs (lncRNAs). Small ncRNAs can be classified into their respective classes by their shared homology and common functions. In contrast, long ncRNA genes represent a disparate set of loci related only by their size, share no conserved sequence homology, and have variable functions (Wright 2014). LncRNA is a broad category which shelters locus biotypes based on their location with respect to protein-coding genes as intergenic lncRNAs, intronic lncRNAs, sense lncRNAs, and antisense lncRNAs (Ponting et al. 2009).

Several ncRNAs have been discovered and annotated in genomes of microorganisms, plants, and mammals those with environmental, agricultural, economical, social and health significance. Many of these reported ncRNAs, which were overlooked in the past and once categorized as "dark matter or junk", are now being vigorously studied due to their physiological and pathological significance. Recent studies have further demonstrated that ncRNAs might act as stable and functional part of a genome, including number of functions like regulation of transcription and translation, RNA modification, and epigenetic modification of chromatin structure, RNA interference, RNA splicing, and RNA editing (Baker 2011; Morris and Mattick 2014). LncRNAs are significantly involved in various functions, such as gene imprinting, immune response, embryonic pluripotency, tumorogenesis, genetic regulation, and body-axis patterning (Hu et al. 2011). *Xist* (X-inactive specific transcript), a well-characterized member of the annotated class of functionally distinct transcripts of the family lncRNA, is involved in X-chromosome inactivation in female placental mammals. Similarly, *HOTAIR* (HOX antisense intergenic RNA) governs epigenetic regulation by the mode of chromatin-modification and gene silencing regulation (Kung et al. 2013). Handful of experimental research and study has implicated participation of lncRNAs in wide range of diseases. For example, lncRNAs have recently been reported to play significant role in the cause of non-Alzheimer's dementia (NAD) and short post-mortem interval Alzheimer's (PMI) disease (Clement et al. 2016). Similarly, the aberrant expression of lncRNAs like, *BIC* (B cell Integration Cluster) and *HIS-1* (Histatin-1) is responsible for causing cancer (Singh and Prasanth 2013). Moreover, lncRNA *MIAT* (Myocardial infarction associated transcript) participates in the single nucleotide polymorphisms (SNPs) associated disease like myocardial infarction, whereas lncRNA *ANRIL* (Antisense non-coding RNA in the INK4 locus) is linked with coronary artery disease and atherosclerosis (Pasmant et al. 2007).

2 LncRNAs and associated databases

Although thousands of lncRNAs have already been reported and profiled on a genome-scale basis, there exists a huge number which are yet to be identified and to be annotated. It is a cumbersome task to distinguish protein-coding transcripts from non-coding transcripts. Despite this fact, the advancement in high throughput research has led to the explosion of lncRNA databases and prediction servers. Few of these databases have been discussed in this section, viz. the long non-coding RNA database (lncRNAdb) which provides comprehensive annotations of eukaryotic functional lncRNAs, whereas LNCipedia is a database for human transcripts and genes with approximately 113,513 human annotated lncRNA transcripts culled from different sources (Volders et al. 2013). Similarly, NONCODE dataset contains collection of lncRNAs from the published literature with integration of RefSeq and Ensembl genome browser (Xie et al. 2014). The Functional lncRNA database stores mammalian long non-protein-coding transcripts including isoforms from human, mouse, and rat (Niazi and Valadkhan 2012), whereas database namely HGNC stores ncRNA genes and pseudogenes (Wain et al. 2004). The LncRNADisease database stores more than 1000 lncRNAs and associated disease information (Chen et al. 2013). Databases like lncRNAtor (Park et al. 2014) and ncRNAimprint (Zhang et al. 2010) store comprehensive information about lncRNAs. Furthermore, the PLncDB deals with the information related to plant lncRNAs (Jin et al. 2013), and lncRNome database has similar information stored for human (Bhartiya et al. 2013). The ncFANs web server is useful for functional annotation of lncRNAs (Liao et al. 2011), whereas DIANA-LncBase hosts elaborated information for both predicted and experimentally verified miRNA–lncRNA interactions (Paraskevopoulou et al. 2013). ChIPBase stores information about transcription factors, their binding sites, expression profiles, and their interaction with other lncRNAs, lincRNAs, miRNAs, snoRNAs, tRNAs, snRNAs, and many more. They also contain information about protein-coding genes from ChIP-seq data (Yang et al. 2013).

3 Related works

Genome-wide searches and screening of ncRNAs are being carried out through various experimental and computational methods. Experimental method includes identification of lncRNAs through cDNA libraries, i.e., these methods are based on the fact that the expression of most of these ncRNAs is lower than other protein-coding transcripts. Unique reported features which distinguish ncRNAs from coding RNAs are sequence length (Prensner and Chinnaiyan 2011), origin (Washietl and Hofacker 2007), biogenesis (Nie et al. 2012), ORF length (Ma et al. 2012), and interactions with other ncRNAs, like miRNA (Amaral et al. 2011).

3.1 Overview of computational analysis tools

However, the reported trivial experimental approaches have certain limitations. The ncRNA species, which exceeds in size range, cannot be directly analyzed. They should be cleaved into smaller pieces prior to the analysis. Similarly, the in vitro cloning of lncRNAs has its own limitations too. It might not always be possible to reverse transcribe a ncRNA into cDNA because of the complex structure of the former, and it is very difficult to identify all such lncRNAs of a cell type or organism, since the cut-off size of such sample often fluctuates in the range of 20–500 nucleotides (Granovskaia et al. 2010).

The computational identification methods complement experimental methods in quickly identifying ncRNAs in new genomes. Various rule-based and supervised learning-based computational approaches have been proposed in the past, to predict ncRNA, utilizing series of features based on sequence statistics. One such rule-based method uses predefined probes from conserved intergenic and intragenic region to identify potential ncRNA transcripts to determine the expression level of lncRNA (Babak et al. 2005). Another method predicts the potential lncRNAs involved in oncogenesis, by performing lncRNA gene expression profile analysis in tumor and adjacent non-tumor (NT) tissues (Zhu et al. 2014). In the run, microarray data have also been used to identify ectopic and eutopic endometrial lncRNA and mRNA expression levels in patients, predicting lncRNAs functions with the help of co-expressed mRNA annotations (Jiang et al. 2014). The Arraystar LncRNA Expression and Promoter microarrays allow profiling of protein binding or methylation sites at gene promoters for both lncRNA and mRNA. Similarly, few databases like ncFANs (Liao et al. 2011) and NRED (ncRNA Expression Database) (Dinger et al. 2009) use the abundant pre-existing microarray data for functional annotation of lncRNA, utilizing the aspects of coding-non-coding gene co-expression (CNC) network and filtering expression data based on probe characteristics values of the expression data, respectively. Large-scale expression profiling datasets, such as SAGE (serial analysis of gene expression) technology, produce large numbers of short sequence tags and is capable of identifying and investigating the expression patterns of polyadenylated lncRNAs in a wide range of human tissues (such as in male germ cell) and cancerous cells (Gibb et al. 2011). Similarly, EST (Expressed Sequence Tag), a short subsequence of cDNA generated from one-shot sequencing of cDNA clone, has discovered novel transcripts in mammalian cell (Furuno et al. 2006) and bovine cell (Huang et al. 2012). Next-generation sequencing technology, RNA-seq is currently the most widely used technology in identifying lncRNAs in different organisms and for the identification of disease-causing lncRNAs (Tripathi et al. 2016). Also, there are numerous databases available for predicting the presence and structure of lncRNAs from RNA-seq data. LncRNA2Function (Qinghua et al. 2015) and NONCODE are comprehensive, user-friendly web interface resources for the functional investigation of human lncRNAs based on RNA-seq data.

RNA-immunoprecipitation (RNA-IP) is a newly developed method to identify lncRNA that interacts with specific protein. *Xist* that interacts with PRCII (polycomb repressive complex 2) was discovered using this approach (Zhao et al. 2010). Various kits available from Active Motif and Millipore rely on RNA-IP method to purify proteins and to identify bound lncRNAs in ribonucleic protein complexes. Life Technologies kits provide TaqMan quantitative PCR (qPCR) assays to precisely evaluate the expression of certain lncRNAs.

Chromatin signature-based approach (ChIP-seq) is widely used to generate genome-wide profiles of chromatin signatures. The transcribed regions are mapped to the genome, where lncRNAs' locations are determined and studied (Marques et al. 2013). On analysis of these ChIP-seq peak of transcription factors (TFs), researchers have developed web-based interface (TF2lncRNA, ChIPBase) to identify which TFs present statistically significant number of binding sites (peaks) within the regulatory region of the input lncRNA genes and to decode the transcriptional regulation of lncRNA and miRNA genes from ChIP-seq data, respectively. LncRNABase (miRNA–lncRNA interaction descriptions) views the predicted miRNA–lncRNA interactions by scanning lncRNA sequences overlapping with CLIP-seq peaks for potential miRNA targets (miRanda/mirSVR) and, then, outputs the detailed information (Yang et al. 2013). Chromatin isolation by RNA purification (ChIRP) allows high-throughput discovery of RNA-bound proteins and DNA in downstream assays using deep sequencing (Chu et al. 2011). The study has revealed that RNA occupancy sites in the genome are numerous,

sequence-specific, and focal. CHART—a hybridization-based technique which has been developed to capture RNA targets to determine the genomic binding sites of specific ncRNAs (Simon 2013).

However, these rule-based approaches discussed above remain computationally challenging. Microarray has it's limitation in the form that it is not sensitive enough to detect RNA transcripts with low-expression level, as well as annotated datasets are required to study the expressions of lncRNA. SAGE is more expensive than microarray and, therefore, is not widely employed in large-scale studies. Compared to traditional microarray technology, RNA-seq has many advantages in studying gene expression. It is more sensitive in detecting less-abundant transcripts, and identifying novel alternative splicing isoforms and novel ncRNA transcripts. The existing methods and techniques must include other signal features that may help in better characterization of lncRNAs (Wang et al. 2009).

Computational methods using stable sequence and less densely structured features have successfully identified highly conserved and low expressed lncRNAs. Learning-based methods based on ORF length strategy, sequence and secondary structure conservation strategy, and machine learning strategies have led to the development of classification tools to identify lncRNAs. The start codons and termination codons in lncRNAs are distributed randomly due to which the length of lncRNA is not more than 300 nucleotides. Using these sequence-derived features, ncRNAs can be discriminated from coding RNAs. Computational prediction approaches, such as FANTOM project (Dinger et al. 2009), CRITICA (coding region identification tool invoking comparative analysis) suite of programs (Badger and Olsen 1999), and CPAT (coding-potential assessment tool) a logistic regression model, incorporate ORF length feature to distinguish lncRNAs from other coding transcripts (Volders et al. 2013). Arraystar uses another computational method, designed in a way to filter transcripts according to known coding RNAs, small ncRNAs, and structural RNAs, at the end retaining multiexonic transcripts (>200 nt) also suggesting that lncRNAs may encode proteins in the form of small or unrecognized ORFs. LncRScan is a pipeline consisting of five steps for detecting novel lncRNAs from a set of candidate transcripts annotated by Cuffcompare. It includes a feature 'extract_ORF' in Step 3 to exclude the assemblies that have long ($\geq$300 nt) putative ORFs (Sun et al. 2012). However, this approach may cause misclassification since some lncRNAs are known to have ORFs longer than 100 codons, while some protein coding genes have fewer than 100 amino acids. In addition to small ORFs, in silico prediction of coding ORFs is further complicated by the existence of non-canonical (non-AUG) start codons. By the implication of sequence and secondary structure conservation strategy, one can also classify these lncRNAs. In contrast to the protein coding genes, lncRNAs are generally less conserved revealed by the sequence features including paucity of introns and low GC content. Codon substitution frequency (CSF) score is one of the criteria to measure the coding potential of lncRNAs (Guttman et al. 2009). Further, combining CSF with reading frame conservation (RFC) can be used to discriminate lncRNAs from mRNAs (Clamp et al. 2007). Other similar methods include PhyloCSF which uses a phylogenetic framework to build two phylogenetic codon models that can distinguish coding from non-coding regions (Lin et al. 2011). There are also methods that explore the conservation of RNA secondary structures to identify lncRNAs, including programs—QRNA, RNAz, and EvoFold. However, these approaches are limited due to lack of common conserved secondary structures specific for lncRNAs, and use of secondary structure feature is not sufficiently statistically robust enough to detect lncRNAs. This is because a random RNA with low GC content can also fold into low-energy structure.

Although the above-described methods have shown their effectiveness in identifying lncRNAs, exceptional cases are still reported. The steroid receptor RNA activator (SRA) was characterized as ncRNA previously, but the coding product was detected later. Such ambiguity will be clarified when more about lncRNA is known. Classic approaches are based on open reading frame (ORF) length, ORF conservation or structural protein domains. Owing to the complex identities of lncRNAs, recently, increasing number of machine learning-based methods are being developed. These machine learning-based methods integrate various sources of data to distinguish lncRNAs from other ncRNAs as listed in Table 1. For instance, a series of protein features, such as amino acid composition, secondary structure, and peptide length, are used to train SVM model to differentiate lncRNAs from mRNAs. Coding-Potential Calculator (CPC) uses SVM to train and classify feature datasets. This method describes long, high-quality ORFs with sequence similarity (BLASTX) to known proteins for modeling and extracting sequence features and the comparative genomics features to assess the coding potential of transcripts (Altschul et al. 1997; Rè et al. 2009). PhyloCSF is a comparative genomics method for distinguishing protein-coding from non-coding regions. An alignment-free tool called PLEK (predictor of lncRNAs and mRNAs based on an improved *k-mer* scheme), uses a computational pipeline based on SVM algorithm to distinguish lncRNAs from mRNAs, in the absence of genomic sequences or annotations (Li et al. 2014).

Table 1 Summary of few machine learning-based methods, features, dataset, dataset types, algorithms, specificity, sensitivity, and accuracy percentage which are being used to train various models for identifying lncRNAs

Method	Features	Dataset and dataset type	Algorithm	Specificity (SP), sensitivity (SN) and accuracy (ACC)	References and web server link
CONC ('coding' or 'noncoding')	Peptide length, amino acid composition, hydrophobicity, secondary structure content, percentage of residues exposed to solvent, sequence compositional entropy, number of homologs obtained by PSI-BLAST, alignment entropy	cDNAs transcripts (SwissProt) ncRNAs (RNAdb&NONCODE)	SVM	SP: 97 % SN: 98 %	(Liu et al. 2006)
CPC	ORF prediction quality, CPC number of homologs obtained by BLASTX, alignment quality, segment distribution	ncRNA datasets (Rfam 7.0, RNAdb) protein-coding RNA dataset (EMBL, UniProt/SwissProt)	SVM	SN: 99.0 %, SP: 19.0 % ACC: 97.0 %	(Altschul et al. 1997) http://cpc.cbi.pku.edu.cn/
incRNA(integrated ncRNA finder)	GC %, DNA conservation, secondary structure free energy, secondary structure conservation, protein sequence conservation, Poly-A + RNA seq (max), small RNA seq (max), Total RNA tilling array (max), Poly-A + RNA tilling array (max)	ncRNA (modENCODE consortium)	RNAz and Dynalign/SVM	SN: 89 % ACC: 97.0 %	(Lin et al. 2011) http://compbio.mit.edu/PhyloCSF
RF-based classifier	SCORE (structure, sequence, modularity, structural robustness, coding potential)	mRNAs (RefSeq), lncRNA (lncRNAdb)	RF	SN: 90.7 %, SP: 93.5 % ACC: 92.1 %	(Lertampaiporn et al. 2014)
PLEK	*k-mer*	mRNAs (RefSeq), lncRNAs (GENCODE)	SVM	SN: 94.7 %, SP: 95.8 % ACC: 95.6 %	(Li et al. 2014)

Reliable identification of lncRNAs interfaces are critical for understanding the structural bases, functional implications, and for developing effective computational methods that offer a fast, feasible as well as cost-effective way to recognize putative lncRNAs. Existing computational tools have tried to predict some ncRNA features by testing against the available experimentally validated high-throughput experiment datasets including physical interactions, genetic interactions, and phylogenetic profiles. A number of functional prediction tools have already been designed for other ncRNAs, such as ComiR: a combinatorial miRNA target prediction tool (Coronnello et al. 2012); miRDeep: an integrated application tool for miRNA (An et al. 2013); Tfold: an efficient in silico prediction of ncRNA secondary structures (Engelen and Tahi 2010); and MAGIA: a web-based tool for miRNA and genes integrated analysis (Sales et al. 2010).

Few in silico methods have been developed for predicting lncRNAs; still, it is a challenging task to mine out the potential functions for this type of large molecules. Also, among thousands of lncRNAs, only a small subset is functionally characterized, and the functional annotation of lncRNAs on the genomic scale remains inadequate. Major shortcomings which concise our knowledge in the particular field are, (1) presence of the functional elements in the primary sequence of non-coding genes; (2) poorly conserved sequence of lncRNAs which hinders genomic comparison of this newly uncharacterized sequence with other coding and non-coding sequences; (3) missing collateral information, such as molecular interaction data and expression profiles related to lncRNAs; (4) less association between the secondary structure and function of lncRNAs; (5) structural identification of lncRNAs based on low-energy structures lacks statistical robustness, because a random RNA with high GC content can also fold into a low-energy structure (Prensner and Chinnaiyan 2011; Goff and Rinn 2015; Rinn 2014; Sacco et al. 2012).

Putting all these features together and using the data as input from the database is a simple and initial step. However, accomplishing the full potential of the data for biologically meaningful interpretations needs algorithms, that can automatically extract regularities from the data.

Machine learning-based methods are well-equipped in solving computational biology problems with higher accuracy. They have potential to deal with large datasets with high-dimensional value, and have—flexibility in modeling diverse sources of data.

4 Model discussion and testing

The study aims to develop a generalized deep neural network (DNN) classifier: DeepLNC for discriminating lncRNAs from coding RNAs (mRNAs). To classify the long member of ncRNA efficiently, we have considered *k-mer* pattern as the sole characteristic of lncRNA and calculated the *k-mer* entropy content for feature set generation. Several classifiers based on variable combinations of *k-mer* features were utilized to decide the final set of features utilizing forward selection-backward elimination (FSBE) feature selection strategy (Sutter and Kalivas 1993). The term *k-mer* is meant for the combination of possible substrings of length *k* in full string of a candidate training and test sequence (Haubold et al. 2005). The main concept behind selecting the *k-mer* information content is to harness the intricacy of *k-mer* pattern occurrence vs. the overall information content of the complete sequence, which is crucial for genomic enrichment of any given sequence (Akhter et al. 2013).

Our method utilizes DNN, a probabilistic machine learning approach to identify and learn the probable *k-mer* combinations ($n = 2, 3, 4, 5$) with a many-to-one target mapping. To reduce the complexity of the data and to produce the counts of all non-unique *k-mer*s, Shannon entropy method was used. Shannon entropy is one of the most important metrics applied in the information theory denoted by H and given as,

$$H = -\sum_i p_i \log_b p_i \quad (1)$$

where p_i is the probability of occurrence of *k-mers* in a transcript. DeepLNC approach reduced the memory requirements for counting *k-mer*s as compared to other traditional machine learning algorithms with its distributed implementation on multi-thread H_2O platform. The current tool was validated to infer its improved performance (accuracy, specificity, sensitivity) over previously used tools such as CONC, CPC, incRNA, and PLEK. We compared our tool DeepLNC with PLEK (which showed higher accuracy than CONC, CPC and incRNA) using our own constructed training dataset. The performance of PLEK showed 75 % accuracy for predicting mRNAs, whereas DeepLNC predicted 82 % of mRNAs correctly. 98 % of lncRNAs were identified using PLEK, whereas 99.8 % of lncRNAs were identified by our proposed method.

4.1 Datasets

Two datasets were taken into account: training and testing as shown in Fig. 1. Training dataset comprised sequences and annotations extracted from the LNCipedia database (http://www.lncipedia.org/db) version 3.1, which is the latest version of lncRNA database containing 111,685 human annotated lncRNAs which provides comprehensive annotations of eukaryotic lncRNAs. Testing dataset comprised of mRNA transcripts from RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) database. The positive training dataset was composed of 80,214 human long-non-coding transcripts from LNCipedia. The negative training dataset included 99,395 protein-coding transcripts from RefSeq.

To remove redundancy in the dataset, Shannon Entropy feature based distance function was calculated. The function measures inter sequence-similarity between the two datasets. Euclidean distance was calculated between each transcript on the basis of entropic information content values obtained by above calculation. Top 30,000 transcripts of each positive and negative dataset were extracted which showed maximum dissimilarity between them, and rest were discarded. Training and testing datasets were taken in 1:1 ratio, each consisting of 30,000 sample datasets in total, where 15,000 sequences are of lncRNA and 15,000 sequences are of mRNA, respectively. A balanced ratio between the number of positive and negative training set is maintained throughout to improve the performance of the classifier.

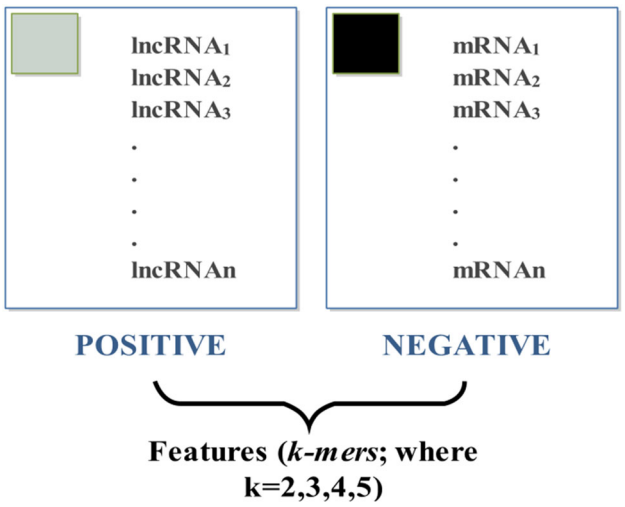


Fig. 1 Feature consists of various ranges of *k-mer* combinations and their information content. Data comprise 30,000 human lncRNAs from LNCipedia and 30,000 mRNA transcripts from RefSeq. Information content was calculated by comparing the *k-mer* values using Shannon entropy

4.2 Features extraction

To further proceed with the result, both datasets were processed, and the information content was calculated for *k-mer* ($k = 2, 3, 4, 5$) patterns for each transcript. The best *k-mer* combination was chosen among the possible four sets [(i) 2, 3; (ii) 2, 3, 4; (iii) 2, 3, 5; and (iv) 2, 3, 4, 5] on the basis of accuracy. The selection of discriminative feature subset was done using 'Forward selection, Backward elimination' (FSBE) method (Thangaiah et al. 2009). While testing the independent dataset using model prepared from all combinations of *k-mer* using FSBE, we found that *k-mer* combinations of 2, 3, 5 exhibited improved and accurate result in lncRNA identification. However, the DeepLNC has limitation in checking the higher-order of *k-mer* (6, 7, 8…) which could lead to tremendous increase in the feature space which is beyond the computational dimension of our machine. The obtained result analysis in the form of accuracy against all the combinations of *k-mer* has been illustrated in Fig. 2 and Table 2 shows the number of feature combinations for each value of *k*.

Table 2 Number of feature combinations for each value of *k*

Value of *k*	Number of possible combinations (4^k)	Total
2	4^2	16
3	4^3	64
4	4^4	256[a]
5	4^5	1024
Total features = 1104		

[a] Excluded from the feature space by applying FSBE method due to poor performance and low accuracy rate, as shown in Fig. 2

4.3 Machine learning algorithm (deep neural network)

We have trained our dataset using DNN as the main classifier. DNN is a set of machine learning algorithms (supervised or unsupervised), heavily dependent on the choice of data representation (or features), on which they are applied to learn various layered model of non-linear operational input. The applications may prove to be multifunctional and involves pattern recognition, statistical clssification, convolutional deep neural networks, and deep belief networks using a mixture of manual annotations, experimental analysis, and computational biology methods. We have hypothesized that certain significant features would improve the characterization of heterogeneous lncRNAs. Therefore, DNN is based on six set of new composite significant features: sensitivity, specificity, precision, ROC score (AUC), Matthews correlation coefficient, and Youden's index (J). We developed a characteristic *k-mer* based algorithm, which is independent of age-old techniques of alignment, to solve the above mentioned problem.

4.4 Implementation of DeepLNC

DeepLNC takes calibrated *k-mer* frequencies of lncRNAs and coding transcript sequences as its computational features. The *k-mer* based-features and implication of DNN algorithm were used to build a binary classification model to separate lncRNAs from mRNAs. The classification model achieved high accuracy rate of (98.07 %) on the training dataset (where *k-mer* combination was 2, 3, 5) with tenfold cross-validation. The proposed DNN algorithm efficiently handles non-linearity in data using fewer parameters and better hierarchical layer-wise function compression. It facilitates global error correction within multiple weight layers with the use of accelerated gradient learning algorithm. The inclusion of advance optimization algorithms, such as, Adaptive Learning—ADADELTA (Zeiler 2012), Dropout (Wager et al. 2013) and Nesterov's Accelerated Gradient (Nesterov 2007) enabled the minimal chance of over fitting, improved rate of fast error minimization and high predictive accuracy through the layers of

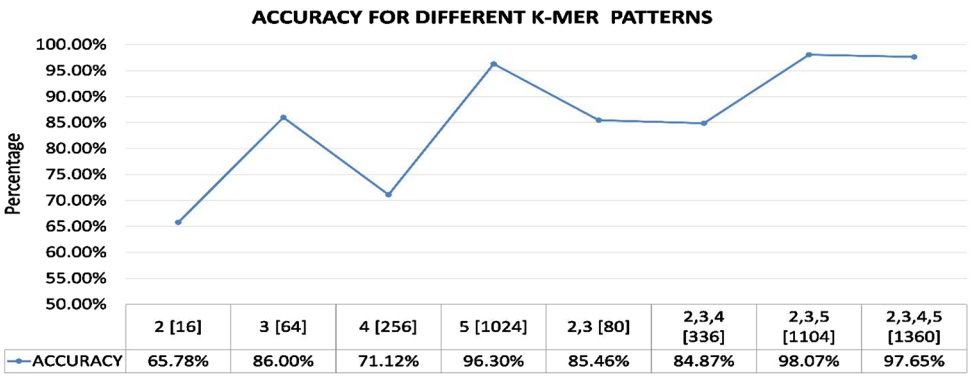


Fig. 2 *Plot* showing the performance of the classifier using different *k-mer* combinations

DNN. The DNN was implemented using Oxdata H_2O. It is an open source predictive analytics platform which uses Hadoop to perform mathematical analysis. Oxdata H_2O was connected with R for processing data using REST API. Deep learning is based on a multi-layer feed-forward ANN that is trained using back propagation with stochastic gradient descent. As our data were non-linear in nature, we used tan h activation function. Advanced features enabled high predictive accuracy. The prediction performance of our classifier DNN was evaluated using standard matrices including several standard performance measures as described below in Eqs. 2, 3, and 4.

The efficiency of the classifier was further evaluated using few quantitative variables: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) which were calculated for the testing data. TP are the correctly predicted lncRNAs, TN are the coding RNAs correctly classified from negative dataset, FP are negative entities which fall falsely into lncRNAs group, whereas FN are the cases where actual lncRNAs get incorrectly classified as non-lncRNAs.

Accuracy (ACC), sensitivity (SN), precision or positive prediction value (PPV), and specificity (SP) metrics, which indicate the accuracy of a prediction system of the classifier to classify lncRNAs, were calculated using Eqs. 2, 3, 4, and 5, and receiver operating characteristic curve (ROC) was plotted for the same. False Discovery Rate (FDR), False Negative Rate (FNR), Fall-out or False Positive Rate (FPR), Negative Predictive Value (NPV) were calculated using the formulas given in Eqs. 6, 7, 8, and 9, respectively.

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Sensitivity (SN) or true positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision or positive prediction value (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Specificity (SP)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{False discovery rate (FDR)} = 1 - \text{positive prediction value} \tag{6}$$

$$\text{False negative rate (FNR)} = 1 - \text{true positive rate} \tag{7}$$

$$\text{Fall} - \text{out or false positive rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{8}$$

$$\text{Negative predictive value (NPV)} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{9}$$

To show overall performance of our classifier system DNN, to discriminate positive dataset from negative dataset and to correctly classify lncRNAs, Matthews correlation coefficient (MCC) and Youden's index (J) were computed as shown in Eqs. 10 and 11.

$$\text{Matthews correlation coefficient (MCC)} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{10}$$

$$\text{Youden's index (J)} = (\text{Sensitivity} + \text{Specificity}) - 1. \tag{11}$$

DNN was implemented on Hadoop, utilizing H_2O Oxdata framework which facilitates the general mathematical analysis using R language. Hadoop is an Apache framework for big data analytics, and it is an integral part of H_2O Oxdata module for doing mathematical computation (http://hadoop.apache.org/). H_2O's DNN was based on a multi-layer feed-forward artificial neural network (ANN) which is trained using back-propagation algorithm. The main idea behind successful implementation of DNN was to train the layers of the network one at a time i.e., network with 1st hidden layer is trained, followed with 2nd hidden layers, and so is done for each layer of the network. We took old network with $n - 1$ hidden layer, and added an additional nth hidden layer at each step. nth layer takes input from previous n-1 layer, which has already been trained. DNN can be used as supervised and non-supervised learning methods. When all layers of the network get trained individually, resulted weight from training layer is used to initialize the weight in the final/overall network. This approach is known as greedy layer-wise training.

For the proposed work, we used tan h (rescaled version of the sigmoid function) as an activation function. In case of tan h, the activation value is symmetric lying within the range of-1 to $+1$ as given in Eq. 12.

$$\tan \text{h } (\gamma) = \frac{e^{\gamma} - e^{-\gamma}}{e^{\gamma} + e^{-\gamma}} \tag{12}$$

Similar to the back-propagation algorithm, the weight and bias were updated based on first-order gradient information as shown below in Eqs. 13 and 14.

$$W_{t+1}^{\ell} = W_t^{\ell} + \mathcal{E}\Delta W_t^{\ell} \tag{13}$$

and,

$$b_{t+1}^{\ell} = b_t^{\ell} + \mathcal{E}\Delta b_t^{\ell} \tag{14}$$

where, W_t^{ℓ} and b_t^{ℓ} are weight matrix and bias vector after tth update at layer ℓ in a network with learning rate ε.

$$\boldsymbol{W}_{t+1}^{\ell} = \rho \boldsymbol{W}_t^{\ell} + (1 - \rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\boldsymbol{W}_{t+1}^{\ell}} j(W, b; O^m, y^m) \tag{15}$$

$$b_{t+1}^{\ell} = \rho b_t^{\ell} + (1 - \rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{b_{t+1}^{\ell}} (W, b; O^m, y^m). \tag{16}$$

The average weight matrix gradient and the average bias vector gradient at iteration $(t+1)$ are given in Eqs. 15 and 16, respectively. Where O and y, are observation and input vector, respectively. $j(W,b;O^m,y^m)$ represents loss function given $\{\boldsymbol{W},\boldsymbol{b}\}$ as model parameters for sample size M, $(0 \leq m < M)$. Convergence can be reached much faster if model update is made based on all previous gradients instead of only current one. ρ is a factor called as momentum, which remains between 0.9 and 0.99.

For binomial classification, as in our case, the output neuron represents a class $i \in \{1,\ldots,C\}$, where C represents number of classes. The probability $P_{\text{dnn}}(i|O)$, classifies the output value for the ith output neuron of class$_i$.

Actual output at perceptron P at any layer ℓ having n perceptron in the previous layer can be calculated as shown in Eq. 17.

$$\text{class}_i = P_{\text{dnn}}(i|O), = O_P^{\ell} = \tanh(\sum_{n=1}^{n=n} y_n^{\ell-1} W_n^{\ell-1,\ell} - b_P^{\ell}) \tag{17}$$

where $y_n^{\ell-1}$ is an output from the previous layer $\ell-1$, $W_n^{\ell-1,\ell}$ is a weight between all perceptrons of $\ell-1$, and perceptron P at layer ℓ.b_n^{ℓ} is a bias at layer ℓ for perceptron P. The proposed DNN classifier was trained using tenfold cross-validation to avoid over-fitting of training dataset. Further, the best model out of cross-validation was chosen as final model for model validation through test dataset. The proposed DNN consists of four hidden layers with 1000, 700, 300, and 150 perceptrons in each layer, respectively. The DNN architecture discussed above was selected through a series of trial and error runs. Outputs were marked as 1 or 0, where 1 indicates lncRNAs, and 0 stands for mRNAs. Threshold parameter of DNN learning was kept at 0.95.

To avoid over-fitting and slow convergence in proposed DNN classifier, dropout algorithm has been implemented, where the input dropout rate is 0.2, and the hidden layer dropout rate is kept at 0.5. Dropout prevents brittle co-adaptation and over-fitting by adding noise to the perceptron in network layers. The dropout technique uses a constant probability for omitting a unit along with its connections from the network. For multi-layer feed-forward network, the activity S in perceptron i of layer h can be expressed as shown in Eq. 18.

$$O_i^h = \tanh(S_i^h) = \tanh\left(\sum_{\ell<h}\sum_j W_{ij}^{h\ell} O_j^{\ell}\right) \quad \text{with} \quad O_J^{\text{O}} = I_j \tag{18}$$

where,W denotes weight, and O_i^h is the output of perceptron i of layer h, and dropout can be defined in Eqs. 19 and 20.

$$O_i^h = \tanh(S_i^h) = \tanh\left(\sum_{\ell<h}\sum_j W_{ij}^{h\ell}\,\delta_j^{\ell} O_j^{\ell}\right) \quad \text{with } O_J^{\text{O}} = I_j \tag{19}$$

$$\text{NWGM}(O_i^h) = \frac{\prod_{\mathcal{N}} O_i^{hP(\mathcal{N})}}{\prod_{\mathcal{N}} O_i^{hP(\mathcal{N})} + \prod_{\mathcal{N}}(1-O_i^h)^{P(\mathcal{N})}} \tag{20}$$

where δ_j^{ℓ} and NWGM are Bernoulli selector variable, normalized weighted geometric mean, respectively. $\mathcal{N}$ spans over all possible subnetworks formed in network as a result of dropout.

$$E\left(O_i^h\right) \approx \text{NWGM}(O_i^h) \tag{21}$$

$$\text{NWGM}(O_i^h) = (\tan h)_i^h [E(S_i^h)] \tag{22}$$

$$E(S_i^h) = \sum_{\ell<h}\sum_j W_{ij}^{h\ell} P_j^{\ell} E(O_j^{\ell}). \tag{23}$$

Equations 21, 22, and 23 are fundamental equations, which represent recursive dropout ensemble in DNN. Since, the numbers O_i^h are non-identical over possible subnetworks $\mathcal{N}$ in our DNN approach, the NWGM provides a good approximation.

Likewise, with every learning algorithm, the learning rate is of prime importance to decide the convergence of system to global minima value. It has been reported that choosing slow learning rate results in delay convergence, and a higher learning rate can cause the system to diverge in terms of objective function. To ensure a smooth learning, we have adopted ADADELTA method, which prevents the continuous decay of learning rate throughout the training, and it automatically adjusts the global learning rate (Duchi et al. 2011).

The learning parameter x is updated over a time to optimize an objective function $f(x)$

$$x_{t+1} = x_t + \Delta x_t \tag{24}$$

by ADADELTA method of updating Δx_t from time $t=0$ to $t=T$ as shown in Eq. 24.

$$\Delta x_t = -\frac{\eta}{\sqrt{\sum_{T=1}^{t} g_T^2}} g_t. \tag{25}$$

As shown in Eq. 25, the denominator accumulates the square gradient for all previous iterations, which is responsible for continuous decay of learning rate η, where g_t is a gradient of parameter at tth iteration. Also, instead of accumulating squared gradient from each iteration from beginning of training, ADADELTA accumulates it over a fixed size window (ω). This prevents denominator to grow infinite and ensure that learning continues to make progress even after many iterations. As ADADELTA uses first-order information, it requires very less computation per iteration

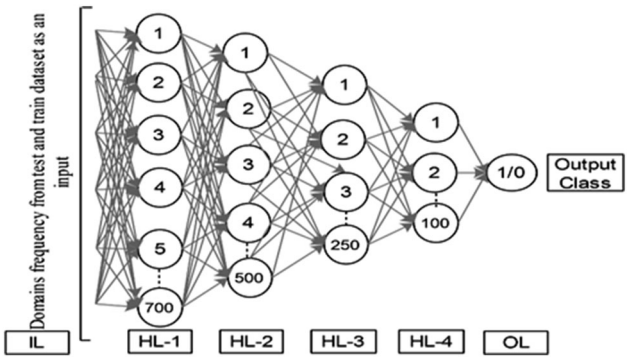


Fig. 3 A deep neural network used for training and testing (IL = input layer, HL = hidden layer, OL = output layer)

over Newton's method or quasi-Newton's methods used in shallow learning algorithm.

Further, we have used Nesterov's accelerated gradient descent, which accelerated the optimization process to reach global minima using less computational complexity over time. The plain gradient descent algorithm has a rate of convergence as $O\left(\frac{1}{t}\right)$, while the Nesterov's accelerated gradient has rate of convergence as $O\left(\frac{1}{t}\right)$.

The computational speed associated with DNN mainly depends on three factors, (1) number of hidden layers used, (2) size of input feature vector and (3) number of output nodes. Even though the proposed DNN architecture uses higher number of input nodes, the sparse nature of data matrix resulted in fewer input values to be used in weight optimization computation. Also, the number of hidden layers was fixed at '4' as given in Fig. 3. It took approximately 20 min for each fold cross-validation for training sample on i7, 2.6 GHz AMD machine with 14 GB RAM.

5 Results

Several studies have identified lncRNAs using different high-throughput techniques. However, the features of lncRNA have not yet been portrayed in well frame. Therefore, it has become more essential to develop an algorithm which is independent of already predicted features and which can be easily applied on massive dataset. Appropriate computational features are very important for classification; hence, we have tried to apply conventional *k-mer* feature in our studies. Finally, it was concluded that the proposed *k-mer* usage frequencies and entropy calculation are comparatively better features to identify lncRNA. Longer *k-mer* strings contain more information than the shorter ones.

DNN classifiers based on ten algorithms trained on the positive and negative dataset and combination of features using tenfold cross-validation on the training datasets with input dropout ratio of 0.2, hidden dropout ratio of 0.5 for each layer were evaluated. Each training datum was iterated for five times on the same training dataset. Accuracy (ACC) was highest to be 98.07 %, Matthews correlation coefficient (MCC) was calculated to be 0.968104, which indicated a good performance as the result is very much close to 1. Youden's Index (J) was found to be 0.968208 which reflects that there are very less false positives (FP) and false negatives (FN), i.e., the test is close to perfect as compared to other algorithms. Our data include a broad range of features, thus, DNN can be said to be more accurate than other machine learning-based classifiers as shown in Fig. 2. Performance and analysis of our classifier are summarized in Table 3.

We evaluated the performance under the receiver operating characteristics (ROC) curve, in which sensitivity (TPR) is plotted against a function of the false positive rate (1 − specificity) at different decisions thresholds as illustrated in Fig. 4. The area under the ROC curve describes

Table 3 Performance and analysis of DeepLNC in detecting lncRNAs

TP	14,572
TN	14,850
FP	150
FN	428
Accuracy	0.980733
Precision	0.971467
Sensitivity	0.989811
Specificity	0.971986
FPR	0.028014
NPV	0.99

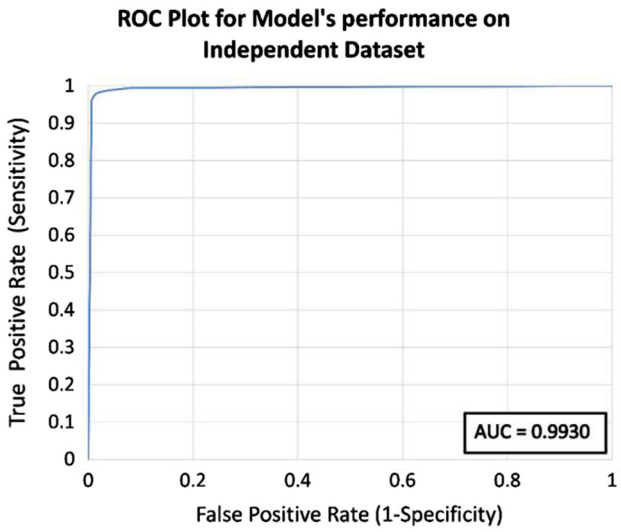
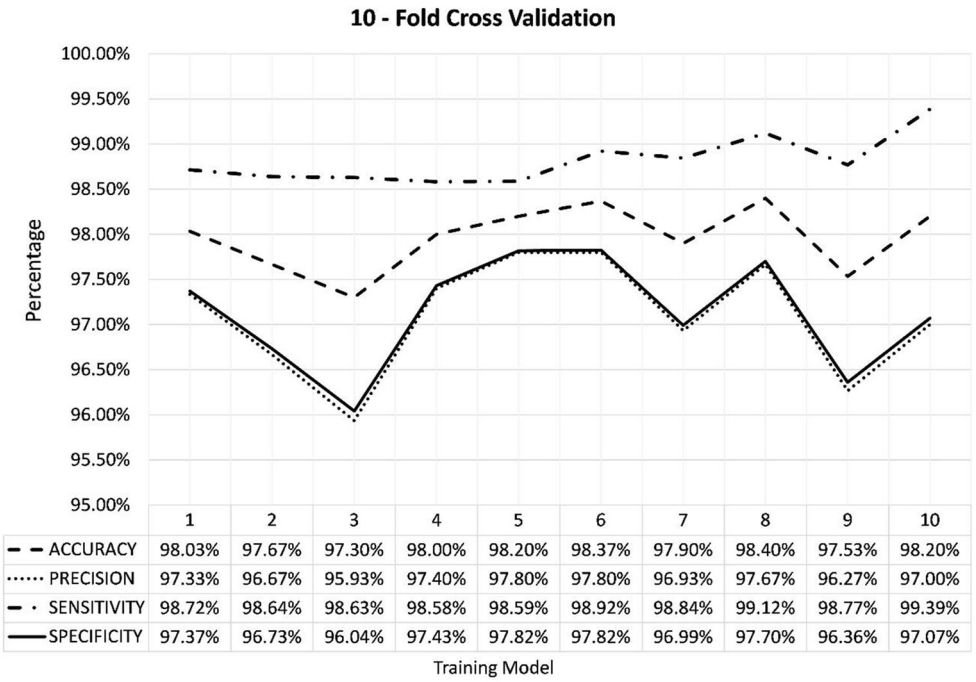


Fig. 4 The receiver operating characteristics (ROC) curve showing performance of the classifier in tenfold cross-validation



	1	2	3	4	5	6	7	8	9	10
ACCURACY	98.03%	97.67%	97.30%	98.00%	98.20%	98.37%	97.90%	98.40%	97.53%	98.20%
PRECISION	97.33%	96.67%	95.93%	97.40%	97.80%	97.80%	96.93%	97.67%	96.27%	97.00%
SENSITIVITY	98.72%	98.64%	98.63%	98.58%	98.59%	98.92%	98.84%	99.12%	98.77%	99.39%
SPECIFICITY	97.37%	96.73%	96.04%	97.43%	97.82%	97.82%	96.99%	97.70%	96.36%	97.07%

Fig. 5 Accuracy, precision, sensitivity, and specificity *curves* showing the performance of DNN in tenfold cross-validation on the training data sample

the overall performance of the method under different thresholds. The ROC value calculated was 0.9930 achieved by DeepLNC, which indicates a good classification result. Accuracy, precision, sensitivity, and specificity curves of our classifier, DNN at 10-Fold cross-validation on the training dataset has been shown in Fig. 5.

6 Discussions and conclusion

We have described a machine learning technique that uses *k-mer* combinations as a comparative feature, DNN-based classification model to detect lncRNAs. The DNN technique is more advance in its approach and functioning over other existing machine learning techniques (such as SVM, Random Forest, Bayesian Network etc.), and can productively sort many problems with more efficiency and ease. It is based on a multi-layer feed-forward ANN that is trained using back propagation with stochastic gradient descent (Bottou 2010). DNN has the advantage of robustness, because the data choice is random and is applied to learn various layered model of non-linear operational input (Yan et al. 2013). The applications may prove to be multifunctional and involve pattern recognition, statistical classification, convolutional deep neural networks, and deep belief networks using a mixture of manual annotations, experimental analysis, and computational data available (Lee et al. 2009; Krizhevsky et al. 2012).

LncRNA is a challenging class of ncRNA because of limited information available about its characteristics. The approach defined in this paper, DeepLNC exhibits high performance of accuracy and prediction rate, as well as it can prove to be helpful while predicting novel lncRNAs. Many lncRNAs have been falsely categorized as coding transcripts; thus, using DNN classifier, one can wipe out all the falsely predicted lncRNAs from the false positive datasets. DNN can also be used to characterize other ncRNAs whose knowledge is limited and conserved. More composite features can be added in this model using newly developed combinations. Genomic sequences are highly complex and cannot be fully explained by any simple statistical method; therefore, in future, we are also interested in applying this classifier on lncRNAs of different organisms particularly focusing on genomes of environmental, social, agricultural importance as well as those which possess pathological significance.

There exists strong association between lncRNA and human disease and disorder (Wapinski and Chang 2011). The understanding may further get more prevalent with the gain in deep knowledge about the characterization of these ncRNAs. Since the characterization lacks fully fledged evidence, the advent of the prediction models can be successful in reconnecting the broken thread, in upcoming years, which may further be appreciated in understanding the disease etiology. Unresolved issues like functional role of lncRNAs in a particular disease ranging from neurodegeneration to cancer, differential regulation of transcripts in condition-specific disease, delocalization, dysregulation, and mutation of lncRNAs in various biological processes can help us in identifying candidate lncRNAs and linked biomarkers for disease diagnosis, treatment, and prognosis (Chen and Gui 2013; Zhou et al. 2015). A newly developed high-quality classification approach can highlight basic

concepts lying behind lncRNA biology that still need limelight to construct a robust framework for lncRNA genetics. Thus, our main goal behind lncRNA prediction and discrimination from coding RNA is to determine whether these lncRNAs can act as useful evidence in disease detection, drug target identification or can even correctly classify the biomolecules which were once tagged as "hypothetical" due to lack of full-proved evidence. The prediction tool can help in understanding the underlying diverse range of mechanisms monitoring the role of lncRNAs in human diseases in depth.

7 Data access

DeepLNC is available as freely accessible web interface at http://bioserver.iiita.ac.in/deeplnc.

Acknowledgments We are thankful to Department of Bioinformatics, Indian Institute of Information Technology-Allahabad, India for providing the computational facility to perform the study.

References

Akhter S, Bailey B, Salamon P, Aziz RK, Edwards R (2013) Applying Shannon's information theory to bacterial and phage genomes and metagenomes. Sci Reports 3:1033

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI BLAST: A new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS (2011) LncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 39(Database issue):D146–D151

An J, Lai J, Lehman ML, Nelson C (2013) MiRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res 41(2):727–737

Babak T, Blencowe BJ, Hughes TR (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. BMC Genom 6:104

Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16(4):512–524

Baker M (2011) Long noncoding RNAs: the search for function. Nat Methods 8(5):379–383

Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry. W H Freeman, New York

Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B et al (2013) LncRNome: a comprehensive knowledgebase of human long noncoding RNAs. Database (Oxford) 2013:bat034. doi:10.1093/database/bat034

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'10, pp 177–186

Chen X, Gui Y (2013) Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics 29(20):2617–2624

Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X et al (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 41(Database issue):D983–D986

Clement C, Hill JM, Dua P, Culicchia F, Lukiw WJ (2016) Analysis of RNA from Alzheimer's Disease Post-mortem Brain Tissues. Mol Neurobiol 53(2):1322–1328. doi:10.1007/s12035-015-9105-6

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell 44(4):667–678

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF et al (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA 104(49):19428–19433

Coronnello C, Hartmaier R, Arora A, Huleihel L, Pandit KV, Bais AS et al (2012) Novel modeling of combinatorial miRNA Targeting identifies SNP with potential role in bone density. PLoS Comput Biol 8(12):e1002830 **(Print)**

Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS (2009) NRED: a database of long noncoding RNA expression. Nucleic Acids Res 37(Suppl. 1):D122–D126

Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 12:2121–2159

Engelen S, Tahi F (2010) Tfold: efficient in silico prediction of non-coding RNA secondary structures. Nucleic Acids Res 38(7):2453–2466

Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC, Bult C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Mattick JS, Suzuki H (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. PLoS Genet 2(4):537–553

Gibb EA, Vucic EA, Enfield KSS, Stewart GL, Lonergan KM, Kennett JY et al (2011) Human cancer long non-coding RNA transcriptomes. PLoS One 6(10):e25915 **(Print)**

Goff LA, Rinn J (2015) Linking RNA biology to lncRNAs. Genome Res. Cold Spring Harbor Laboratory Press 25(10):1456–1465

Granovskaia MV, Jensen LJ, Ritchie ME, Toedling J, Ning Y, Bork P, Wolfgang H, Steinmetz LM (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. Genome Biol 11(3):R24

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D et al (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458(7235):223–227

Harries LW (2012) Long non-coding RNAs and human disease. Biochem Soc Trans 40(4):902–906

Haubold B, Pierstorff N, Moller F, Wiehe T (2005) Genome comparison without alignment using shortest unique substrings. BMC Bioinform 6(1):123

Hu W, Yuan B, Flygare J, Lodish HF (2011) Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. Genes Dev 25(24):2573–2578

Huang W, Long N, Khatib H (2012) Genome-wide identification and initial characterization of bovine long non-coding RNAs from EST data. Anim Gene 43(6):674–682

Hüttenhofer A, Schattner P, Polacek N (2005) Non-coding RNAs: hope or hype? Trends Genet 21:289–297

Jiang Q, Wang J, Wang Y, Ma R, Wu X, Li Y (2014) TF2LncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-seq data. BioMed Res Int 2014:317642. doi:10.1155/2014/317642

Jin J, Liu J, Wang H, Wong L, Chua NH (2013) PLncDB: plant long non-coding RNA database. Bioinformatics 29(8):1068–1071

Krizhevsky A, Sutskever I, Hinton GE (2012) Image net classification with deep convolutional neural networks. Advances in neural information processing systems, pp 1–9

Kung JTY, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193(3):651–669

Lasda E, Roy P (2014) Circular RNAs: diversity of form and function. RNA (New York, N.Y.) 20(12):1829–1842

Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning ICML 09, pp 1–8

Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchatchawalya M (2014) Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. Nucleic Acids Res 42(11):e93. doi:10.1093/nar/gku325

Li A, Zhang J, Zhou Z (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinform 15:311

Liao Q, Xiao H, Bu D, Xie C, Miao R, Luo H et al (2011) NcFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res 39(Suppl):2

Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27(13):i275–i282

Liu J, Gough J, Rost B (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2:529–536

Ma H, Hao Y, Dong X, Gong Q, Chen J, Zhang J, Tian W (2012) Molecular mechanisms and function prediction of long noncoding RNA. Sci World J 2012(1):541786

Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP (2013) Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol 14(11):R131

Morris KV, Mattick JS (2014) The rise of regulatory RNA. Nat Rev Genet 15(6):423–437

Nesterov Y (2007) Gradient methods for minimizing composite objective function. Core discussion paper. ReCALL 76.2007076 (2007): 2007/76

Niazi F, Valadkhan S (2012) Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3'UTRs. RNA 18(4):825–843

Nie L, Wu HJ, Hsu JM, Chang SS, LaBaff AM, Li CW, Wang Y, Hsu JL, Hung MC (2012) Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. Am J Transl Res 4(2):127–150

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, Hatzigeorgiou AG (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. Nucleic Acids Res 41(D1):D239–D245

Park C, Yu N, Choi I, Kim W, Lee S (2014) lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. Bioinformatics 30(17):2480–2485

Pasmant E, Laurendeau I, Héron D, Vidaud M, Vidaud D, Bièche I (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res 67(8):3963–3969

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136(4):629–641

Prensner JR, Chinnaiyan AM (2011) The emergence of lncRNAs in cancer biology. Cancer Discov 1(5):391–407

Qinghua J, Rui M, Jixuan W, Xiaoliang W, Shuilin J, Jiajie P, Tan R, Zhang T, Li Y, Wang Y (2015) LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. BMC Genom 16(3):S2

Rè M, Pesole G, Horner DS (2009) Accurate discrimination of conserved coding and non-coding regions through multiple indicators of evolutionary dynamics. BMC Bioinformatics 10:282. doi:10.1186/1471-2105-10-282

Rinn JL (2014) LncRNAs: linking RNA to chromatin. Cold Spring Harb Perspect Biol 6(8). pii: a018614. doi:10.1101/cshperspect.a018614

Sacco LDA, Baldassarre A, Masotti A (2012) Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. Int J Mol Sci 13(1):97–114

Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C (2010) Magia, a web-based tool for miRNA and genes integrated analysis. Nucleic Acids Res 38(2). **(Print)**

Simon MD (2013) Capture hybridization analysis of RNA targets (CHART). Curr Protoc Mol Biol. doi:10.1002/0471142727.mb2125s101

Singh DK, Prasanth KV (2013) Functional insights into the role of nuclear-retained long noncoding RNAs in gene expression control in mammalian cells. Chromosome Res Int J Mole Supramole Evolut Aspects Chromosome Biol 21(6–7):695–711

Sun L, Zhang Z, Bailey TL, Perkins AC, Tallack MR, Xu Z, Liu H (2012) Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. BMC Bioinform 13:331

Sutter JMJ, Kalivas JHJ (1993) Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. Microchem J 47:60–66

Thangaiah PR, Shriram R, Vivekanandan K (2009) Adaptive hybrid methods for Feature selection based on Aggregation of Information gain and Clustering methods. Int J Comput Sci Netw Secur 9(2):164–169

Tripathi R, Sharma P, Chakraborty P, Varadwaj PK (2016) Next-generation sequencing revolution through big data analytics. Front Life Sci. doi:10.1080/21553769.2016.1178180

Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdaghet P (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 41(Database issue):D246–D251

Wager S, Wang S, Liang PC (2013) Dropout training as adaptive regularization. NIPS, pp 1–11

Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S (2004) Genew: the human gene nomenclature database. Nucleic Acids Res 32:255–257

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21:354–361

Washietl S, Hofacker IL (2007) Identifying structural noncoding RNAs using RNAz. Curr Protoc Bioinformatics. doi:10.1002/0471250953.bi1207s19

Wright MW (2014) A short guide to long non-coding RNA gene nomenclature. Human genomics. BioMed Central Ltd 8(1):7

Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res 42(Database issue):D98–D103

Yan ZJ, Huo Q, Xu J (2013) A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. International Speech and Communication Association, pp 104–108

Yang JH, Li JH, Jiang S, Zhou H, Qu LH (2013) ChIPBasea database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res 41(D):177–187

Zeiler MD (2012) ADADELTA: an adaptive learning rate method. eprint http://arXiv.1212.5701

Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH (2010) ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA 16(10):1889–1901

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song J, Kingston R, Borowsky M, Lee JT (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 40(6):939–953

Zhou M, Wang X, Li J, Hao D, Wang Z, Shi H, Han L, Zhou H, Sun J (2015) Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Mol BioSyst 11(3):760–769

Zhu J, Liu S, Ye F, Shen Y, Tie Y, Zhu J, Jin Y, Zheng X, Wu Y, Fu H (2014) The long noncoding RNA expression profile of hepatocellular carcinoma identified by microarray analysis. PLoS One 9(7):e101707. doi:10.1371/journal.pone.0101707