

# SCIENTIFIC REPORTS



OPEN

## Prediction of Immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants

Received: 14 June 2015

Accepted: 31 December 2015

Published: 10 February 2016

Kumardeep Chaudhary\*, Gandharva Nagpal\*, Sandeep Kumar Dhanda & Gajendra P. S. Raghava

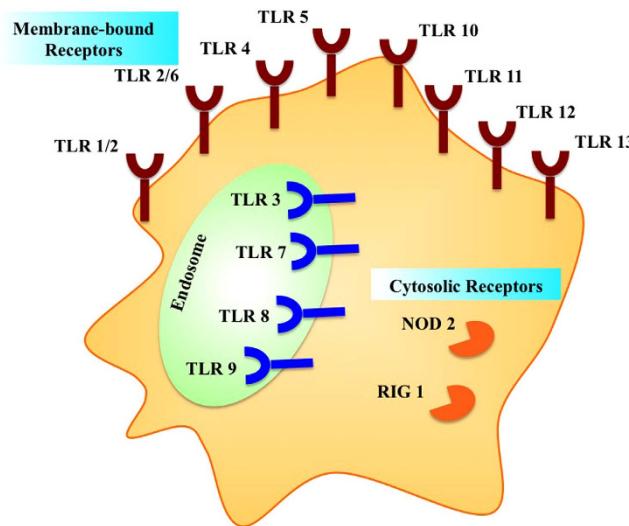
Our innate immune system recognizes a foreign RNA sequence of a pathogen and activates the immune system to eliminate the pathogen from our body. This immunomodulatory potential of RNA can be used to design RNA-based immunotherapy and vaccine adjuvants. In case of siRNA-based therapy, the immunomodulatory effect of an RNA sequence is unwanted as it may cause immunotoxicity. Thus, we developed a method for designing a single-stranded RNA (ssRNA) sequence with desired immunomodulatory potentials, for designing RNA-based therapeutics, immunotherapy and vaccine adjuvants. The dataset used for training and testing our models consists of 602 experimentally verified immunomodulatory oligoribonucleotides (IMORNs) that are ssRNA sequences of length 17 to 27 nucleotides and 520 circulating miRNAs as non-immunomodulatory sequences. We developed prediction models using various features that include composition-based features, binary profile, selected features, and hybrid features. All models were evaluated using five-fold cross-validation and external validation techniques; achieving a maximum mean Matthews Correlation Coefficient (MCC) of 0.86 with 93% accuracy. We identified motifs using MERCI software and observed the abundance of adenine (A) in motifs. Based on the above study, we developed a web server, imRNA, comprising of various modules important for designing RNA-based therapeutics (<http://crdd.osdd.net/raghava/imrna/>).

Mammalian innate immunity has dedicated sensor and effector proteins to combat viral invasion. Interferons were the first class of molecules to be discovered, acting as armor for extracellular defense against viruses<sup>1</sup>. Much later, Janeway postulated the theory of pattern recognition, propounding the existence of receptors in the host cells that recognize molecular structures associated with pathogens named pathogen-associated molecular patterns (PAMPs). To be recognized as 'foreign', the PAMPs should be present in the pathogens but not in the host cells<sup>2</sup>. These PAMPs are recognized by dedicated receptors called pattern recognition receptors (PRRs)<sup>3</sup>. In mammalian systems, the PRRs have been broadly categorized into two classes. The first category is that of membrane-bound receptors that includes the Toll-like receptors (TLRs) and C-type lectin receptors (CLRs), while the second category is composed of the cytoplasmic sensors, for example, NOD-like receptors (NLRs), RIG-I-like receptors (RLRs) and a growing list of cytosolic nucleic acid sensors<sup>4</sup> (23846113). A schematic depicting the localization of major mammalian PRRs within the cell has been shown in Fig. 1.

In case of a few viral infections, cellular defense system recognizes the pathogen from its RNA sequence. In other words, RNA sequences also serve as PAMP for viral infections. It was shown that a firm innate defense against bacterial RNA involved cytosolic detection of bacterial mRNA<sup>5</sup>. This study showed that the mammalian innate immune system can detect microbial viability by virtue of the viability-associated PAMPs (vita-PAMPs) present in the pathogen and recognized prokaryotic mRNA as a vita-PAMP. The authors thus suggested the inclusion of vita-PAMPs like mRNA in vaccine formulations to supplement the safety of dead vaccines with the superior attributes of live vaccine<sup>5</sup>.

Numerous receptors for RNA in the cytosol of human cells have been discovered. In addition to the diversity of types of RNA ligands of these receptors, the overlap of the Pathogen-associated molecular patterns in the

Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to G.P.S.R. (email: [raghava@imtech.res.in](mailto:raghava@imtech.res.in))



**Figure 1.** A diagrammatic representation of major mammalian pattern recognition receptors (PRRs) with their localization within the cell.

ligands of these receptors is also enigmatic<sup>6</sup>. Single-stranded RNA from microbial sources is known to bind endosomal Toll-like receptors 7/8 (TLR-7/8)<sup>7,8</sup> and the cytosolic receptors RIG-I<sup>9</sup> and NOD2<sup>10</sup>. Despite the knowledge that ssRNA binding with TLR-7/8 is sequence-dependent, sequence characteristics of such binding are not well defined<sup>11</sup>. Proclivity to the uridine or guanine-rich sequences including the motifs UG and UGU has been reported in case of TLR-7/8 ligands<sup>7,8</sup>. Another study demonstrates the concomitance of immunostimulatory property of RNA with AU-richness in the sequence<sup>12</sup>.

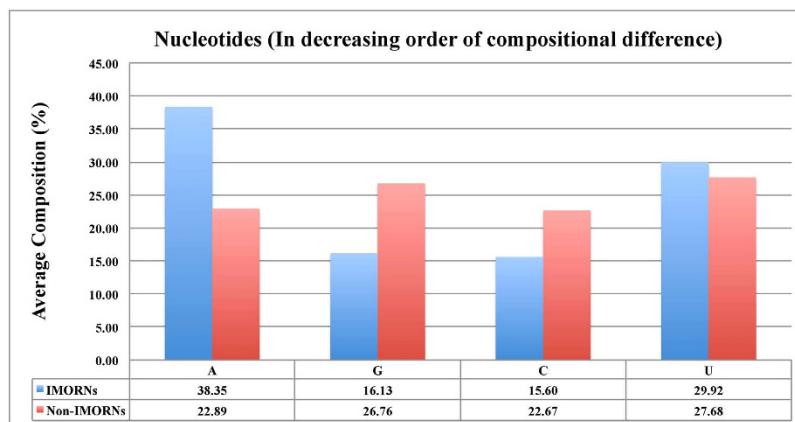
siRNAs are RNA oligonucleotides that spearhead the RNA Interference (RNAi) technology. Our cellular sensors may recognize these siRNAs as foreign RNA and may activate the immune system, which is undesirable and may cause immunotoxicity<sup>13,14</sup>. If such off-target immunotoxicities are uncontrolled, the clinical results of administering such RNAi molecules could be catastrophic<sup>15</sup>. On the contrary, in diseases like cancer and viral infections, immunostimulatory properties of siRNA could augment the therapeutic effect that relied solely on gene silencing ability. Such bifunctionality has been achieved using sequence-based modifications<sup>16,17</sup>. These studies suggest importance of single-strand RNA (ssRNA) motifs in mediating immune response due to RNA species. Previously, a method has been patented (EP1764107A1), which classifies the ssRNA based on its immunostimulatory activity.

In this study, we collected and compiled ssRNA oligoribonucleotide sequences (ORNs) whose immunomodulatory potential has been tested experimentally. Using these experimentally validated siRNAs, we developed sequence-based models for discriminating immunomodulatory oligoribonucleotides (IMORNs) and non-immunomodulatory oligoribonucleotides (non-IMORNs). The experimentally validated IMORNs have been picked up from literature, patents, and RNAimmuno<sup>18</sup> database. Finally, A web server has been developed for providing comprehensive information about a given RNA sequence in terms of its immunomodulatory potential. This server will facilitate the scientific community in designing an ssRNA sequence with desired properties.

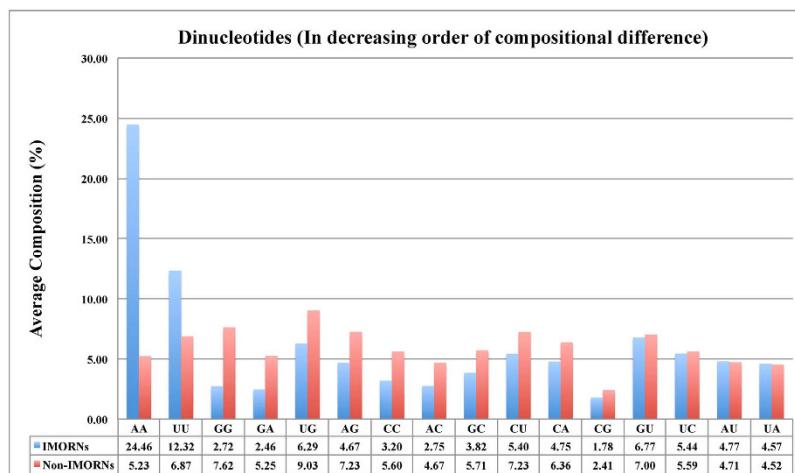
## Results

**Composition Analysis.** In order to understand the nucleotide bias in IMORNs and non-IMORNs, we computed and compared the nucleotide compositions of two types of sequences. Student's t-test between mononucleotide composition of IMORNs and non-IMORNs revealed significant compositional differences in the cases of Guanine (G) and Adenine (A) (Fig. 2 and Table S1). This was further reflected in the dinucleotide composition analysis where 3 out of top 4 motifs (AA, UU, GG, GA), abundant in IMORNs, contain these two mononucleotides (Fig. 3 and Table S2) when sorted on the basis of absolute value of difference in the mean compositions in IMORNs and Non-IMORNs. Similarly, we identified trinucleotide motifs (e.g., AAA, UUU, UAA, GGG and AAG) that have maximum difference in mean values between the negative and positive sequences (Fig. 4 and Table S3); majority of them have G and/or A. We extended our analysis to higher-order composition and identified tetranucleotides and pentanucleotides with maximum difference in mean values among immunomodulators and non-immunomodulators (Tables S4 and S5). The top 5 tetranucleotides identified in this manner were AAAA, CCCC, AAAG, CAAA and UAAA while the top 5 pentanucleotides were AAAAA, CCCCC, AAAGC, CAAAA and GCAAA. In both of these cases, most of the topmost motifs were A-rich. These observations indicated that the composition of RNA sequences could be used to classify IMORNs and non-IMORNs.

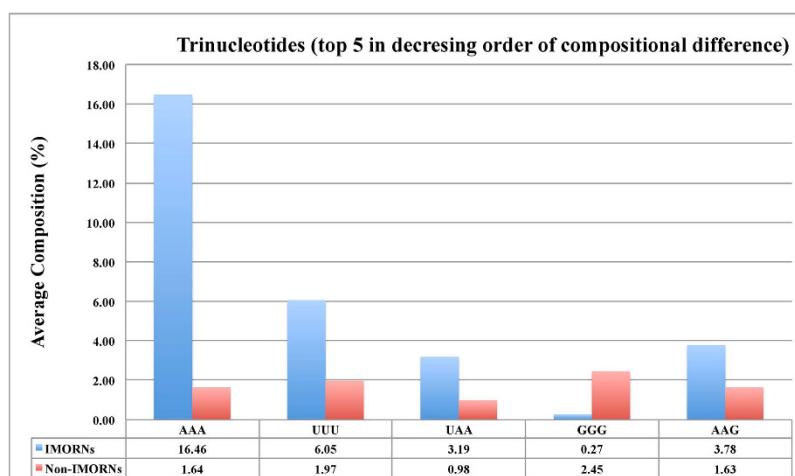
**MERCi motifs.** The role of motifs and patterns is crucial to discriminate and classify the biological sequences. Therefore, we have also tried to identify exclusive motifs in IMORNs and non-IMORNs. The Motif—EmeRging and with Classes—Identification (MERCi) program<sup>19</sup> is a powerful software that can discover motifs by comparing two types of datasets. In this study, we used MERCi with different gap lengths (no gaps, 1, 2 or 5 gap maximum length) for discovering motifs in our positive and negative datasets. It was observed that most discriminative motifs were found when maximum gap length allowed was 5 (Table S6). The best motif was “AAA-AA-AA-A” in



**Figure 2. Comparison of Mononucleotide Composition between immunomodulatory and non-immunomodulatory RNAs.**



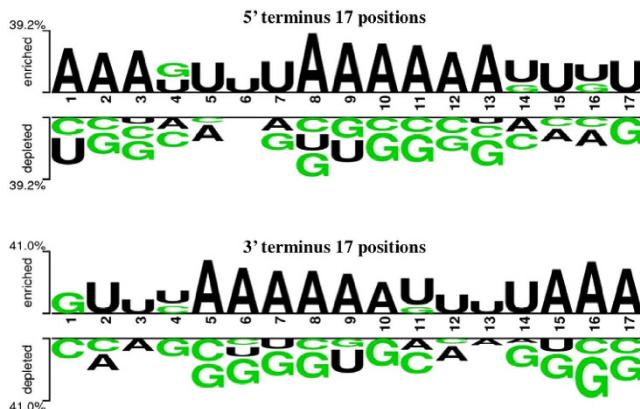
**Figure 3. Comparison of Dinucleotide Composition between immunomodulatory and non-immunomodulatory RNA.**



**Figure 4. Comparison of Trinucleotide Composition between immunomodulatory and non-immunomodulatory RNA.**

Positive motifs	Maximum Gap length	Coverage	Negative motifs	Maximum Gap length	Coverage
AAA-AA-AA-A	5	270	C-G-G-G-GG-G	5	81
AA-AAA-AA-AA	5	268	G-C-G-GGG	5	66
A-AAAA-AA-A	5	268	C-G-G-GGG	5	66
AAAAAA	0	267	C-C-C-GGG	5	62
AAA-G-A-AAA-A-G	5	266	G-C-GGG	2	61
AA-AAA-A-G-AA	2	194	C-G-GG-G	1	59
A-A-G-C-AAA	1	141	A-GC-G-GG-G	5	59
AA-G-C-A-AA	1	139	C-G-GGG	2	55
A-AAA-G-G-A-A-A	1	103	G-C-G-GG	1	54
G-AAAAA-A-A	1	98	AG-GC-G-AG	5	53

**Table 1. Distribution of the positive and negative motifs with their coverage and maximum number of gaps allowed.** The motifs were extracted using MERCI program.



**Figure 5. Two-Sample Logo of the 17 positions in the 5' and 3' termini of the immunomodulatory (enriched) and non-immunomodulatory (depleted) RNA sequences.**

the positive dataset that was observed in 270 IMORNs (Table 1). On the other hand, the best motif in negative dataset was “C-G-G-G-GG-G” with the coverage of 81 non-IMORNs.

**Two Sample Logos.** In order to visualize the position-specific preference of mononucleotides in IMORNs, we obtained the Two Sample Logos (TSL)<sup>20</sup> depiction using the tool available at <http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi>. Creation of TSL requires fixed length of all the sequences. Since the sequence length in the dataset varied from 17 to 27, we created a 17mer TSL separately for 5' and 3' termini of the sequences (Fig. 5). TSL depiction (Fig. 5) clearly shows the dominance of Adenine (A) in the IMORNs while Guanine (G) is more abundant in non-IMORNs. This corroborates well with the compositional as well as MERCI motif analysis, where motifs rich in A were found to be significantly different in occurrence between the positive and negative datasets.

**Prediction of Immunomodulatory RNA.** *Composition-based models.* The compositional analysis revealed that the average composition of certain mononucleotides, dinucleotides, trinucleotides, tetranucleotides and pentanucleotides is significantly higher in IMORNs in comparison to non-IMORNs. This means composition can be used as a feature for discriminating these two classes of RNA sequences. Thus, we developed support vector machine (SVM)-based models using composition of RNA sequences as input feature, for predicting IMORN sequences. To evaluate the performance of these models, we used internal validation and external validation techniques. Additionally, we also used a sampling approach for calculating the average performance of models on internal and external datasets. In the case of internal validation, we computed the performance of models on training-testing datasets (IMrnaTT) using five-fold cross-validation technique. The average performance of our models on internal datasets is shown in Table 2. The performance of models improved with the order of composition (from mononucleotide to pentanucleotide) and achieved the maximum average Matthews Correlation Coefficient (MCC) of 0.86 with 93.00% average accuracy using the model based on pentanucleotide composition.

**Models developed using binary vector.** In case of binary representation, we represented a mononucleotide with a vector of length four, it means a sequence of length  $N$  will be represented by a vector of length  $N \times 4$ . In the present study, we have RNA sequences of varied length thus it is difficult to represent them using fixed length vector (a requirement of SVM). To overcome this limitation, we used fixed length of mononucleotides from the termini

Features (Composition)		Thres	Average Performance of Models			
Type	Vector size		Sen	Spec	Acc	MCC
MNC	4	0	78.01 ± 1.87	77.98 ± 1.09	78.29 ± 1.45	0.57 ± 0.03
DNC	16	0	88.84 ± 1.06	89.73 ± 1.11	89.25 ± 0.84	0.78 ± 0.02
TNC	64	0	91.83 ± 0.81	91.71 ± 1.09	91.83 ± 0.83	0.84 ± 0.02
TetNC	256	0	91.85 ± 0.51	90.70 ± 1.31	91.32 ± 0.71	0.83 ± 0.02
PNC	1024	0	93.96 ± 0.67	91.88 ± 1.41	93.00 ± 0.83	0.86 ± 0.02
5' bin	68	0	89.00 ± 1.06	88.87 ± 0.79	88.94 ± 0.69	0.78 ± 0.01
3' bin	68	0	89.48 ± 1.09	88.39 ± 1.13	88.82 ± 1.24	0.78 ± 0.02
5' - 3' bin	136	0	91.70 ± 0.60	88.99 ± 1.52	90.44 ± 0.89	0.81 ± 0.02
hybrid	200	0	91.20 ± 0.84	92.48 ± 0.85	91.79 ± 0.70	0.84 ± 0.01

**Table 2. Average performance obtained from internal five-fold cross-validation of the SVM models developed for predicting immunomodulatory potential of RNA sequence using different types of features.**  
**Thres:** Threshold, **Sen:** Sensitivity, **Spec:** Specificity, **Acc:** Accuracy, **MCC:** Matthews Correlation Coefficient, **MNC:** Mononucleotide Composition, **DNC:** Dinucleotide Composition, **TNC:** Trinucleotide Composition, **TetNC:** Tetranucleotide Composition, **PNC:** Pentanucleotide Composition, **5' bin:** Binary profile 5', **3' bin:** Binary profile 3', **5'-3' bin:** Binary Profile 5'-3', **hybrid:** Binary Profile 5'-3' + Trinucleotide Composition.

Features (Composition)		Thres	Average Performance of Models			
Types	Vector size		Sen	Spec	Acc	MCC
MNC	4	0	78.42 ± 5.26	79.71 ± 3.38	79.02 ± 2.44	0.58 ± 0.04
DNC	16	0	90.50 ± 3.60	89.42 ± 2.03	90.00 ± 2.43	0.80 ± 0.05
TNC	64	0	92.58 ± 2.27	88.65 ± 2.02	90.80 ± 0.82	0.82 ± 0.02
TetNC	256	0	92.83 ± 2.52	90.58 ± 2.47	91.79 ± 1.56	0.84 ± 0.03
PNC	1024	0	95.00 ± 2.39	91.73 ± 2.09	93.48 ± 1.49	0.87 ± 0.03
5' bin	68	0	89.83 ± 2.88	88.85 ± 2.91	89.38 ± 1.83	0.79 ± 0.04
3' bin	68	0	91.25 ± 3.00	85.96 ± 3.15	88.79 ± 1.57	0.78 ± 0.03
5'-3' bin	136	0	93.92 ± 3.47	90.67 ± 3.82	92.41 ± 3.09	0.84 ± 0.06
hybrid	200	0	93.50 ± 3.23	91.44 ± 3.19	92.54 ± 2.34	0.85 ± 0.05

**Table 3. Average performance obtained from the external validation of SVM models on independent datasets developed using different types of features.**

of RNA sequences for developing binary-based models. First, we used 17 nucleotide positions from 5' terminus of RNA sequences for developing the model using binary profile of these mononucleotides; these mononucleotides were represented by a vector of length 68 ( $17 \times 4$ ). Similarly, we developed models using 3' end of RNA sequences. Finally, we developed models using both the ends of the sequences; our final model used both 5' and 3' termini sequences, achieving a mean accuracy and MCC of 90.44% and 0.81, respectively.

**Model based on hybrid features.** We also checked a hybrid model having all features of trinucleotide composition along with the binary profile of 5' and 3' termini taken together for performance. The accuracy, in this case, was 91.79% while MCC was 0.84, both being mean values of those obtained in each of the 10 sampling cases.

**Performance of models on independent dataset.** The performances of the composition, binary as well as hybrid models were evaluated on the independent datasets (IMrnaID) also called external validation datasets. It is important to validate any model on independent datasets as these have not been used in training or testing of the models. As shown in Table 3, our models performed equally well on independent dataset and achieved the mean accuracies in the range of ~79% to ~93% using different types of nucleotide compositions. Similarly, our hybrid model achieved a mean accuracy of 92.54% on independent dataset. This means our models are not over optimized, as the performance of these models is comparable in both, the internal and the external validation.

**Feature Selection.** WEKA<sup>21</sup> was used for selecting relevant features that may significantly contribute to the prediction. We selected the best features from various types of nucleotide compositions and used these selected features for developing models (Table S7). In the case of trinucleotides, WEKA selected 20 best features out of 64 and the models based on best features achieved the mean MCC of 0.75 and 0.76 on training-testing and independent datasets respectively. In case of pentanucleotides, WEKA selected 29 features out of the 1024 features and achieved poor performances on both, IMrnaTT as well as IMrnaID datasets. We also selected best features from all composition-based features (e.g., mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide) and got 81 features. These 81 features were used for developing prediction models and achieved the

Length Bin	IMORNs				non-IMORNs			
	17–20	21–23	24–27	Total	17–20	21–23	24–27	Total
# of Sequences	425	68	109	602	44	454	22	520
# of Extended Linear Structure Sequences	273 (64.23%)	32 (47.05%)	48 (44.04%)	353 (58.64%)	11 (25.00%)	115 (25.33%)	2 (9.09%)	128 (24.62%)
# of Stem Loop Structure Sequences	152 (35.76%)	36 (52.95%)	61 (55.96%)	249 (41.36%)	33 (75.00%)	339 (74.67%)	20 (90.91%)	392 (75.38%)
Total # of loops in sequences	152	36	65	253	33	340	23	396
# of loops with at least one Uridine	81 (53.29%)	31 (86.11%)	61 (93.84%)	173 (68.37%)	26 (78.79%)	247 (72.65%)	16 (69.56%)	289 (72.97%)

**Table 4. Lengthwise structure analysis of IMORNs and non-IMORNs as predicted by the RNAfold program.**

maximum MCC of 0.84 and 0.76 on training-testing (IMrnaTT) and independent datasets (IMrnaID) respectively. It is clear from the above analysis in this study that feature selection was not advantageous in this study.

**Structure Analysis.** Previous studies have implied the role of RNA secondary structure in immunostimulatory activity<sup>16,22</sup>. We performed a comparative analysis of predicted RNA secondary structures of IMORNs and non-IMORNs in our dataset using the RNAfold program available in the ViennaRNA package<sup>23</sup>. First we predicted the secondary structure of both the IMORNs as well as the non-IMORNs and then we compared the structures between the two classes. The RNAfold program predicts extended open structure (no stem-loop) for 353 out of 602 IMORNs (58.64%) and 128 out of 520 non-IMORNs (24.61%). The observation that more than half of the IMORNs have linear structure while only about a quarter of the non-IMORNs have extended linear structure might indicate that linear structures are preferred in IMORNs. For observing the dependence of RNA secondary structure on the length of the RNA sequences in our datasets, we divided the positive and the negative datasets into the bins of lengths 17–20, 21–23 and 24–27 as shown in Table 4. We observed that most of the IMORN sequences have length in the range 17–20 while the non-IMORNs are mostly 21–23 nucleotides long. Moving from the bin of shorter length to the longer ones in IMORNs, the proportion of sequences with regular secondary structure is increasing as compared to the sequences with linear structure. In contrast, the proportion of regular structures is greater across all the length bins in the non-IMORNs. In the length range 24–27, only 2 out of 22 non-IMORNs (9.09%) have linear structures while 48 out of 109 IMORNs (44.04%) are in linear conformation. Thus, looking at the proportion of sequences having stem-loop structures in both IMORNs as well as non-IMORNs, it appears that the tendency to form stem-loop secondary structure is proportional to the length of the RNA sequence. Yet, the inclination to form secondary structure appears to be more in non-IMORNs compared to the IMORNs as evident from the length bin 24–27 (Table 4). Next, we evaluated the occurrence of uridines in the stem-loop structures in IMORNs and non-IMORNs. In both the cases and across all the length bins, more stem-loops have uridines than those lacking, though their proportion seems to be increasing with length in case of IMORNs while in non-IMORNs, uridine-containing stem-loops are decreasing as the length of the sequences is increasing.

**Web interface for imRNA.** In order to provide service to the scientific community, we developed a web server that implements the above prediction models. This web server is available for free use at <http://crdd.osdd.net/raghava/imrna/>. It has a number of modules; here we provide a brief description of two major modules of this server.

**Designing of vaccine adjuvants.** This module allows the users to design RNA-based vaccine adjuvants with desired immunomodulatory potentials. It has three sub-modules developed for performing various predictions on ssRNA. First sub-module allows the users to perform virtual screening of RNA oligonucleotides where the server predicts immunomodulatory potential of each oligonucleotide in a given library. Secondly, the server allows the users to design analogs by generating all the possible analogs (single site substitutions) of an RNA oligonucleotide and predicting the class (IMORN or non-IMORN) of each analog. This way, the user can select the most appropriate analogs with desired immunomodulatory potentials. Thirdly, the server facilitates the users in the identification of immunomodulatory regions (RNA oligonucleotides) in a given RNA sequence. In order to provide useful information about predicted RNA oligonucleotides, the server also computes the structure of these RNAs using RNAfold program available from the ViennaRNA package 2.0<sup>23</sup> visualized in the VARNA<sup>24</sup> applet.

**Immunotoxicity of siRNAs.** For the users working in the field of siRNA-based therapeutics, we developed a module that facilitates the users in predicting toxicity of siRNA sequences. In cases where the immunomodulatory ability of siRNA proves to be a disadvantage, the user might be in search of siRNAs with minimal immunotoxicity. First sub-module allows the users to submit a library of siRNA sequence for predicting immunotoxicity or immunomodulatory capability of each siRNA. The server also computes siRNA efficacy of query siRNAs using desiRm<sup>25</sup> software developed previously by our group. Thus, the user may select siRNAs with desired efficacy and toxicity from a given set of siRNAs. Another sub-module allows the users to design siRNA with desired efficacy and immunotoxicity. Under this sub-menu, the server generates a large number of analogs corresponding to the siRNA submitted by a user followed by the prediction of efficacy and immunotoxicity of each analog.

## Discussion

In the past, meticulous efforts have demonstrated the importance of understanding the immunostimulatory potential of RNA sequence for designing better therapeutics. It has been shown previously that combining gene

silencing with virus-like immunostimulation of RNAs (two-pronged strategy) can be used for effective treatment of cancer and viral infections<sup>26</sup>. Gantier *et al.*<sup>16</sup> have demonstrated that sequence-based rational approaches can confer immunomodulatory potential to a siRNA without affecting its gene silencing activity. On the other hand, Judge *et al.*<sup>27</sup> identified the immunomodulatory motifs in siRNAs and showed that altering the nucleotide sequence could subside the immunomodulatory effect considered to be a siRNA-associated immunotoxicity. Thus, while the immunomodulatory potential of siRNA is a desirable therapeutic effect in some cases, it is an associated immunotoxicity in others. imRNA provides an *in silico* platform that has modules for both the purposes – enhancing immunological therapeutic ability of ORNs as well as diminishing the siRNA immunotoxicity.

In diseases like cancer and viral infections, immunostimulatory properties of siRNA could augment the therapeutic effect that relied solely on gene silencing ability. Such bifunctionality has been achieved using sequence-based modifications<sup>16</sup>. The study showed an increase in immunostimulatory effect on introducing a non-pairing U-rich (uridine rich) bulge within the passenger strand of siRNA enhanced the immunostimulatory effect remarkably. In another study, a thermodynamic investigation into the hybridization strength between two strands of siRNA revealed a negative correlation between duplex strength and immunostimulation<sup>17</sup>. The authors suggest that increasing the duplex strength by introducing CG or GC in the sequence retards the release of immunostimulatory single strand and have proposed a model in which the U-rich strand of the siRNA duplex and the ease of its release as a single strand is the major source of causing immunostimulation mediated by TLR-7/8. The detailed investigation into the mechanism of siRNA immunostimulation proposed a model in which the siRNA duplex undergoes endosomal denaturation leading to two single strands of RNA and the U-rich sequences among these single-stranded species activate TLR 7/8<sup>17</sup>. A comparative study also showed that siRNA sequences having immunostimulatory motifs are more effectively recognized by the innate immune system as single-stranded molecules rather than their duplex counterparts<sup>28</sup>. Such studies provide a strong indication towards the importance of ssRNA motifs in mediating immune response due to RNA species. This forms the theoretical basis for the need of methods for predicting whether a given ssRNA oligonucleotide would be immunostimulatory or not. Such a method could also help in selecting or rejecting siRNA sequence alternatives that are immunomodulatory or non-immunomodulatory depending on the case whether this attribute is considered to be a therapeutic potential or a toxic effect.

In our study, we found that the positive (immunomodulatory) sequences are enriched in A-rich motifs as compared to the negative sequences where the G-rich motifs prevail. The primary reason for this result is that our negative dataset is made from the circulating miRNAs found in the body fluids. On comparison of the nucleotide composition between the positive and the negative dataset, the difference in the U composition between positive and negative sequences was found to be least significant (greatest p-value, Table S1) among the four types of nucleotides. Previous literature firmly propounds the importance of U-richness in RNA immunostimulation. In contrast, Jurk *et al.*<sup>29</sup> have reported non-uridine rich sequences that have a strategically located single uridine in a particular sequence-context mediating immunostimulation through TLR 7. Even Gantier *et al.* 2008<sup>22</sup> also found that the localization of uridines in the stem-loop of ssRNA structure is important in addition to the previous claims of U-richness for RNA immunostimulation. Thus, rather than being contrary to previous studies reporting U-richness, our result of A-rich motifs found in the IMORNs might be referring to the sequence context accompanying the uridines in the sequences. More than 98% (593 out of 602) of the positive sequences have atleast one uridine. Additionally, it has also been proposed that TLR 8 preferentially senses loose regions of single strand secondary structures containing uridines<sup>16</sup> that might be provided by A-rich motifs as sequence context.

Gantier *et al.* 2008<sup>22</sup> in their experimental procedure showed that introduction of uridines in the stem-loop structure (predicted by mFOLD program) of an ssRNA transformed it from a non-stimulatory sequence to a TNF- $\alpha$  inducing sequence thus suggesting a possible role of RNA secondary structure in immunostimulation. In our analysis, we found that more than half of the IMORN sequences were predicted to have open linear structures while this was the case in only a quarter of the non-IMORNs. Gantier *et al.* 2010<sup>16</sup> have proposed that TLR 8 preferentially recognizes loose secondary structure but this recognition is affected by the neighboring intramolecular secondary structure. When we distributed the IMORNs into lengthwise bins and observed the secondary structure occurrence, we found that starting from more number of linear structures in shorter sequences, increase in length has a tendency to favor stable secondary structure (stem-loop) formation partly in agreement with the experimental study. Similarly, the proportion of uridine containing stem-loop structures out of the total stem-loops in IMORNs (68.37%) does not seem to be very different from their proportion in non-IMORNs (72.97%). Yet, their occurrence in lengthwise distribution bins shows an opposite trend. With increase in length, IMORNs seem to favor uridine-containing stem-loops while the non-IMORNs show decreasing proportions of such stem-loops. Thus, the RNA secondary structure of IMORNs seems to be length-dependent.

One of the challenges in this study was to compile data from the literature for developing prediction models. The quality of any model depends on datasets used for developing models. Thus, we systematically collected 602 immunomodulatory ssRNA sequences from patents and RNAimmuno database<sup>18</sup>. Similarly, we collected 520 miRNAs that have been found to be in circulation in the human body from the database miRandola<sup>30</sup>. We assigned these miRNAs as non-immunomodulatory RNA sequences based on the assumption that being in circulation in the human body fluids, these have little or no immunomodulatory effect. One may argue why these miRNAs have been assigned non-immunomodulatory without experimental validation. We agree with argument, but we do not have better assumption for creating negative dataset, which is must for developing classification model. Another limitation is the moderate size of the positive dataset; a bigger dataset is needed to develop more reliable models. With larger datasets, we believe in future better and more reliable models will be developed. This study may be considered an initiative towards the development of models for predicting immunomodulatory ssRNA sequences.

## Conclusion

There are numerous resources and *in silico* tools for predicting immunogenicity and antigenicity potential of peptides<sup>31-38</sup>. In contrast, limited efforts have been made to understand the immunomodulatory potential of RNAs. This study is an attempt in this direction where *in silico* models have been developed to predict the immunomodulatory effect of RNA sequences. These models can be used for designing RNA sequences with desired immunostimulatory effects. It means that the user may increase or decrease immunomodulatory property of an RNA sequence by mutating certain nucleotide positions. The choice remains with the scientific community as to how and for what purpose they use our web server, either for designing better adjuvants or for reducing the toxicity of their RNA sequences. A major application of these methods is the design of RNA sequences that manifests gene silencing and elicits immune responses concomitantly. To implement this application, a separate module has been provided in the web server that suggests whether the query siRNA sequences would be immunomodulatory or not, apart from predicting its gene-silencing efficacy.

## Methods

**Datasets.** In the present study, we took ssRNA sequences that were experimentally shown to be immunomodulatory. The positive sequences taken have been adjudged to be the immunomodulatory ssRNAs in different studies through various assays like release of cytokines IFN- $\alpha$  and TNF- $\alpha$  from human PBMCs (Peripheral Blood Mononuclear Cells)<sup>8,22</sup> upon administration as a single-strand complexed with DOTAP. Sioud, 2006 has found that adherent PBMC respond more to single-stranded siRNAs (ss siRNAs) in comparison to the siRNA duplexes and addition of ss siRNAs to siRNA duplexes enhanced the immune response<sup>28</sup>. Administration with DOTAP ensured delivery of ss siRNAs into the endosome which is a compartment harboring the sensors of ssRNAs namely TLR 7 and 8.

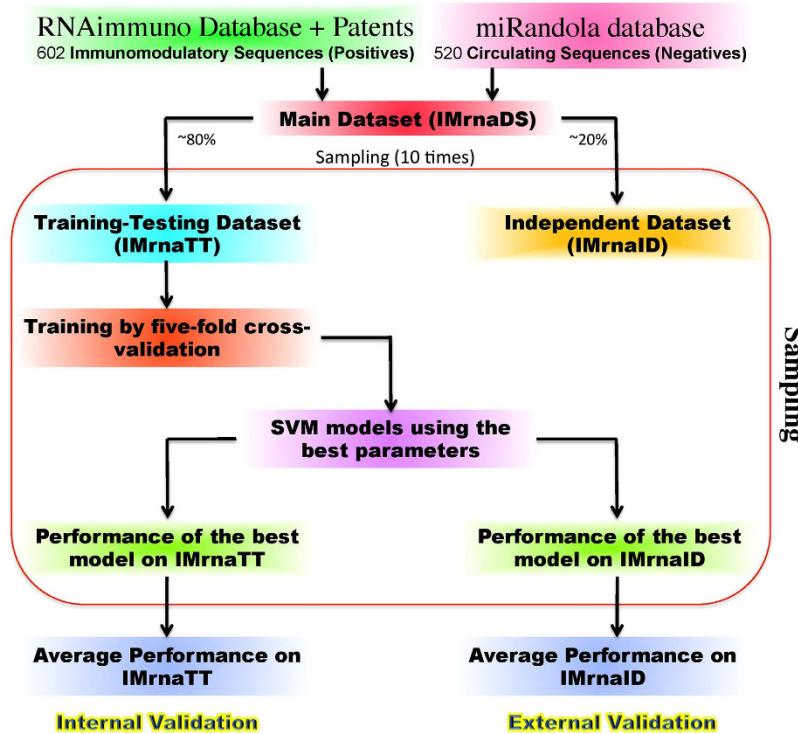
We found 617 such sequences of ssRNA in 4 patents available in the literature of which 493 sequences had length from 17 to 27 nucleotides. RNAImmuno<sup>18</sup> is a collection of immune side effects of RNA molecules. From this database, 232 ssRNA sequences were retrieved that were reported for immunological side effects of which 166 sequences were in the length range of 17–27. From these two sources taken together, we obtained 602 unique sequences having nucleotide length range of 17–27, which were used for building the positive dataset. Only those studies from the RNAImmuno database and only those patents where the RNA species examined was single-stranded including the ss siRNAs were taken for constituting the positive dataset. Further, only the nucleotide sequences were considered and this study did not include the chemical modifications at the termini or the backbone of the sequences to build the prediction models. In the absence of RNA sequences experimentally shown to be non-immunomodulatory, we collected the miRNAs from the miRandola<sup>30</sup> database. miRandola is a database of extracellular circulating miRNAs in human biological fluids like serum, plasma, saliva and urine. Since the source of these miRNAs is the extracellular fluids of healthy and cancerous human tissues, these RNA sequences were assumed to be non-immunomodulatory and were used to constitute the negative dataset. A total of 520 miRNAs in miRandola were mapped onto the miRBase<sup>39,40</sup> database to retrieve the sequences. Our main dataset IMrnaDS contains 602 positive (or immunomodulatory) sequences called IMORNs, and 520 negative (non-immunomodulatory) sequences called non-IMORNs.

In this study, we performed both internal and external cross-validation, as well as sampling approach for computing unbiased performance of our models. First we created a training-testing dataset called IMrnaTT that constitutes ~80% sequences (482 positive and 416 negative sequences) of main dataset IMrnaDS. We used the remaining sequences for creating independent dataset IMrnaID containing ~20% (120 positive and 104 negative sequences) of the main dataset IMrnaDS (Fig. 6).

**Internal Cross-Validation.** The training-testing dataset, IMrnaTT, was subjected to five-fold cross-validation in which the dataset is divided into five parts. In each of the five rounds of evaluation, the model is trained on four parts and tested on the left out fifth part. In this manner, each part has been tested once. Finally, we computed overall performance of a model on whole IMrnaTT dataset. To optimize our SVM models, we tried various combinations of SVM parameters (like kernel, cut-off values) for getting maximum performance. We tried most of the SVM parameters or combination of SVM parameters for developing and evaluating the performance of each model on IMrnaTT dataset. Finally, we identified the SVM parameters that gave best performance of an SVM model on our training-testing dataset IMrnaTT. The best performance of the model was found from the above process called internal validation.

**External Validation.** In case of internal validation, we optimized the performance of our models by trying all possible combinations of SVM parameters and identified the best SVM parameters. We developed an SVM model on training-testing dataset IMrnaTT using these best SVM parameters. We evaluated the performance of this SVM model on the independent dataset IMrnaID, this process being called external validation of a model.

**Sampling.** The process of creating the training-testing dataset (IMrnaTT) and the independent dataset (IMrnaID) from the main dataset (IMrnaDS) was repeated ten times. Each time, the sequences for IMrnaTT were randomly selected from IMrnaDS, and the remaining sequences were included in IMrnaID (a procedure similar to *Monte-Carlo Sampling*). Finally, we evaluated the performance of our models using various features on training-testing and independent dataset as described in the above sections. This process gave 10 performance values using internal and 10 performance values using external validation from 10 rounds of sampling. We computed the mean and standard deviations of these performance values on IMrnaTT and IMrnaID datasets provided in the Tables 2 and 3.



**Figure 6.** Flow diagram showing the sampling method used for evaluating the prediction models.

**Compositional analysis.** Compositional features of nucleotide sequences have been used for analysis and prediction of different classes of non-coding RNAs<sup>41</sup>. In our study too, we used the compositional features of oligonucleotides such as mono-, di-, tri-, tetra- and pentanucleotide compositions. The vector size for mono-, di-, tri-, tetra- and pentanucleotide compositions comes out to be 4, 16, 64, 256 and 1024, given that the mononucleotide types are 4 in number.

**Motif-based analysis.** Motif identification can provide functionally important aspects of biological sequences<sup>19</sup>. MERCi program was used to find motifs that could effectively distinguish positive sequences from the negative ones or vice versa. The program can identify motifs exclusive to any of the positive and negative dataset or those more prevalent in one of these and allowed to a defined lower extent in the other. In our case, the MERCi program compared immunomodulatory RNA (positive) sequences with non-immunomodulatory RNA (negative) sequences for finding the motifs.

**Implementation of support vector machine.** In this study, we developed SVM-based classifiers, which are efficient classifiers heavily used in previous studies including the work on RNA classification<sup>41</sup>. Compositional features calculated in these categories, *viz.* mononucleotide, dinucleotide, trinucleotide, tetranucleotide and pentanucleotide composition were calculated for each of the sequences in the dataset. These features were used for developing SVM-based prediction models separately for each of the categories using the software SVM<sup>light</sup><sup>42</sup>. Apart from this, models based on the binary composition were also used for prediction. Binary notations of A (1,0,0,0), C (0,1,0,0), G (0,0,1,0) and U (0,0,0,1) were used to generate the concatenated profile derived from 17 nucleotide positions taken from the 5' end and 17 nucleotides from the 3' end; and were used for SVM predictions. Additionally, hybrid models using trinucleotide as well as binary composition or by using the combination of trinucleotide composition and binary composition were also developed to investigate the contribution of nucleotide positions in a sequence to the immunomodulatory property.

**Feature Selection.** WEKA<sup>21</sup> is a package that helps to identify the features that contribute most significantly to the prediction models using various classifiers. In the categories, trinucleotide and pentanucleotide compositions, WEKA was used for identifying the important features (trinucleotide and pentanucleotide motifs in this study).

## References

1. Isaacs, A. & Lindenmann, J. Virus interference. I. The interferon. By A. Isaacs and J. Lindenmann, 1957. *J Interferon Res* 7, 429–438 (1987).
2. Janeway, C. A., Jr. Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb Symp Quant Biol* 54Pt 1, 1–13 (1989).
3. Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A., Jr. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature* 388, 394–397, doi: 10.1038/41131 (1997).

4. Broz, P. & Monack, D. M. Newly described pattern recognition receptors team up against intracellular pathogens. *Nat Rev Immunol* **13**, 551–565, doi: 10.1038/nri3479 (2013).
5. Sander, L. E. *et al.* Detection of prokaryotic mRNA signifies microbial viability and promotes immunity. *Nature* **474**, 385–389, doi: 10.1038/nature10072 (2011).
6. Vabret, N. & Blander, J. M. Sensing microbial RNA in the cytosol. *Front Immunol* **4**, 468, doi: 10.3389/fimmu.2013.00468 (2013).
7. Diebold, S. S., Kaisho, T., Hemmi, H., Akira, S. & Reis e Sousa, C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* **303**, 1529–1531, doi: 10.1126/science.1093616 (2004).
8. Heil, F. *et al.* Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science* **303**, 1526–1529, doi: 10.1126/science.1093620 (2004).
9. Pichlmair, A. *et al.* RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. *Science* **314**, 997–1001, doi: 10.1126/science.1132998 (2006).
10. Sabbah, A. *et al.* Activation of innate immune antiviral responses by Nod2. *Nat Immunol* **10**, 1073–1080, doi: 10.1038/ni.1782 (2009).
11. Jackson, A. L. & Linsley, P. S. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov* **9**, 57–67, doi: 10.1038/nrd3010 (2010).
12. Forsbach, A. *et al.* Identification of RNA sequence motifs stimulating sequence-specific TLR8-dependent immune responses. *J Immunol* **180**, 3729–3738 (2008).
13. Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* **21**, 635–637, doi: 10.1038/nbt831 (2003).
14. Fedorov, Y. *et al.* Off-target effects by siRNA can induce toxic phenotype. *RNA* **12**, 1188–1196, doi: 10.1261/rna.28106 (2006).
15. Whitehead, K. A., Dahlman, J. E., Langer, R. S. & Anderson, D. G. Silencing or stimulation? siRNA delivery and the immune system. *Annu Rev Chem Biomol Eng* **2**, 77–96, doi: 10.1146/annurev-chembiomed-061010-114133 (2011).
16. Gantier, M. P. *et al.* Rational design of immunostimulatory siRNAs. *Mol Ther* **18**, 785–795, doi: 10.1038/mt.2010.4 (2010).
17. Goodchild, A. *et al.* Sequence determinants of innate immune activation by short interfering RNAs. *BMC Immunol* **10**, 40, doi: 10.1186/1471-2172-10-40 (2009).
18. Olejniczak, M., Galka-Marciniak, P., Polak, K., Fligier, A. & Krzyzosiak, W. J. RNAimmuno: a database of the nonspecific immunological effects of RNA interference and microRNA reagents. *RNA* **18**, 930–935, doi: 10.1261/rna.025627.110 (2012).
19. Vens, C., Rosso, M. N. & Danchin, E. G. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**, 1231–1238, doi: 10.1093/bioinformatics/btr110 (2011).
20. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536–1537, doi: 10.1093/bioinformatics/btl151 (2006).
21. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–2481, doi: 10.1093/bioinformatics/bth261 (2004).
22. Gantier, M. P. *et al.* TLR7 is involved in sequence-specific sensing of single-stranded RNAs in human macrophages. *J Immunol* **180**, 2117–2124 (2008).
23. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26, doi: 10.1186/1748-7188-6-26 (2011).
24. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975, doi: 10.1093/bioinformatics/btp250 (2009).
25. Ahmed, F. & Raghava, G. P. Designing of highly effective complementary and mismatch siRNAs for silencing a gene. *PLoS One* **6**, e23443, doi: 10.1371/journal.pone.0023443 (2011).
26. Schlee, M., Hornung, V. & Hartmann, G. siRNA and isRNA: two edges of one sword. *Mol Ther* **14**, 463–470, doi: 10.1016/j.mt.2006.06.001 (2006).
27. Judge, A. D. *et al.* Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nat Biotechnol* **23**, 457–462, doi: 10.1038/nbt1081 (2005).
28. Sioud, M. Single-stranded small interfering RNA are more immunostimulatory than their double-stranded counterparts: a central role for 2'-hydroxy uridines in immune responses. *Eur J Immunol* **36**, 1222–1230, doi: 10.1002/eji.200535708 (2006).
29. Jurk, M. *et al.* Immunostimulatory potential of silencing RNAs can be mediated by a non-uridine-rich toll-like receptor 7 motif. *Nucleic Acid Ther* **21**, 201–214, doi: 10.1089/nat.2011.0298 (2011).
30. Russo, F. *et al.* miRandola: extracellular circulating microRNAs database. *PLoS One* **7**, e47786, doi: 10.1371/journal.pone.0047786 (2012).
31. Lata, S., Bhasin, M. & Raghava, G. P. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes* **2**, 61, doi: 10.1186/1756-0500-2-61 (2009).
32. Saha, S., Bhasin, M. & Raghava, G. P. Bcipep: a database of B-cell epitopes. *BMC Genomics* **6**, 79, doi: 10.1186/1471-2164-6-79 (2005).
33. Lata, S. & Raghava, G. P. PRRDB: a comprehensive database of pattern-recognition receptors and their ligands. *BMC Genomics* **9**, 180, doi: 10.1186/1471-2164-9-180 (2008).
34. Peters, B. *et al.* The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* **3**, e91, doi: 10.1371/journal.pbio.0030091 (2005).
35. Singh, H., Ansari, H. R. & Raghava, G. P. Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One* **8**, e62216, doi: 10.1371/journal.pone.0062216 (2013).
36. Gupta, S., Ansari, H. R., Gautam, A., Open Source Drug Discovery, C. & Raghava, G. P. Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. *Biol Direct* **8**, 27, doi: 10.1186/1745-6150-8-27 (2013).
37. Dhanda, S. K., Vir, P. & Raghava, G. P. Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct* **8**, 30, doi: 10.1186/1745-6150-8-30 (2013).
38. Dhanda, S. K., Gupta, S., Vir, P. & Raghava, G. P. Prediction of IL4 inducing peptides. *Clin Dev Immunol* **2013**, 263952, doi: 10.1155/2013/263952 (2013).
39. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140–144, doi: 10.1093/nar/gkj112 (2006).
40. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**, D68–73, doi: 10.1093/nar/gkt1181 (2014).
41. Panwar, B., Arora, A. & Raghava, G. P. Prediction and classification of ncRNAs using structural information. *BMC Genomics* **15**, 127, doi: 10.1186/1471-2164-15-127 (2014).
42. Joachims, T. in *Advances in kernel methods* (eds Scholkopf, B. *et al.*) 169–184 (MIT Press Cambridge, 1999).

## Acknowledgements

Authors are thankful to projects, Open Source Drug Discovery (OSDD) and GENESIS BSC0121. GN and KC are thankful to CSIR and SKD is thankful to ICMR for providing fellowships.

The Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology, Government of India, supported this work.

### Author Contributions

G.N. collected the data. K.C. and G.N. organized the data. K.C., G.N. and S.K.D. performed the experiments. G.N. and K.C. developed the web interface. G.N., K.C., S.K.D. and G.P.S.R. analyzed the data. G.N., S.K.D. and G.P.S.R. prepared the manuscript. G.P.S.R. conceived the idea and coordinated the project.

### Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

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

**How to cite this article:** Chaudhary, K. *et al.* Prediction of Immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants. *Sci. Rep.* **6**, 20678; doi: 10.1038/srep20678 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>