frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 December 2014
doi: 10.3389/fbioe.2014.00071

ncPred: ncRNA-disease association prediction through tripartite network-based inference

***Salvatore Alaimo*[1]*, Rosalba Giugno*[2]*[†] *and Alfredo Pulvirenti*[2]*[†]**

[1] *Department of Mathematics and Computer Science, University of Catania, Catania, Italy*
[2] *Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy*

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Helder I. Nakaya, Emory University, USA
Mikhail P. Ponomarenko, Russian Academy of Sciences, Russia

***Correspondence:**
Rosalba Giugno and Alfredo Pulvirenti, Department of Clinical and Experimental Medicine, c/o Department of Mathematics and Computer Science, University of Catania, Viale A. Doria 6, 95125 Catania, Italy
e-mail: giugno@dmi.unict.it; apulvirenti@dmi.unict.it

[†] Rosalba Giugno and Alfredo Pulvirenti have contributed equally to this work.

Motivation: Over the past few years, experimental evidence has highlighted the role of microRNAs to human diseases. miRNAs are critical for the regulation of cellular processes, and, therefore, their aberration can be among the triggering causes of pathological phenomena. They are just one member of the large class of non-coding RNAs, which include transcribed ultra-conserved regions (T-UCRs), small nucleolar RNAs (snoRNAs), PIWI-interacting RNAs (piRNAs), large intergenic non-coding RNAs (lincRNAs) and, the heterogeneous group of long non-coding RNAs (lncRNAs). Their associations with diseases are few in number, and their reliability is questionable. In literature, there is only one recent method proposed by Yang et al. (2014) to predict lncRNA-disease associations. This technique, however, lacks in prediction quality. All these elements entail the need to investigate new bioinformatics tools for the prediction of high quality ncRNA-disease associations. Here, we propose a method called *ncPred* for the inference of novel ncRNA-disease association based on recommendation technique. We represent our knowledge through a tripartite network, whose nodes are ncRNAs, targets, or diseases. Interactions in such a network associate each ncRNA with a disease through its targets. Our algorithm, starting from such a network, computes weights between each ncRNA-disease pair using a multi-level resource transfer technique that at each step takes into account the resource transferred in the previous one.

Results: The results of our experimental analysis show that our approach is able to predict more biologically significant associations with respect to those obtained by Yang et al. (2014), yielding an improvement in terms of the average area under the ROC curve (AUC). These results prove the ability of our approach to predict biologically significant associations, which could lead to a better understanding of the molecular processes involved in complex diseases.

Availability: All the *ncPred* predictions together with the datasets used for the analysis are available at the following url: http://alpha.dmi.unict.it/ncPred/

Keywords: ncRNAs-diseases association predictions, lncRNAs functional characterization, network-based inference, tripartite networks, resource transfer algorithm

1. INTRODUCTION

In recent years, great efforts have been employed in the study of non-coding RNAs (ncRNAs), a class of genes involved in a wide variety of biological functions. Small ncRNAs, such as siRNA, miRNA, and piRNA, are highly conserved in different species and have a key role in transcriptional and post-transcriptional silencing of genes. Long ncRNA (transcribed RNA molecules whose length is greater than 200 nucleotides) instead are poorly preserved and have the task of regulating gene expression through mechanisms still largely unknown (Mercer et al., 2009; Ponting et al., 2009; Wilusz et al., 2009). It has been shown that these molecules are involved in the regulation of gene expression by acting as controllers of processes such as RNA maturation or transportation, or altering chromatin structure. ncRNAs have great variety in structure and in gene regulation outcomes, however, several similarities can be identified in the way they act (Wang and Chang, 2011).

The connection between diseases and de-regulation of small ncRNAs has been established for years. However, recent studies show that mutations and de-regulations of lncRNAs are heavily involved in the development or progression of several diseases (Wapinski and Chang, 2011). Alterations in the structure (primary or secondary), or in the expression levels are the main underlying causes of diseases, from cancer to neurodegenerative disorders (Wapinski and Chang, 2011).

Pasmant et al. (2011) highlight how the expression of the lncRNA *ANRIL*, antisense transcript to *INK4b* gene, is correlated with the epigenetic silencing of *INK4a*, or *p16 protein*, which is involved in the regulation of cell cycle. High levels of *ANRIL* were found in prostate cancer tissues (Yap et al., 2010). Yap et al. (2010), also, hypothesizes that this transcript is an initiating factor in tumor formation due to its silencing action on the *INK4b/ARF/INK4a* locus. Other experimental evidence link *ANRIL* de-regulation to a number of pathologies, including coronary disease, intracranial aneurysm, and type II diabetes (Pasmant et al., 2011).

Another example of correlation between lncRNAs and diseases is the *HOTAIR* transcript, which is involved in the progression of breast cancer by chromatin landscape remodeling (Burd et al., 2010). In particular, increased expression of tHOTAIR is an index of poor prognosis and tumor metastasis. Gupta et al. (2010) show that *HOTAIR* is also responsible for invasiveness and metastasis in epithelial cancer cells and its inhibition may lead to a reduction of invasiveness in cells where *PRC2 complex* is highly activated.

Further evidence of lncRNAs-diseases correlation is the transcript called *MALAT-1*, an RNA of more than 8000*nt* present in chromosome 11*q*13, whose over-expression is related to bad prognosis in patients with non-small cell lung cancer (Ji et al., 2003). In addition, the antisense transcript of β-*secretase-1* (*BACE1-AS*) has been identified in high concentrations in subjects with Alzheimer's disease and in amyloid precursor protein transgenic mice (Faghihi et al., 2008).

Therefore, despite the enormous importance that ncRNAs show in connection with several diseases, the number of entities, which somehow has been functionally characterized and associated to diseases, is extremely small (Wapinski and Chang, 2011). For this purpose, the developing a methodology that is able to predict ncRNA-disease interactions is crucial in order to formulate new hypotheses on the molecular mechanisms underlying complex diseases, and to identify potential new biomarkers for their diagnosis, treatment and prevention. Despite the use of such a methodology could be very helpful by making the search for new associations more focused and less costly, it must be emphasized that the task of determining, which are beneficial remains a responsibility of bio-physicians. They, indeed by identifying appropriate patient groups and properly documenting such cases, can establish the actual relationship, while also allowing a broader understanding of the underlying phenomena.

In this direction, Yang et al. (2014) developed a method, which exploits a bipartite network and a propagation algorithm to predict new associations that can be evaluated through appropriate *in vitro* experiments. Yang et al. (2014) based their method on the database assembled by Chen et al. (2013): a collection of approximately 1028 experimentally validated interactions among 322 lncRNAs and 221 diseases. The database has been further extended, through deep literature mining, to include additional interactions. The database includes also 478 experimentally validated interactions among 126 lncRNAs and 236 protein coding genes. For such genes a modulation in expression values is known to be carried out by such ncRNAs.

In this paper we present *ncPred*, a resource propagation methodology, which uses a tripartite network to guide the inference process of novel ncRNA-disease associations. The tripartite network allows the introduction of two levels of interaction: ncRNA-target and target-disease. Here, we call targets a group of biomolecules (i.e., genes, microRNAs, proteins) whose activity is modulated by a ncRNA (e.g., regulation of expression, binding to improve the efficiency of its activity, or binding to help the formation of complexes). In this way, we can exploit the greater quantity of known interactions between targets (i.e., proteins and miRNAs) and diseases to build a wider knowledge base and obtain a greater number of high quality predictions.

To perform a proper evaluation of our method, we applied a k-fold Cross-Validation procedure to the (Chen et al., 2013) database, remodeled to include information on targets. A further analysis uses a database of experimentally verified interactions between ncRNAs and miRNAs shown in Helwak et al. (2013).

2. MATERIALS AND METHODS

2.1. ALGORITHM

Let $O = \{o_1, o_2, \ldots, o_n\}$ be a set of non-coding RNAs (ncRNAs), let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of targets (i.e., genes, microRNA), and let $D = \{d_1, d_2, \ldots, d_p\}$ be a set of diseases. The ncRNA-target and target-disease interactions can be represented in a tripartite graph $G(O, T, D, E)$, where E is the set of interactions (edges) between nodes in O and T and nodes in T and D. Such a graph, can be represented by using a pair of adjacency matrices $A^{OT} = \left\{a_{ij}^{OT}\right\}_{n \times m}$ and $A^{TD} = \left\{a_{rs}^{TD}\right\}_{m \times p}$ where $a_{ij}^{OD} = 1$ if o_i is connected to t_j in G, and $a_{rs}^{TD} = 1$ if t_r is connected to d_s in G.

Our technique is based on the concept of resources transfer within the network. We refer to Alaimo et al. (2013) for details of resources transfer (drug-targeting) in bipartite networks. The bipartite network carries a prior knowledge which can be used to infer novel interactions. Starting from such a network, it computes weights between each pair of target. Those weights can be seen as the likelihood by which we can affirm that if a drug is associated with a target then it may be associated with another one. For each prediction, the algorithm also associates a score indicating the degree of certainty of the interaction.

In this paper, due to the tripartite network, we developed a multi-level transfer approach that at each step takes into account the resource transferred in the previous one (see **Figure 1** for an example). In the first level of the transfer, the resource is moved from the nodes in T (targets) to nodes in O (ncRNAs) and vice versa. In the second level, the resource is moved from D nodes to T nodes and it is combined with the resource of the previous step. Then, the resources are moved back to the D nodes. In this way, we define a methodology for the computation of a combined weight matrix $W^C = \left\{w_{ij}^c\right\}_{m \times p}$, where w_{ij}^c corresponds to the likelihood allowing us to claim that if a ncRNA interacts with a target t_i then it may be associated with the pathology d_j.

To compute such a matrix, we start by defining two partial weight matrices corresponding to the intermediate levels of transfer. These two matrices are then used to obtain the combined weight matrix and, therefore, compute the recommendations.

Let $k'(x)$ be the degree of node x in the ncRNA-target sub-network and $k''(y)$ the degree of node y in the target-disease sub-network.

The matrix $W^T = \left\{w_{ij}^T\right\}_{m \times m}$, associated with the first level of transfer, can be defined as:

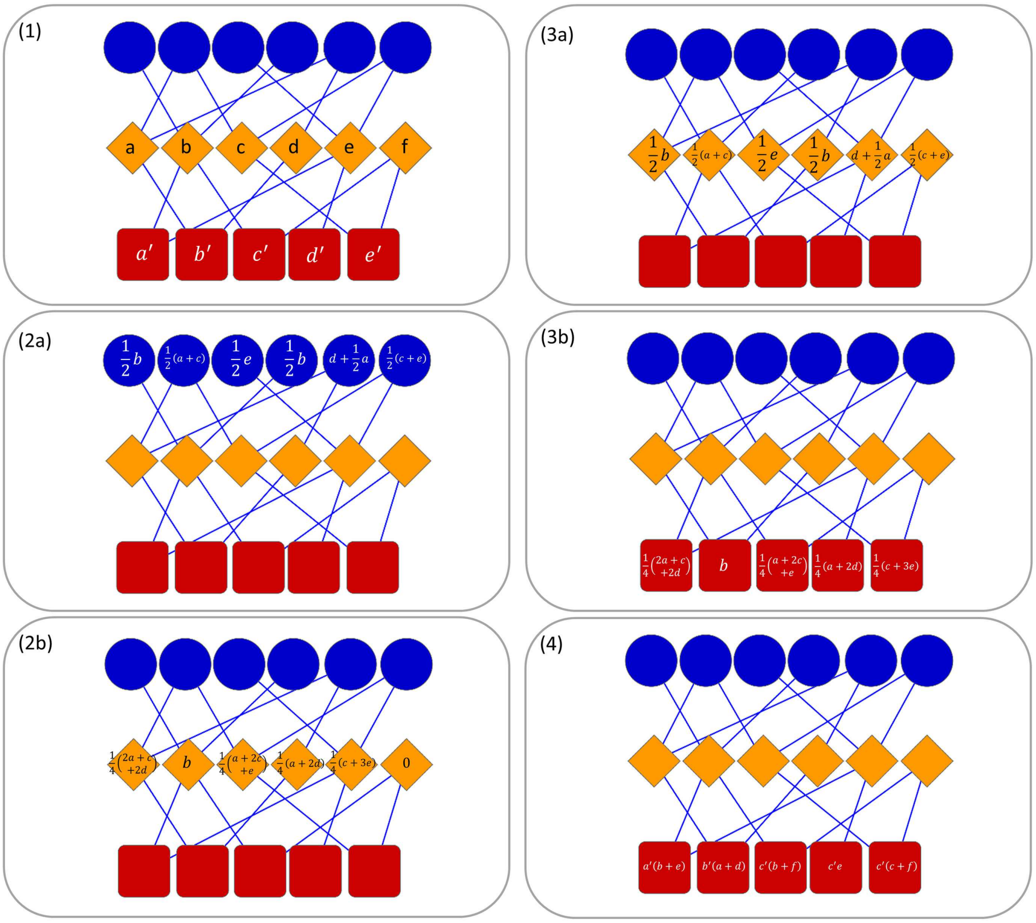

FIGURE 1 | Operating principle of ncPred in a tripartite network. Here, we represent ncRNAs in blue, targets in orange, and diseases in red. Without loss of generality, and in order to simplify the reading of the image, we decided to put λ_1 and λ_2 to 1, so as to obtain a uniform distribution of resources in the network. In the first step, a resource is assigned to each target and disease node (1). Thereafter, two separate transfer process are launched to compute the resource in target nodes (2a, 2b) and disease nodes (3a, 3b). Finally, resources are combined to obtain the total quantity in each disease node (4). In (4), the literals are used only for example purposes due to lack of space. They are to be replaced with the values computed in steps (2b) and (3b).

$$w_{ij}^{T} = \frac{1}{k'(t_i)^{(1-\lambda_1)}\,k'(t_j)^{\lambda_1}} \sum_{l=1}^{n} \frac{a_{li}^{OT} a_{lj}^{OT}}{k'(o_l)}, \tag{1}$$

where w_{ij}^{T} corresponds to the likelihood that given a ncRNA interacting with target t_i, then it may also interact with target t_j. By using such an equation, we assign higher weights to the pairs of targets that share many ncRNAs, rather than those who share only a few.

The same applies to $W^D = \left\{w_{ij}^{D}\right\}_{p\times p}$, matrix associated with the second level of the transfer, where:

$$w_{ij}^{D} = \frac{1}{k''(d_i)^{(1-\lambda_2)}\,k''(d_j)^{\lambda_2}} \sum_{l=1}^{m} \frac{a_{li}^{TD} a_{lj}^{TD}}{k''(t_l)}. \tag{2}$$

In equation 2, w_{ij}^{D} indicates whether we can assert that given a target associated with the disease d_i, it may also be linked to the disease d_j. w_{ij}^{D} is higher for the disease pairs, which are associated to many common targets with respect to those with fewer common targets.

In equations 1 and 2, the $\lambda_1 \in [0, 1]$ and $\lambda_2 \in [0, 1]$ parameters are used to tune the quality of the predictions. Parameter values close to zero indicate that the resource of a node is computed as the average of those in its neighborhood, while values close to one indicate that the resource is uniformly distributed among the nodes of its neighborhood. In terms of predictions, lambda values close to zero correspond to conservative predictions, while values close to one correspond to a larger number of predictions.

Therefore, the combined weight matrix $W^C = \left\{w_{ij}^{c}\right\}_{m\times p}$ can be obtained as:

$$w_{ij}^{C} = \sum_{t=1}^{m} \left[w_{it}^{T} \sum_{r=1}^{p} \left(a_{tr}^{TD} \cdot w_{rj}^{D} \right) \right]. \tag{3}$$

In equation 3, the weight of a target-disease pair is computed by taking into account both the targets with a similar neighborhood and the diseases with a similar neighborhood. In this way, a larger weight is assigned to those pairs for which more frequently there is a path, which passes through them.

Given the above weights, it is now possible to compute the recommendation matrix $R = \{r_{ij}\}_{n \times p}$ as:

$$R = A^{OT} \cdot W^{C}. \tag{4}$$

We call each r_{ij} prediction score for the pair (i, j). For each ncRNA o_i, its list of predictions R_i can be obtained by selecting those disease-prediction score pairs for which there is no path with o_i in the tripartite network. Such a list is sorted in descending order with respect to the value of r_{ij}, as the higher the score, the greater the belief that the ncRNA will have some connection with that particular disease.

Table 1 | Description of the datasets: number of ncRNAs, targets and diseases together with the count of interactions, average degree, density, modularity, number of connected components, and average path length.

Metrics	Chen et al. (2013)	Helwak et al. (2013)
ncRNAs	119	338
Targets	110	179
Diseases	514	134
ncRNAs–targets interactions	247	1699
Targets–diseases interactions	1005	1572
Average degree	1.572	5.025
Density	0.002	0.008
Modularity	0.609	0.274
Number of connected components	24	1
Average path length	1.572	1.734

2.2. DATASETS AND BENCHMARKS

We evaluated our method using two datasets containing experimentally verified interactions between ncRNAs, targets, and diseases. The first data set (Figure S1 in Supplementary Material) was built by collecting from (Chen et al., 2013) 478 interactions between lncRNAs and genes. These interactions were mapped by converting each target identifier to its Entrez Id. This allowed us to remove about 230 duplicates or superseded interactions. From the remaining targets, we then extracted 1005 experimentally validated gene-disease associations by searching in DisGeNET (Bauer-Mehren et al., 2010).

The second data set (Figure S2 in Supplementary Material) was obtained by collecting about 4000 lncRNA-miRNA interactions found by Helwak et al. (2013) by applying the CLASH methodology (Kudla et al., 2011). Each association indicates that a lncRNA contains one or more binding sites for miRNAs. From such a list, we removed all targets not present in miR2Disease database (Jiang et al., 2009), obtaining 1699 lncRNA-miRNA associations. Finally, using Jiang et al. (2009), we recovered 1572 miRNA-disease associations. **Table 1** provides a summary of the two datasets together with some metrics that can further elucidate their characteristics. Moreover, in **Figure 2**, we calculated the degree distribution of the two networks. These show that they can be considered scale-free networks.

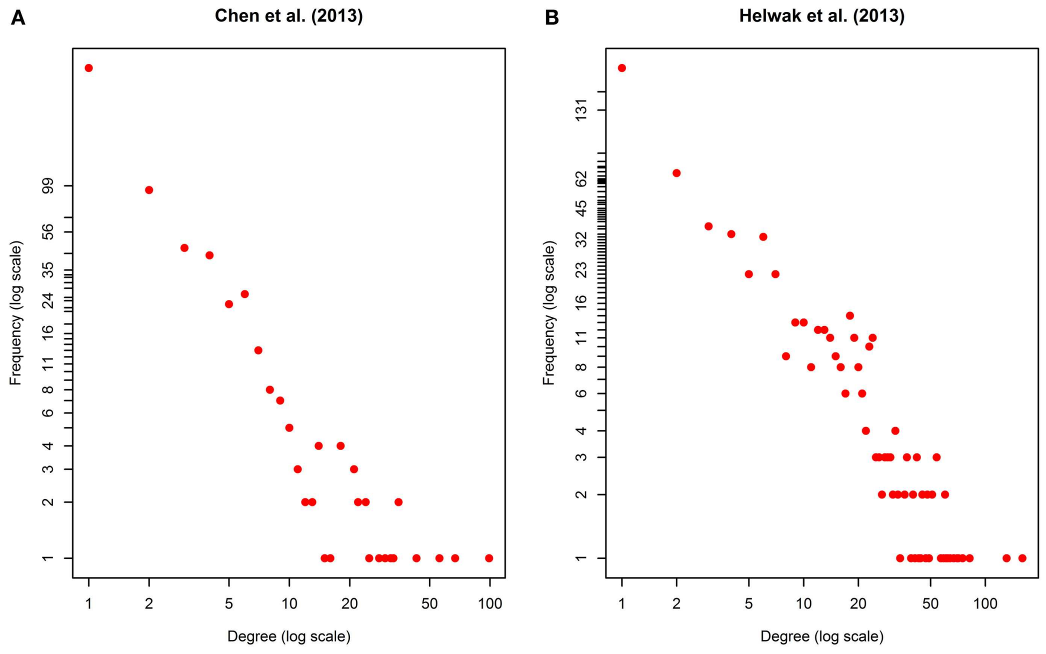

FIGURE 2 | Degree distribution of the two networks used as datasets: **(A)** Chen et al. (2013), **(B)** Helwak et al. (2013). The two plots are in log-log scale. As can be seen the degree distribution for the two networks can be approximated to an exponential one. We can therefore assume that the two networks are scale-free.

For the evaluation of our method, we applied a 10-fold cross-validation procedure repeated 30 times to obtain more reliable results. Each fold is built in the following way. Given the tripartite graph, we selected all possible pairs of ncRNA-disease interactions. Then, we randomly partitioned them into each fold. We make sure that the tripartite network generated from each fold is not disconnected. *ncPred* makes predictions only on connected networks. We considered the following four metrics (Alaimo et al., 2013) to assess the performance of our method: precision and recall enhancement, recovery, personalization, and Surprizal. The first two establish the ability of the method to recover the interactions of the test set, therefore, obtaining biologically relevant predictions. The other two measure the ability of the method to propose unexpected interactions, which may lead to novel insights onto ncRNA functions. Special care should be given to the precision and recall enhancement metrics. They measure the reliability of the prediction algorithm by comparing the standard precision and recall with a null model. Such a model is defined as a methodology that randomly assigns ncRNA-disease pairs. This implies that values greater than one are to be considered synonymous of higher quality and, therefore, reliability.

3. RESULTS

As stated earlier, to evaluate the power of our method, we applied a 10-fold cross-validation procedure repeated 30 times and averaged results to obtain more reliable estimates. In **Table 2**, we illustrate the behavior of *ncPred*, comparing it with Yang et al. (2014), in terms of precision and recall enhancement. The results demonstrate that *ncPred* clearly outperforms its competitor. In particular, we can see that while Yang et al. (2014) obtains a recall close to the null model, *ncPred* has much better results. This is crucial since the recall measures the ability of the algorithm to recover existing interactions in the network, and is therefore a sign of their reliability, namely their biological relevance.

In **Figure 3**, we report the receiver operating characteristic (ROC) curves computed on both datasets. The simulations were repeated 30 times and their results were averaged to obtain a more accurate evaluation. Both methods show a high true positive rate against low false positive rate, although *ncPred* is clearly able to achieve better results. This is also shown in **Table 2**, where we can see a significant increase in the average area under the ROC curve (AUC). Such a significance is further proved by the results shown in **Table 3**. By applying the Friedman rank sum test, we determined that the performance improvement achieved by our algorithm is

Table 2 | Comparison of ncPred and Yang et al. (2014) through the precision and recall enhancement metric, and the average area under ROC curve (AUC) calculated for each of the two datasets listed in Table 1.

Dataset	$e_P(20)$		$e_R(20)$		$AUC(20)$	
	Yang et al. (2014)	ncPred	Yang et al. (2014)	ncPred	Yang et al. (2014)	ncPred
Chen et al. (2013)	5.5113	12.3290	0.7297	1.6636	0.6217 ± 0.0178	0.7566 ± 0.0218
Helwak et al. (2013)	1.8654	5.8197	1.6509	5.6572	0.7069 ± 0.0084	0.7669 ± 0.0093

*The results were obtained using the optimal values for λ_1 and λ_2 parameters as shown in **Table 3**.*

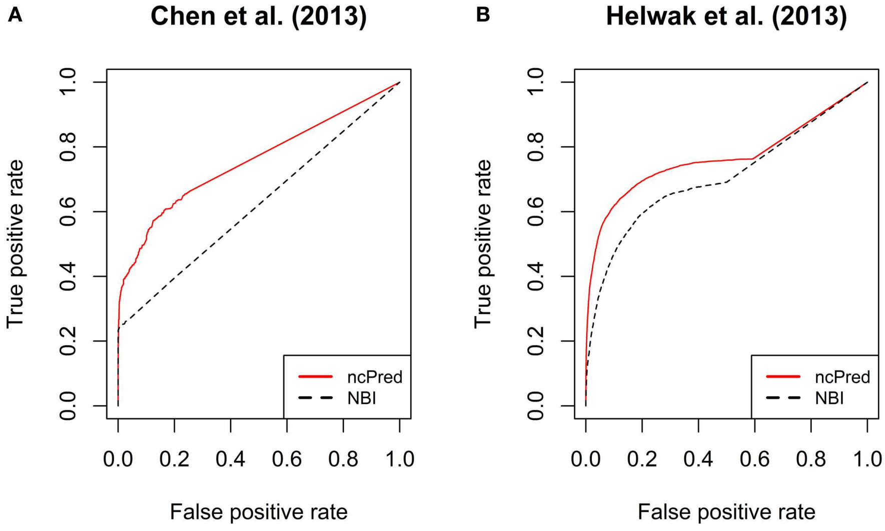

FIGURE 3 | Comparison between ncPred and Yang et al. (2014) by means of receiver operating characteristic (ROC) curves, computed for the recommendation lists built on our two datasets. Such curves measure the quality of the algorithms in terms of false positives rate against true positives rate. **(A,B)** are independent since computed on two separate datasets. The significance of the difference highlighted between ncPred and Yang et al. (2014) was measured by applying the Friedman rank sum test as assessed in **Table 4**.

statistically significant (i.e., the p-value is close to zero on both datasets).

Regarding the parameters λ_1 and λ_2, we performed a comprehensive analysis to establish the relationship between them and the prediction quality. In the supporting materials, we report the results of such analysis. The results indicate that there is no specific law, which governs their behavior. The peculiar characteristics of each dataset greatly affect the performances and, consequently, the parameters. It is, therefore, necessary to perform an *a priori* analysis in order to determine, which values give the best results. In our experiments, we used such an analysis to determine the best parameters in terms of precision and recall enhancement (see **Table 4** for details on their values). By looking at the characteristics of our data sets, the values obtained from such an analysis allowed us to suppose that the two parameters are close to zero in Helwak et al. (2013) dataset because of the greater density. This implies that to maintain high quality predictions it is necessary to reduce their number to avoid the introduction of noise. On the other hand, the Chen et al. (2013) dataset has a lower density. This allows us to produce a higher number of predictions before they start losing quality. Therefore, this explains the lambda values closer to one. It is important to point out that in order to determine the best parameters an analysis was performed considering only precision and recall enhancement, since they are closely related to the biological significance of the predictions. In this context, we report in **Table 2** only precision and recall enhancement and the AUC, ignoring the other metrics, which are available in the supporting materials.

Finally, assuming that the number of targets dominates the ncRNA one, we can state that the computational complexity of our method is $O(m^2p)$. However, it is quite straightforward to implement parallelization and optimization techniques to make the computation faster.

3.1. CASE STUDIES

The analysis of the predictions for each non-coding showed that *ncPred* is able to find exactly the same predictions provided by Yang et al. (2014). The main difference between the two algorithms lies in the different scores given to each prediction. As highlighted in the previous section, ncPred is clearly able to provide more substantially accurate predictions.

Table 3 | Friedman rank sum test applied to establish the statistical significance in the performance improvement of ncPred compared to Yang et al. (2014).

Dataset	Friedman χ^2	p-Value
Chen et al. (2013)	1026.315	$<2.2 \times 10^{-16}$
Helwak et al. (2013)	6537.915	$<2.2 \times 10^{-16}$

Table 4 | Optimal values of λ_1 and λ_2 parameters for the datasets used in our experiments.

Dataset	λ_1	λ_2
Chen et al. (2013)	0.5	1
Helwak et al. (2013)	0.2	0.2

To further demonstrate the ability of our method, we reviewed in detail the results of five diseases (i.e., Alzheimer's Disease, Myocardial Infarction, Pancreatic Cancer, Parkinson's disease, and Gastric Cancer) as case studies. The top 10 predicted genes for each case are listed in **Table 5**. **Table 5** also shows the rank obtained by applying on our dataset, the Yang et al. (2014) method. In this context, the two datasets were taken together in order to start from a wider knowledge base.

3.1.1. Alzheimer's disease

Alzheimer's disease (AD) is one of the most common forms of dementia (Hebert et al., 2003). Recent studies indicate that it affects approximately 0.40% of the world population (Brookmeyer et al., 2007). The disease is, at present, untreatable, and it is characterized by a progressive loss of mnemonic, cognitive, and intellectual capacity, which ultimately leads to the death of the patient. Among the first 10 ncRNAs, we find *PVT1* a lncRNA,

Table 5 | List of top 10 predictions computed by ncPred and their rank obtained with Yang et al. (2014) for five case studies (Alzheimer's Disease, Myocardial Infarction, Pancreatic Cancer, Parkinson's Disease, and Gastric Cancer).

ncRNA	ncPred rank	Yang et al. (2014) rank	ncRNA	ncPred rank	Yang et al. (2014) rank
ALZHEIMER'S DISEASE					
PVT1	1	3	B2 SINE RNA	6	28
MEG3	2	19	TP53TG1	7	22
TUG1	3	32	WRAP53	8	23
lincRNA-p21	4	21	Kcnq1ot1	9	48
CDKN2B-AS1	5	20	Evf2	10	35
MYOCARDIAL INFARCTION					
H19	1	43	Kcnq1ot1	6	23
SRA1	2	24	PVT1	7	47
TUG1	3	26	CDKN2B-AS1	8	25
7SL	4	29	B2 SINE RNA	9	17
BDNF-AS1	5	34	Airn	10	18
PANCREATIC CANCER					
HOTAIR	1	16	PCAT1	6	40
LINC00312	2	15	ncRNACCND1	7	9
Kcnq1ot1	3	25	Six3OS	8	45
Xist	4	43	Airn	9	14
TERRA	5	10	RepA	10	47
PARKINSON'S DISEASE					
PVT1	1	11	LINC00312	6	24
MEG3	2	16	TP53TG1	7	20
TUG1	3	26	WRAP53	8	21
BACE1-AS	4	23	CDKN2B-AS1	9	27
lincRNA-p21	5	19	B2 SINE RNA	10	40
GASTRIC CANCER					
PTENP1	1	38	Evf2	6	60
LINC00312	2	15	Airn	7	13
Xist	3	1	TERRA	8	18
PCAT1	4	29	B2 SINE RNA	9	40
Six3OS	5	39	RepA	10	37

which regulates the transcription of *MYC* on the long distance (Carramusa et al., 2007). In Jiang et al. (2013), *MYC* has been characterized as the source of the main pathway substantially active in AD, thus having an important role in disease progression. Such a discovery confirms that *PVT1* could play a key role in the progress of AD. We have also identified the lncRNA *MEG3* that activates *TP53* and improves its binding affinity to target gene promoter (Liao et al., 2011). *TP53* was identified in Tan et al. (2012) as potential biomarker for AD. Therefore, further analysis to confirm *MEG3* role in AD are needed.

3.1.2. Myocardial Infarction

Myocardial infarction (MI) is a heart condition that occurs when the proper flow of blood to a part of the heart stops, and the heart muscle is damaged due to lack of sufficient oxygen. Genome-wide association studies have identified 27 epigenetic factors that are associated with an increased risk of MI (Feero et al., 2011). For example, the genomic locus 9p21 has one of the strongest associations with the pathology (Feero et al., 2011). The majority of such factors have been identified in regions implicated in other heart diseases (Feero et al., 2011). Among our predictions, we identified the lncRNA *SRA1* that Friedrichs et al. (2009) found crucial in cardiomyopathies. This leads us to assume a possible link with MI. In the top 10 predictions we also found the lncRNA *7SL*, which, by hybridizing to the *reverse-Alu-element-containing 3′ UTR of MnSOD* gene, represses its expression (Lipovich et al., 2010). Overexpression of *MnSOD* has been identified as a possible protection against MI in transgenic mice (Chen et al., 1998). This could be a cue for further investigations to understand the role such a lncRNA.

3.1.3. Pancreatic cancer

Pancreatic cancer is an aggressive disease whose 5-year survival rate is extremely low (Amundadottir et al., 2009). The analysis of the predictions obtained by our algorithm has provided the association with lncRNA *HOTAIR*, whose overexpression has been associated with a poor prognosis in pancreatic cancer, as well as show a pro-oncogenic activity (Kim et al., 2012). A further lncRNA is *Airn*. The deletion of its promoter in paternal allele results in aberrant activation of *IGF2R* (Nagano and Fraser, 2009), whose polymorphisms are associated with an increased risk of pancreatic cancer (Dong et al., 2012).

3.1.4. Parkinson's disease

Parkinson's disease (PD) is a degenerative disorder of the central nervous system. The main cause of the disease is the death of dopamine-generating cells in the substantia nigra. The cause of this death is still unknown, nevertheless, the process of aging and metabolic stress are its common triggers (Parlato and Liss, 2014). It is interesting to note that the response to stress conditions and mechanisms for quality control are compromised in patients with PD. The reduction in the transcription of rRNA (ribosomal ribonucleic acid) is an important strategy to maintain cellular homeostasis under stress. An altered transcription is associated with neurodegenerative disorders. There are many triggers for nucleolar stress, but they seem to depend on the extitp53 protein (Parlato and Liss, 2014). Our algorithm is able to identify two probable lncRNA associated with this function: *PVT1*, also associated with AD, whose gene locus is a target of *p53* (Barsotti et al., 2012), and *MEG3* that promotes the expression of *Tp53* and increases the binding affinity to the promoters of its target (Liao et al., 2011).

3.1.5. Gastric cancer

Gastric cancer is a disease typically characterized by an overall 5 years survival rate lower than 10%, mainly due to the plurality of common symptoms that lead to treatments only in advanced disease stages (Orditura et al., 2014). Among our predictions, we find the lncRNA *Xist*. In Weakley et al. (2011), it was identified as differentially expressed in stomach preneoplastic cells, which could be a symptom of gastric cancer. Another factor could be the lncRNA *Evf2*, which is a direct putative positive regulator of transcription factor *Dlx-2* (Lipovich et al., 2010). Increased expression of *Dlx-2* was correlated with more advanced stages of the disease (Tang et al., 2013).

4. DISCUSSION

In this paper, we propose *ncPred* to predict novel associations between ncRNAs and diseases. The aim is to compute ncRNA-disease association's prediction starting from a tripartite network. Such a network integrates information on ncRNAs, targeting (i.e., those genes, microRNAs, proteins whose activity is affected by non-coding RNA), and their associations with diseases in order to improve prediction quality and accuracy.

Our experimental analysis shows that our approach predicts more biologically significant associations with respect to Yang et al. (2014). This assertion is confirmed by the results obtained in terms of recall, which as described above measures biological quality of results. The use of Friedman rank sum test also showed that the difference between our predictions and those of Yang et al. (2014) is not random but due to a better interpretation of available information. The results showed that our method could provide interesting suggestions in the study of the implications between ncRNA and pathologies. However, as stated in the introduction, the method can only help to make such a search more targeted and less expensive, offering a ranking of associations from more probable to less probable. Determine whether those associations are useful still remains within the competence area of bio-physicians that can provide conclusive evidence by identifying suitable patients and documenting such cases.

Despite what stated earlier, our method still has some limitations that should be taken into account. Firstly, ncRNA-target associations are still too small in number. It may be necessary to resort to additional targeting prediction techniques so as to expand original knowledge base. Secondly, the methodology does not use the biological information accompanying each association (e.g., type of ncRNA-target interaction, conditions in which the target-disease association was detected, tissues in which associations have significance). For this reason, it may be useful to further expand the methodology by using such additional information, which could make the methodology more reliable in terms of significant predictions.

SUPPLEMENTARY MATERIAL

In the Supplementary Material (Data Sheet 1.pdf) we report the ncPred parameter tuning further details concerning the comparison with Yang et al. (2014). The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00071/abstract

REFERENCES

Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. (2013). Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 29, 2004–2008. doi:10.1093/bioinformatics/btt307

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R. Z., Fuchs, C. S., Petersen, G. M., Arslan, A. A., et al. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* 41, 986–990. doi:10.1038/ng.429

Barsotti, A. M., Beckerman, R., Laptenko, O., Huppi, K., Caplen, N. J., and Prives, C. (2012). p53-Dependent induction of PVT1 and MIR-1204. *J. Biol. Chem.* 287, 2509–2519. doi:10.1074/jbc.M111.322875

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics* 26, 2924–2926. doi:10.1093/bioinformatics/btq538

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement.* 3, 186–191. doi:10.1016/j.jalz.2007.04.381

Burd, C. E., Jeck, W. R., Liu, Y., Sanoff, H. K., Wang, Z., and Sharpless, N. E. (2010). Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6:e1001233. doi:10.1371/journal.pgen.1001233

Carramusa, L., Contino, F., Ferro, A., Minafra, L., Perconti, G., Giallongo, A., et al. (2007). The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *J. Cell. Physiol.* 213, 511–518. doi:10.1002/jcp.21133

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, D983–D986. doi:10.1093/nar/gks1099

Chen, Z., Siu, B., Ho, Y.-S., Vincent, R., Chua, C. C., Hamdy, R. C., et al. (1998). Overexpression of MnSOD protects against myocardial ischemia/reperfusion injury in transgenic mice. *J. Mol. Cell. Cardiol.* 30, 2281–2289. doi:10.1006/jmcc.1998.0789

Dong, X., Li, Y., Tang, H., Chang, P., Hess, K. R., Abbruzzese, J. L., et al. (2012). Insulin-like growth factor axis gene polymorphisms modify risk of pancreatic cancer. *Cancer Epidemiol.* 36, 206–211. doi:10.1016/j.canep.2011.05.013

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., et al. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730. doi:10.1038/nm1784

Feero, W. G., Guttmacher, A. E., O'Donnell, C. J., and Nabel, E. G. (2011). Genomics of cardiovascular disease. *N. Engl. J. Med.* 365, 2098–2109. doi:10.1056/NEJMra1105239

Friedrichs, F., Zugck, C., Rauch, G.-J., Ivandic, B., Weichenhan, D., Müller-Bardorff, M., et al. (2009). HBEGF, SRA1, and IK: three cosegregating genes as determinants of cardiomyopathy. *Genome Res.* 19, 395–403. doi:10.1101/gr.076653.108

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi:10.1038/nature08975

Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A., and Evans, D. A. (2003). Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* 60, 1119–1122. doi:10.1001/archneur.60.8.1119

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041. doi:10.1038/sj.onc.1206928

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA desregulation in human disease. *Nucleic Acids Res.* 37(Suppl. 1), D98–D104. doi:10.1093/nar/gkn714

Jiang, W., Zhang, Y., Meng, F., Lian, B., Chen, X., Yu, X., et al. (2013). Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease. *Bioinformatics* 29, 2596–2602. doi:10.1093/bioinformatics/btt423

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., et al. (2012). HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* 32, 1616–1625. doi:10.1038/onc.2012.193

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10010–10015. doi:10.1073/pnas.1017386108

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878. doi:10.1093/nar/gkq1348

Lipovich, L., Johnson, R., and Lin, C.-Y. (2010). MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA. *Biochim. Biophys. Acta* 1799, 597–615. doi:10.1016/j.bbagrm.2010.10.001

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi:10.1038/nrg2521

Nagano, T., and Fraser, P. (2009). Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mamm. Genome* 20, 557–562. doi:10.1007/s00335-009-9218-1

Orditura, M., Galizia, G., Sforza, V., Gambardella, V., Fabozzi, A., Laterza, M. M., et al. (2014). Treatment of gastric cancer. *World J. Gastroenterol.* 20, 1635. doi:10.3748/wjg.v20.i7.1635

Parlato, R., and Liss, B. (2014). How Parkinson's disease meets nucleolar stress. *Biochim. Biophys. Acta* 1842, 791–797. doi:10.1016/j.bbadis.2013.12.014

Pasmant, E., Sabbagh, A., Vidaud, M., and Bièche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–448. doi:10.1096/fj.10-172452

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641. doi:10.1016/j.cell.2009.02.006

Tan, M., Wang, S., Song, J., and Jia, J. (2012). Combination of p53 (ser15) and p21/p21 (thr145) in peripheral blood lymphocytes as potential Alzheimer's disease biomarkers. *Neurosci. Lett.* 516, 226–231. doi:10.1016/j.neulet.2012.03.093

Tang, P., Huang, H., Chang, J., Zhao, G.-F., Lu, M.-L., and Wang, Y. (2013). Increased expression of DLX2 correlates with advanced stage of gastric adenocarcinoma. *World J. Gastroenterol.* 19, 2697. doi:10.3748/wjg.v19.i17.2697

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends Cell Biol.* 21, 354–361. doi:10.1016/j.tcb.2011.04.001

Weakley, S. M., Wang, H., Yao, Q., and Chen, C. (2011). Expression and function of a large non-coding RNA gene XIST in human cancer. *World J. Surg.* 35, 1751–1756. doi:10.1007/s00268-010-0951-0

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi:10.1101/gad.1800909

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* 9:e87797. doi:10.1371/journal.pone.0087797

Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., et al. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674. doi:10.1016/j.molcel.2010.03.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 September 2014; accepted: 25 November 2014; published online: 12 December 2014.

Citation: Alaimo S, Giugno R and Pulvirenti A (2014) ncPred: ncRNA-disease association prediction through tripartite network-based inference. Front. Bioeng. Biotechnol. ***2****:71. doi: 10.3389/fbioe.2014.00071*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.