Indian Statistical Institute Kolkata

M.Tech (Computer Science) Dissertation

miRNA Target Prediction in Humans

A dissertation submitted in partial fulfillment of the requirements for the award of M.Tech.(Computer Science) degree

Author:
Amit Yadav
Roll No:MTC1217

Supervisor:
Dr. Pradipta Maji
Machine Intelligence Unit

M.TECH. (CS) DISSERTATION THESIS COMPLETION CERTIFICATE

Student : Amit Yadav (MTC1217)

Topic : miRNA Target Prediction in Humans

Supervisor : Pradipta Maji

This is to certify that the thesis titled miRNA target prediction in humans submitted by Amit Yadav in partial fulfillment for the award of the degree of Master of Technology is a bonafide record of work carried out by him under our supervision. The thesis has fulfilled all the requirements as per the regulations of this Institute and, in our opinion, has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other university for the award of any degree or diploma.

Dr. Pradipta Maji

Date : 11^{th} *July*, 2014

Acknowledgements

I would like to take this opportunity to thank **Dr.Pradipta Maji** for his guidance and support through out the project.

I would also like to thank all my friends and classmates for their motivation and support to complete the project.

Contents

Chapter 1

Introduction

microRNAs(miRNAs) are endogenous $\sim$ 22 nt RNAs that play important gene-regulatory roles in animals and plants by pairing to the mRNAs of protein-coding genes and resulting in post-transciptional repression. Each miRNA is believed to regulate thousands of genes.The study of regulation of mRNA by miRNA becomes important from the fact that deregulation of miRNAs can lead to specific disease phenotypes and it pushes for the investigation of miRNA's role in diagnosis,prognosis and therapeutic application for disease treatment.[1].

Since miRNA potentially targets a large number of genes,elucidating its function using only wet-lab experiments is extremely laborious and economically infeasible.The more favourable approach involves screening of candidates using computational methods.The present state of algorithms employ a number of strategies,based on prior knowledge and high-throughput data.But still the algorithms are far from being perfect.This calls the need for searching new patterns and developing algorithms to tackle the problem.

1.1 Discovery of miRNA

MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene lin-14 in C. elegans development.[2] They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the lin-4 gene. A 61-nucleotide precursor from the lin-4 gene matured to a 22-nucleotide RNA that contained sequences partially complementary to multiple sequences in the 3' UTR of the lin-14 mRNA. This complementarity was both necessary and sufficient to inhibit the translation of the lin-14 mRNA into the LIN-14 protein. Retrospectively, the lin-4 small RNA was the first microRNA to be identified, though at the time, it was thought to be a nematode idiosyncrasy. Only in 2000 was a second RNA characterized: let-7, which repressed lin-41, lin-14, lin-28, lin-42, and daf-12 expression during developmental stage transitions in C. elegans. let-7 was soon found to be conserved in many species[3][4], indicating the existence of a wider phenomenon.

1.2 Biogenesis of miRNA

MicroRNAs are produced from either their own genes or from introns.The majority of miRNA genes are transcribed as independent units.However, in some cases a microRNA gene is transcribed together with its host gene; this provides a means for coupled regulation of miRNA and protein-coding gene[5]. As much as

40 percentage of miRNA genes may lie in the introns of protein and non-protein coding genes or even in exons of long nonprotein-coding transcripts[6].

Following are the steps involved in the biogenesis of miRNA

1. The first step is the nuclear cleavage of the pri-miRNA,which liberates a 60-70 nt stem loop intermediate,known as the miRNA precursor,or the pre-miRNA[7][8][9].This processing is performed by the Drosha RNaseIII endonuclease,which cleaves both strands of the stem at sites near the base of the primary stem loop[10].

2. The pre-miRNA is actively transported from the nucleus to the cytoplasm by Ran-GTP and the export receptor Exportin-5[8][11].

3. Dicer first recognizes the double-stranded portion of the pre-miRNA,perhaps with particular affinity for a 5'phosphate and 3'overhang at the base of the stem loop. Then,at about two helical turns away from the base of the stem loop,it cuts both strands of the duplex.

4. This cleavage by Dicer lops off the terminal base pairs and loop of the pre-miRNA,leaving the 5'phosphate and aroung 2nt 3' overhang characteristic of an RNase III and producing an siRNA-like imperfect duplex that comprises the mature miRNA and similar-sized fragment derived from the opposing arm of the pre-miRNA.The fragments from the opposing arm are called the miRNA* sequences[12].

5. Although either strand of the duplex may potentially act as a functional miRNA, only one strand is usually incorporated into the RNA-induced silencing complex (RISC) where the miRNA and its mRNA target interact.

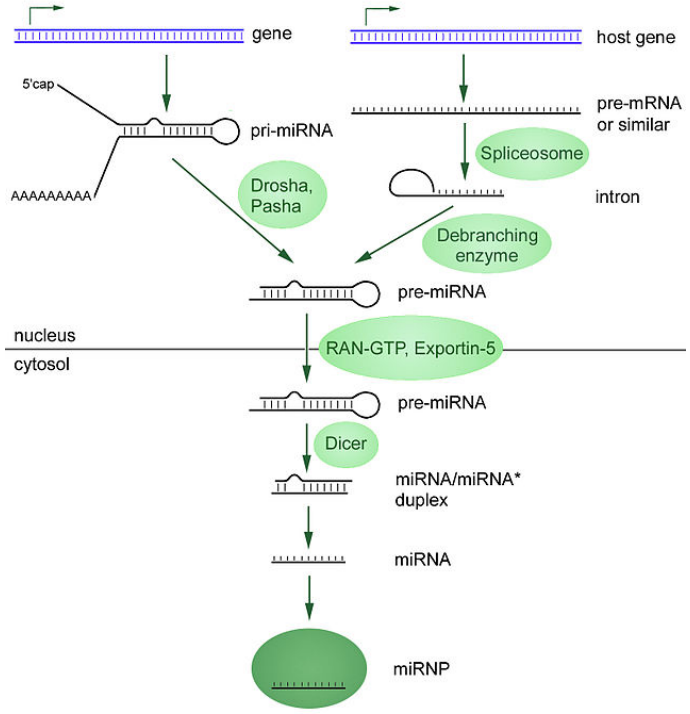

Figure 1.1: miRNA biogenesis

1.3 Why study miRNA ?

miRNA regulate the gene expression by pairing to the mRNAs of protein-coding genes. Failure to regulate targets in this way can have severe consequences or subtle effects,depending on the nature of the targets.Dysregulation of miRNA has been associated with diseases.A manually curated,publicaly available database,**miR2Disease**,documents known relationships between miRNA dysregulation and human disease[14].miRNA has been associated with cancer,dna repair,heart disease,nervous system,obesity and inherited diseases.

1.4 Factors responsible for targeting

In case of plants, targets(mRNA) can be predicted with confidence simply by searching for messages with extensive complemenatary to the miRNA[15].But in case of animals,it is not the case.
The pairing of miRNA with the mRNA usually takes place in the 3'UTR region of the mRNA.

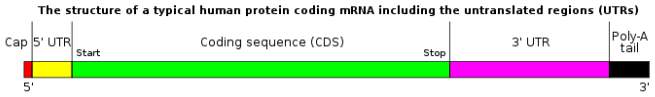

Figure 1.2: Structure of mRNA

Following are the factors affecting targeting in animals

1. Perfect pairing of miRNA nucleotides from 5'end 2-7,known as the miRNA seed with the 3'UTR region of mRNA have been found to be important for the recognition of miRNA targets.This has been corroborated by a wide range of methods,including comparative sequence analysis,site-directed mutagenesis,genetics,mRNA profiling,coimmunoprecipitation and proteomics.

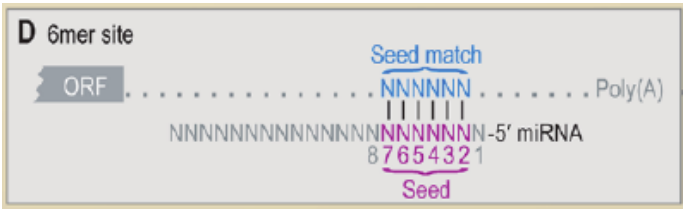

Figure 1.3: 6mer Sites

2. The above 6 length nucleotide can be augmented by either a match with miRNA nucleotide 8(7mer-m8 site) ,an A across from nucleotide 1(7mer-A1 site) or both (8-mer site).

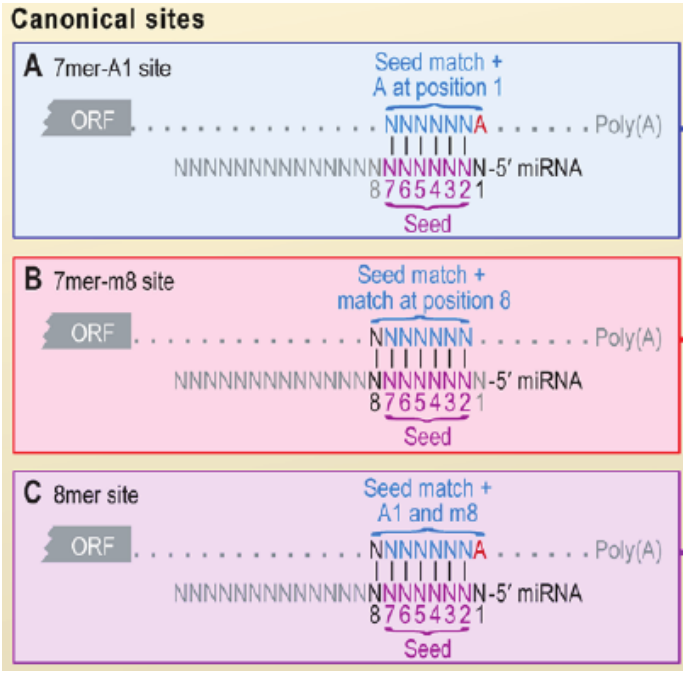

Figure 1.4: 7mer/8mer Sites

3. Also,it has been observed that pairing on 3'miRNA nucleotides 13-16 and the UTR region directly opposite this miRNA segment augments seeds pairing.Such sites are called 3'-supplementary sites.

4. Pairing to the 3'portion of the miRNA can not only supplement a 7-8mer match ,but it can also compensate for a single nucleotide bulge or mismatch in the seed region.These sites are called "3'-compensatory sites" .Although,very less such sites have been found.

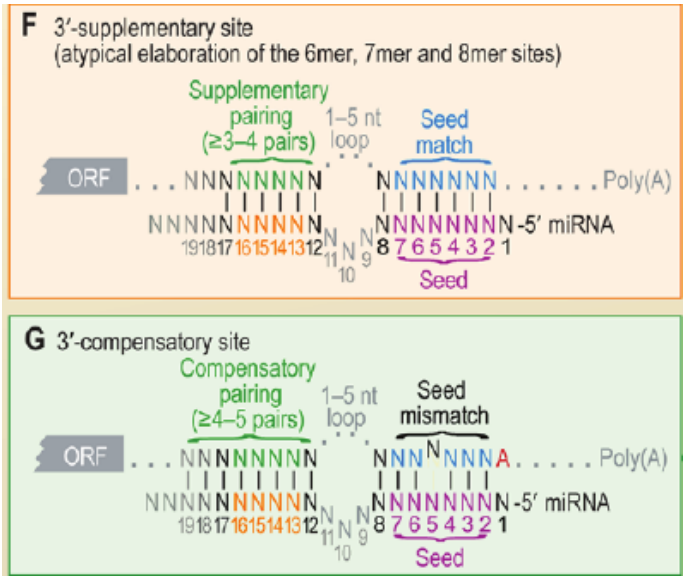

Figure 1.5: Supplementary and Compensatory Sites

5. The site in 3'UTR of mRNA is effective only if it is at least 15nt away from the stop codon.The

reason being that the first 15 nt of the 3'UTR area are cleared of silencing complexes when they enter the ribosome as the translation machinery approaches the stop codon[15].

6. Genome-wide analysis of site conservation,site efficacy and site depletion all indicate that 7-8 nt sites within the 3'UTR and out of the path of the ribosome tend to be most effective if they do not fall in the middle of long UTRs[15].One explanation for these results is that sites in the middle of long UTRs might be less accessible to the silencing complex because they would have opportunities to form occlusive interactions with segments from either side,whereas sites near the UTR ends would not.

7. AU-rich nucleotide composition near the site works in favour of targeting the mRNA,since it places the site within a more accessible UTR context[15].

8. Seed pairing stability(SPS)-It has been observed the (A+U)-rich seed regions, could lower the stability of seed pairing interaction and a minimum duration association between miRNA and target is required for effective regulation [16].

9. Target Abundance(TA) - miRNAs with (A+U)-rich seed regions have more 3'UTR-binding sites,a consequence of the (A+U)-rich nucleotide composition of 3'UTRs,which could dilute the effect on each target message[16].So, it suggests that miRNAs having more Target Abundance(TA) are less likely to target a mRNA effectively.

10. A site within a mRNA which is conserved across the related species is more likely to be an actual site as compared to the non-conserved sites.

11. The thermodynamic stability of miRNA-mRNA duplex which is assessed by calculating the free energy of the putative binding[17].

12. Although most investigation has been for sites in 3'UTR region,experiments using artificial sites have show that targeting can occur in 5'UTR sites and open reading frames(ORFs).Overall ,endogenous ORF targeting appears to be less frequent and less effective than 3'-UTR targeting but still much more frequent than 5'UTR targeting. One reason for 5'UTRs and ORFs being less hospitable for targeting is that silencing complexes bound to these regions would be displaced by the translation machinery as it translocates from the cap-binding complex through the ORF[18].

13. Recently,a class of miRNA target sites that lack both perfect seed pairing and 3' compensatory pairing and instead have 11-12 contiguous Watson-Crick pairs to the center of the miRNA have been found. Such sites are called "centered sites"[19].

Chapter 2

Existing target prediction algorithms

1. Miranda

   The algorithm is based on the fact that there is some complementarity between the miRNA and the 3'UTR region of the mRNA.First,using dynamic programming miRNA and UTR sequences are aligned.The alignment between the two is scored,where a CG/AT match is given a score of +5,GT match has a score of +1,gap opening score is -9 ,gap extending score is -4 and a mismatch has a score of -5.Also,the score for 8 nucleotides of 5' end of miRNA are scaled by a factor of 4.The non-overlapping alignments which are above a threshold score(140) are selected. For these alignments,gibbs free energy is calculated using RNAfold program of ViennaRNA package.The alignments which are below the energy threshold of 1.0 are finally reported as the sites present in the mRNA for targeting by miRNA[20].

2. TargetScanHuman

   This method first searches for the presence of 8mer and 7mer sites(these sites have been defined above),that match the corresponding region of the miRNA.The method gives option for finding both conserved and non-conserved sites. The found sites are extended to cover the whole of the miRNA i.e 3' end of the miRNA and are finally reported,alongwith the 3'pairing contribution,local AU-contribution ,position contribution and site-type contribution.A final context score is given which is the weighted sum of the above factors and the two other factors,target abundance(TA) and seed pairing stability(SPS).The weights are found by carrying out multiple linear regression on the 11 microarray data sets[16].

3. RFMirTarget

   It is a machine learning based approach.The method involves running the Miranda algorithms on the miRNA-UTR pairs and selecting those which cross a threshold score of 140 and an energy threshold of 1.From the obtained alignments,a feature set consisting of a total 34 features which comprises of alignment,thermodynamic,structural,seed and position based features is obtained.After this,random forest is used to classify targets and non-targets.[21]

4. TargetSpy

   The method first finds the candidate sites in the 3'UTR region of mRNA by searching for areas in the target sequence where the predicted Gibbs free energy of the microRNA-target du-

plex is below a certain microRNA-specific energy threshold.Once such sites are found,compositional features,bulge-related features,position specific features,general extent of the microRNA-mRNA binding features,compactness feature are created.After this,learning scheme based on boosting called MultiBoost with decision stumps as base learners is applied for classification.[22]

5. Pictar

This method takes multiple alignments of RNA sequences(typically 3'UTR) and a search set of mature(coexpressed)microRNA sequences. The program nuclMap locates all perfect nuclei(length 7,starting at position 1 or 2 of the 5'end of the microRNA) and imperfect nuclei in 3'UTR sequences.Highly probable nuclei that survive the optimal free energy filter and fall into overlapping positions in the alignments for all species under consideration are called anchors.If a 3'UTR multiple alignment has a minimal(user-defined) number of anchors,each UTR in the alignment will be scored by the central PicTar maximum likelihood procedure.Scores for the individual UTRs in an alignment are combined to obtain the final PicTar score which can be used to obtain a ranked list of all sets of orthologous trancripts.For, pictar scoring Pictar tallies all segmentations of the RNA sequence(3'UTR) into binding sites and background sequences.Pictar computes the maximum likelihood score that the RNA sequence is targeted by combinations of microRNAs from the search set when compared to background and the individual probability p(i) for each subsequence of the RNA sequence to be bound by a microRNA.These posterior probabilitiAes are different from the probability that a single subsequence is a microRNA binding site and reflect the competition of microRNAs and background for binding in the UTR.[23]

Discussion

Miranda searches for sites in the 3' UTR region of miRNA where there is good enough complementarity.Also, Miranda scores the seed region of UTR by a factor of 4.This results on a average of around 7 percentage of all predicted sites which show a mismatch to 7mer seeds[22].So, in effect the method is able to predict seed sites and supplementary sites.But, the method suffers from the fact that a seed site may be ineffective depending on the UTR context of the site. On the other hand,TargetScanHuman searches for these seed sites and reports various UTR context features and the total context score.And leaves it to the user decide whether to consider the reported site as a target or non-target.It doesn't directly tell that a given site is target or non-target.

Talking about machine learning approach to the problem, we need to generate features from the examples.In the target examples,the exact location on mRNA is not known where miRNA attaches itself.During initial days,when it was found that miRNA regulates mRNA's, some 7-8 nt length complemenatarity was observed between miRNA and UTR region of mRNA(which are now known as seed sites).But later on examples were found,where even though seed sites are present in the mRNA ,still it is not targeted by the corresponding miRNA.So,we have both the examples of targets and non-targets,where seed sites are present.Recently,centered sites (some 10-12 nt length in the centre of the miRNA) have been reported[19],but there number is few.But,there are examples of targets,where none of these sites are present.So,there are still some sites or rules needed to be discovered,which can explain these target examples.

Motivated by the fact there is some complementarity between miRNA and 3'UTR region of mRNA, RFMirTarget uses Miranda algorithm to find locations in UTR region where there is a minimum

complemenatarity with miRNA (above a threshold 0f 140) and then use these locations to generate the feature set for the given examples.But using this method,there are certain examples (both targets and non-targets),which don't have this minimum complemenatarity.So,for such examples we don't have a corresponding feature set.In effect, we won't be able to classify certain examples.

TargetSpy tries to solve the above problem. TargetSpy works by searching for areas in the target sequence where the predicted Gibbs free energy of the microRNA-target duplex is below a certain energy threshold.To cover large functional binding sites, a conservative cut-off is kept.Using this method,feature set can be generated for all the examples. But since it is not known whether Gibbs free energy is the deciding factor,we might be including examples from targets which have characteristics of being a non-target and examples from non-targets which have characteristics of being a target.

Chapter 3

Proposed Method

Since,both the targets and non-targets can contain 8mer,7mer and 6mer sites.The proposed algorithm classifies such examples.

The steps involved in the proposed algorithm are as a follows:

1. Input to the algorithm consists of pairs of miRNA and 3'UTR sequences,both for targets and non-targets.
2. Three different types of sites (8mer,7mer and 6mer sites) are searched in the 3'UTR region of the mRNA for the corresponding miRNA.
3. Once the site is found, the site is extended to cover the whole of miRNA as proposed in the TargetScan algorithm.
4. Once the complete miRNA is aligned with the 3' UTR,a feature set is generated.
5. The set of best features for classification are selected based on average decrease in gini impurity during creation of random forest classifier.
6. Random Forest is applied to classify the examples using the three features selected in the above step.

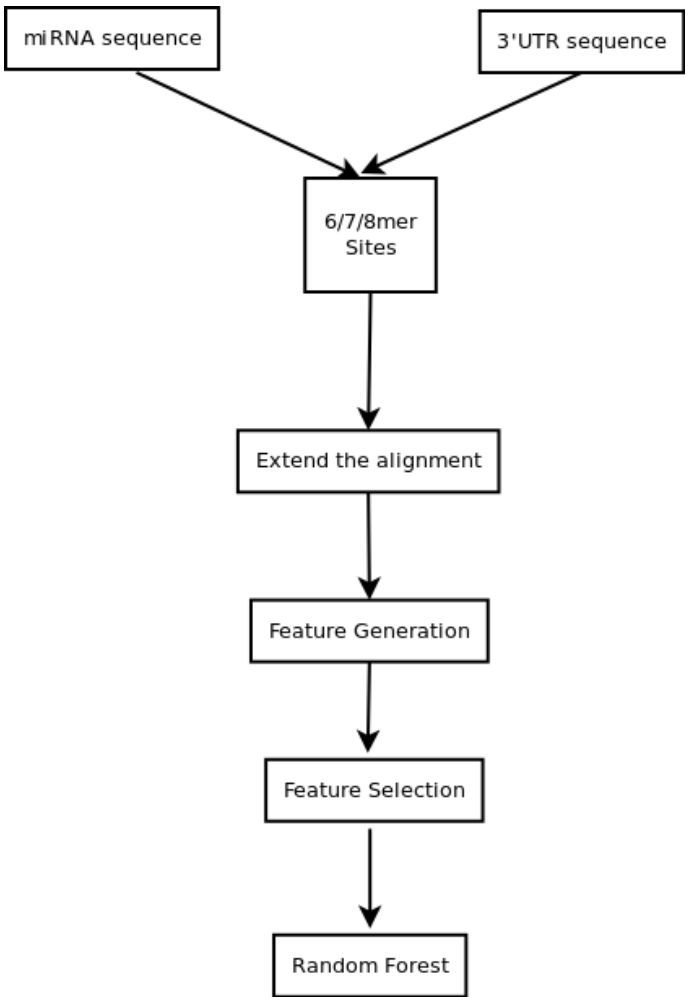

Figure 3.1: Proposed Approach

3.1 Feature Set

1. Type of site
   (a) 7mer-1a represented as 1
   (b) 7mer-m8 represented as 2
   (c) 8mer-1a represented as 3
   (d) 6mer represented as 6
2. Length of the miRNA
3. Length of the UTR sequence
4. Distance of the site from the middle of the UTR sequences
5. Normalized distance of the site from the middle of the UTR sequence.It is calculated by dividing the distance of site from the middle of the UTR sequence by the length of the UTR sequence.

6. Local AU contribution - 30 nt upstream and 30 nt downstream sequences in the 3'UTR are considered for AU contribution.The contribution of AU at any particular position decreases inversely as the distance from the site.If i is the distance of a position(either upstream of downstream) from the site,then the contribution of that position to the AU score is 1/i.

7. 3' end pairing score as obtained by the TargetScan algorithm in the step 3 of the proposed algorithm.

8. 20 position based features - The alignment obtained above between miRNA and UTR is used for taking position based feature.The alignment length considered for these features is 20.CG pair,AU pair,GU pair,mismatch pair and gap are represented by 1,2,3,5,4 respectively.

9. GC content of the seed region(position 2-8)

10. AU content of the seed region(position 2-8)

11. Total number of gaps,total CG count,total AU count,total GU count and total mismatch count in the 20nt miRNA-mRNA duplex.

12. If the position of the site is less than 15 from the starting of the UTR sequence, the feature value is set as 1,otherwise it is 0.

13. Minimum free energy of the full alignment as calculated by using RNAVienna package[24]

14. Minimum free energy of the seed region alignment as calculated by using RNAVienna package

15. Target Abundance(TA) score as provided by the TargetScan Algorithm.It is calculated by considering the number of sites(7mer) in a curated set of distinct 3' UTRs.

3.2 Random Forest

Random-forest is a classifier consisting of a collection of tree-structured classifiers.
The random forest algorithm is as follows:

(a) Draw n bootstrap samples from the original data,where n is the number of trees to be created.

(b) For each of the bootstrap samples,grow an unpruned classification tree. At each node,randomly sample m of the features and choose the best split among these features based on the gini impurity.

(c) Predict new data by aggregating the predictions of the n trees i.e majority votes for classification.

In the first step,bootstrap samples are created.This results in an average of about one-third samples being left out.These left out samples are called "out of bag" data.This OOB data is used for calculating error,on the tree trained with sampled data.All the OOB errors can be averaged to calculated the error rate.
There are two variables,n and m involved in the above algorithm.The default value for n is 500.But,the number of trees should be increased as the number of features or the data points increase. The value of m is found by iteratively increasing the value of m from 1 to total number of features, and selecting the value of m for which we get the best results.This process has been implemented in the caret R package.

3.3 Gini Impurity

While constructing a tree,the feature used for splitting a node is decided based on the gini impurity.Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.Gini impurity is computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item.It reaches its minimum(zero) when all cases in the node fall into a single target category. The feature which is used for splitting the node is the one which results in maximum decrease in the gini impurity.
Decrease in gini impurity = Gini impurity of the parent node - Weighted sum of gini impurities of its two descendents
The importance of individual feature is calculated by taking average of gini decreases over all the trees in the random forest.

Chapter 4

Results and Discussion

4.1 Dataset

The dataset is the one which is used in the research paper titled "RFMirTarget :Predicting Human MicroRNA Target Genes with a Random Forest Classifier" published in PLOS ONE in July 2013.This dataset contains 274 examples of targets and 468 examples of non-targets.
After searching for 6mer-7mer and 8mer sites,237 targets and 264 non-targets are obtained for feature generation.Since there can be multiple sites present in a target or non-target, a total of 780 examples of target class and 535 examples of non-target class are generated,each containing a total of 38 features.

4.2 Classification

On the above prepared dataset,five repeatitions of 10 fold cross-validation was performed using SVM(default radial basis kernel),C4.5 and Random-forest.LIBSVM[25] sofware was used for SVM.For RandomForest,randomForest[26] and caret[27] R packages were used. Among the three,SVM predicted 86.36 percentage targets and 82.74 percentage non-targets correctly.C4.5 had 93.46 percentage sensitivity and 93.64 percentage specificity.But,RandomForest performed the best with a sensitivity of 95.38 percentage and 96.79 percentage specificity.In terms of standard deviation also,RandomForest performed best with 2.43 and 2.52 percentage standard deviation for both targets and non-targets.

Table 4.1: Classifier Results

Classifier	True-targets	True Non-Target	Sd(Targets)	Sd(Non-Targets)
SVM	86.36	82.74	4.157	5.47
RandomForest	95.38	96.79	2.43	2.52
C4.5	93.46	93.64	3.33	3.32

4.3 Feature Selection

Next the features used above were ranked using feature selection algorithms Relief and Infogain.Also,features were ranked using average decrease in the Gini impurity during the construction of trees in the Random-

Forest.The top 15 features estimated using the above methods are given in the table below.

Table 4.2: Feature Ranking

Relief	Infogain	Gini Impurity
Mfe Seed	Length UTR	Length UTR
Target Abundance	Mfe Seed	Mfe Seed
Length UTR	AU contribution	Target Abundance
Pos 7	Pos 7	AU Contribution
Pos 2	Pos 2	Pos 2
Pos 5	Mfe Full	Pos 7
Pos 6	Seed GC	Mfe Full
Pos 16	Pos 8	Seed GC
Seed GC	Pos 5	Pos 4
Mfe Full	Pos 6	Pos 5
Pos 3	Pos 4	Pos 6
Pos 4	Target Abundance	Dist Mid Normal
Seed AU	Distance Middle	AU count
AU contribution	Pos 15	Dist Middle
Pos 1	Dist Mid Normal	Seed AU

Length of the UTR is ranked at the top position by both the Infogain and the average decrease in the Gini Impurity.One possible explanation for the length of UTR to be ranked at the top is that the longer regions have a higher probability of possessing more miRNA binding sites and this can act as a distinguishing factor between targets and non-targets.Minimum free energy of the seed region is ranked at the second position by average decrease in Gini impurity and it supports that fact there has to be a minimum stable interaction between the seed region and the mRNA for effective targeting. Target Abundance ranks amongst the top and it is corroborated by the fact that miRNAs with (A+U)-rich seed regions have more 3'UTR-binding sites,a consequence of the (A+U)-rich nucleotide composition of 3'UTRs, which could dilute the effect on each target message.AU contribution shows the imporatance of the region flanking the sides of the target site in the UTR region. Other position based features in the seed region are ranked in the top 15 and this shows the importance of seed position in miRNA targeting.
Based on the featuring ranking for the above three methods,1.Average results for 5 repeatitions was calculated. For the increasing number of features,sensitivity and specificity was calculated. Three plots below show the same for Information Gain,Relief and Average decrease in Gini Impurity

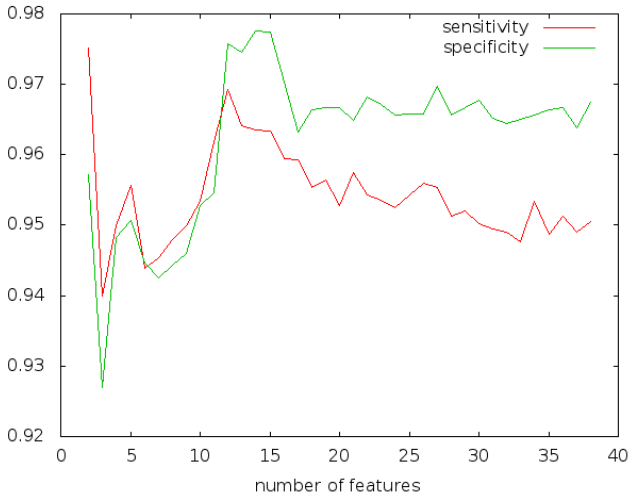

Figure 4.1: Information Gain

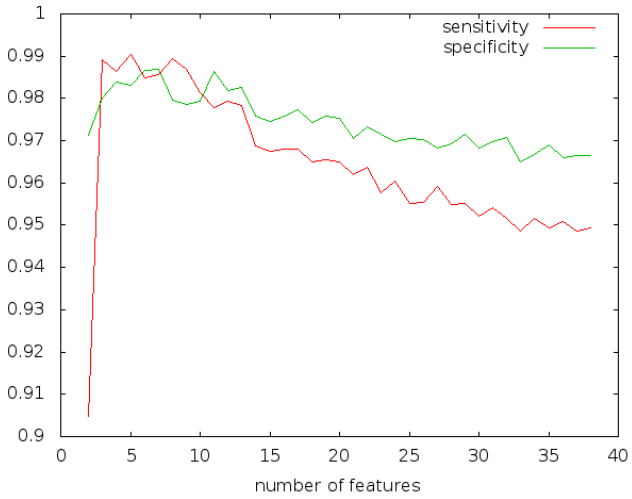

Figure 4.2: Relief

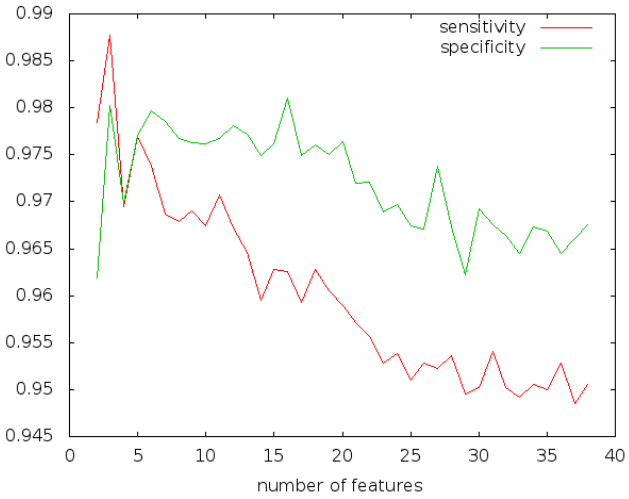

Figure 4.3: Avg decrease in Gini Impurity

Table 4.3: Feature Selection

Method	No of features	Sensitivity	Specificity	Sd(Targets)	Sd(Non-Targets)
Information Gain	12	96.92	97.57	1.96	2.15
Relief	6	98.48	98.66	1.44	1.68
Avg decrease Gini Impurity	3	98.76	98.02	1.15	2.37

From the above table,it can be observed that the best results with a good balance between sensitivity and specificity are obtained when we consider the top 3 features as predicted by "the average decrease in gini impurity".For these top three features,sensitivity and specificity valus are 98.76 and 98.02 percentage respectively. The values for the top 15 features for the average decrease in the gini impurity are as shown in the table below

Table 4.4: Feature score for top 15 features

Feature	Score
Length UTR	118.60
Mfe Seed	90.97
Target Abundance	88.59
AU Contribution	28.12
Pos 2	24.90
Pos 7	20.83
Mfe Full	19.38
Seed GC	15.06
Pos 4	12.61
Pos 5	10.40
Pos 6	10.37
Dist Mid Normal	9.74
AU count	9.54
Dist Middle	9.45
Seed AU	9.22

Since,there is a huge difference in the feature scores after the top 3 features.To find the impact of top three features on the classification,first the top three features from each method i.e Information Gain,Relief,Avg decrease in Gini Impurity were eliminated and using the remaining features,the examples were classified. RandomForest was used as a classifier and average results of five repeatitions of ten fold cross validation was calcuated.Sensitivity and Specificity is plotted for increasing number of features based on their ranking in the three methods.

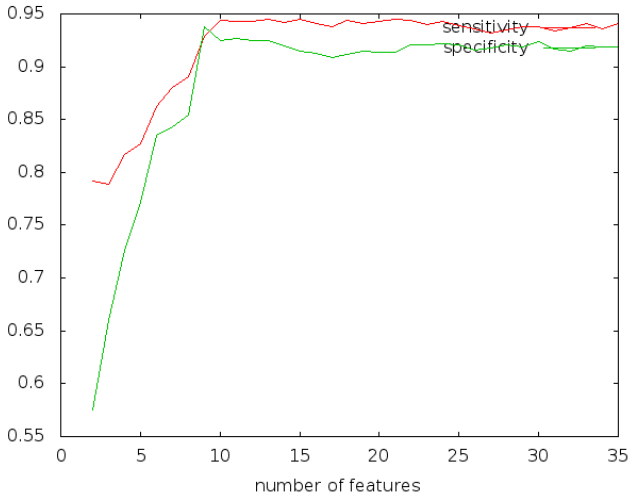

Figure 4.4: Information gain

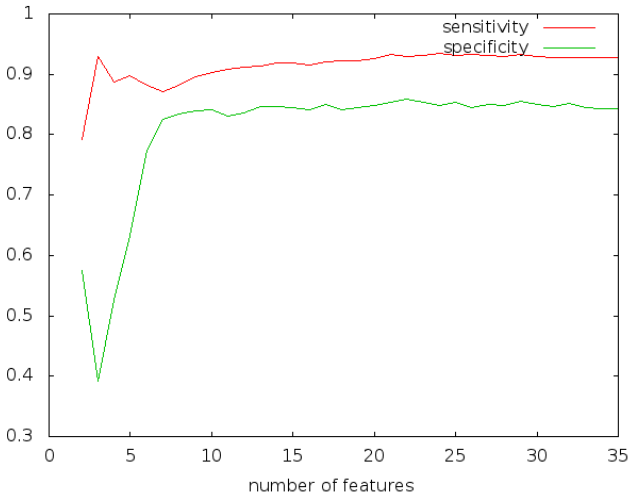

Figure 4.5: Relief

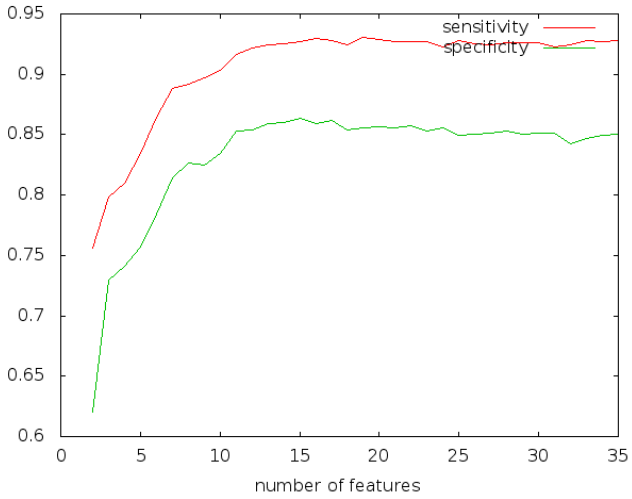

Figure 4.6: Avg decrease in Gini Impurity

Considering the average decrease in Gini impurity graph,it can be seen that sensitivity is around 92 percentage,which is lesser as compared to the 98.76 percentage for the top three features. In terms of specificity, there is marked reduction from 98.02 to 86 percentage. This shows that the top three features play a bigger role in influencing the specificity.Length of UTR and target abundance is common in top three features of all the three methods.Target Abundance is common in Relief and Gini Impurity in the top three and AU contribution is the remaining third feature in Infogain method. Among the three methods,infogain reports the best sensitivity and specificity of around 94 and 92 percentage respectively,after removing the top 3 ranked features. This shows the importance of Target Abundance as a deciding factor.

4.4 Comparison

The original dataset was taken and the results were obtained for the available programs i.e.Miranda,TargetScanHuman conserved),TargetSpy(seed),TargetSpy(non-seed) and RFMirTarget were run.

Table 4.5: Different algorithms

Method	Sensitivity	Specificity
Miranda	74.81	54.27
TargetScanHuman(non-conserved)	72.62	61.32
TargetSpy(seed)	32.11	85.68
TargetSpy(non-seed)	58.75	65.59
RFMirTarget	89.64	89.46

Considering Miranda,TargetScanHuman and TargetSpy algorithms,Miranda performs best with respect to sensitivity but performs worst in terms of specificity.On,the other hand TargetSpy(seed) performs best with respect to specificity but gives bad results in terms of sensitivity.RFMirTarget reports sensitivity and specificity for only those input miRNA-UTR pairs which cross the score threshold of 140 on running the miranda algorithm.
Since,the proposed algorithm works only for the input pairs which have 6/7/8mer sites in the UTR sequence,the existing algorithms were also run on only such pairs and the results obtained are as shown in the table.

Table 4.6: Different Methods(6/7/8mer sites)

Method	Sensitivity	Specificity
Miranda	81.85	32.19
TargetScanHuman(non-conserved)	83.96	31.43
TargetSpy(seed)	33.75	84.09
TargetSpy(non-seed)	58.64	61.36
RFMirTarget	90.38	86.04
Proposed Method	98.76	98.02

Clearly,the proposed algorithm performs better than Miranda,TargetScanHuman and TargetSpy.The reported results for RFMirTarget are on a subset of the input pairs (those which cross the threshold of 140) whereas the proposed algorithm reported results are for all the input pairs.It means that the proposed algorithm works better than RFMirTarget.

4.5 Tarbase 6.0

The second dataset is the Tarbase 6.0[28]. This dataset contains the maximum number of experimentally verified targets and non-targets.Using biomaRt R package[29][30],miRNA and 3'UTR sequences were collected from different databases for humans. But,since a gene can be transcribed in multiple mRNA by the method of alternative splicing ,there are multiple 3'UTR sequences for a given geneId.And it is not known which 3'UTR sequence is the desired one.So, all the 3'UTR sequences were considered and if any one of the UTR's for a given miRNA was found to be predicted as target,the corresponding example was considered as

target.Considering the above criteria,results were obtained for methods Miranda and TargetScanHuman. Miranda reported a 45.26 percentage sensitivity and 56.32 percentage specificity. TargetScanHuman gave a 49.33 and 51.7 percentage ,sensitivity and specificity respectively.TargetScanHuman classifies almost with a 50:50 probability and Miranda also almost classifies targets and non-targets with equal probability.The dataset reports the kind of method used for experimental verification alongwith each example.The results for Miranda and TargetScanHuman separated on the basis of the method used were as shown in the tables below

Table 4.7: Miranda

Method	Predicted as targets(No of targets)	Predicted as targets(No of non-targets)
Reporter Gene	377(514)	68(89)
Northern Blot	4(4)	1(1)
Western Blot	109(167)	3(10)
qPCR	68(141)	12(22)
Proteomics	1247(3017)	30(81)
Microarray	3839(10500)	64(206)
Sequencing	2274(3355)	0(0)
Others	186(207)	5(10)

Table 4.8: TargetScanHuman

Method	Predicted as targets(No of targets)	Predicted as targets(No of non-targets)
Reporter Gene	415(514)	72(89)
Northern Blot	4(4)	1(1)
Western Blot	128(167)	4(10)
qPCR	89(141)	16(22)
Proteomics	1214(3017)	32(81)
Microarray	4241(10500)	68(206)
Sequencing	2553(3355)	0(0)
Others	189(207)	9(10)

The above two tables show that the reporter gene method has the highest sensitivity,but performs badly in terms of specificity.Since,we know that TargetScan effectively reports the 7/8mer sites,it is obvious for the sensitivity to be high but it is not able to classify those 7/8mer sites which are in fact non-targets. Also,reporter gene is a direct method of verification as compared to the other methods which are indirect.In effect,the results obtained from reporter gene have more probability of being correct in comparison to the indirect methods.One reason for high number of 7/8mer sites in the non-targets could be in the approach of finding targets.Generally,the mRNA's which have 6/7/8mer sites present in them are experimentally verified first.The other methods are high throughput methods, in which results can be affected by various other factors.

Since the reporter gene method is the most dependable among all the methods,RFMirTarget and the proposed method was run on reporter gene dataset.For each miRNA-gene pair,the largest UTR sequence was considered.The reason being that shorter UTR sequences are generally contained in the longest UTR sequence. RFMirTarget reported 96.90 percentage sensitivity and 2.02 percentage specificity,whereas the proposed algorithm reported 98.47 percentage sensitivity and 0.40 percentage specificity. This implies that all the examples were classified as targets by both the methods and this is not what is expected.Next,all the

UTR sequences were considered for each miRNA-gene pair,and a sensitivity of 98.47 percentage and specificity of 55.625 percentage was reported by the proposed method. Whereas,RFMirTarget again performed badly with 97.01 percentage sensitivity and 4.05 percentage specificity. In the second approach,specificity showed a marked improvement by the proposed algorithm.One important thing which the above two experiments suggest is that the length of the UTR sequence plays an important role in targeting and care should be taken in considering the largest UTR sequence while preparing the dataset. TargetSpy(seed) reported a sensitivity of 59.33 and a specificity of 43.82,whereas TargetSpy(non-seed) reported a sensitivity of 66.24 and specificity of 56.17 for the reporter gene method by considering all the UTR pairs. But,since the exact UTR sequence corresponding to a miRNA-Gene pair is not known,it can't be said conclusively that the proposed method is the best. The above results only suggest that the proposed algorithm would work better.

Chapter 5

Conclusion and Future Work

The proposed method attempts to solve the problem of classifying examples which have 6mer/7mer/8mer sites and it was shown that the length of UTR sequence plays a significant role in target recognition.But still,the whole problem is far from being solved.As, new type of sites such as centered sites are being reported,it only suggests that there are still many more sites to be found and rules to be defined which can draw a clear line between targets and non-targets.We may look for other criterias others than complemenatarity and free energy to look for potential target sites in the miRNA.Also,the other regions of mRNA i.e 5' UTR and ORF could be investigated more for any concrete conclusions.A new emerging field of genomic signal processing can be explored for a new perspective to the problem.

References

1. Dong Yue,Jia Meng,Mingzhu Lu,C.L.Philip Chen,Maozo Guo and Yufei Huang "Understanding microRNA regulation: A computational perspective" IEEE Siganl Processing Magazine ,January 2012

2. Lee RC, Feinbaum RL, Ambros V "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". Cell,December 1993

3. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G . "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans". Nature 403,February 2000

4. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA". Nature 408,November 2000.

5. Lee RC, Ambros V "An extensive class of small RNAs in Caenorhabditis elegans". Science,October 2001

6. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A "Identification of mammalian microRNA host genes and transcription units". Genome Research ,October 2004

7. Yoontae Lee,Kipyoung Jeon,Jun-Tae Lee,Sunyoung kim,V.Narry Kim " MicroRNA maturation:stepwise processing and subcellular localization " EMBO Journal,2002

8. Rui Yi,Yi Qin,Ian G.Macara,et al. "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs" Genes Development ,2003

9. Bryan R.Cullen "Nuclear RNA export " Journal of Cell Science ,2003

10. Yoontae Lee et al. "The nuclear RNaseIII Drosha initiates microRNA processing " Nature,September 2003

11. Elsebet Lund,Stephan Guttinger,Angelo Calado,James E.Dahlberg,Ulrike Kutay "Nuclear Export of MicroRNA Precursors" Science ,January 2004

12. David P Bartel "MicroRNAs:Genomics,Biogenesis,Mechanism and Function " Cell, January 2004

13. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. "miR2Disease: a manually curated database for microRNA deregulation in human disease" Nucleic Acids Research ,January 2009

14. Matthew W.Rhoades,Brenda J.Reinhart,Lee P.Lim,Christopher B.Burge,Bonnie Bartel,David P.Bartel "Prediction of plant microRNA targets" Cell, August 2002

15. Andrew Grimson,Kyle Kai-How Farh,Wendy K.Jonston,Philip Garrett-Engele,Lee P.Lim,David P.Bartel "MicroRNA targeting specificity in Mammals:Determinants beyond seed pairing " Molecular Cell ,July 2007

16. David M Garcia,Daehyun Baek,Chanseok Shin,George W Bell,Andrew Grimson David P Bartel "Weak seed-pairing stability and high target-site abudance decrease the proficiency of lsy-6 and other microRNAs" Nature structural and molecular biology ,September 2011

17. Hyeyoung Min and Sungroh Yoon "Got target?:computational methods for microRNA target prediction and their extension" Experimental and molecular medicine , April 2010

18. David P.Bartel "MicroRNAs: Target recognition and regulatory functions " Cell, January 2009

19. Chanseok Shin,Jin-Wu Nam,Kyle Kai-How Farh,H.Rosaria Chiang,Alena Shkumatava,David P.Bartel "Expanding the MicroRNA Targeting Code:Functional sites with centered pairing" Molecular Cell,June 2010

20. Enright AJ, John B, Gaul U, Tuschl T, Sander C and Marks DS "MicroRNA targets in Drosophila" Genome Biology ,2003

21. Mariana R.Mendoza,Guilherme C.da Fonseca,Guilherme Loss-Morais,Ronnie Alves,Rogerio Margis,Ana L.C.Bazzan "RFMirTarget:Predicting Human MicroRNA Target Genes with a Random Forest Classifier" PLOS ONE,July 2013

22. Martin Sturm,Michael Hackenberg,David Langenberger,Dmitrij Frishman "TargetSpy:a supervised machine learning approach for microRNA target prediction " BMC Bioinformatics,2010

23. A Krek,D Grun,M N Poy,R Wolf,L Rosenberg,E J Epstein,P Machmenamin,I da Piedade,K C Gunsalus,M Stoffel,N Rajewsky "Combinatorial microRNA target predictions" Nature Genetics ,May 2005

24. Lorenz, Ronny and Bernhart, Stephan H. and Höner zu Siederdissen, Christian and Tafer, Hakim and Flamm, Christoph and Stadler, Peter F. and Hofacker, Ivo L "ViennaRNA Package 2.0" Algorithms for Molecular Biology ,November 2011

25. Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines" ACM Transactions on Intelligent Systems and Technology 2:27:1–27:27, 2011

26. A. Liaw and M. Wiener " Classification and Regression by randomForest." R News 2(3),18–22(2002).

27. Max Kuhn "Building Predictive Models in R Using the caret Package" Journal of Statistical Sofware ,November 2008

28. Vergoulis, T. I. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, AG Hatzigeorgiou "Tarbase 6.0: Capturing the Exponential Growth of miRNA Targets with Experimental Support" Nucleic Acids Research,December 2012

29. Durinck S, Spellman P, Birney E and Huber W (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." Nature Protocols, 4, pp. 1184–1191.

30. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A and Huber W (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." Bioinfor-

matics, 21, pp. 3439–3440.