

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800

Open access books available

122,000

International authors and editors

135M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.

For more information visit [www.intechopen.com](http://www.intechopen.com)



# *In-silico* Approaches for RNAi Post-Transcriptional Gene Regulation: Optimizing siRNA Design and Selection

Mahmoud ElHefnawi<sup>1</sup> and Mohamed Mysara<sup>2</sup>

<sup>1</sup>National Research Center

<sup>2</sup>The University of Nottingham

<sup>1</sup>Egypt

<sup>2</sup>UK

## 1. Introduction

RNA interference (RNAi) is a naturally occurring endogenous biological post-transcriptional cellular mechanism that regulates against foreign genetic elements such as viruses and inserted gene transcripts as well as in-house gene expression regulation. Small interfering RNA (siRNA) molecules utilize this mechanism to promote homology dependent messenger RNA (mRNA) degradation.

The utilization of siRNA as a molecular target to silence gene expression has been used extensively as a research tool in functional genomics. The unprecedented advantage of siRNA molecules, which is mainly related to the ability of effective and specific inhibition of disease causing genes, elicited great expectations in therapeutic applications and drug discovery. siRNAs' potential as a drugs was investigated in viral and cancer models, and showed successful results with diseases such as HIV, HCV and several types of cancer; as most of these diseases have no cure. One advantage of siRNA-based drugs is their feasibility in clinical trials following approval of phase 1. Moreover, they do not rely on an intact immune system which give the advantage over other long double stranded RNA (dsRNA). However, several factors challenge the design of selective siRNA molecules with highly guaranteed silencing efficiency. Therefore, careful selection of siRNAs complying with all necessary properties is crucial for efficient functional performance.

This Chapter discusses RNA interference using small interfering RNS (siRNA) starting with the biological nature of mRNA and siRNA. Then it tackles factors contributing to siRNA-mRNA silencing from both biological and bioinformatics aspects that should affect siRNA effectiveness. Then, it represents step wise workflow for rational siRNA design considering state of the art tools and algorithms. By the end of this chapter, various tools are presented for siRNA evaluation phases that are used to predict siRNA efficiency and efficacy, with a practical example applying the proposed methodology.

## 2. Small interfering RNA

Small interfering RNAs 'siRNAs' are one of the cell defence mechanisms that act against not only exogenous genetic materials like virus genes but also against cell endogenous genes as

one of the post-transcription regulation method (Ullu et al. 2002). These natural-occurring siRNAs target mRNAs (whether they are over expressed or abnormal) in a manner, so selective and potent, that they became the core of interest of many biologists in the last decade. Although siRNAs are not the only layer responsible for post-transcriptional regulation, they have the advantage of hardly invoking the innate immune response (Interferon-response) in contrast to long double stranded RNA (Stark et al. 1998) . In addition, siRNAs, have shown to be very promising new therapeutic agents in various diseases especially in Cancer, Aids and Neurodegenerative disorders as most of these diseases have no cure (Hutvagner & Zamore 2002; Surabhi & Gaynor 2002; Xia et al. 2004). That is why siRNA has been used as a drug for cancer clinical trials on human producing the efficient and specific effect on human as it was expected (Davis et al. 2010).

## 2.1 siRNA mechanism of action

The mechanism pathway of siRNA is as follows: long dsRNA is cleaved by "DICER" a ribonuclease III-type enzyme into the short molecules of siRNA duplexes, being homologous to the mRNA targeted for silencing, siRNA triggers the formation of RNA-induced silencing complex (RISC) in which the double stranded siRNA is incorporated cutting the long double-stranded RNA molecules to double stranded small interfering RNA (ds-siRNA), as illustrated in the [Fig. 1]. Then it is unwound leading to single stranded siRNA that binds to the target mRNA sequence resulting in its cleavage, and according to the type of the RISC complex the RNAi action is directed through mRNA degradation, action arrest or chromatin modification. [5]. This is detailed below:



Fig. 1. Small interfering RNA formed of two short stranded RNA sequences complementary to each other.

Due to the homology (similarity) between the double stranded siRNA (ds-siRNA) and the targeted messenger RNA (mRNA), the aggregation of a complex called RNA induced silencing complex (RISC) is triggered. After binding with ds-siRNA, RISC acts to separate (unwind) the strand making the sense and the antisense strands (passenger and guide strand). After siRNA unwinding into small single strand, it could produce its action with three different mechanisms [Fig 2].

### 2.1.1 Direct cleavage method

The single stranded RNA together with RISC bind to the targeted mRNA and induce its degradation by the Ago-2 degradation (protein triggered by RISC-siRNA complex acts to break the targeted mRNA). The degraded mRNA is finally digested with, what is called, cellular lysosomes. This is the main mechanism by which siRNA causes selective and potent gene silencing, but this only occurs in case of high level of similarity between siRNA and the targeted mRNA region (Birmingham et al. 2006).

### 2.1.2 Seed-mediated translational attenuation

The complementation between the siRNA seeding region hexamer (from the second to the seventh position) and the 3'UTR (untranslated region) of the mature mRNA has been identified capable of inhibition of that mRNA's translation and causing its degradation (E. M. Anderson et al. 2008; Birmingham et al. 2006).

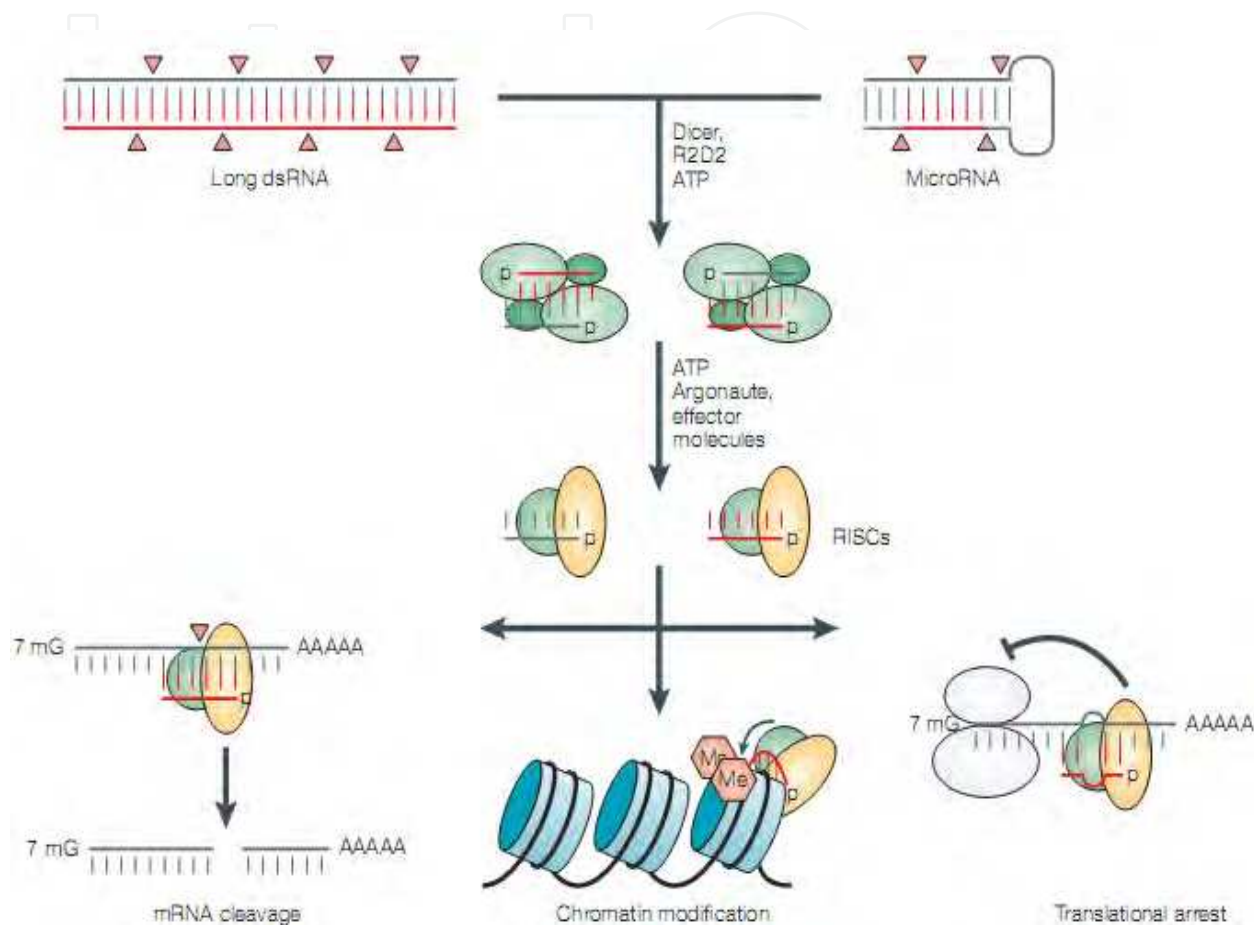


Fig. 2. Naturally occurring siRNA synthesis pathway and three its possible mechanisms of action. Endogenous (naturally occurring) siRNA are produced from either microRNA or long double strand RNA after their cleavage by the Dicer enzyme so they produce double strand siRNA. Both endogenous and exogenous (introduced by researchers) ds-siRNA pass through the activation process starting with unwinding and RNA induced silencing complex (RISC) to the lead single strand siRNA. Then RISC- single strand siRNA complex silence the targeted gene either by one of the three mechanisms: 1) Binding to the mRNA leading to their breakage through Age2 mechanism. 2) Binding to the 3' end and mediate translational attenuation of the mRNA. 3) Gene silencing through chromatin modification [Figure from the work of (Dorsett & Thomas Tuschl 2004)].

### 2.1.3 Chromatin modification

siRNA has another mechanism of interference by chromatin modification as illustrated by Dorsett and Tuschl in their description of Scherer work that siRNA is one of the three major nucleic-acid-based gene silencing mechanisms (Dorsett & Thomas Tuschl 2004).

### 3. Factors that affect siRNA design

In order to understand the interaction between siRNA and the targeted mRNA, several factors have been known to affect the design of effective and specific siRNA. These factors can be further sub classified into four major classes design as illustrated by Birmingham (Birmingham et al. 2007). Firstly, Targeted region or what is called “sequence space”, this section handles the identification of regions in the mRNA to be targeted by the designed siRNA. This step is highly critical as targeting the wrong region would abolish the effect of all designed siRNAs. Sequence space is affected by several factors: Transcript region, Transcript size, mRNA multiple splicing, Orthologs consensus and Single nucleotide polymorphism. Secondly, siRNA sequence space preparation, here we discuss internal repeats, positional preferences, and other desirable/undesirable words/motifs are discussed. Thirdly, siRNA thermodynamic properties and both siRNA and mRNA target accessibility. It includes parameters like GC content, palindromes, in addition to thermodynamic stability and differential ends stability which have been identified to be highly important factors in siRNA selection. Forthly, siRNA specificity describing mechanisms through which siRNA could invoke immune reaction or has off-target effect. Each of these factors can greatly affect siRNA selection and therefore they should be studied thoroughly.

#### 3.1 Target sequence space [Targeted region preprocessing]

Targeted regions (or what is called “sequence space”) are areas of the mRNA that should be assigned for targeting by the designed siRNA. There are five factors affecting the selection of the proper sequence space summarized in (Birmingham et al. 2007).

##### 3.1.1 Transcript regions and size

siRNA should target regions in the mRNA that is not affected by the maturation process, hence targeting 3’UTR, 5’UTR and (most importantly) open reading frame (ORF) [Fig 3]. Normally both 3’UTR and 5’UTR could be excluded from targeted sequence space, unless sequence space needs to be widened. If the mRNA length is < 500 nucleotides, 3’UTR and 5’UTR should be included in target space selection.

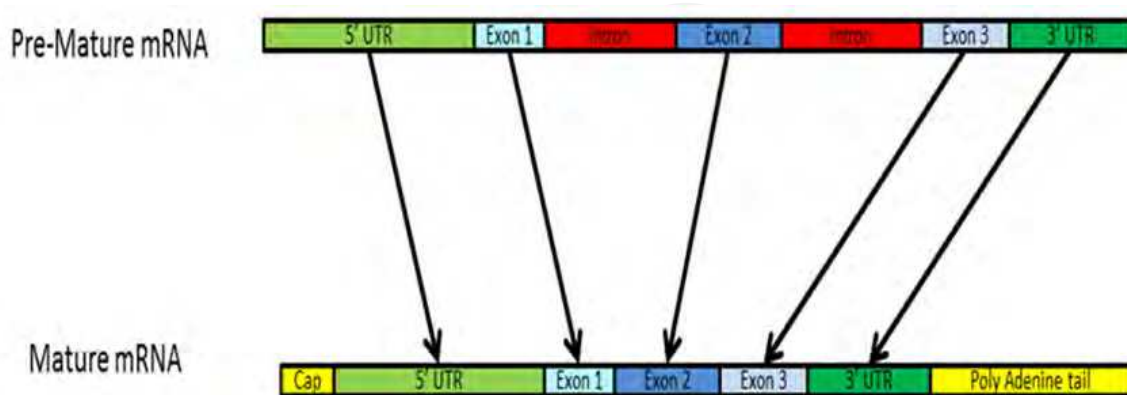


Fig. 3. Maturation process of premature to mature mRNA. This figure illustrates different regions that vary due to omission and insertion during the maturation process. In the maturation process omission of the introns (non coding areas) and addition of 5’ cap and 3’ tail) takes place (Mysara 2010).

### 3.1.2 Multiple splicing and orthologs consensus

One mRNA could be coding for several proteins as the process of splicing is accompanied by rearrangement of exons. There are several mechanisms of alternative (differential) splicing as exon insertion or deletion but the main mechanism; as described in the work of Black; is exon skipping (Black 2003). This phenomena form a huge obstacle if there is a need to target all the mRNA transcripts; therefore, regions in common among them should be recognized and targeted [Fig 4]. All the mRNA's transcripts should be included in target space selection. In case of handling multiple organisms (as in global vaccines or rapidly mutated species as virus) the consensus between different targeted mRNAs should be considered in the target space selection.

### 3.1.3 Single Nucleotide Polymorphism (SNPs)

Single Nucleotide polymorphism (SNP) is very crucial in siRNA design where single (several) Nucleotide(s) difference could cause dramatic shift in the produced protein (or in its regulation) or could have a non-sensible effect in this case it is named silence polymorphism. There are two main locations for SNPs existence non-coding and coding regions [Fig 5]. The first region is the **non-coding region**, where SNP existing in the Introns will not affect the mature mRNA, thus the siRNA targeting it. However, if the SNP is located in the 3' UTR or 5' UTR, caution should be taken in cases where the siRNA is designed to target them. The second region is **coding region**, where SNP exists in the protein coding region (ORF or Exons), there are two possibilities: SNPs will not affect the produced protein due to degeneracy of the genetic code, or it could cause changes in the produced protein, hence siRNA targeting this region should be excluded.

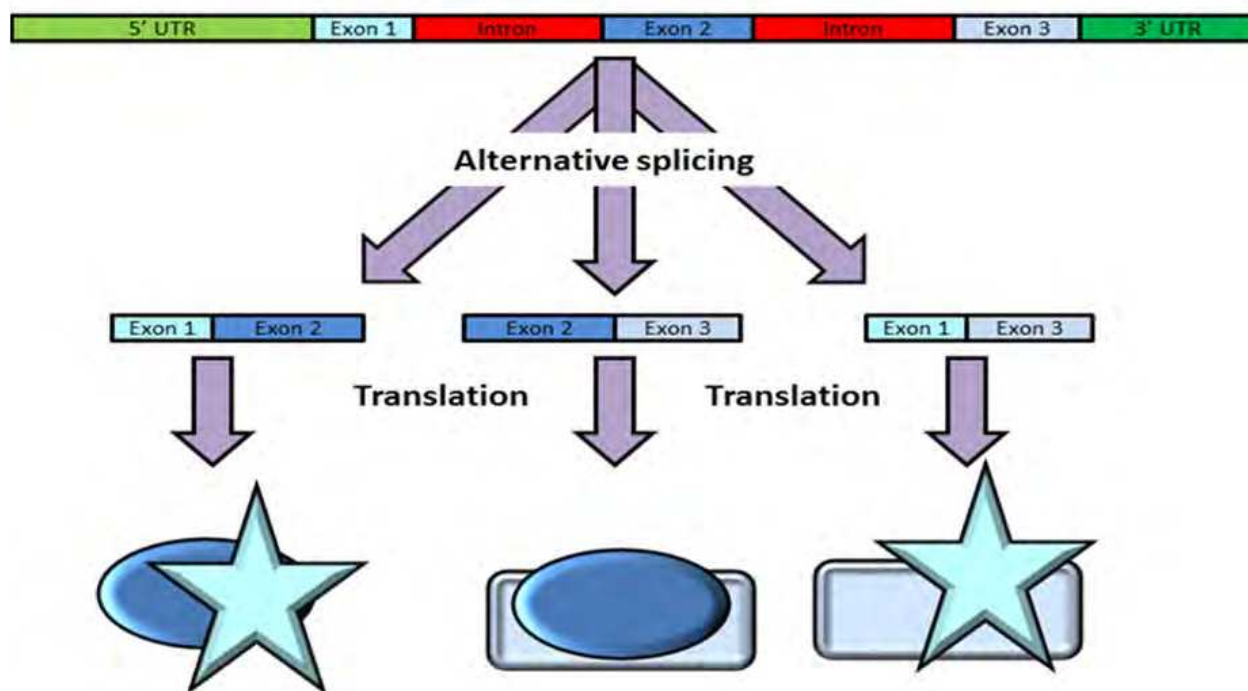


Fig. 4. mRNA alternative splicing phenomena results in several transcripts from the same gene. Each of these transcripts is later translated in a different protein. These proteins functions could be similar or non-similar to each other.

### 3.2 SiRNA sequence space [Positional/word preferences]

Positional/word preferences in the sense/antisense strand of the siRNA are a crucial determinant of siRNA functionality. Several position dependant preferences were identified from analysis of siRNA experimental dataset, which can affect siRNA selection process. Among those preferences within the sense strand (Ui-Tei et al. 2004): (i) A/U at the 5' end of the antisense strand;(ii) G/C at the 5'end of the sense strand; (iii) at least five A/U residues in the 5' terminal one-third of the antisense strand; and (iv) the absence of any GC stretch of more than 9 nt in length. (Reynolds et al. 2004): (I) At least 3 'A/U' bases at positions 15–19 (sense strand). (II) Absence of internal repeats. (III) An 'A' base at position 19 (sense strand). (IV) An 'A' base at position 3 (sense strand). (V) A 'U' base at position 10 (sense strand). (VI) A base other than 'G' or 'C' at 19 (sense strand). (VII) A base other than 'G' at position 13 (sense strand). (Mohammed Amarzguioui & Prydz 2004): asymmetry in the stability of the duplex ends (measured as the A/U differential of the three terminal basepairs at either end of the duplex) and the motifs S1, A6, and W19. The presence of the motifs U1 or G19 was associated with lack of functionality.

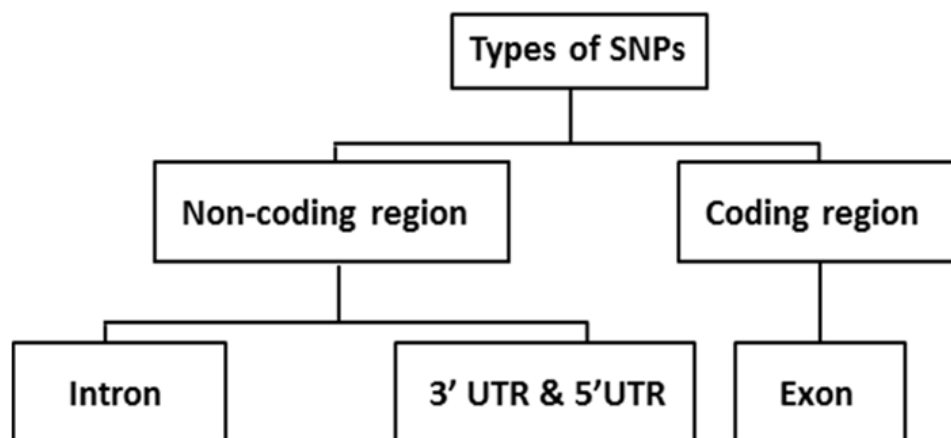


Fig. 5. Classification of SNPs according to region of occurrence in the mRNA.

Several positions in siRNA duplex that could affect their efficiency, as in (Birmingham et al. 2007), candidate duplexes with five or more of any single base in a row, should be removed. Although less detrimental than G/C stretches, repeated bases have also been shown to reduce functionality. Stretches of repeated base-containing sequences are less selective, and A or U/T stretches may additionally target regulatory motifs. Moreover, candidate duplexes with more than six consecutive G's and/or C's stretches of G's and C's have been shown to be one of the strongest negative determinants for siRNA activity that should be removed. Such regions have pronounced local stability, greatly inhibiting duplex dissociation. In addition, GC- rich stretches are not compatible with some synthetic nucleic acid chemistries utilized in vector-based expression.

### 3.3 The target accessibility evaluation

Several studies have been done to illustrate the structural and sequence features affection siRNA functionality, all of these aspects affect siRNA and mRNA accessibility (Patzel et al. 2005; Ladunga 2007). Target accessibility evaluation is crucial for proper designing of efficient siRNA, as mRNA tends to form secondary structure that affects its accessibility and hence reduces the capability to design siRNA targeting certain regions of mRNA. Therefore,

target accessibility evaluation represents where the mRNA is more likely be accessed by short oligomers as siRNAs, it involves not only mRNA secondary structure evaluation, but also energetic calculation of siRNA and mRNA. For interaction between two RNA sequences (siRNA and mRNA) two types of energies are needed: first energy required for opening the binding site, second energy required to gain hybridization the summation of these three energies is defined as interaction energy. The energy required for opening siRNA duplex and mRNA should have lesser than the hybridization energy between siRNA and the mRNA. There are evidence based correlation between siRNA inhibition efficiency and siRNA-mRNA binding energy (Mückstein et al. 2006), that strengthens the findings of Ladunga in which target accessibility information was found to provide the most predictive feature among the 142 features studied and improves the prediction of highly efficient siRNA (Ladunga 2007). Other parameters affecting target accessibility are presented below:

### 3.3.1 GC content

GC content represent the percentage of Guanine and Cytosine (two of the four nucleotides types that build the mRNA) should not be too high in order not to impair the double strand siRNA unwinding and enables the ease of RISC protein entrance.

### 3.3.2 Palindrome

Palindrome should be addressed in target accessibility evaluation where region(s) in one strand binds to another region in the same strand due to reverse complementation. Therefore, palindromes should be avoided in siRNA design as they tend to make intramolecular structure (2ry structure) which impairs RISC binding [Fig 6].

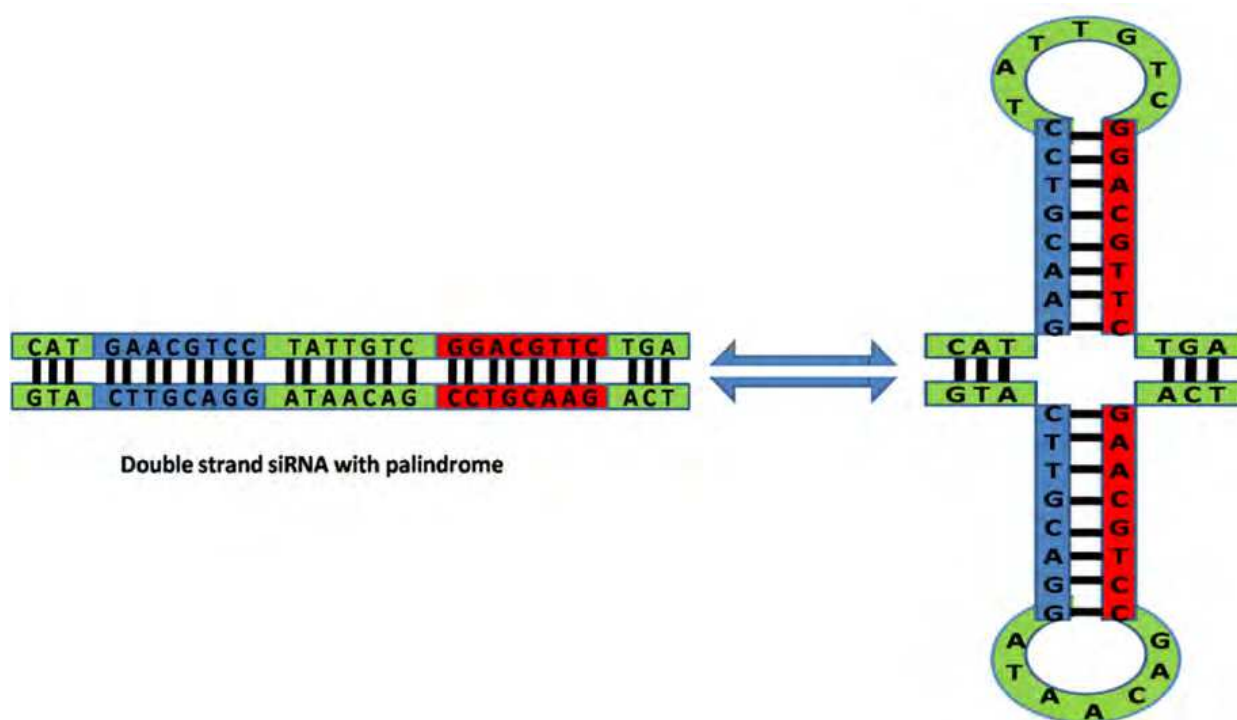


Fig. 6. Palindrome patterns and their affect on siRNA binding to RISC and the targeted mRNA. Palindromes lead to changing double stranded siRNA secondary structure which in turn affects their ability to bind to RISC and the targeted mRNA (Mysara 2010).



### 3.3.3 Thermodynamic stability

It is important to keep/introduce relative thermodynamic stability at both ends of the siRNA (3'UTR and 5'UTR) and low stability at the central zone as these facilitate ds-siRNA cleavage.

### 3.3.4 Differential end stability

Differential end stability is considered one of the most important features that affect siRNA functionality (Schwarz et al. 2003) [Fig 7]. RISC binds to either sense or the antisense strand, but with different ratio. This ratio depends on "Differential stability" between the first couple of bases of the 5' end from both strands. As these couple of bases affect what is called Thermo-Dynamic stability (TDS), so the lower the stability the better it binds to RISC. It has been found that only the antisense (leading) strand is capable of causing gene silencing (Dorsett & Thomas Tuschl 2004). Therefore, it is essential design siRNA with low TDS at 5' end of the antisense than TDS of the 3' end, to have a better binding with RISC and better efficiency in silencing the target mRNA.

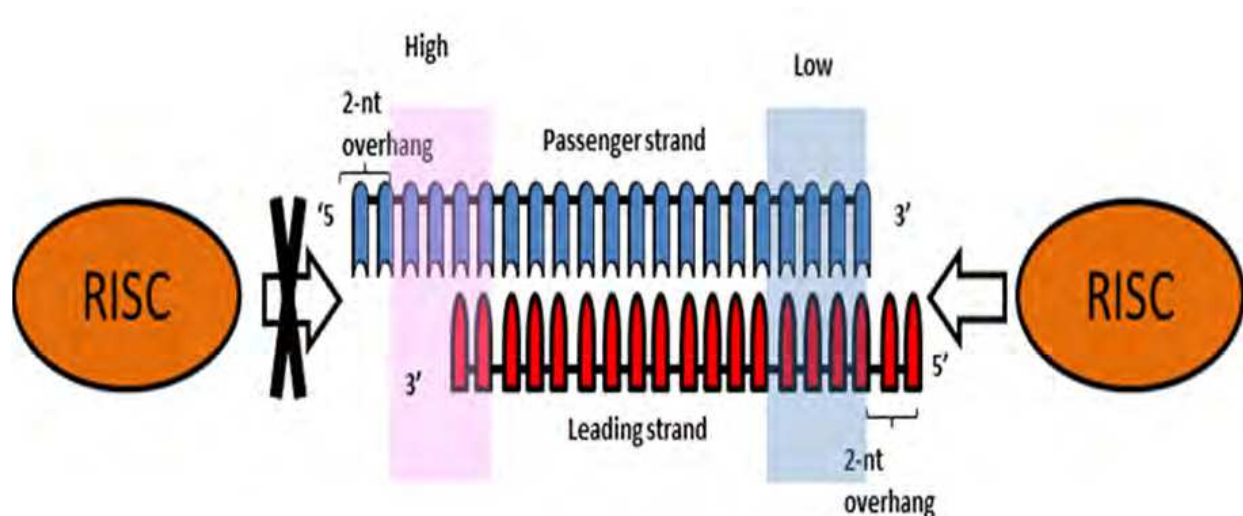


Fig. 7. RISC and Differential end stability. This figure illustrates the effect of differential end stability of RISC annealing with ds siRNA. Therefore, it is very important to ensure the 5' end of the siRNA lead-strand is less stable than the 5' end of the passenger-strand. This way RISC would form complex with only the lead-strand that is designed to bind to the targeted mRNA.

### 3.3.5 The number of single-stranded base pairs at the 5' and 3' ends of the target mRNA

It has recently been shown to significantly contribute to the effectiveness of siRNAs by Patzel and Kaufman in their recent (S. H. E. Kaufmann & Patzel 2008). (Patzel, Rutz et al. 2005) (Patzel, Rutz et al. 2005) (Patzel, Rutz et al. 2005) (Patzel, Rutz et al. 2005) The same conclusion was reached by Gredell and co-workers in July 2008.

### 3.4 siRNA specificity [The off-targeting effect]

"Ideally, the siRNA must not cause any effects other than those related to the knock down of the target gene" (Semizarov et al. 2003). It is essential that the designed siRNA affects only the

targeted mRNA. In other words, siRNA should not invoke innate immunity nor has any off-targeted mRNA.

### **3.4.1 Innate immunity effect**

Concerning the innate immunity effect, by rational selection of appropriate length of the siRNA (21-23 nucleotides) the innate immunity will not be triggered (Birmingham et al. 2006). Although, duplexes of less than 30 nt are short enough to evade immunorecognition by cytosolic double-stranded RNA (dsRNA) receptors, but are long enough to trigger Toll-like receptor 7 sequence-dependent recognition (Patzel et al. 2005). Recognition of motifs as 5'-GUCCUCAA-3', 5'-UGUGU-3' and tetrad-forming poly(G) stretches and avoidance of their presence in the sensitized siRNA, help over coming Toll-like receptor recognition. There was several works using chemical modification in order to mask the innate immunity response (interferon response) (Patzel 2007).

### **3.4.2 Off-target effect**

Apart from that, comes the problem of off-targets which is one of the most important factors for siRNA selection and filtration. "siRNA off-target" is mainly any target that is affected by siRNA other than the assigned target. It is very common for siRNA to have a multi-target as they are only 21-23 nucleotide length; therefore, there is a good chance siRNA could match with more than one mRNA. In fact, as observed in the work of Jackson et al, both sense and antisense and know to have an off-target effect with several mRNA transcripts (Jackson et al. 2003; Jackson & Linsley 2010). There are different mechanisms through which siRNA can trigger off-targeting actions:

#### **I) Complete or near complete off-target matches (siRNA-like effects)**

This mechanism is triggered whenever the designed siRNA is completely identical (or with one mismatch) with a region in the off-targeted mRNA. This complete (near complete) matches between siRNA and mRNA leads to the destruction of that mRNA with the same mechanism that siRNA silences the targeted mRNA as described before.

#### **II) Partial off-target (miRNA-like effects through Seed matching off-target)**

If the designed siRNA seeding region (second to seventh position) matches with 3'UTR of off-targeted mRNA, this will result in affecting the off-target translation as illustrated by (E. M. Anderson et al. 2008). Therefore, these siRNAs are considered as partial off-targets and should be excluded. Chemical modifications have been applied here to reduce off-target and increase the specificity (Birmingham et al. 2007). These off-target effects (complete, near complete or partial) are responsible for loss of specificity as they make this unwanted silencing with other proteins synthesising genes. Moreover, they also cause loss of siRNA potency as the unwanted off-targeting of other mRNA could lead to unavailability of these siRNAs at the original targets (Semizarov et al. 2003; Vert et al. 2006).

In addition to those types of off-target effects, there is the protein interaction. As siRNAs are known to bind to different cellular proteins and alter them, which is known as "Aptamer Effect" as described in (Semizarov et al. 2003). Moreover, avoidance of sequence motifs interfering with RNA synthesis and purification should be considered, as Guanine-rich RNA sequences and sequences containing consecutive stretches of more than three G bases (Patzel et al. 2005).

### 3.5 siRNA duplex chemical modification

Several chemical modifications could be introduced to the designed siRNA in the aim of enhancing its tolerability, improving its stability, limiting its off-target effect and conjugation with tracking agent properly. There are multiple types of chemical modifications that are typically introduced into siRNAs, as summarized in the work of (Birmingham et al. 2007):

I) Sense strand disabling: it is done to increase the specificity and efficiency of siRNA designed. Various approaches were used as 2' ribose modifications including 2'-OR where R<sup>1/4</sup> fluoro, alkyl, O-alkyl<sup>40-43</sup>; LNA modifications at the 5' end of the sense strand.

II) Stabilization: Chemical modifications of the phosphate backbone (e.g. phosphorothioate linkages), the ribose (e.g. locked nucleic acids, 2'-deoxy-2'-fluorouridine, 2'-O-ethyl), and/or the base (e.g. 2'-fluoropyrimidines) increase the resistance of siRNA to nuclease. Stability of siRNAs in biological fluids needs various modifications as 2'-halogen, 2'-alkyl and/or 2'-O-alkyl modifications of one or both strands of the siRNA as well as stabilizing internucleotide modifications of the overhangs. Care should be taken when addressing those modifications not to interfere with siRNA efficiency.

III) Specificity: Chemical modifications in the aim of increasing siRNA specificity and decrease its off-target activity, include includes 2'-O-alkyl modification of unique positions of the sense and/or antisense strand. These modification patterns severely limit sense and antisense off-target effects by disrupting seed-mediated off-target activity.

IV) Conjugations: siRNAs have been conjugated with lipophilic derivatives of cholesterol, lauric acid or lithocholic acid to enhance their cellular uptake and specificity (Lorenz et al. 2004). The safest sites for conjugation are the 5' and 3' termini of the sense strand.

## 4. Guidelines for siRNA rational design

After have discussed factors influencing the siRNA efficacy, here we present our methodology and phases for efficient siRNA rational design. Originally, this methodology was inspired by the repeated Influenza pandemics, and our trials to design a novel siRNA therapy that would work for any new pandemic (ElHefnawi, Alaidi et al. 2011). There are seven phases that should be considered for proper designing of siRNA with high specificity and efficiency [Fig 8].

### 4.1 Targeted gene selection

"Targeted gene" selection is extremely critical for siRNA design as the purpose of gene silencing is to stop the expression of specific abnormal proteins most commonly be involved in the biological pathways as "cancer pathways". Therefore, the protein of interest should play a key role in this pathway, in order to produce the desired therapeutic effect from the silencing process. Therefore, there is a need to search the key regularity protein annotated in various biological pathway databases as "Reactome" and "KEGG" (<http://www.reactome.org/> & <http://www.genome.jp/kegg/pathway.html>) and design siRNA capable of targeting them.

### 4.2 Targeted sequence specification and filtration

After selecting the gene of interest, as gene itself is not targeted but rather its transcript(s), all the available transcripts should be located. In some instance only one transcript should be targeted; in this case all other transcripts should be excluded as targeting them is considered as lack of specificity (off-target). But on the other hand, if there is a need to silence all of the gene's transcripts (which is very common), several options are available for handling such situation [Fig 9] either by mapping the transcripts on the genome, as

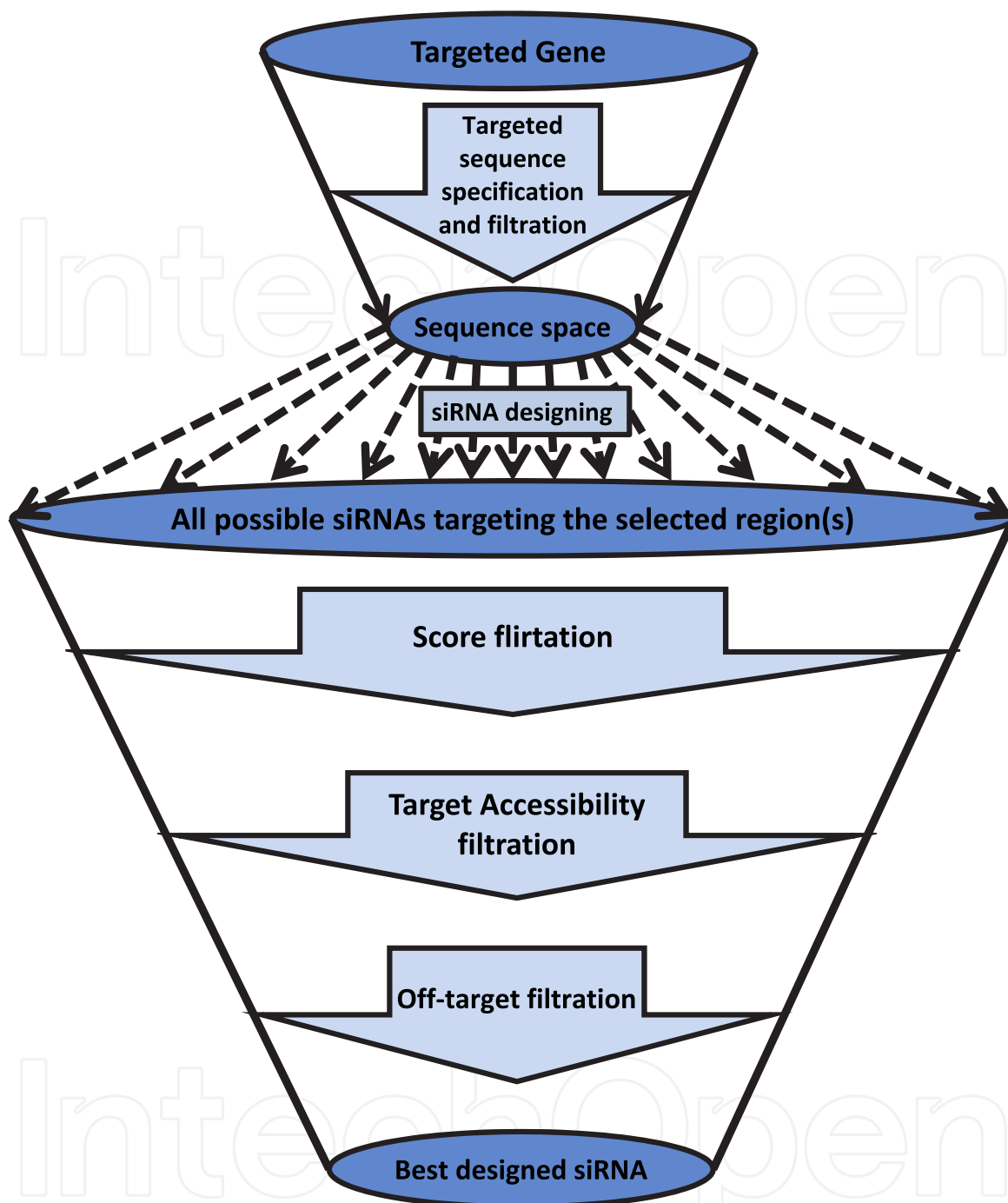


Fig. 8. Different phases for designing siRNA with high efficiency & sensitivity. There are seven distinguished phases for siRNA design: 1st choosing the targeted gene for silencing. 2nd identifying the proper target sequence space that represent all gene's transcripts and doesn't have any unstable regions. 3rd designing all possible siRNA with nineteen nucleotides length with both sense and antisense strand. 4th these potential siRNAs are scored and evaluated according to several scoring mechanisms and criteria and then filter them according to produced scores. 5th siRNA are filtered according to target accessibility. 6th off-target filtration of the remaining siRNA is performed excluding siRNAs with unwanted off-target effect. 8th select the best designed siRNAs that pass all the previous filtration phases and achieve the highest predicted efficiency (Mysara 2010).

proposed in the work of (Y.-kyu Park et al. 2008), or via using multiple sequence alignment (MSA). Multiple sequence alignment is performed either by aligning different gene transcripts and selecting the transcript with the highest identity to the alignment profile, in other word; select the transcript that is more capable of representing all the other transcripts. Another manoeuvre is by considering all transcripts' regions in common (conserved).

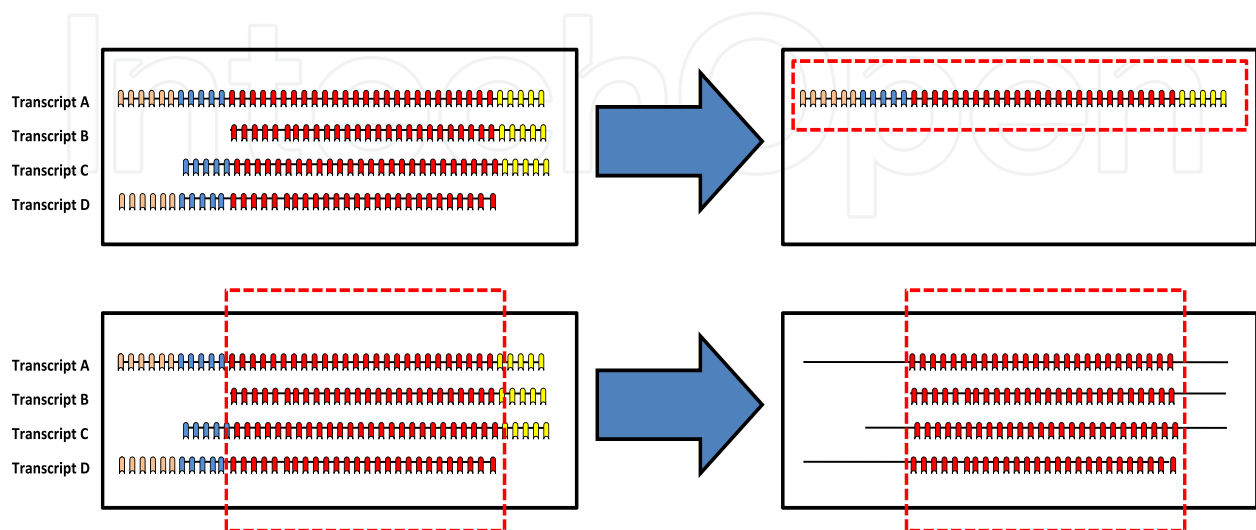


Fig. 9. Different approaches of handling multiple gene transcripts. There are two proposed approaches 1st by aligning between different gene transcripts in order to get the alignment profile and choose the closest transcript to the alignment profile. 2nd way is to get the gaped consensus between these transcripts and choose the regions that are 100% conserved between them (Mysara 2010).

The latter method ensures designing siRNA targeting all the gene's transcripts, this is very important as one mismatch between the target mRNA (transcript) and siRNA could dramatically affect siRNA efficiency (Czaderna et al. 2003) there was noticeable decrease in the efficiency of designed siRNA when induced central single nucleotide variation between the siRNA and targeted mRNA, all together with the findings in the following works (M. Amarzguioui 2003; Sayda M Elbashir et al. 2002). However, one of the main disadvantages of this approach is that it narrows the target sequence space and it is possible that no active siRNA will pass the multi-filtration phases described in [Fig 8]. In this occasion, sequence space should be widened via inclusion of (3'UTR and 5'UTR) or using the first approach. After locating the targeted sequence space, both SNPs and unstable (highly variable) positions should be identified (if any) and any designed siRNAs targeting these residues/regions should be rejected. This way the targeted sequence space will be limited to mRNA (or the conserved region among different gene transcripts) either representing the ORF or (ORF + 3'UTR + 5'UTR) free from any SNPs or unstable regions.

#### 4.3 Designing all possible siRNA targeting the selected regions

This section illustrates the proper siRNA length that ensures high efficiency and stability having neutral effect on host innate immunity. Then it discusses how to select siRNA from the mRNA sequence space.

### 4.3.1 Selection of the appropriate siRNA length

Using siRNA (with its short length) has better advantages over using long double stranded RNAi as they do not trigger immune response and they also silence the targeted mRNA more efficiently. However, siRNA with length equal to thirty nucleotides were found to be inactive (S M Elbashir, Lendeckel, et al. 2001). After that the selection of proper siRNA length was heavily studied, upper and lower limits have been assigned. It was found that shortening the length from nineteen to seventeen affected the siRNA capability to silence the targeted gene; as at least nineteen nucleotides are required for RISC binding. (Czauderna et al. 2003). To establish the upper length limit, it was found that siRNAs with length from (18 to 23) are at least eight folds more effective than other lengths. In addition the 24-25 nucleotide length siRNAs were completely inactive (S M Elbashir, J Martinez, et al. 2001). Therefore, siRNA with length 19-21 plus a 2-nt overhang is the appropriate length for siRNA design and any further deviation above or below this length threshold will have a direct effect on the siRNA activity. It was also demonstrated that using 2-3mer nucleotides dT 3' UTR overhangs increases the efficiency of antisense strand loading to the RISC complex.

### 4.3.2 Picking up siRNA from sequence space

After establishing the desired siRNA length (19 nucleotides), all possible siRNA molecules should be considered using one nucleotide shift per time [Fig 10] till reaching the end of sequence space. Although the ideal case is that each gene would have only one transcript with no SNPs nor highly variable regions, the vast majority of the gene's sequence space would be separate pieces (not intact) as shown in [Fig. 9]. Therefore, the selection of the 19 nucleotide length siRNA should only be restricted to those sequence spaces free from any gaps.

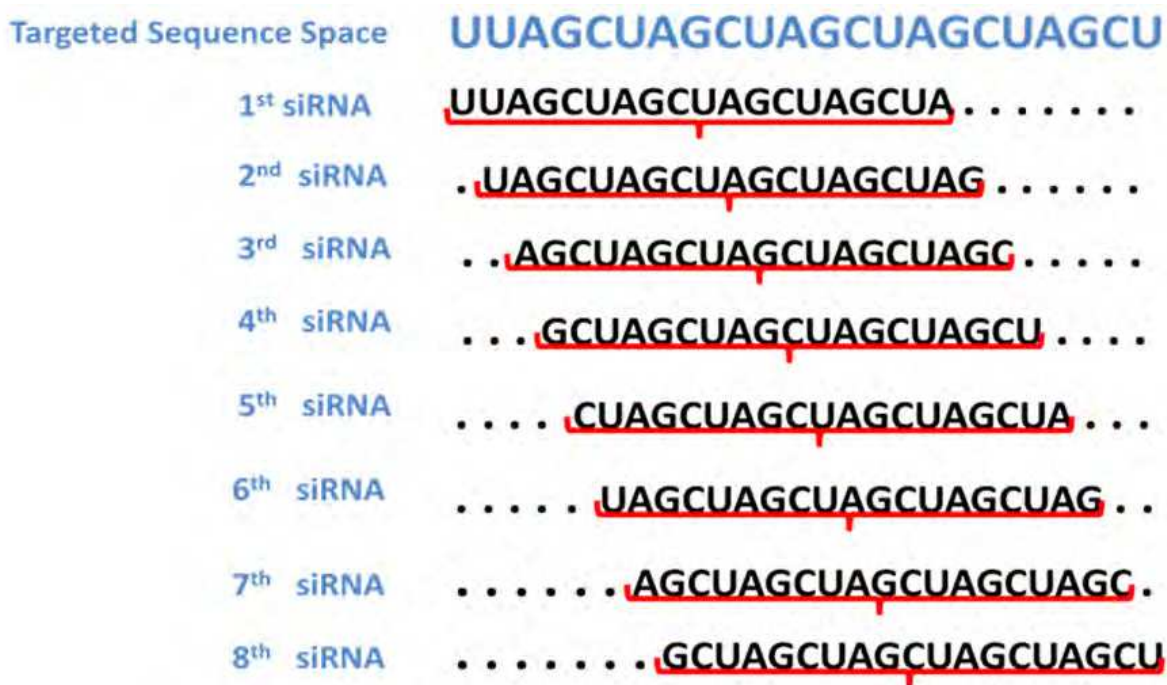


Fig. 10. Designing of all possible siRNA using fixed frame shift. After choosing the appropriate length (most propably 19 nucleotides + 2 overhangs) the target sequence space is scanned and all possible siRNAs with one nucleotide shift at a time(Mysara 2010).

#### 4.4 siRNA scoring and scores filtration

This stage is the most important stage for siRNA design as proper scoring and evaluation of siRNA activity assist the time and cost consumption. Moreover, developing siRNA scoring tools with enhanced specificity and sensitivity would also serve a lot in that regard. As normally single mRNA would produce thousand potential siRNAs, these siRNAs need to be evaluated in order to filter them to smaller number suitable for experimental testing. There are several tools have been developed to predict siRNA activity; these tools differ a lot in these prediction capabilities in the terms of specificity and sensitivity. They use several rules and trained with various datasets, therefore, careful evaluation and picking up the right tool is essential for proper siRNA scoring phase. The details regarding siRNA scoring is further explained in the next section of the chapter.

#### 4.5 siRNAs target accessibility filtration

For interaction between two RNA sequences (siRNA and mRNA) two types of energies are needed: first energy required for opening the binding site, second energy required to gain hybridization. There are several programs that is used to calculate each energy among them *RNA duplex* is capable of calculating duplex energy and *RNAplfold* capable of calculating opening energy for ds-siRNA and targeted mRNA (target site accessibility energy). Both *RNA duplex* and *RNAplfold* belong to Vienna RNA package <http://www.tbi.univie.ac.at/~ivo/RNA/>. There are two more tools that are able to provide better advantages, *RNAup* and *RNAxs*.

##### 4.5.1 RNAup

*RNAup* (that also belongs to Vienna RNA package) is capable of calculating all the three energies required for assisting the interaction energy (Mückstein et al. 2006). *RNAup* starts with calculating the probability that the sequence intervals (after splicing the sequence in small subsequences) are unpaired. Then, it computes the interaction energy, ending with choosing the ones with the least free energy (i.e. the highest stability). However, it cannot handle sequences longer than 5000 nucleotides as it needs a lot of memory.

##### 4.5.2 RNAxs

*RNAxs* program (modification of the older *RNAplfold*) is one of the programs used to evaluate siRNA efficiency according to target accessibility evaluation; it combines *RNAplfold*, *RNAfold* and *RNA duplex* (Tafer et al. 2008). *RNAxs* is able to provide two major advantages: Time reduction and single phased process, which has been shown in the work of (Hofacker & Tafer 2010) the comparative experiment done between *RNAxs* and 2 other target accessibility based programs (*OligoWalk* & *Sirna*). It was found that only *RNAxs* is able to identify siRNAs with inhibition efficiency >50% and to classify 50% of experiment siRNA producing prediction capability higher than the other two programs.

#### 4.6 siRNAs off-target filtration

All the siRNAs that pass the assigned thresholds for each scoring tools are filtered by their tendency to trigger off-target effect. As described in section later (under siRNA specificity), first siRNA having complete matches (or near complete) with the off-target mRNA should be excluded (19/19 or 18/19 or 18/18). Next, the rest of the siRNAs are filtered according to

the presence of partial off-target by excluding siRNAs with matches between their seeding regions (second to seventh position) and the 3' UTR of the off-targeted mRNAs. This way the selected siRNA candidates would have the required specificity [Fig 11].

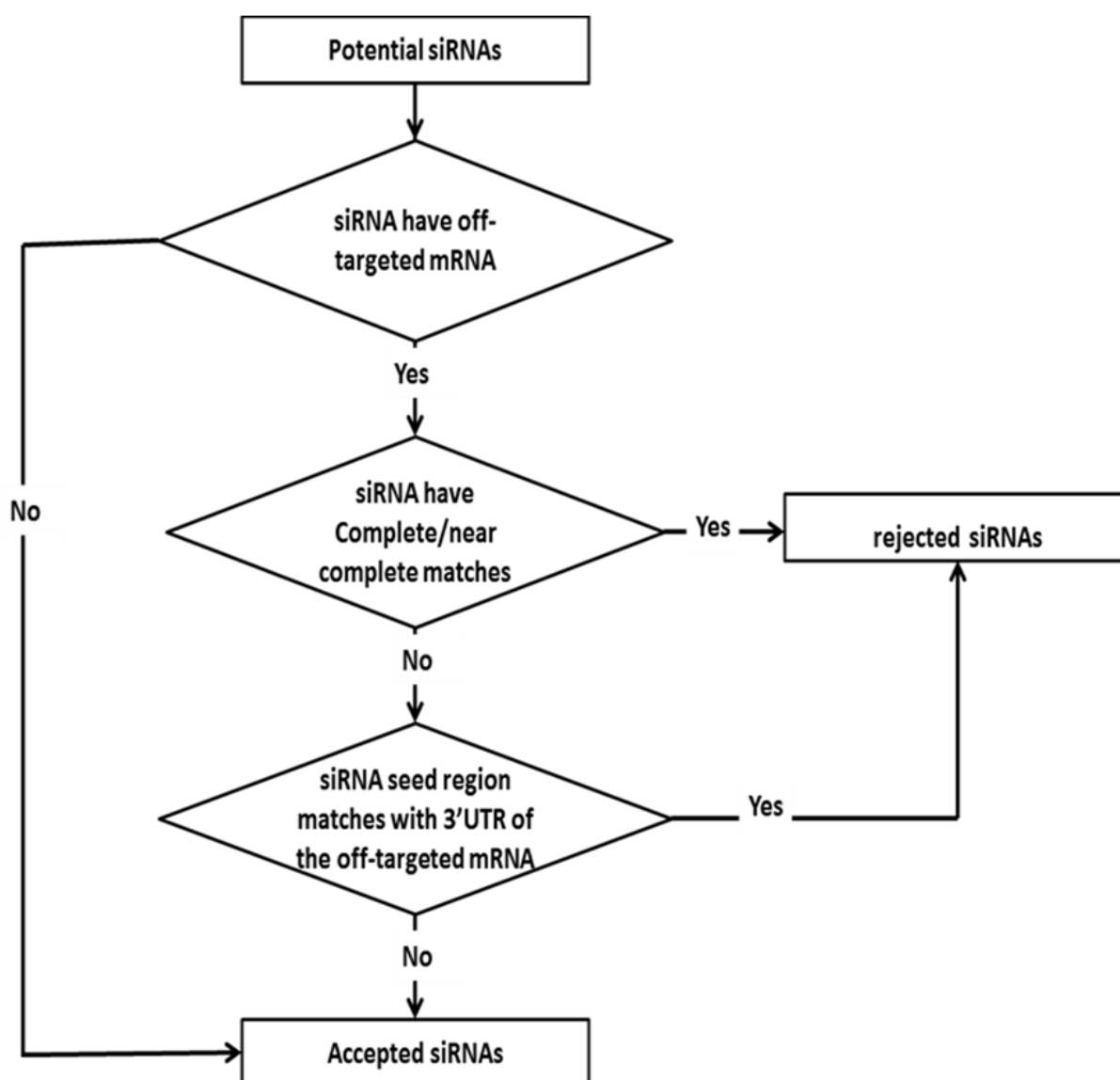


Fig. 11. Off-target filtration workflow describing decision making process for siRNAs off-target filtration.

#### 4.7 Selecting the best designed siRNA

The final step is sorting the acceptable siRNAs candidates according to the predicted inhibition score. Taking the top 10-50 (if applicable) and order them for synthesis by adding UU or dTdT to the 3' ends. Final result is a double strand siRNAs containing leading strand and antisense strand with two 3'end overhangs. There are also several chemical modifications could be applied to the ds-siRNA would serve in increasing the stability, efficiency or neutralizing the immune response by various possible modifications (Birmingham et al. 2007).



#### 4.8 Automation of siRNA design

This stepwise approach for designing siRNAs with acceptable target accessibility properties, passing predicted score, SNPs and off-target filtration could be automated using various programs and tools. The extend of considering those steps varies from program to another according to the used algorithm and the state of the art protocol available at its time, in our previous work, we managed to develop MysiRNA-Designer, siRNA design tool that implements all of the steps presented above [Table 1](Mysara, J. Garibaldi, et al. 2011).

Tools name	Multi-transcripts Consideration	Conserved Region Analysis	SNPs Evaluation	Multi- algorithms Scoring	2ry structure Evaluation	Target accessibility	Full Homology Off-target	Seed Region off-target	Server Based
MysiRNA-Designer	+	+	+	+	+	+	+	+	-
siDESIGN Center *1	+	+	+	-	-	-	+	+	+
Asi-Designer *2	+	-	+	-	+	-	+	-	+
RNAXs *3	-	-	-	-	+	+	-	-	+
siDRM *4	-	-	-	-	-	-	+	+	+

Table 1. Comparison between MysiRNA-Designer and several programs used for siRNA full automation designing. This Comparison involves tools ability to perform alignment between different transcripts, conserved regions consideration, all together with siRNA candidate evaluation using several algorithms and target accessibility. siRNAs filtration by the presence of Single Nucleotide Polymorphisms and off-targets (both full homology and seed regions)(Mysara, J. Garibaldi, et al. 2011, submitted). \*1 siDESIGN Center at <http://www.dharmacon.com/designcenter/DesignCenterPage.aspx>. \*2 Asi-Designer available at <http://sysbio.kribb.re.kr:8080/AsiDesigner/menuDesigner.jsf>. \*3 RNAXs available at <http://rna.tbi.univie.ac.at/cgi-bin/RNAXs>. \*4 siDRM available at <http://sidrm.bioclead.org/index.php>.

#### 5. Models used for predicting siRNA activity

There are several methods for scoring and predicting designed siRNA activity, some of them are more accurate than the others; however, they are classified into two groups (Ichihara et al. 2007): (i) Huesken dataset non-dependant [first generation]. (ii)Huesken dataset dependant [second generation]

### 5.1 Huesken dataset non dependant [First Generation]

These tools were developed to select the most efficient siRNAs, and they depend on differential thermodynamic stability measures, mRNA secondary structure and base preferences specific position target uniqueness. Example of these rules: Reynolds (Reynolds et al. 2004), Amarzguioui (Mohammed Amarzguioui & Prydz 2004), Takasaki (Takasaki et al. 2004), Katoh (Katoh & Suzuki 2007), Ui-Tei (Ui-Tei et al. 2004), Hsieh (Hsieh et al. 2004). However, these first generation scoring techniques have shown to have low accuracy, as up to 65% of the siRNAs predicted as active (by these tools) failed to achieve 90% inhibition when tested experimentally and up to 20% of them were false positive, as described by (Ren et al. 2006). Therefore, there was a need for another approach that does not only take the site-specific position into consideration but also implement data mining techniques to interpret the experimentally obtained data.

### 5.2 Huesken dataset dependant [Second Generation]

This class has been developed mainly through experimental data observation, as the existence of a dataset with fully annotated experimentally siRNAs with their different efficiency enabling sophisticated data mining handling of this data, was not available until the dataset of Novartis that was introduced by Huesken (Huesken et al. 2006) and used for training of several scoring tools as: *Biopredsi* (Huesken et al. 2006), *DSIR* (Vert et al. 2006), *ThermoComposition21* (S.A. Shabalina et al. 2006), *i-Score* (Ichihara et al. 2007) and *Scales* (Matveeva et al. 2007).

These scoring techniques predict siRNA efficiency more accurately than the older tools. Although they use completely different algorithms to evaluate siRNA efficiency, they have very close accuracy compared to the rest of the second generation algorithms, as described by (Ichihara et al. 2007). As in the comparative study done in the Ichihara's work all the second generation (except for Scales) and only Reynold and Katoh from the first generation achieved 90% successful prediction. Moreover, sensitivity of the second generation compared to Reynold and Katoh (which appear to have approximate accuracy) was at least 8 fold lower than the second generation sensitivity (this also supports Ren's findings mentioned earlier). Here, we handle the basic information of each member of this group of tools and provide comparison between them [Table 2].

#### 5.2.1 Biopred

In Biopred, artificial neural network (ANN) was trained using huge number of records (2,182 training and 259 test), considering not only single nucleotide residue but certain patterns (as di-nucleotides). This work is considered the start of the second generation siRNA approaches and noticeable shift in the scoring accuracy. Although ANN used in this work provided ambiguity to the model and prevented further development (due to the complexity of the model), it was considered, at that time period, the best way to handle all these different parameters. The server based Biopred model was later simulated and released as Biopredsi excel-based tool together with i-Score, which is going to be illustrated later on (Ichihara et al. 2007).

#### 5.2.2 DSIR

In DSIR, they used the exact training and test data as Biopredsi but with simplified linear regression model to give prediction based on two main sequence features and three main parameters with Pearson Correlation coefficient = 0.67. The main sequence feature is A/U presence at the first position of the 5' end guidance strand and the absence of Cytosine from

both positions seven and eleven. The main parameters that have been used to build the model are: sprase21, spectra21, composition representation. These three parameters divide the siRNA by different manners and calculate the total score of all of them providing a very representative and interpretable method to evaluate a siRNA sequence, see Table 2.

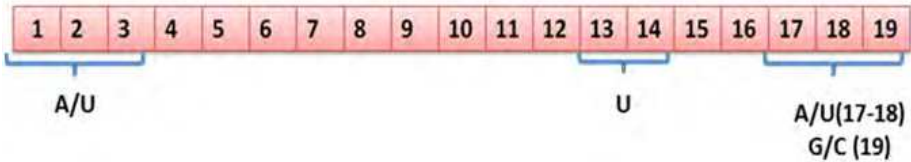
Parameter	Description
<b>Position dependant consensus</b>	<p>As several positions has been identified to be conserved between effective siRNA, so the are scored out of 11 for the presence of desirable residues and out of 10 (in -VE charge) for the presence of undisirable residue in specific position.</p> 
<b>Dinucleotide Content Index</b>	<p>As the occurrence of some dinucleotide combinations have exceeded the random distribution, so by combining these unique pairs with the level of effectiveness of these siRNA, it was found (or precisely confirmed) the low frequency of G/C dinucleotide pairs accompany high siRNA efficacy.</p>
<b>Thermodynamic Profile &amp; Free Energy (<math>\Delta G</math>)</b>	<p>It was found that the difference between 5' and 3' in free energy (or its oppose "stability") especially at the last 2,3,4,5 from each side plays a crucial role in not only distinguishing the sense and antisense but in efficiency evaluation.</p>

Table 2. Description of parameters considered by ThermoComposition (Mysara 2010).

### 5.2.3 Thermo composition

Here a small number of parameters have been used in to train neural network using 653 siRNA-records as a training set. These parameters have been carefully selected from 18 parameters leading to this small number of parameters (three parameters), that had provided the advantage of simplicity over other neural network as no need for huge number of training dataset is required, and that opened the door for any further development. The uniqueness in this work is that it combines "the position dependant features" with "Thermodynamic features" [Table 2].

### 5.2.4 i-Score

In i-Score, linear regression model was built on identifying the nucleotide that is preferred in each position and calculated the inhibition score (i-score) working on Huesken dataset (2431) with Pearson correlation coefficient = 0.635. Also in this work they pointed out a very important threshold as the exclusion of Thermostable siRNA (with stacking energy (whole  $\Delta G$ ) < -34.6 k.cal) improved the score accuracy of not only i-Score but also DSIR, Biopredsi and ThermoComposition21 (Ichihara et al. 2007).

### 5.2.5 Scales

Linear regression model fitting with local stability of siRNA duplex and other parameters was the way Matveeva’s team managed to score siRNA in “siRNA Scales”, using Huesken dataset for training and three other dataset from various pharmaceutical companies for validation. The use of linear regression provided additional advantages over neural network, as it enabled the introduction of relevant importance to the same parameter at different positions which cannot be applied to the same node parameter in the neural network. In “scales” the linear regression was build on two sets of parameters: the first group covers the stability of siRNA ends especially the 1st and last two base pair of the siRNA the second group depends on evaluation of certain nucleotide at specific positions. A comparison between all second generation tools is provided in [Table 3] (Mysara 2010).

However, all these models have limitations in performance. There are recent efforts to enhance the siRNA scoring functionality through applying a second artificial intelligent layer that depends on the predicted scores of other second generation tool, as in MysiRNA model. It is siRNA functionality/efficacy prediction model that was developed by combining two existing scoring algorithms (ThermoComposition21 and i-Score), together with the whole stacking energy ( $\Delta G$ ), in a multi-layer artificial neural network. It was found that this kind of combination increases the correlation coefficient of the prediction accuracy from (0.5 to 0.6) between scales and MysiRNA models (Mysara, M. Elhefnawi, et al. 2011, submitted).

## 6. Experimental section

Here we present an example about working with the previously mentioned protocol for proper siRNA design for targeting human TP53 gene that has been identified as oncogenes. We start with finding P53 mRNA, by searching the NCBI Nucleotide dataset; we will find mRNA refseq id “NM\_000546.4” for Homo sapiens tumor protein p53 (TP53), transcript variant 1. Knowing that we need to target all the gene’s transcripts, we should find all available transcripts. One way to do that is by blasting the mRNA refseq database, searching for mRNA sharing the same name and organism, using NCBI remote Blast. Seven different mRNAs were identified as following: NM\_000546, NM\_001126112, NM\_001126115, NM\_001126117, NM\_001126116, NM\_001126114, NM\_001126113. All of these transcripts were later alignment together, as an approach to identify conserved regions. We used ClustalW to align those 7 transcripts with their different lengths 2586, 2583, 2271, 2331, 2404, 2719 and 2646 respectively. The resulted alignment file, was the treated with “cons” tool to find the consensus between those transcripts, using 100% conservation.

Tools	Model Technique	Training Dataset used	Tool available at	Disadvantages
Biopredsi (Huesken et al. 2006)	Neural network	2,431 records from Huesken dataset.	<a href="http://www.biopredsi.org">http://www.biopredsi.org</a>	Possible over estimation due to over fitting of training set with test (S.A. Shabalina et al. 2006)



CTTTTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAGGTTGGCTCTGACTGTAC  
CACCATCCACTACAACACTACATGTGTAACAGTTCCTGCATGGGCGGCATGAACCGGA  
GGCCCATCCTCACCATCATCACTGGAAGACTCCAGTGGTAATCTACTGGGACGG  
AACAGCTTTGAGGTGCGTGTTTGTGCCTGTCTGGGAGAGACCGGCGCACAGAGGA  
AGAGAATCTCCGCAAGAAAGGGGAGCCTCACCACGAGCTGCCCCCAGGGAGCACT  
AAGCGAGCACTGCCCAACAACACCAGCTCCTCTCCCCAGCCAAAGAAGAAACCAC  
TGGATGGAGAATATTTACCCCTTCAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NN  
NN  
NNNNNNNNNNNCCGTGGGCGTGAGCGCTTCGAGATGTTCCGAGAGCTGAATGAGGC  
CTTGGAACTCAAGGATGCCCAGGCTGGGAAGGAGCCAGGGGGGAGCAGGGGCTCAC  
TCCAGCCACCTGAAGTCCAAAAAGGGTCAGTCTACCTCCCGCCATAAAAACTCAT  
GTTCAAGACAGAAGGGCCTGACTCAGACTGACATTCTCCACTTCTTGTTCCCCACTG  
ACAGCCTCCCACCCCATCTCTCCCTCCCCTGCCATTTTGGGTTTTGGGTCTTTGAAC  
CCTTGCTTGCAATAGGTGTGCGTCAGAAGCACCCAGGACTTCCATTGCTTTGTCCCC  
GGGCTCCACTGAACAAGTTGGCCTGCACCTGGTGTGTTGTTGGGGAGGAGGATGGG  
GAGTAGGACATAACCAGCTTAGATTTTAAGGTTTTTACTGTGAGGGATGTTTGGGAGA  
TGTAAGAAATGTTCTTGCAAGTAAAGGGTTAGTTACAATCAGCCACATTCTAGGTAG  
GGGCCACTTCACCGTACTAACCAGGGAAGCTGTCCCTCACTGTTGAATTTTCTCTA  
ACTTCAAGGCCATATCTGTGAAATGCTGGCATTGTCACCTACCTCACAGAGTGCAT  
TGTGAGGGTTAATGAAATAATGTACATCTGGCCTTGAAACCACCTTTTATTACATGG  
GGTCTAGAACTTGACCCCTTGAGGGTGTGTTCCCTCTCCCTGTTGGTTCGGTGGGT  
TGGTAGTTTCTACAGTTGGGCAGCTGGTTAGGTAGAGGGAGTTGTCAAGTCTCTGCT  
GGCCAGCCAAACCCTGTCTGACAACCTCTTGGTGAACCTTAGTACCTAAAAGGAA  
ATCTCACCCCATCCCACACCCTGGAGGATTTTCATCTCTTGTATATGATGATCTGGATC  
CACCAAGACTTGTTTTATGCTCAGGGTCAATTTCTTTTTTCTTTTTTTTTTTTTTTCT  
TTTTCTTTGAGACTGGGTCTCGCTTTGTTGCCAGGCTGGAGTGGAGTGGCGTGATCT  
TGGCTTACTGCAGCCTTTGCCTCCCGGCTCGAGCAGTCTGCCTCAGCCTCCGGAG  
TAGCTGGGACCACAGGTTTCATGCCACCATGGCCAGCCAACCTTTGCATGTTTGTAG  
AGATGGGGTCTCACAGTGTGCCCAGGCTGGTCTCAAACCTCCTGGGCTCAGGCGATC  
CACCTGTCTCAGCCTCCCAGAGTGCTGGGATTACAATTGTGAGCCACCACGTCCAGC  
TGGAAGGGTCAACATCTTTTACATTCTGCAAGCACATCTGCATTTTACCCCCACCCTT  
CCCCCTTCTCCCTTTTTATATCCCATTTTATATCGATCTCTTATTTTACAATAAAA  
CTTTGCTGCCACCTGTGTGTCTGAGGGGTG

Then, we used this consensus to evaluate its target accessibility using RNAs, finding all possible regions to be targeted by siRNA. 1033 possible siRNA were designed using RNAs. Those siRNAs were evaluated using 10 siRNA efficiency prediction tools as Reynolds (Reynolds et al. 2004), Amarzguoui (Mohammed Amarzguoui & Prydz 2004), Takasaki (Takasaki et al. 2004), Katoh (Katoh & Suzuki 2007), Ui-Tei (Ui-Tei et al. 2004), Hsieh (Hsieh et al. 2004), *Biopredsi* (Huesken et al. 2006), *DSIR* (Vert et al. 2006), *ThermoComposition21* (S.A. Shabalina et al. 2006) and *i-Score* (Ichihara et al. 2007). Selecting siRNA passing 90% or 0.90 predicted score. 111 siRNAs passed these filtration processes, those siRNAs were searched to identify SNPs occurrence residues. All of those 111 siRNAs were found to be targeting SNPs free regions. The last step was to filter those siRNAs against mRNA dataset, to identify those having off-targets. Any siRNA with either complete or partial off-target should be excluded. 85 siRNAs were found to be off-target free candidates. Finally they were filtered and only siRNA with inhibition efficiency above 90%, according to MysRNA model, were accepted.

siRNA position	Sense	Antisense	Predicted efficiency
800	GCGUGUGGAGUAUUUGGAU	AUCCAAAUACUCCACACGCaa	90.6%
822	AGAAACACUUUUCGACAU	UAUGUCGAAAAGUGUUUCUgu	92.6%
883	GUACCACCAUCCACUACAA	UUGUAGUGGAUGGUGGUACag	91.5%
1330	CCCGCCAUAAAAACUCAU	AUGAGUUUUUAUGGCGGGag	91.4%
1842	GAAACCACCUUUUAUUACA	UGUAAUAAAAGGUGGUUUCaa	92.3%
1915	GGUGGGUUGGUAGUUUCUA	UAGAAACUACCAACCCACCga	92.1%
1919	GGUUGGUAGUUUCUACAGU	ACUGUAGAAACUACCAACCca	90%
2016	CCUUAGUACCUAAAAGGAA	UUCUUUUAGGUACUAAGGuu	95.4%
2111	GCUCAGGGUCAAUUUCUUU	AAAGAAUUGACCCUGAGCau	92.9%
2499	CCCUCCUUCUCCCUUUUUA	UAAAAGGGAGAAGGAGGGga	92%
2530	CUCCUUUUUAUAUCCCAU	AUGGGUAUAAAAAGGGAGaa	91.5%
25030	AUAUCGAUCUCUUAUUUUA	UAAAUAAGAGAUCGAUAUaa	93.1%

Table 3. Final siRNA candidates after all stages of design and filtration.

In ElHefnawi et. Al., other examples of optimal siRNA design and selection as silencers for difficult targets such as the Hepatitis C virus (HCV), and the Influenza a virus that have been experimentally tested for verifications of the methodology are under publication (Mahmoud ElHefnawi1 2011) (Mahmoud ElHefnawi 1 2011)).

## 7. Conclusion

In this chapter we provide a comprehensive foundation of the underlying bioinformatics methodology for optimal design and selection of siRNA molecules. We address factors affecting siRNA interference, covering both siRNA and mRNA sides. These factors can be classified into four major classes, **the first class of factors, “targeted region”** or “target sequence space”, addresses how to identify regions in the mRNA that should be targeted by the designed siRNA; and discusses five factor affecting target sequence space: transcript region, transcript size, mRNA multiple splicing, single nucleotide polymorphism and orthologs consensus. **The second class of factors, “siRNA sequence space”**, addresses positional/word preferences in the sense/antisense strand of the siRNA. siRNA sequence space is affected by several factors including nucleotide positional preferences Protocol, GC content, and palindrome. In addition, thermodynamic stability and differential ends instability have been identified to be highly important factors in siRNA functionality. **The third class of factors, is the “target accessibility”**, and how the targeted mRNAs tend to form secondary structure that affect their accessibility hence reduce the capabilities of the designed siRNA to target certain regions of mRNA. Target accessibility is considered as the sum of the energy required to open mRNA and siRNA duplex and the energy required to stabilize siRNA-mRNA duplex. **The fourth class of factors, “off-target matches”**, that influence siRNA specificity via perfect-match, and partial off targets & sequence motifs that invoke immune reaction. Each of these classes can greatly affect siRNA selection and therefore are studied thoroughly in this chapter.

We present a step wise protocol for designing siRNA with the highest specificity and sensitivity in seven different phases, Targeted gene assignment, targeted sequence specification and filtration, designing all possible siRNAs targeting the selected regions, siRNAs scoring and scores filtration, siRNAs target accessibility filtration, siRNAs off-target

filtration, selecting the best designed siRNA. We cover state of the art tools for siRNA efficiency prediction, in two generations: the **first generation tools** select the most efficient siRNAs depending on differential ends thermodynamic stability measures, mRNA secondary structure and base preferences specific position target uniqueness. **The second generation tools** have been developed by applying sophisticated data mining techniques to handle huge annotated records of siRNAs with their experimental inhibition, as in *Biopredsi*, *ThermoComposition21* and *Scales's* artificial neural network model and *DSIR* and *i-Score's* linear regression model. By the end of the chapter, we design siRNA targeting human P53 protein, as a practical example of the proposed protocol. Future directions would be to find additional factors that affect shRNA (siRNAs inserted into expression vectors) that further decrease the efficacy of the expressed siRNAs from them, and extending this methodology latter for miRNA target recognition predictions.

## 8. References

- Amarzguioui, M., 2003. Tolerance for mutations and chemical modifications in a siRNA. *Nucleic Acids Research*, 31(2), pp.589-595.
- Amarzguioui, Mohammed & Prydz, H., 2004. An algorithm for selection of functional siRNA sequences. *Biochemical and biophysical research communications*, 316(4), pp.1050-8.
- Anderson, E.M. et al., 2008. Experimental validation of the importance of seed complement frequency to siRNA specificity. *RNA (New York, N.Y.)*, 14(5), pp.853-61.
- Birmingham, A. et al., 2006. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature methods*, 3(3), pp.199-204.
- Birmingham, A. et al., 2007. A protocol for designing siRNAs with high functionality and specificity. *Nature protocols*, 2(9), pp.2068-78.
- Black, D.L., 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry*, 72, pp.291-336.
- Czauderna, F. et al., 2003. Structural variations and stabilising modifications of synthetic siRNAs in mammalian cells. *Nucleic acids research*, 31(11), pp.2705-16.
- Davis, M.E. et al., 2010. Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. *Nature*, 464(7291), pp.1067-70.
- Dorsett, Y. & Tuschl, Thomas, 2004. siRNAs: applications in functional genomics and potential as therapeutics. *Nature reviews. Drug discovery*, 3(4), pp.318-29.
- Elbashir, S M, Lendeckel, W. & Tuschl, T, 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & development*, 15(2), pp.188-200.
- Elbashir, S M et al., 2001. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *The EMBO journal*, 20(23), pp.6877-88.
- Elbashir, Sayda M et al., 2002. Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods (San Diego, Calif.)*, 26(2), pp.199-213.
- Hofacker, I.L. & Tafer, H., 2010. Designing optimal siRNA based on target site accessibility. *Methods in molecular biology (Clifton, N.J.)*, 623, pp.137-54.
- Hsieh, A.C. et al., 2004. A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic acids research*, 32(3), pp.893-901.
- Huesken, D. et al., 2006. Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 23(8), pp.995-1002.



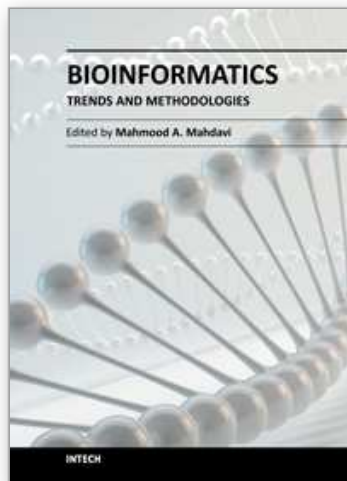
- Hutvagner, G. & Zamore, P.D., 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science (New York, N.Y.)*, 297(5589), pp.2056-60.
- Ichihara, M. et al., 2007. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. *Nucleic Acids Research*, pp.1-10.
- Jackson, A.L. et al., 2003. Expression profiling reveals off-target gene regulation by RNAi. *Nature biotechnology*, 21(6), pp.635-7.
- Jackson, A.L. & Linsley, P.S., 2010. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature reviews. Drug discovery*, 9(1), pp.57-67.
- Katoh, T. & Suzuki, T., 2007. Specific residues at every third position of siRNA shape its efficient RNAi activity. *Nucleic acids research*, 35(4), p.e27.
- Kaufmann, S.H.E. & Patzel, V., 2008. Structures of Active Guide Rna Molecules and Method of Selection.
- Ladunga, I., 2007. More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. *Nucleic acids research*, 35(2), pp.433-40.
- Lorenz, C. et al., 2004. Steroid and lipid conjugates of siRNAs to enhance cellular uptake and gene silencing in liver cells. *Bioorganic & medicinal chemistry letters*, 14(19), pp.4975-7.
- Matveeva, O. et al., 2007. Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Access*, 35(8), pp.1-10.
- Mysara, M., 2010. *MysiRNA: Automation of siRNA Design Considering Multi-score Filtration*.
- Mysara, M. et al., 2011. MysiRNA: Improving siRNA Efficacy Prediction Using a Machine-Learning Model Combining Multi-tools and Whole Stacking Energy ( $\Delta G$ ). *Journal of Biomedical Informatics*, pp.1-23.
- Mysara, M., Garibaldi, J. & Elhefnawi, M., 2011. MysiRNA-Designer : a Workflow for Efficient siRNA Design. *PLoS One*, pp.1-14.
- Mückstein, U. et al., 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics (Oxford, England)*, 22(10), pp.1177-82.
- Park, Y.-kyu et al., 2008. AsiDesigner : exon-based siRNA design server considering alternative splicing. *Knowledge Creation Diffusion Utilization*, 36(May), pp.97-103.
- Patzel, V., 2007. In silico selection of active siRNA. *Drug Discovery Today*, 12(3-4), pp.139-48.
- Patzel, V. et al., 2005. Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nature biotechnology*, 23(11), pp.1440-4.
- Ren, Y. et al., 2006. siRecords : an extensive database of mammalian siRNAs with efficacy ratings. *Access*, pp.1-10.
- Reynolds, A. et al., 2004. Rational siRNA design for RNA interference. *Nature biotechnology*, 22(3), pp.326-30.
- Schwarz, D.S. et al., 2003. Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell*, 115(2), pp.199-208.
- Semizarov, D. et al., 2003. Specificity of short interfering RNA determined through gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 100(11), pp.6347-52.
- Shabalina, S.A., Spiridonov, A.N. & Ogurtsov, A.Y., 2006. Computational models with thermodynamic and composition features improve siRNA design. *BMC bioinformatics*, 7(1), p.65.
- Stark, G.R. et al., 1998. How cells respond to interferons. *Annual review of biochemistry*, 67, pp.227-64.

- Surabhi, R.M. & Gaynor, R.B., 2002. RNA interference directed against viral and cellular targets inhibits human immunodeficiency Virus Type 1 replication. *Journal of virology*, 76(24), pp.12963-73.
- Tafer, H. et al., 2008. The impact of target site accessibility on the design of effective siRNAs. *Nature biotechnology*, 26(5), pp.578-83.
- Takasaki, S., Kotani, S. & Konagaya, A., 2004. An effective method for selecting siRNA target sequences in mammalian cells. *Cell cycle (Georgetown, Tex.)*, 3(6), pp.790-5.
- Ui-Tei, K. et al., 2004. *Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference.*
- Ullu, E. et al., 2002. RNA interference: advances and questions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 357(1417), pp.65-70.
- Vert, J.-P. et al., 2006. An accurate and interpretable model for siRNA efficacy prediction. *BMC bioinformatics*, 7(1), p.520.
- Xia, H. et al., 2004. RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature medicine*, 10(8), pp.816-20.
- ElHefnawi, M., O. Alaidi, et al. (2011). "Identification of novel conserved functional motifs across most Influenza A viral strains." *Virol J* 8: 44.
- BACKGROUND: Influenza A virus poses a continuous threat to global public health. Design of novel universal drugs and vaccine requires a careful analysis of different strains of Influenza A viral genome from diverse hosts and subtypes. We performed a systematic in silico analysis of Influenza A viral segments of all available Influenza A viral strains and subtypes and grouped them based on host, subtype, and years isolated, and through multiple sequence alignments we extrapolated conserved regions, motifs, and accessible regions for functional mapping and annotation. RESULTS: Across all species and strains 87 highly conserved regions (conservation percentage  $\geq 90\%$ ) and 19 functional motifs (conservation percentage = 100%) were found in PB2, PB1, PA, NP, M, and NS segments. The conservation percentage of these segments ranged between 94-98% in human strains (the most conserved), 85-93% in swine strains (the most variable), and 91-94% in avian strains. The most conserved segment was different in each host (PB1 for human strains, NS for avian strains, and M for swine strains). Target accessibility prediction yielded 324 accessible regions, with a single stranded probability  $> 0.5$ , of which 78 coincided with conserved regions. Some of the interesting annotations in these regions included sites for protein-protein interactions, the RNA binding groove, and the proton ion channel. CONCLUSIONS: The influenza virus has evolved to adapt to its host through variations in the GC content and conservation percentage of the conserved regions. Nineteen universal conserved functional motifs were discovered, of which some were accessible regions with interesting biological functions. These regions will serve as a foundation for universal drug targets as well as universal vaccine design.
- Mahmoud ElHefnawi<sup>1</sup>, Rania Siam<sup>3</sup>, Nafisa Hassan<sup>2</sup>, Mona Kamar<sup>2</sup>, Marco Sgarbanti<sup>4</sup>, Annalisa Rimoli<sup>4</sup>, Iman El-Azab<sup>5</sup>, Osama AlAidy<sup>6</sup>, Giulia Marsiliin Marco Sgarbanti<sup>4</sup> (2011). "The design of optimal therapeutic small interfering RNA molecules targeting diverse strains of influenza A virus." *Bioinformatics* OUP under revision.

- Mahmoud ElHefnawi 1, TaeKyu Kim<sup>3</sup>, Mona A. Kamar 2, Nafisa M. Hassan 2, Iman A El-Azab<sup>4</sup>, Suher Zada<sup>5</sup>, Marc P. Windisch<sup>3\*</sup> (2011). "Novel DESIGN AND SELECTION OF EFFICIENT SPECIFIC UNIVERSAL SMALL INTERFERING RNA MOLECULES tested in Hepatitis C Virus replicon cell lines." *PLOS1* submitted.
- Patzel, V., S. Rutz, et al. (2005). "Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency." *Nat Biotechnol* 23(11): 1440-1444.

IntechOpen

IntechOpen



## **Bioinformatics - Trends and Methodologies**

Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1

Hard cover, 722 pages

**Publisher** InTech

**Published online** 02, November, 2011

**Published in print edition** November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

### **How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mahmoud ElHefnawi and Mohamed Mysara (2011). In-silico Approaches for RNAi Post-Transcriptional Gene Regulation: Optimizing siRNA Design and Selection, Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.), ISBN: 978-953-307-282-1, InTech, Available from:

<http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/in-silico-approaches-for-rnai-post-transcriptional-gene-regulation-optimizing-sirna-design-and-selec>

**INTECH**  
open science | open minds

### **InTech Europe**

University Campus STeP Ri  
Slavka Krautzeka 83/A  
51000 Rijeka, Croatia  
Phone: +385 (51) 770 447  
Fax: +385 (51) 686 166  
[www.intechopen.com](http://www.intechopen.com)

### **InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai  
No.65, Yan An Road (West), Shanghai, 200040, China  
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元  
Phone: +86-21-62489820  
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the [Creative Commons Attribution 3.0 License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen