Ravi Sachidanandam works in Cold Spring Harbor Laboratory and is a physicist who is currently interested in computational biology. He is working on several computational aspects in RNAi and its application as a tool for biological research. He is also working on various aspects of splicing, as well as modelling and simulating systems involved in apoptosis, splicing, RNAi and live-cell imaging.

Keywords: RNAi, miRNA, PTGS, TGS, gene silencing, chromatin

RNAi as a bioinformatics consumer

Ravi Sachidanandam

Date received (in revised form): 17th March 2005

Abstract

RNAi has shown great potential for use as a tool for biological discovery, analysis and therapeutics. The involvement of the RNAi pathway in post-transcriptional silencing, transcriptional silencing and epigenetic silencing as well as its use as a tool for forward genetics and therapeutics throws up several bioinformatics challenges. This paper delineates several areas of research and reviews work that has already been done, the tools that are available and the challenges that lie ahead.

INTRODUCTION

RNAi, or RNA interference, is the disruption of a gene’s expression by a double stranded RNA (dsRNA) in which one strand is complementary (either perfectly or imperfectly) to a section of the gene’s mRNA. A dsRNA can enter the cytoplasm, through the expression of a hairpin (or inverted repeats), through viral gene expression or through artificial constructs that enter the cell through the cell membrane.

The disruption can take the form of mRNA degradation, translational repression or transcriptional repression through epigenetic modifications. The pathway can also lead to message modification by cleavage of mRNAs. The RNAi pathway is conserved across eukaryotic species, from yeasts to humans. Interestingly, in yeasts, Schizosaccharomyces pombe has portions of the RNAi pathway while Saccharomyces cerevisiae seems to have lost most of it. Thus, it is a universal mechanism in biological systems and is amenable to use as a tool for forward genetics, where the expression of genes can be regulated by external triggers and phenotypic consequences can be observed.

The RNAi pathway and its various roles, depicted in Figure 1, are post-transcriptional gene silencing (PTGS) and transcriptional gene silencing (TGS). PTGS is the degradation of transcribed and processed mRNA in the cytoplasm. It also includes translational silencing.

- Degradation of mRNA. Short interfering RNAs (siRNAs) are 22 nt double-stranded RNA with 2 nucleotide 3’ overhangs on both strands and a phosphate on the 5’ end. siRNAs in the cytoplasm get unwound and a single strand gets incorporated into the RNA-induced silencing complex (RISC). The RISC along with the siRNA is guided to the complementary mRNA, by an unknown process. The mRNA can get silenced through degradation by Argonaute 2 slicing when the siRNA strand shows perfect complementarity.

- Inhibition of translation. In plants and animals, short (22 nt) non-coding RNAs called microRNAs (miRNAs) regulate developmental pathways by the regulation of protein expression. miRNAs usually block translation. Certain viruses have also been shown to use miRNAs to control gene expression, using the hosts’ machinery.

- Cleavage of RNA. mRNAs get cleaved by miRNAs that have near-perfect matches to them.
TGS is the silencing of gene expression in the nucleus through changes in chromatin structure directed in a sequence-specific manner by dsRNA. This is a heritable effect and probably shares some of the pathway with the PTGS mechanism. Two forms of TGS are heterochromatin formation and methylation of promoters:

- Heterochromatin formation. DNA in the nucleus is wrapped around histone proteins in a periodic arrangement, which is then organised as a condensed structure called chromatin. This helps in packaging the DNA into a small volume and allows for control of expression through the modification of the histones via methylation, acetylation, ubiquitination or phosphorylation of specific residues. Heterochromatin (as opposed to euchromatin) is condensed and transcriptionally silent chromatin. One pathway for its formation involves the methylation of the residue lysine 9 of the histone H3 (H3K9). In *S. pombe* it has been shown that H3K9 methylation is initiated by the association of a strand of siRNA in the nucleus with the RNA-induced transcriptional silencing (TGS) complex, which includes Chp1, Tas3 and Ago. The RITS complex is involved in TGS. The RNAi pathway can also induce promoter silencing through DNA methylation (a process not fully understood). miRNAs are also implicated in programmed DNA elimination in organisms such as *Tetrahymena*.

**Figure 1:** RNAi pathway and its role in post-transcriptional and transcriptional gene silencing. Hairpin RNAs, from either microRNA (miRNA) genes or short hairpin RNAs (shRNAs), constructs, are transcribed and then processed in the nucleus by the Rnase III enzyme Drosha. These stem loop products contain a two-nucleotide 3' overhang and a 5' phosphate at the base of the stem and are exported to the cytoplasm by exportin 5. In the cytoplasm, the Drosha product is further processed by Dicer into a 22-nucleotide RNA duplex with two-nucleotide 3' overhangs and 5' phosphates on both ends. One strand of these 22 nucleotide products (siRNAs or miRNAs) is incorporated into the RISC, composed of proteins including Argonaute (Ago), FXRP, VIG, Tudor SN and Gemin 3 and 4. RISC, with a single 22-nucleotide siRNA strand, can slice/cleave or block translation of an mRNA complementary to the siRNA strand. Chemically synthesised siRNAs and shRNAs can also be introduced through the cell membrane and become incorporated in the appropriate processing step. In transcriptional gene silencing (TGS), double-stranded RNA transcribed from inverted repeats in repetitive DNA induces heterochromatin formation and epigenetic silencing through H3K9 methylation. In *S. pombe*, siRNAs are components of the RNA-induced transcriptional silencing (RITS) complex, which includes Chp1, Tas3 and Ago. The RITS complex is involved in TGS. The RNAi pathway can also induce promoter silencing through DNA methylation (a process not fully understood). miRNAs are also implicated in programmed DNA elimination in organisms such as *Tetrahymena*.
siRNA and shRNA can silence genes

- Methylation of promoters. The cytosine (C) of CpG (C followed by G) motifs on the sequence can get methylated by DNA methyl transferase (DMTase) enzymes. This can silence the underlying region of the genome. RNAi can initiate the methylation of gene promoters, thereby blocking transcription.

The exact mechanism of this process is not understood.

The introduction of large dsRNA into mammalian cells results in a general response (interferon or protein kinase PKR response) that leads to cell death. It was discovered that using shorter dsRNA (<29 nt) can bypass this response. Short double-stranded RNA with 2 nt 3' overhangs and a 5' phosphate group, called siRNAs, that mimic the product of Dicer activity, can get incorporated directly into the RISC and result in silencing activity. This is a popular method of silencing genes in cells.

Another method of inducing RNAi is to insert hairpin constructs into the genome using vectors which can get expressed stably (Figure 2). The expressed hairpins are processed by Drosha and exported to the cytoplasm, where Dicer acts on them to create siRNAs, which then get incorporated into the RISC. These constructs are called short hairpin RNAs or shRNAs. shRNAs can also be chemically synthesised and introduced into the cytoplasm, in which case, it is important to mimic the product of Drosha, which has a 2 nt 3' overhang.

RNAi is also being proposed for therapeutic uses, for example, by targeting viral genes for silencing. Applying this to a disease such as HIV or polio is problematic since the viruses mutate very rapidly and escape message degradation by RNAi. There are trials using siRNAs in the human eye against vascular endothelial growth factor.

Figure 2: A shRNA hairpin construct. Additional elements in the construct are used for engineering purposes. The U6 promoter ensures that the gene is expressed and the terminator ensures that only the hairpin gets expressed, that is, there is no transcriptional run through. The barcode at the end is a random 60 mer that is unique to each hairpin allowing identification of the hairpin, either via microarrays or via the use of polymerase chain reaction (PCR). The oligo is synthesised as a DNA, and then cloned into a plasmid using PCR, grown and then inserted into a retroviral vector. For transient effects, the plasmid can be transfections into the cell with the effects seen about 48 hours after transfection. For permanent insertion, the vector is used to infect the cell. There are two possible orientations of the hairpin sense loop antisense (SAS) or antisense loop sense (ASS). If a 29 mer is used for the hairpin as shown in the figure, then the first 22 mer from the end is believed to be used for creating the siRNA. siRNA design considerations are used to design 19 mers that are then extended nine base pairs at the 5' end and one at the 3' end to obtain the 29 mer. Variations on this construction are possible, depending on the biological needs.
(VEGF) to treat wet age-related macular degeneration (wet AMD). The exact details of the biology delineated here are bound to change as this is a rapidly evolving field, but some of the broad themes and bioinformatics concerns are bound to be relatively longer lasting. Listed below are several areas in RNAi research and the role of bioinformatics in them. Each section will conclude with a short list of bioinformatics challenges that are definitely skewed by the author's preferences.

**RNAi-RELATED PROTEIN CODING GENE/FUNCTION DISCOVERY**

The RNAi pathway has not been fully elucidated. Many more proteins than are currently known may be involved in the RNAi pathway. Experimentally, many of the genes have been discovered using mutational studies in *Caenorhabditis elegans*, as well as screening all genes of the genome using RNAi against them. Comparative genomics has helped with RNAi research in several ways. The discovery that Argonaute exists in *S. pombe* but not in *S. cerevisiae* led to the discovery of RNAi-controlled heterochromatin silencing through histone H3 lysine 9 methylation. The prediction of an Argonaute homologue in the archae *Pyrococcus furiosus* allowed the crystallisation of the protein. This, in turn, led to the determination of its crystal structure, as well as the identification of the function of the PIWI domain.

A computational approach to predicting novel RNAi genes involves looking for domains that are peculiar to proteins involved in RNAi. For example, Argonaute has the PAZ and PIWI domains, and Dicer has a DexH box helicase domain, two RNaseIII domains, a PAZ domain and a dsRBD (double-stranded RNA binding) domain. Thus proteins with PAZ and PIWI domains are worthy of experimental investigation. This kind of analysis usually involves searching for motifs using hidden Markov models (HMMs) and phylogenetic analysis. Phylogenetic analysis has also helped with understanding the functions of various homologues of Argonaute. The study of protein families involved in RNAi can help reveal new members of these families and uncover new mechanisms of the pathway.

**Bioinformatics challenges**

- Identification of all the counterparts to RNAi proteins in each of the model organisms.
- Understanding the evolution of the pathway through phylogenetic analysis, taking into account structural elements of the RNAi proteins (such as the differences in catalytic activity between various forms of Argonaute). The double-stranded RNA binding domain is used in several other pathways: how did it evolve to be used in the RNAi pathway?
- Identifying critical components of the RNAi pathway through the analysis of the topology of the pathway. There are several experiments where RNAi is used to silence parts of the pathways. This process needs to be understood and its temporal behaviour explained (or predicted, since there are no data out there).

**miRNA PATHWAYS**

miRNA pathways are developmentally important in almost all eukaryotic organisms, but may also have a major role to play in the nervous systems of mammals. miRNA genes are small non-coding genes (<150 bp) whose transcripts form hairpins called primary miRNA (pri-miRNA). The pri-miRNAs get processed by an RNaseIII enzyme called Drosha in the nucleus into precursor miRNAs (pre-miRNAs), which have a two-nucleotide 3′ overhang. The pre-miRNAs get exported into the cytoplasm, where they are processed by Dicer and then get associated with the RISC (Figure 1).
miRNA genes

This RISC-associated single-strand miRNA blocks translation of the complementary mRNA. The level of match to target need not be very high, as shown in Figure 3. miRNAs that have near-perfect matches to an mRNA can also direct cleaving (or message destruction) of the target.\textsuperscript{16,17}

MicroRNAs control the timing of gene expression, especially during development.\textsuperscript{44,48}

Occurrence, expression and prediction of miRNA genes

Many miRNA genes have been experimentally identified in a variety of organisms and can be accessed at the miRNA registry website.\textsuperscript{49,50} The registry was constructed by sequencing small RNAs (\(\approx 22\) nt) and mapping them to the genome, and delineating the hairpins that contain them.

miRNAs can occur in intergenic regions, in introns of protein coding regions or in exons and introns of non-coding genes.\textsuperscript{51} The intronic miRNAs probably rely on Pol II mediated pre-miRNA transcription and splicing\textsuperscript{52} for expression, while the intergenic ones are believed to be transcribed independently by Pol II polymerases.\textsuperscript{53}

miRNAs can also occur in clusters

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{miRNA_hairpins}
\caption{The first three hairpin structures are let-7 miRNA hairpins from \textit{C. elegans}, \textit{D. melanogaster} and \textit{H. sapiens}. The shaded portions are the mature sequences that result from processing of the miRNA gene. The last two alignments are for mature let-7 against two targets in the 3’ UTR of lin-41 in \textit{C. elegans}\textsuperscript{44}}
\end{figure}
(separated by a few kilobases) that can be identified by mapping the miRNA genes to the genome. These clusters can be transcribed together and are believed to be miRNAs that control miRNAs from pathways that are involved in related functions. Not much is known about their promoters, though in the case of C. elegans some conserved motifs have been identified. Identification of such promoter sequences can help with the prediction of miRNA genes.

It is almost impossible to identify miRNA genes by scanning the genome of any organism, as the number of putative hairpins far outnumbers the number of known genes. Additional signals are necessary. One method of analysis uses the known biases in targeting strands, but this may not be sufficient to make reasonably useful predictions. Another method that has been used is the conservation of sequences between closely related species such as rice (O. sativa) and Arabidopsis (A. thaliana). Discovery of new miRNA genes can help identify critical features that control the silencing behaviour of dsRNAs. MiRscan and MiRseeker are programs that use the conservation of stem loop structures across species to identify novel miRNA genes.

**Target prediction**

The binding sites of miRNAs are presumed to be in the 3' UTRs and usually occur in multiple copies (Figure 3). Combinatorial control, with multiple miRNAs targeting a single gene, is probably an important feature of miRNA targeting, very similar to the mode of transcription factor control of genes. One of the first miRNAs to be identified is let-7, which targets the gene lin-41 (Figure 3). The alignments of let-7 and its target regions on the 3' UTR of lin-41 show that several bulges as well as G:U wobble base pairing are allowed. Thus, the number of potential targets of any miRNA is quite large, unless some restrictions can be placed on these matches.

Some rules for miRNA target matches have been inferred from a mutational study, which found that the strong binding of the 5' end (the first eight base pairs) of the mature miRNA to the 3' UTR sequence was very important and G:U wobble pairing reduced the silencing efficiency. In addition, it is thought that having multiple binding sites for an miRNA on the 3' UTR increases the efficiency of silencing. This implies multiple binding sites for different miRNAs on the same UTR allows for combinatorial control of the silencing process. The mechanism of targeting has not been settled. There have been several attempts (for example, Rhoades et al., John et al.) at predicting targets, but the agreement between the different attempts at this for the same set of miRNAs is not as good as one would expect, presumably because there could be a large number of targets for each miRNA and each group uses their own criteria to limit results. Models have been developed for miRNA target interactions, based on experiments. miRNAs could also target other miRNAs for silencing. As discussed below, miRNAs show tissue-specific expression patterns; thus correlation of miRNA with miRNA expression can help narrow down targets. It should also be possible to cluster miRNAs and miRNAs based on expression patterns. There are a number of websites and papers that deal with miRNA target predictions.

**miRNA microarrays**

Microarrays that can detect miRNAs have been constructed to study the expression of miRNAs. These experiments have shown that miRNA expression shows tissue and temporal specificity, being
predominantly found in developmental pathways and brain tissue. Some of these studies also show that there is a correlation between mRNA and miRNA expression in various tissues. This is an important piece of evidence for validating miRNA target predictions.

**miRNA libraries**

Precursor miRNAs can be artificially synthesised and either transfected or stably expressed in the genome (as explained in the caption for Figure 2) which can then be expressed using inducible promoters. This can allow study of miRNA function. The design of such libraries is not difficult, but analysis of results can be interesting. The analysis of mRNA networks due to silencing of components can help elucidate the topology of pathways.

**Message modification, DNA suppression, viral control, DNA methylation**

These constitute a class of miRNA-directed activities that have been recently discovered, or seen only in a few cases so far. The developmental gene HOXB8 is cleaved by miRNA mir 196b in mouse. This seems to imply that target mRNAs that have a high degree of complementarity to miRNAs are likely to be cleaved by them, and the location of the target sequence need not be in the UTRs. Viruses can use miRNAs and the host’s RNAi machinery to silence genes. There have also been reports of HIV virus containing miRNAs.

Genomic rearrangements in Tetrahymena are believed to be RNAi-controlled and are probably the result of miRNA action. miRNAs have also been shown to be involved in DNA methylation.

**Bioinformatics challenges**

- Identification and prediction of miRNA promoters.
- Identification of miRNA targets. This is not a solved problem, many underlying assumptions in miRNA target discovery (such as homologies, UTR location of target sites) can help identify good targets, but may not cover all possible sites, some of which may be unique to certain species and may not follow these rules.
- Genomes of viruses can be scanned for possible miRNAs that can target host genes.
- miRNAs must act in concert on members of pathways, so it might be useful to consider function in analysing targets of miRNA. Thus study of the topology of genetic networks and tissue specificity of genes can help with miRNA target studies.
- Identification of miRNA-directed cleavage targets. These might be an important feature of biological systems and can easily be probed by bioinformatics means.

**SHORT INTERFERING RNA (siRNA)**

Short interfering RNAs are short (22 nt) double-stranded RNAs (dsRNA) with two nucleotide 3' overhangs on both strands and a phosphate on each 5' end. siRNAs are created by the action of Dicer on dsRNAs. These can get incorporated into RISC and silence genes that are complementary to the RISC incorporated strand.

**Design principles for functional siRNAs**

It was discovered that the strand whose 5' end is easier to open, or more unstable, is favoured for inclusion in the RISC. RISC has a putative helicase activity of unknown origin which probably unwinds the dsRNA from the unstable end. This probably explains the strand bias of RISC.

Several studies, including studies of miRNA hairpins, have revealed biases of base composition along the active
strand of the siRNA. Using these has allowed for rational design of siRNAs for more efficient silencing. This is done using a weight matrix, very similar to how splice sites are scored.

The partial explanation for the weight matrix is that it agrees with observations on strand usage bias. The compositional bias at different positions on functional siRNAs are derived from studies on large populations of siRNAs. It is not clear how the underlying physical phenomena are related to this. This is reminiscent of the rules for detection of good splice sites in genomic sequences.

There could be other unknown factors that affect siRNA performance that have not been discovered yet. In addition, it has been observed that siRNAs tend to show off target effects, that is targeting of genes for silencing that were not the intended targets. These factors can confound good designs in actual laboratory use.

siRNAs are synthesised with 2 nt overhangs (usually TT) on the 3' ends of each strand and transfected into cells. siRNAs can also be produced by the action of Dicer on dsRNA and by the transfection of siRNA vectors into cells for constant production.

**Construction of an siRNA library**

siRNAs are now routinely used to silence genes *in vitro*. They can be ordered off the shelf for common genes and a lot of effort has gone into understanding their action. Building an siRNA library involves designing silencing constructs for each mRNA using the biases in base composition at each position as discussed above.

The constructs have to be unique, in that they cannot target other genes. To ensure this, a database of non-redundant mRNAs (excluding splice variants) needs to be built, so that the siRNA constructs can be checked for uniqueness. In addition, designs that do not have GC content within 30–60 per cent or contain the sequences AAAA, TTTT, GGGG or CCCC are eliminated. siRNAs can be ordered from companies, which synthesise them with TT overhangs on the 3' ends of each strand. Using several siRNAs against each gene of interest ensures that you have different off-targets for each siRNA and there is a chance that at least one might work. There are several websites that allow design of siRNAs.

Since siRNAs cause silencing by binding to targets that are complementary, specificity is achieved by low concentrations of siRNAs in the cell. At higher concentrations there is a greater likelihood of binding to targets that do not show exact complementarity and, hence, off-target effects can occur more frequently.

Well-characterised siRNAs for several genes can be found in the literature. Several companies sell well-characterised off-the-shelf siRNAs for genes. This takes the guesswork from experiments. However, siRNAs are expensive and the use of high dosage (each cell usually receives several siRNAs) makes induction of other effects, such as off-target silencing, more likely.

**Use of siRNA libraries**

In *C. elegans*, several genome-wide screens have been used to identify pathways important in fat metabolism and in the role of mitochondrion in longevity. In higher organisms, siRNA screens can be done, but they are prohibitively expensive. In addition, this work is most easily done in cell cultures, though some medical applications of siRNAs have been proposed in living organisms.

**Bioinformatics challenges**

- A better algorithm for siRNA design is still an open problem. There is room for new ideas, consideration of factors hitherto unconsidered, such as secondary structure.

- Designing an algorithm to find off-targets would be a very useful tool in designing siRNAs for genome-wide
studies. There have been several attempts, for example Yamada and Morishita, but there is work to be done here.

- Experiments to mimic effects of drugs can be analysed using siRNAs. A perturbative analysis to understand the effects on genetic networks is essential. This field is just starting, as the expense of large-scale screens comes down.

- A resource that would list all the data on siRNAs that are known to function, the conditions under which they function and the design failures, from the literature as well as anecdotal information. It should allow easy access to such data and would definitely help improve design algorithms. This would at least save the labour of wasted effort on siRNAs that do not work under certain conditions.

- siRNA designs can also be improved by using knowledge gained from miRNA studies and analysis of miRNA structures. A lot more work needs to be done in this area, which would allow fine-tuning siRNA designs.

**SHORT HAIRPIN RNA (shRNA)**

Short hairpin RNAs result from the expression of inverted repeats or constructs that can fold into short hairpins. shRNAs or short hairpin RNAs are artificial constructs that can be inserted into a genome and expressed endogenously. The expressed hairpins can then fold to form dsRNA, and Drosa and Dicer can act on these hairpins to create mature sequence, used by the RISC to target the genes. The hairpin constructs can be synthesised and inserted into the genome, or transfected into the cell, as explained in Figure 2.

**Design principles**

The basic idea is to use the lessons learnt from the siRNA designs to ensure that the desirable strand is picked. For a 29 mer based design, the 22 mer from the 3' end of the shRNA hairpin is believed to be picked by the action of Dicer on the shRNA. The siRNA based design is used to design the 21 mer, extended by 1 base in the 3' direction and 7 base pairs in the 5' direction to make a 29 mer, which is then synthesised as a hairpin with a 4 base pair loop.

Another method of constructing these is to use the context of a known miRNA. An miRNA with a target strand of length 22 is picked, and the target sequence is replaced with the antisense strand from the design above. The complementary strand is also replaced, taking care to preserve the bulges, loops and types of mismatches. This will probably ensure that the hairpin gets processed by Drosa, is properly and efficiently processed by the Dicer and gets associated properly with the RISC.

These ideas have been implemented and are available as a web-based resource for public use. shRNAs are cheaper than siRNAs for large-scale studies. shRNAs do not suffer from dosage problems as they are expressed in small numbers in each cell in contrast to most applications of siRNAs where a large number of siRNAs are transfected into any given cell. They can be inserted into germlines, allowing for stable suppression of gene expression in organisms, and are easier to manipulate than knockouts. However, constructing and validating libraries is labour intensive and time consuming.

**Building a library of shRNAs**

The considerations for shRNA libraries are identical to the ones for siRNA libraries. There are several libraries that are now commercially available.

**Use of the shRNA libraries**

The libraries allow large-scale studies that are prohibitively expensive with siRNAs.
We describe below one method of using an shRNA library.

- Design the study using the annotation of the genes. For example, to design a study on the effect of kinases on cancer cells, find all the shRNAs that silence kinases and get them from the library.

- The list of shRNAs will form the pool that can be obtained from the laboratories that have constructed the libraries. The relevant shRNAs are infected into cells at a dilution that ensures that, statistically, there is only one shRNA per cell.

- The cells are grown for a few generations.

- The shRNAs are extracted from the cells using restriction enzymes, and microarrays designed to detect either the barcodes or the shRNAs are used to find shRNAs that have been depleted.

- Find the mRNAs that are targets of the shRNAs that have been depleted and design further studies which may then involve traditional biological work (knockouts etc).

**Bioinformatics challenges**

- Simulate and model behaviours of libraries of shRNAs when used in populations of cells. This can help understand outcomes of library screens.

- A central repository of shRNA constructs. Such a resource can act as a clearing-house that can track results, identify patterns in shRNA performance and allow users to find constructs from a variety of sources. There exists one now called the RNAi Codex, but more work needs to be done.

- Design microarray probes for more efficient study of shRNA expression in cell populations. Since the shRNA constructs can either be probed with the hairpin, or with barcodes that maybe inserted into the shRNAs, there is not much leeway in terms of sequence length and freedom for design, so designing multiple probes per construct that will allow distinguishing different shRNAs is crucial.

- Perturbative analysis of genetic networks. There is a need to develop techniques that can analyse results of silencing of genes either on a one by one basis or in groups. Such techniques have been developed to study the action of drugs to infer genetic networks, but a lot of work remains to be done.

**EPIGENETICS**

Epigenetics is the study of heritable modifications of the genome that leave the DNA sequence unchanged, but can have phenotypic consequences. A consequence of this is the phenomenon of imprinting, which is the dependence of the expression of an allele of a gene on the parent from whom it was inherited.

Epigenetic modifications can silence regions of the genome, and can be of two types:

- **Histone modifications.** Chromatin consists of DNA wrapped around histones, which have long tails. Residues in these tails can be modified by attachment of various groups, such as acetyl groups (acetylation), methyl groups (methylation), ubiquitin groups (ubiquitination) and phosphate groups (phosphorylation). DNA is negatively charged, while histones are positively charged. Addition of positive acetyl groups can prevent heterochromatin formation while adding a methyl group to lysine 9 on the histone H3 tail (H3K9) reduces
the positive charge which can then lead to compaction of DNA and the formation of heterochromatin. Thus a 'histone code', involving different combinations of histone modifications, has been postulated to regulate the state of the chromatin.  

- **DNA methylation.** DNA methyl transferases (DMTases) attach methyl groups to C residues that are followed by Gs (CpG dinucleotides). DNA methylation occurs in plants, mammals and fungi but is absent or nearly absent in *Drosophila, S. cerevisiae, S. pombe* and *C. elegans*. There is a connection between DNA methylation and histone modification as Methyl CpG binding protein (meCP2) binds to methylated regions, which can then recruit histone de-acetylase complexes (HDAC) which can remove acetyl groups from histone tails and hence change the state of the chromatin. In *Arabidopsis*, miRNAs have been shown to be involved in DNA methylation.

The RNAi pathway plays a critical role in the regulation of epigenetic silencing by establishing heritable sequence specific patterns of DNA methylation and histone modifications. We have already discussed possible mechanisms in the introduction. These modifications are replicated on newly synthesised DNA; thereby they become heritable. The epigenome is the combination of the genome along with its epigenetic state. The big challenge is to find the connection, if there is one, between the sequence and the underlying epigenetic state.

Histone modifications can be identified using chromatin immunoprecipitation (ChIP). In the ChIP assay DNA is linked to its associated proteins using formaldehyde and then sheared into small fragments. The DNA protein complex can then be probed for particular protein modifications using antibodies (immunoprecipitation). The DNA associated with these proteins can be delinked and hybridised against microarrays (ChIP on chip), or analysed using the polymerase chain reaction (PCR).

DNA methylation can be identified using a methylation-sensitive restriction enzyme (mcrBC) that can cut methylated DNA but leaves unmethylated DNA untouched, or else by bisulphite treatment that will convert cytosine to uracil and leave methylated Cs untouched. Both techniques in combination with PCR and microarray-based hybridisation can help detect methylation states. Comparison of genomic regions between wild type cells and cells with DMTases knocked out can also be used to identify regions that are methylated.

In mammals, the occurrence of the CpG dinucleotide is lower than expected, probably because of the conversion of methylated Cs into Ts, but there are regions that show enrichment of CpGs called CpG islands. CpG islands are usually not methylated. About 60 per cent of promoters in the mouse and human genome contain CpG islands, signifying their euchromatic state. CpG islands have been cloned and sequenced and these have been used to recognise CpG islands computationally (using HMMs or other pattern recognition techniques).

In maize and other plants, the genome has a high repeat content, which is usually methylated. Thus, using an *Escherichia coli* host strain that harbours a mcrBC system, that cuts methylated DNA, methylated DNA can be eliminated and reads enriched in genes, called methyl-filtered reads, can be obtained. The reads can be computationally analysed to predict genes.

**Bioinformatics challenges**

- Evolution of DNA methylation. Since DNA methylation is not present in yeasts, but is present in other organisms and there seems to be a connection between methylation and histone modification in other organisms, it will be useful to understand the evolution of the proteins and the pathway.
• **Signals in genomic sequence for methylation states.** CpG islands are an example of such signals, but there might be other signals. It is known that there are correlations between the nature of the sequence and the functional state of chromatin. For example, tandem repeats and transposons are silenced.

• **Insulators.** These are sequence elements that isolate or insulate a region of the chromatin from a neighbouring region and perhaps could block the spread of methylation. For example, analysis of CpG islands may yield binding motifs for factors that block DNA methylation. The zinc finger-binding protein CTCF (CCCTC binding factor) regulates the genomic imprinting of the H19 and Igf2 genes. CTCF binds the differentially methylated domain (DMD) on the maternally inherited chromosome preventing its methylation and subsequently the expression of Igf2.

• **Better identification of CpG islands.** This problem was suggested by Prof. Denise Barlow, who has observed that certain experimentally verified CpG islands in mouse are not annotated in the mouse genome viewer on Ensembl. There have been some efforts directed towards genome-wide CpG island annotations, such as on the human chr 21 and chr22. Current computational techniques based on HMMs or average sequence statistics cannot identify short CpG islands and there might be a need for species-specific CpG island predictors, which in turn might point to differences in the mechanisms by which such islands are generated.

• **Dynamics of epigenetic processes and modelling.** Heterochromatic DNA is compact and silent but not static. Transcription from heterochromatic DNA does occur and this transcription is important for maintaining the silenced state. The dynamics of this process is not well known and resolving this is one of the big challenges. There are experimental challenges, but bioinformatics challenges involve modelling the dynamics and making predictions that can direct biology. There is also histone exchange going on, and it is important to observe these processes in vivo and measure rate constants in a quantitative manner. There are systems available to study these processes and these promise to allow real-time visualisation, which will then allow modelling. Modelling these systems would also allow ranking these processes in terms of relative importance in any pathway.

• **Data collection and curation.** Several epigenomic projects are underway; in addition an exploration of the role of epigenetic states, such as DNA methylation, in cancer is being carried out. Collecting these data in a central repository and allowing intuitive browsing of the data is an interesting problem to solve. Having multiple centres with diverse methods of analysis and display would be the key to solving this problem. Data on genes and regions that are maternally or paternally imprinted would also need to be collected. This will allow discerning patterns in a location-specific, or gene/tissue-specific, manner. Patterns of this nature will not be obvious at first glance, but will become clear when looked at in a genome-wide manner. It is quite possible that there are no universal rules when the epigenome is concerned, but different rules in different regions of the genome.

Acknowledgments

Susan Janicki discussed the manuscript critically, edited it and pointed out a number of references.
Leemor Joshua Tor and Jidong Liu kindly shared figures from their papers and talks. Mike Ronemus and Matt Vaughan gave critical comments and freely shared their knowledge of the field. Dhruv Pant helped with Figure 3. The anonymous reviewers helped with several comments, corrections and suggestions.

References

RNAi as a bioinformatics consumer


35. Acuity Pharmaceuticals, 'siRNA use for AMD treatment' (URL: http://www.acuitypharma.com).


coding RNAs, have lower folding free energies than random sequences', Bioinformatics, Vol. 20(17), pp. 2911–2917.


57. Marsow, M., Lim, L. P. and Lau, N. (2003), MiRscan website (URL: http://genesmitedu/mirscan/).

58. Tomancak, P. (2003), miRseeker website (URL: http://toyiblogov:9050/cgi bin/miRseekerpl/).


68. MSKC Center, miRanda website (URL: http://www.microrna.org).


83. Ambion Company website (URL: http://


84. Dharmacon Company website (URL: http://www.dharmacon.com).

83.﹗(URL: http://codex.cshl.org).


81.﹗(URL: http://www.dharmacon.com/tds_design_center/html).


64.﹗(URL: http://www.dharmacon.com/catalog/supp/shri_summary.html).


43.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).

42.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).

41.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).

40.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).


38.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).


34.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).

33.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).


17.﹗(URL: http://www.dharmacon.com/catalog/shri_summary.html).


0.ivantion, Vol. 37(6), pp. 961–969.


