EMERGING TOOLS FOR RNA STRUCTURE ANALYSIS IN POLYMORPHIC DATA

RADHAKRISHNAN SABARINATHAN[1,2], STEFAN E SEEMANN[2] AND JAN GORODKIN[2]

[1]RESEARCH PROGRAM ON BIOMEDICAL INFORMATICS, IMIM HOSPITAL DEL MAR MEDICAL RESEARCH INSTITUTE AND UNIVERSITAT POMPEU FABRA, BARCELONA, CATALONIA, SPAIN. [2]CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH, DEPARTMENT OF VETERINARY CLINICAL AND ANIMAL SCIENCES, UNIVERSITY OF COPENHAGEN, DENMARK. GORODKIN@RTH.DK

The RNA structure of protein coding as well as non-(protein) coding genes plays essential roles for their functions. Recent research suggests that mutations responsible for disease and phenotypic traits exist in far greater numbers outside the protein coding part of the genome than in the coding part. Simultaneously, the transcriptome holds a large potential for folding in RNA structures that are relevant for the transcript's stability and functionality. To analyze this part of the non-coding mutations, new classes of bioinformatics tools that can predict the impact on the structure as a result of the mutations are emerging.

POLYMORPHIC SEQUENCE

Mutation is a process in which a building block of the DNA (nucleotides A,C,G,T) is replaced by another, hence causing a small change in the DNA sequence. In evolution, mutations that are favorable will remain and stay fixed in the entire species or in sub population, which then might show a distinct phenotype relative to the remaining population. Some mutations might even be carried from parent to progeny and only in some cases be "activated" as disease causing mutations. An example of a disease causing mutation is shown in Fig 1. The DNA of the LMNA gene is changed by a *single nucleotide polymorphism* (SNP) causing a change in the protein building blocks from the amino acid argenine to leucine. This substitution might alter the stability of the mutated protein lamin A, and it has been linked to the rare autosomal recessive syndrome mandibuloacral dysplasia (Al-Haggar et al., 2012). Mutations that change the protein product are called *nonsynonymous* and their effect is often destabilization of the protein folding or interference with binding to other molecules. Whereas such SNPs are genetically inherited, other types of mutations can spontaneously emerge in the DNA, but only in some cells. Many types of cancer are governed by such driver mutations.

Besides the mutations resulting in SNPs, a variety of other events might cause diseases such as chromosome breaking or duplications of DNA regions. Here, we focus on the SNPs. Besides resulting in changes in proteins, the SNPs can induce changes in a variety of other aspects. These include the so-called regulatory elements, which recruit molecular complexes involved in turning genes on and off. If these regions at the genomic level hold SNPs, the regulated gene products, such as proteins, might appear in the wrong number, which can lead to a disease.

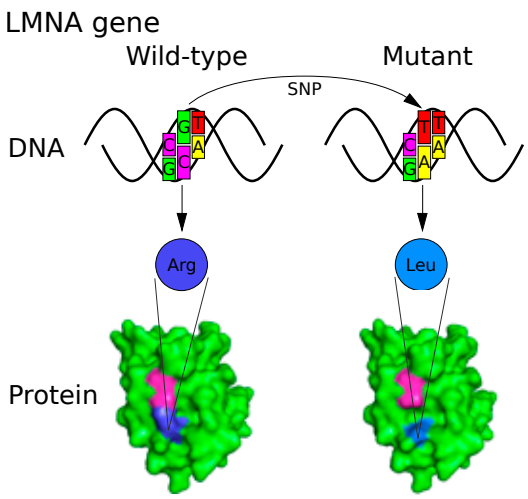

Figure 1: *DNA sequence of LMNA and its protein product lamin A IG-like domain (pdb id: 1IFR). The wild-type DNA on the left side has the building blocks (and codon) CGT, which encodes for the amino acid arginine (Arg), whereas in the mutant DNA on the right side a SNP occurs changing the building blocks to CTT, which encodes for leucine (Leu). The surface plots of the lamin A protein in the bottom highlight the interaction between arginine (blue) and glutamate (magenta) in the wild-type, whereas leucine in the mutant does not interact with glutamate and, thus, cause a less compact protein fold. The lamin A mutant is linked to the rare disease mandibuloacral dysplasia (Al-Haggar et al., 2012), causing a variety of abnormalities including underdevelopment of the lower jaw. In the protein surface plot of the mutant arginine 527 has been replaced by leucine in PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.) for illustration. However, the correct crystal structure of the mutant is not known.*

POLYMORPHIC SEQUENCE IN NON PROTEIN CODING REGIONS

SNPs can, not surprisingly, be located outside protein coding regions while showing a phenotype by a disease or a trait. Other regulators such as the so-called microRNAs, which are small non-coding RNA genes, are involved in down regulation of proteins by a mechanism where they interact with the gene transcript. However, SNPs in these binding sites can destroy or alter them (to bind another microRNA) and entire novel binding sites can be induced by SNPs elsewhere. An example is the Texel sheep, whose trait is the extended muscle growth on the hind legs as a result of a SNP that induces a new microRNA binding site. Thereby the microRNA is repressing the muscle growth regulator myostatin that is involved in repressing muscle growth (Clop et al. 2006). When this repressor itself is repressed the muscle growth is increased.

Interestingly there have also been reports on disease or trait associated SNPs causing disruption of the RNA structure. For example, a mutation in human mitochondrial $tRNA^{Met}$ (involved in protein synthesis) has been identified to cause drastic structural change that affects the tRNA function and is associated to mitochondrial myopathy such as muscle weakness and exercise intolerance (Jones et al., 2008). .

THE POTENTIAL FOR SNPS IN THE NON-CODING GENOME

Today probably most SNPs are still searched for in protein coding regions with the aim of finding those that can give rise to a change of protein sequence or structure and thus potentially affect protein function. However, not only does the protein coding sequence make up only about 1.2% of the human genome, several studies have concordantly shown that most disease and trait associated SNPs fall outside the protein coding regions (Welter et al., 2014). A large fraction (40%) of this, however, is located within introns, which are large stretches of DNA sequence inside genes, but intervening the coding sequence itself.

Furthermore, recent studies have shown that most (approximately 75%) of the human genome is transcribed. Thus besides the protein coding genes (usually transcribed in far higher numbers) many other transcripts might function as non-coding RNA (ncRNA) genes, which often function through the (RNA) structure they fold into. Interestingly, another recent study has shown about 15% of transcribed SNPs altering the RNA structure (Wan et al., 2014). Hence, as more and more mutations become associated with diseases and traits, the need for efficient methods rapidly increases.

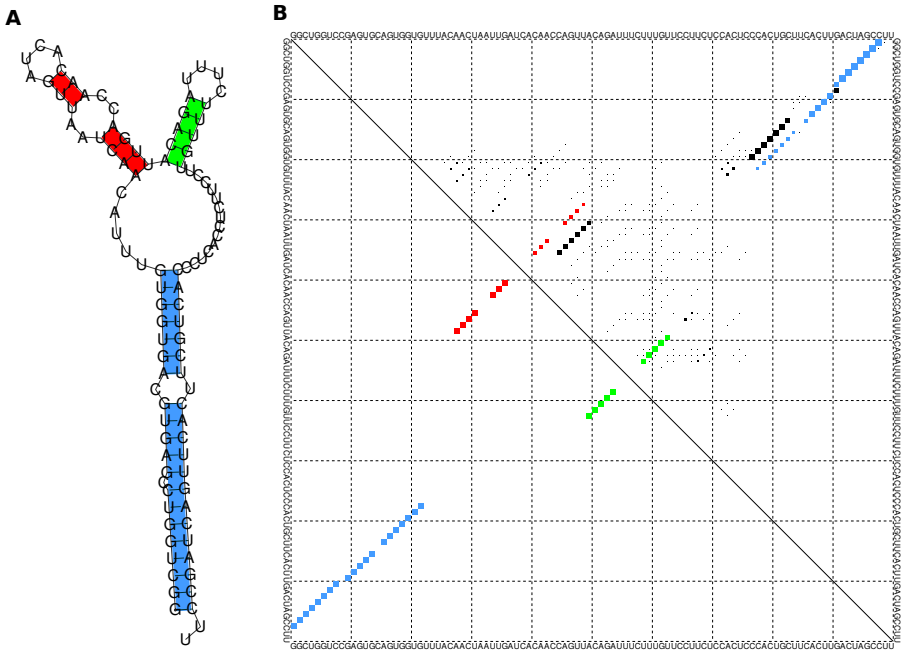

Figure 2: *The ncRNA gene Y RNA is involved in chromosomal DNA replication for which the structure is essential. (A) The MFE structure is shown with its base pairs and unpaired nucleotides. (B) The dot-plot on the upper triangle compares the base pairing probabilities between any two positions over all possible structures, that is how often two positions are observed to pair. The dot-plot on the lower triangle shows the base pairs corresponding to the MFE structure. The figures were made by the Vienna RNA fold server (Lorenz et al., 2011; http://rna.tbi.univie.ac.at/).*

FOLDING RNA SEQUENCES

To analyze the structural impact of mutations on RNA sequences, we first need to explore how the sequences are folded into a structure. Even though rapid progress is made for full three-dimensional structure predictions, much folding is carried out at the level of secondary structure. The basic concept of secondary structure is an energy model based on thermodynamic stability of the RNA molecules. Each structural component holds a precomputed minimum free energy and can be used in an optimized fashion to piece together the optimal structure. Such structure (as depicted in Fig 2A) is in particular based on the so-called canonical base pairs in RNA, A-U, U-A, C-G, G-C, G-U and U-G, where U in RNA sequence correspond to "T" in DNA sequence. The folding of RNA sequence predicts the structure with the lowest or minimum free energy (MFE). An example of the MFE structure for the RNA sequence of the ncRNA gene Y RNA is shown in Fig 2A.

However, more interestingly the folding algorithms allow for computation of the entire *ensemble* of possible structures. One can consider each possible structure to appear with some probability. Whereas the MFE structure is the structure with the highest probability, many other structures might have almost equally high probabilities. This information can be converted into *dot-plots*, such as for Y RNA in Fig 2B, in which the probability of any two positions in an RNA sequence to base pair are displayed (by increasing dot size with probability). As noticed, a series of dots going anti-diagonal corresponds to a stretch of base pairs in an RNA structure. In the upper triangle we can see that there are a number of possible alternative conformations, although they come with different probabilities.

It should be noted that searching for RNA structure is not only done by using energy folding, but indeed comparative information, in particular by considering

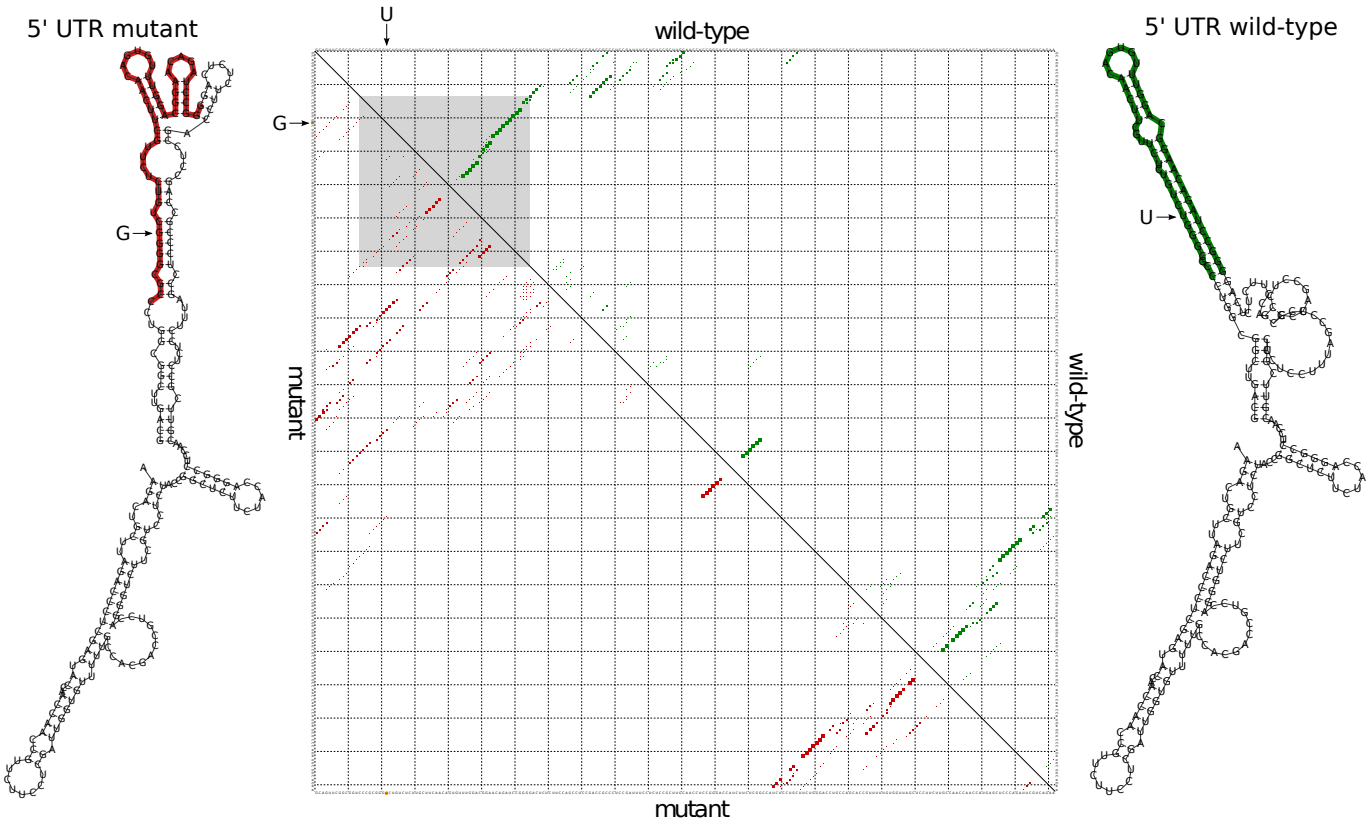

Figure 3: *The mutation T to G at the DNA level in the so-called 5' UTR region of FTL gene causes local structural change that affects structure of a functional element, called iron-responsive element (IRE) (Martin et al., 2012). Note that T is shown as "U" in plot since the RNA version of the gene UTR region is depicted here. RNAsnp detects this local region (highlighted in gray background of the dotplot) containing the functional element with significant structural change (P-value = 0.0518) due to the mutation. The upper and lower triangle of the dot plot contains the base pair probabilities for the wild-type and mutant structure, respectively. The MFE structure of the wild-type and mutant are shown on the right and left side of the dot plot. The region highlighted in the MFE structures corresponds to the detected local region with significant structural change.*

compensating base pairs where for example a structure preserving A-U base pair in human can correspond to a structure preserving C-G base pair in mouse. This strategy combined with elements of MFE folding is typically the preferred way, and in particular folding single sequences is not a reliable approach for full structure determination. For an in depth introduction to RNA folding and RNA bioinformatics methods, we refer to (Gorodkin & Ruzzo, 2014).

DETECTING STRUCTURAL CHANGES IN RNA SEQUENCES

Detecting the impact of mutations on RNA structure is a young scientific discipline and the number of methodologies span approximately a dozen. The approaches span from direct comparison of changes in the MFE structures by different techniques (tree-edit distance and hamming distances on base pairs) to more refined versions taking the whole dot-plot into account. One such approach is the SNPfold program (Halvorsen et al., 2010; http://ribosnitch.bio.unc.edu/snpfold), which makes use of all the information in the dot-plot. For a given position in the sequence the probabilities for base pairing to all other positions are summed together to obtain a score for the position's base pair preference. This is done for wild-type and mutant sequences and the position profiles are subsequently compared. Large differences will then indicate if there is a difference in structure.

One challenge with SNPfold is that it only works globally and hence if the sequence becomes large, then the overall changes between wild-type and mutant might not be significant on global scale, but still there can well be local changes that are missed. Such local elements could be essential for the function of the RNA. Therefore other methods were introduced to cope with this. Our method, RNAsnp (Sabarinathan et al., 2013; http://rth.dk/resources/rnasnp) can detect such local changes regardless of the size of the sequence. The basic concept is to search for the local regions between wild-type and mutant showing the largest differences in structure. This difference can be computed both as a local version of the scoring in SNPfold, but also through a direct comparison (Euclidean distance) between the dot-plots of the wild-type and mutant. In Figure 3, we show an example of mutation in the 5' untranslated region (UTR) of Ferritin light chain (FTL) mRNA that is associated to the hereditary hyperferritinemia cataract syndrome.

PERSPECTIVES

Clearly, with the increasing wealth of data (SNP, genomes and transcriptomes) as well as the increasing awareness of the importance of RNA structure, the need for efficient methods which can screen the RNA sequences

for altering structure as a result of SNPs is rapidly increasing. Besides the ability to conduct local structure comparison, RNAsnp further meets the demands as it can also be employed for genome and transcriptome-wide screens for RNA structure changes by testing for all possible SNPs on the sequence. RNAsnp comes as a stand-alone software and as a webserver available at http://rth.dk/resources/rnasnp. A range of other RNA bioinformatics tools are also available at Center for non-coding RNA in Technology and Health, http://rth.dk/resources.

ACKNOWLEDGEMENTS

The work presented here was made possible due to funding from the Danish Innovation Foundation (Programme Commission on Strategic Growth Technologies), The Danish Council for Independent

Research (Technology and Production Sciences), Danish Center for Scientific Computing (DCSC, DeIC), The Lundbeck Foundation, and The Danish Cancer Society.

REFERENCES

Al-Haggar M, Madej-Pilarczyk A, Kozlowski L, Bujnicki JM, Yahia S, Abdel-Hadi D, Shams A, Ahmad N, Hamed S, Puzianowska-Kuznicka M. A novel homozygous p.Arg527Leu LMNA mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome. Eur J Hum Genet. 20(11):1134-40, 2012.

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C, Georges M. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet. 38:813-8, 2006.

Gorodkin J, Ruzzo WL. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Methods in Molecular Biology, Springer Protocols, 2014.

Halvorsen, M., Martin, J. S., Broadaway, S., & Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble. *PLoS genetics*,*6*(8), e1001074.

Jones, C. N., Jones, C. I., Graham, W. D., Agris, P. F., & Spremulli, L. L. (2008). A disease-causing point mutation in human mitochondrial tRNAMet results in tRNA misfolding leading to defects in translational initiation and elongation. *Journal of Biological Chemistry*, *283*(49), 34445-34456.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 6:26, 2011.

Martin, J. S., Halvorsen, M., Davis-Neulander, L., Ritz, J., Gopinath, C., Beauregard, A., & Laederach, A. (2012). Structural effects of linkage disequilibrium on the transcriptome. *RNA*, *18*(1), 77-87.

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. Hum Mutat. 34:546-5, 2013.

Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, Chang HY. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 505:706-9, 2014.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42(Database issue), D1001-6, 2014.