Advances and challenges in Protein-RNA:

recognition, regulation and prediction

Yael Mandel-Gutfreund (Technion-Israel Institute of Technology, Haifa, Israel)

Gabriele Varani (University of Washington, Seattle, WA , USA)

Jun 07 - Jun 12, 2015

1. Overview of the field and open questions

RNA-protein recognition is central to all biological events in both normal and disease cellular states. The meeting focused on a common discussion of recent development and challenges in the computational, structural and experimental investigation of this important class of biological recognition events.

Immediately following transcription, in fact while a gene is still transcribed, RNA-binding proteins (RBPs) associated with primary transcripts and regulate the fate, cellular localization, stability and translational efficiency for that gene product. The efficiency and location of RNA processing events (splicing, 3'-end processing, editing) is dictated by RBPs alone and in competition with RNA structure. Not surprisingly, many genetic diseases map to the disruption of RNA-protein recognition events. Both coding RNAs (e.g. gene products that contain an open reading sequence that us translated on the ribosome) and non-coding RNAs (sequences that function as primary transcripts and are not translated) associate with a myriad of cellular proteins, many of which are known and more that are still being identified. Understanding the molecular mechanism of protein-RNA recognition and regulation to the point of predicting and rationally altering protein-RNA interactions is a major challenge of experimental and computational biology. In fact, these major questions are only beginning to be addressed quantitatively through recent progress in structure determination and through the development of experimental and computational methods to study RNA-protein complexes in a high throughput manner.

A significant fraction of the genome, at least 5% and perhaps more than 10%, of all eukaryotes code for RBPs and many more genes code for RNAs which are not translated (non coding RNAs) but function by acting on other RNAs or chromatin and perform functions analogous to those of RNA-binding proteins in regulating RNA metabolism. As organisms and cell types become more complex and evolved (e.g. neurons; higher eukaryotes), regulation of gene expression at the post-transcriptional stage becomes comparatively more prominent. In these tissues and organisms, complex gene regulatory networks depend on the RNA/protein interplay, where an RNA-binding proteins act upon an RNA and/or vice versa. Modeling, predicting and altering these cellular networks require a quantitative understanding of individual protein-RNA interaction events, but also a broader genome-wide catalogue of the activity and concentration of RBPs. A significant activity within the meeting was dedicated to new approaches to quantitative mathematical modeling of the activity of these proteins.

Many RBPs can be identified directly from their sequence and are well-annotated. Typically they belong to distinct structural classes that are reasonably well known (e.g. the RRM, the dsRND, Puf proteins etc), but there are potentially many more proteins which could bind to

RNA (e.g. many metabolic enzymes). Identifying the complete complement of RNA-binding proteins in any eukaryotic organism remains a significant challenge of computational and experimental methods, yet the identification of a protein as belonging to the RBP class is not sufficient to address its function, because it does not provide any information on which RNAs it acts upon. Furthermore, the activity of RBPs can be highly specific (e.g. one protein acting on a small number of RNA targets) or diffuse, when a protein can act on a large variety of cellular RNAs. An understanding of the specificity of RBPs is central to biology and requires the joint application of both computational and experimental techniques. Determining RNA-binding specificity was a major theme of the meeting

The first structural and biochemical information on RNA-binding proteins, at atomic detail, dates from the late 1980s. The last 20 years have seen very rapid progress in the structural (NMR and x-ray primarily) and biochemical characterization of RNA-protein complexes. These studies have investigated in considerable detail how many classes of RNA-binding proteins recognize RNA and have dissected the contribution of different molecular forces (e.g. electrostatics) to binding and specificity. However, as highlighted at the meeting, significant gaps in knowledge remain. Most significantly, even when the specificity of a protein is established structurally and/or biochemically, the effect of sequence variation cannot generally be predicted from this knowledge, yet are critical to understand the full complement of RNAs targeted by a protein in the cell and to understand how sequence variation might cause disease or drive species evolution. Despite this limitation, the traditional structural/biochemical approach to understanding RNA-protein interactions remain a central tool to interpret genome wide studies, but structure determination of RNPs still lags the comparable problem of protein-DNA or protein-protein specificity considerably.

More recently, since approximately 2000 but especially since 2005, genome-wide methods have been introduced to interrogate the specificity of RBPs not one sequence at a time, as was traditionally done, but by sampling the complete sequence landscape recognized by an RNA-binding protein in vitro and in vivo. A major component of the meeting was dedicated to the discussion, presentation and critical examination of high throughput experimental methods to address this important problem. The application of these methods, and their interpretation through structural principles on the one hand, and the subsequent generation of mathematical models of these interfaces promise to generate much better understanding of the causes and consequences of variation in RNA-based gene expression on organisms and disease.

The concept behind the workshop arose from the realization that the integration of different experimental and computational approaches is needed to understand RNA-protein recognition with the level of sophistication and depth required to answer fundamental biological processes. It is not sufficient for structural biologist or biochemists to understand in depth what other structural biologists are doing: the greatest opportunity for progress lies in the merging of different experimental and computational approaches to tackle this problem. Thus, the workshop was designed to bring together structural biologists/biochemists that focus on individual RNA-protein complexes, with genome biologists who have developed powerful experimental methods to investigate RNA-protein interaction across genomes, as well as computational biologists who seek to model and develop predictive tools based on the confluence of these experimental advances. The workshop was designed to foster the exchange of ideas between experimental and computational biologists and catalyze the development of new and improved technologies that merge experimental analysis with novel mathematical and computational techniques to better understand the rules of protein-RNA recognition with the ultimate goal of generating a better quantitative understanding of RNA-based biological regulation.

2. Presentation highlights

2.1 High throughput approaches to study protein-RNA interactions and the impact on downstream genes

The most significant advance in the field of RNA-protein recognition in the last 10-15 years has been the development of genome-wide approaches to investigate the RNA population targeted by an RBP and to establish its specificity. The objective of all of these methods is to derive an RBP code, i.e. the system of RNA determinants and protein partners that instruct gene expression at the RNA level. Exhaustive array-based methods and related approaches interrogate the specificity of RBPs in vitro in purified form, while methods such as CLIP (cross-linking and immunoprecipitation) isolate the RNA-binding sites of the RBP in a specific cellular context. The two classes of methods are of course related but investigate RBPs in a different context. While the latter method would naturally seem to be more valuable and closer to provide physiologically relevant results, it also suffers from limitations of the experiment (RNA structural context; over- or under-representation of expressed targets in the cell; false positives due to non-specific interactions or over-expression; false negatives from under-expression; etc). An important and very fruitful theme of the meeting was the open and frank exchange of information and debate about the different limitations and strengths of these methods and how to best interpret and analyze the results and compare the outcomes of different experimental approaches.

Cell-based methods to address the question of which RNA a given protein associates with were introduced about 10 years ago by Ule, building upon earlier methods developed by Keene and Steitz. In short, these approaches use cross-linking of a specific protein to all the possible RNA it is associated with under particular cellular conditions, followed by deep sequencing to identify the target RNA. Keene provided a very interesting historical survey, starting with older biochemical methods that first identified the RRM as an RBP over 25 years ago, following with description of update high-throughput methods to address the same questions in a cellular context. Landthaler described in considerable detail an example of one particular regulatory system and highlighted the remarkable number of sites targeted by any specific RBP (>3,500 in that example) and the structural complexity of binding sites that can be identified by a systematic investigation of these motives.

Keene and Friedersdorf presented their more recent approach (RIP-Seq) to identity the RNA binding sites for RBPs and measure quantitative binding strength. An emphasis of their presentation was the presence of overlapping binding site for different RBPs, which raised the issue of cooperativity and anti-cooperativity in RBP function. These are important questions that remain to be addressed satisfactorily in the current paradigm of one protein/one binding site used by essentially all biochemical and genome wide approaches.

Ule presented recent advances in the CLIP technology that provides nucleotide resolution, as opposed to broader mapping of targeted sequences that was possible in the past, and that addresses the issue of repetitive sequences (e.g. poly-pyrimidine tract binding protein binding to pyridimidine rich splicing signals) as well as proteins that bind to highly structured RNAs (e.g. Staufen). This last problem is particularly important because addressing structural context (i.e. which RNA secondary structure provides a binding framework for sequence specific and even more for structure-specific RBPs) remains a challenging and highly pursued problem for array-based methods that investigate protein specificity in vitro. Yeo presented an update on efforts connected to the ENCODE project to generate a genome-wide analysis of RNA-binding protein networks. He descried highly standardized approaches to map the targets of >300 RBPs using a combination of CLIP-related approaches (CLIP-SEQ; ChiP-Seq; Bind-N-Seq) in cells. He openly described efforts to remove artifacts from the data, improved positive sensitivity (e.g. normalization for RNA copy number, validation of antibodies) and obtain maps at nucleotide resolution. These data will be provided world-wide

through a widely accessible server. In a departure from the focus on eukaryotic proteins of most of the meeting, Margalit focused instead on the equally complex problem of mapping at the transcriptome-wide level the universe of protein-RNA interactions involving small non coding RNAs in bacterial organisms.

2.2 In vitro and computational approaches to assigning RNA binding motifs

The different high throughput approaches to study protein-RNA interactions can generally be broken down into in-vitro and in-vivo methods. While the in-vivo approaches, which were extensively discussed at the meeting (described above in section 2.1) can give a snapshot of the binding sites of a given RBP at a given cell type in a given condition, in-vitro approaches are equally important for determining the specificity of RBPs while controlling for cellular parameters which could influence the binding such as interacting proteins and other cellular factors. Given the well-established knowledge that the RNA structure plays an important role in determining the binding specificity of RBPs there has been a great effort in the field to detect the combined sequence and structural binding preferences of RBPs. At the meeting, Hughes presented a recent collaborative effort between his group and the group of Morris to determine the binding specificities of large cohort of RBPs. In a published study in Nature (2013), the two groups reported systematic analysis of the RNA motifs of 205 RNA-binding proteins that were extracted from a high throughput in-vitro selection experiment. At the meeting Hughes presented the main computational challenges in extracting binding motifs from the large scale experiments, specifically referring to long motifs which may represent co-binding of several proteins or different binding preferences of the same protein. In addition Hughes discussed the challenges and approaches that should be employed for detecting the combined sequence and structural preferences of RBPs. Morris then presented an overview of computational approaches developed over the years by his group and others for extracting combined sequence and structural RBP motif preferences. Morris concentrated on the RNAcontext algorithm, which is a probabilistic model that uses both sequence and structure parameters inferred from the data to extract the most probable motifs which reflect both the structural and sequence preferences of the RBP. Backofen presented a graph-kernel based algorithm named GraphProt which uses an advanced machine learning approach to predict the combined sequence and structural binding preferences of RBPs extracted from in-vivo data and employs it to predict missing binding sites. He presented an example from collaborative work with the Landthaler group where they employed the GraphProt algorithm to identify the composite structure-sequence motif recognized of a zinc finger RBP, which could not be detected by standard computational methods for motif finding.

Extracting the binding preferences from in-vivo experiments adds many different challenges, such as predicting true binding sites which have been miss identified by the experimental tests. Eyras presented an original computational approach for predicting binding motifs of novel RBPs by correlating the differential gene expression of RBPs in cancer vs normal tissues to alternative splicing events altered in the same tissues. By employing this approach on data from the TCGA they were able to recapitulate the binding motif of the well characterized RBP QKI. Dror addressed a major computational challenge: how to distinguish true binding sites from all sites that contain the binding motif of a nucleic acid binding, that yet are not bound by the protein. She presented her studies of DNA binding motifs, emphasizing that very similar protocols can be employed for identifying cognate DNA or RNA binding sites, and showed that the main features contributing to predicting true binding sites are the sequence content around the motif and the similarity of the motif to its neighborhood.

2.3 Approaches for detecting novel RBPs and defining their function

Great advances have recently been made in the development of high-throughput screens to identify RBPs in cultured cell lines. Such methods, known broadly as the “interactome capture”, take advantage of the poly-A tails of primary transcripts RNAs as a bait to be

captured by magnetic poly-T beads. The interactome capture technology has contributed dramatically to the field of RNA-protein interactions, increasing the number of experimentally identified RBPs as well as suggesting novel RBPs and specifically new, yet un explored, cellular mechanisms for these proteins. Milek from the Landthaler group, which were among the first groups to develop the interactome capture methodology, presented an advance study where they employed the methodology to specifically identify RBPs that bind RNA transcript upon ionizing irradiation in MCF-7 cells. This study demonstrated the great advantage of the in-vivo interactome capture approach over the standard approaches for predicting RBPs in vivo and in-silico, enabling the presenter to show the dynamics of RBPs in the cell and specifically to quantify the differential binding of RBPs to mRNAs in the cells under different conditions. Among the great advantages of the new technology to capture RBPs in cells is the ability to conduct comparative experiments to reveal the evolution of the RNA-binding proteome across species to better understand the origin of RNA regulation. Beckmann presented exciting results from his postdoctoral work in the Hentze group, together with computational approaches employed for discovering the commonalities and differences in the RNA binding proteome of human and yeast. Among the unexpected findings presented were inherited differences in the sequences and predicted structures of the RBPs. Gerber presented a comparative interactome study of the RBPs in yeast *S. cerevisiae* and in the nematode *C. elegans*. RBPs were detected again with similar techniques but, in both cases, RBPs were identified from living organisms and not from cultured cell lines. One of the most exciting and unexpected findings presented by both speakers was that among the novel RBPs found in both human and yeast were several well characterized enzymes involved in central metabolic pathways in the cell, such as the carbon metabolism. The discovery that some of the basic enzymes involved in the most conserved and essential metabolic pathways in the cell have also RNA binding capacity may suggest a novel mechanism by which cells sense their metabolic status and provide fine-tuned feedback to the gene expression regulation.

An interesting discussion was conducted concerning the limitations of high throughput methods, which suffer from false detection rate which is in many cases hard to evaluate. One of the main aims of the meeting was to bring together people who develop these technologies with computational experts, to ponder together ways to both evaluate the reliability of the results and propose way for improving the technologies. Mandel-Gutfreund presented new computational advances for predicting RBPs solely based on the physiochemical and electrostatic properties without relying on sequence homology to known RBPs. The algorithm combines modeling the domain structure from the protein sequence with a machine learning approach to define whether the domain binds to RNA was tested on data extracted from the interactome capture experiments and yielded promising results. The main challenges yet to be overcome are related to predicting the reliably of the high throughput experimental results.

RBPs have many diverse function in the cell, and understanding the function of all different RBPs is probably a nonrealistic task, but considerable efforts continue to be dedicated to this necessary task. At the meeting, Maquat presented fascinating evidence for the role of the RBP Staufen, the funding member of a class of proteins involved in subcellular RNA localization, in mediating mRNA-mRNA cross talk via binding to Alu repeats at the 3'UTRs of the mRNAs. As is true for the majority of proteins in the cell, many RBPs undergo alternative splicing generating different protein isoforms, adding further complexity to the problem of predicting RBP function. At the meeting, Fagg presented various biochemical approaches to determine the function of specific isoforms of the well-studied RBP Quaking, demonstrating interesting autoregulation between the different protein isoform, fine tuning the expression level of the different functional isoforms in the different compartments of the cell.

2.4 Structural and biochemical approaches to study protein-RNA interactions and binding specificities

Structural and biochemical analysis of the structure and specificity of RNA-binding proteins remain a mainstay of the field; a requirement to interpret and analyze the results of high-throughput methods, and to translate that knowledge into computational models that are not fully dependent on sequence analysis. Although the complete RNA-binding proteins proteome is large and structurally diverse, the majority of RBPs in all eukaryotic organisms belong to relatively few structural classes that are well characterized structurally and whose mode of binding to RNA has been established in most cases. These proteins are also the most common subjects of high-throughput genomic investigations described before and provide the best subject for computational modeling of these interfaces. A complication, however, is that very often individual domains bind to RNA with only poor sequence specificity and modest affinity, and biological targeting is achieved by either utilizing multiple domains on the same protein or, combinatorially, by forming complexes containing multiple proteins that cooperate to bind to a specific RNA sequence and structure. The analysis and structural investigation of these modes of recognition were a major theme of the workshop and coincides with the state of the art and frontier of understanding of paradigmatic and very abundant RNA-binding domains.

The most common RNA-binding domain is the RRM, which is found in approximately 300 human proteins and represented by >10,000 sequences in Pfam. While the structural basis for recognition of RNA by single RRMs is understood (although not specificity), how multiple domains within a protein combine to generate sequence specific recognition is much less well understood, and very difficult to analyze with high-throughput methods for technical reasons. Allain provided a comprehensive review of the structures solved from his laboratory focusing on RBPs with multiple RRM domains bound to RNA. Rather than providing a unifying theme, it was clear that the binding modes of tandem RRMs on RNAs are extraordinarily diverse, a theme that builds upon and reinforces similar observations made on single RRMs by the same group in the course of the last 15 years. A similar theme was discussed by Tolbert who talked about regulation of HIV splice site by the tandem RRMs of hnRNP A1, where the role of one of the domain was purely structural, yet essential to define the specificity of targeting of these proteins, and dependent on the formation of a specific RNA structural context that mediated interdomain interactions, raising the possibility that RNA structure itself regulates allosterically the protein binding and its downstream biological effects. Clearly, atomic models of these interfaces must somehow account for these observations, the diversity of orientation and recognition principles, and the role of RNA structure on recognition, all features that make it much more difficult to predict the specificity of an RRM from its sequence alone.

A complementary problem was discussed by Sattler, who employs a combination of different structural methods to investigate how specific regulatory complexes assemble on pre-mRNAs. Themes that were observed were cooperativity in binding between different proteins mediated by protein-protein interactions; dynamic rearrangements of the protein and RNA structure; multi-register binding. Not only were these phenomena essential to understand molecular recognition, but also correlated to the biological function of these complexes, as illustrated in the most spectacular way in the study of the complex of proteins responsible for recognition of 3'-splice sites during pre-mRNA splicing. How are these complexes to be modeled? How are high-throughput experiments designed to capture the 'true' specificity of these proteins in the correct functional, multiprotein complex? These are outstanding questions that will require new experimental and computational developments in the near future.
Ramos discussed a paradigmatic KH domain protein, a second very common class of RBPs, called KSRP, that contains four such domains that are used to bind to RNA. Each domain has its own specificity, which was mapped, yet they cooperatively target a specific RNA in a

manner that is as of yet unclear. Murn presented structural data on a less abundant binding domain, the CCCH zinc finger domain. Based on x-ray structural analysis, he showed how six CCCH-zf domains recognize a bipartite recognition site on a mRNA, forming a unique topology of interactions between the protein and the RNA which is highly crucial for the protein function. Similarly, Leeper presented studies of the interaction of a multidomain RRM protein containing four domains and a long non coding RNA, a scarcely studied but biologically very important class of RNA-protein recognition events that deserve much greater attention in the near future.

An important question to be addressed is how specific is an RBP. Sattler illustrated the case of Roquin, which binds conserved stem-loop structures regardless of their sequence and in a manner that depends only on certain structural features (the size of the loop, the length of the base paired region). Most spectacularly, Jankowsky presented his investigation of the C5 E.coli protein, that binds pre-tRNA within a conserved structural context in a manner that is independent of sequence, or so it was believed. Using a clever high-throughput method that couple selection for binding/processing with deep sequencing, he was able to define not just a few high affinity sequences for this protein, but to map the complete landscape, the complete thermodynamic profile for all sequences of five nucleotides bound by this protein. Although functionally the protein is non-specific, the profile is not that dissimilar from that of highly specific transcription factors, with certain sequences in the high affinity tail of the distribution recognized with high specificity. Thus, there is not such a thing as unique sequence preference or non-specific proteins, but a continuum of affinity or affinity distribution. It just so happens that biology does not utilize those sequences, because naturally occurring tRNA substrates are found only in the non-specific center of the distribution. How common is this phenomenon of hidden specificity? How many RBP's utilize sub-optimal sequences that do not coincide with the specific tail of the distribution of affinities? A more subtle point allowed by the generation of such an extensive profile was whether each position was recognized independently and, not surprising; it was found that there were correlation especially involving neighboring nucleotides. Given this cooperativity/anticooperativity, how likely to be successful are computational models based on the independent recognition of each nucleotide within a sequence?

2.5 Novel approaches for designing new RBP's and RNA-binding ligands

The reverse problem of specificity prediction is the redesign of specificity. In fact, a physicist would state that a problem is not satisfactorily understood until successful predictions are made and experimentally verified. While this task has been accomplished successfully for zinc finger proteins binding to DNA, it has been far more difficult to do the same with RBPs.

So far, the only significant success has been obtained with Pumilio proteins. These are multidomain RBPs, each containing typically eight structurally identical repeats that recognize a single nucleotide, A, G, U or C, in a manner that can be specified by changing 2-3 amino acids within each domain. Hall has demonstrated the recognition principle eloquently in the span of about 10 years and how this can be applied to design proteins that bind to single stranded RNA specifically. At the workshop, she illustrated the expansion of her structural analysis to other less canonical members of this protein family, unusual proteins that are more distantly related to the classical Puf motifs, while Henn provided a thorough biophysical analysis of a particular classical human Puf proteins, to thoroughly understand the thermodynamic and biophysical signature of this protein class, suggesting the presence of a non-classical mode of binding to even the classical protein.

Progress in structure determination promises nonetheless to allow other classes of proteins to be designed, which would generate more diverse and interesting tools to interrogate biological processes but, most importantly from the perspective of the workshop, would provide an exacting test of our understanding of the molecular basis of specificity and of computational

programs aimed at predicting, calculating or controlling RBPs. Thus, Ramos shows that a single base change in an RNA can be compensated by a single amino acid change in the protein, in a manner that affects the biological activity of this protein. Similarly, Varani showed the successful redesign of the specificity of an RRM protein, a long sought after goal that has so far escaped successful execution, based on the structural and computational analysis of two binding pockets that resulted in the generation of a protein with altered biological activity. These two examples are idiosyncratic and lack, so far, the systematic power of Puf proteins, but they are necessary steps to expand our understanding of this important protein family and our ability to utilize RBPs as tools.

A related approach that also addresses the issue of specificity (or not) of RBPs deals with the design of so-called arginine-rich motif proteins. This is a common class of domains, containing short, 7-10 amino acids stretches of Arg and Lys residues often found in phage and viral RNA binding proteins. Typically, this protein bind non-specifically, in the absence of cellular factors, but Varani showed how by rigidifying the peptide and providing a cyclic framework, he was able to obtain very high (pM) affinity and specificity. These design projects would be considerable facilitated by better molecular modeling of these complexes, based on atomistic models, as illustrated by Carloni.

2.6 Computational approaches for predicting RNA binding interfaces and protein-RNA docking

The technological advances discussed at the meeting have greatly enhanced our ability to identify new RBPs and find probable RNA targets of selected proteins, but high throughput approached cannot provide the details on the specific mode of interaction between a given protein and its target. As extensively discussed at the meeting, the critical information regarding the pairwise interactions between proteins and RNA can only be derived from structural methods (currently low throughput). However, due to the enormous amount of effort, high cost and time needed to solve the structures of protein-RNA complexes, computational methods have been developed to bridge the gap between the extensive information derived from the high throughput experimental technology and the detailed highly desired but rare structural information.

At the meeting Eric Westhof discussed the different RNA structural features (defined as RNA modules) that can be predicted from sequence alone and a new computational approach for predicting the DNA and RNA pairwise probabilities in proteins, which are directly derived from physicochemical properties (learnt from low throughput structural methods) and evolutionary features (learnt from high throughput sequencing methods). Dobbs presented machine learning and homology based approaches for predicting RNA-protein pairs as well as methods for predicting the specific protein residues which are probable to be involved in the direct interaction with the RNA.

The next extremely challenging computational tasks in the field of protein-RNA interactions discussed at the meeting is predicting the detailed interaction between a protein and an RNA, even when the protein structure is known or can be predicted from close homology. Bujnicki gave an overview of the different strategies and computational methods employed for modeling proteins and RNA and for docking nucleic acids (DNA and RNA) on proteins. While docking methods are widely employed to study protein-protein interactions, very few methods are currently available to model protein-nucleic acid complexes and most are sparsely tested. Bujnicki presented some examples of successful docking predictions, while Tuszynska demonstrated a dedicated software for protein-nucleic acid docking (NPDOCK) developed by the same group. Clearly the field of protein-RNA docking is at early stages and many challenges remain to be overcome before computational modeling can provide nearly atomic resolution structures of protein-RNA complexes.

2.7 The interplay between coding and non-coding RNPs

Cellular RNAs can be partitioned between coding (mRNAs) and non-coding RNAs, including rRNA, tRNA, snoRNA and a very large group of mainly uncharacterized long non coding RNAs (lncRNAs), promoter associated RNAs, antisense transcripts etc. One of the topics discussed at the meeting is the regulation of different classes of RNAs by RBPs. Ohler and Neelanman showed that lncRNAs are generally less stable RNAs and undergo more post-transcription regulations, yet much less is known, compared to mRNAs, about the RBPs which regulated these posttranscriptional events. Interestingly, it was discussed that lncRNA contain short coding region (Open Reading Frames; ORFs), but accurate detection of these short ORFs is a considerable computational challenge. Ohler presented an algorithm to analyses high throughput data derived from a relatively new experimental methodology developed by the Weissman group for mapping ribosome footprints, known as Ribosomal Profiling. The algorithm which is based on Fourier transform approach, evaluates the likelihood that a region codes for a peptide based on the 3-nucleotide periodicity of the signal. Employing this algorithm, they were able to identify hundreds of new putative ORFs in lncRNA. These results again demonstrate the power of combining computational methodologies with high throughput experimental data.

3. Outcome of the meeting

The meeting was an astounding scientific success, according to all speakers, for three reasons.

1. Its format, intimacy and small number of speaker, which provided the meeting with the feel of a workshop, almost group meeting-like, where problems with techniques, approaches and ideas were openly addressed and discussed without hesitation, in a context that was not dominated by any group of speakers
2. The presence of several young speakers, graduate students and post-docs, about 1/3 of all attendants, who, by virtue of giving full presentations, were fully integrated in the community without subjection to more senior speakers
3. The design of the meeting to bring together speakers coming from different communities (structural biologists and biochemists; computational biologists and modelers; genome scientists who apply high-throughput methods), who know of each other and their work, but do not often communicate so closely in such a small workshop setting and with the opportunity to engage freely and extensively with members of the other communities.

As a result of these positive elements, as summarized brilliantly by a set of closing remarks and discussion led by Ares, the meeting highlighted several essential elements that the community believe would push the field further forward.

1. There is a continuing need to increase the number of structures of protein-RNA complexes, which provide the absolutely necessary basis to interpret genome-wide dataset and inform computational prediction methods. The number and quality of structure of RBPs lags significantly behind the equivalent problem of protein-DNA recognition and this severely hinders progress in the field.
2. Computational prediction of the RNA target of an RBP and of RNA structure and the interplay between structure-sequence and recognition lag behind comparable advances in protein-DNA recognition as well. Exploiting the database generated by high-throughput methods and the growth in structures would undoubtedly provide progress in the next few years, but close communications between the communities will be key to advances. Sequence based models should find significant increase in importance in the next few years as high throughout methods grow in scope. Progress in protein design would provide more exacting tests of computational atomic models of interfaces.

3. In addition to technical challenges with the reduction of false positive and false negatives, there is a need to further improve high-throughput methods to better account for the interplay of RNA structure and sequence in RNA recognition as well as the role of multiple protein (and RNA) domains in dictating specificity. These methods need to be expanded to biological systems other than traditional cultured cells. Methods that investigate the landscape of RBP in vitro should be expanded and more closely connected with in cell high throughput methods
4. The role of new RNA-binding proteins, especially metabolic enzymes in cellular function must be better understood. How many unknown RBPs still exist? What is their functional role and which RNAs they interact with? Conversely, the new universe of non-coding RNAs must be characterized with regards to its association with RBPs.
5. Ultimately, this information should be fed into computational models of cellular regulatory circuits. Although the combinatorial complexity of RNA-based regulation is stunning and daunting, efforts should be initiated to establish programs to mathematically model these circuitry.