

On MicroRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology

Richard M. Burian

*Department of Philosophy
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061, USA*

ABSTRACT – This paper is devoted to an examination of the discovery, characterization, and analysis of the functions of microRNAs, which also serves as a vehicle for demonstrating the importance of exploratory experimentation in current (post-genomic) molecular biology. The material on microRNAs is important in its own right: it provides important insight into the extreme complexity of regulatory networks involving components made of DNA, RNA, and protein. These networks play a central role in regulating development of multicellular organisms and illustrate the importance of epigenetic as well as genetic systems in evolution and development. The examination of these matters yields principled arguments for the historicity of the functions of key biological molecules and for the indispensability of exploratory experimentation in contemporary molecular biology as well as some insight into the complex interplay between exploratory experimentation and hypothesis-driven science. This latter result is not only important for philosophy of science, but also of practical importance for the evaluation of grant proposals, although the elaboration of this latter claim must be left for another occasion.

KEYWORDS – microRNA, exploratory experimentation, molecular biology, post-genomic biology, complexity of organisms, limitations on theoretical explanation

Introduction

In the last twenty years, the experimental and conceptual bases for work in molecular biology have undergone enormous change. In this paper, I present a foreshortened account of some recent work illustrating distinctive features of “post-genomic” molecular biology, focusing on the discovery, announced formally in 2001, of microRNAs (standard abbreviation: miRNAs) and on some of the work that has gone into characterizing them. This work exhibits some features of genomics-influenced biology relevant to employment of exploratory experimentation in molecular biology. The biology involved is extremely complex and must be described sketchily here, but there is a simple conceptual core to the story that yields strong morals for philosophy of science.

The miRNA case study is thoroughly intertwined with the philosophical purpose of the paper. The discovery of miRNAs and characterization of their structures and functions depended heavily on exploratory experimentation (hereafter EE), a style or category of investigation that has received some attention in the last decade (see, e.g. Burian 1997; Franklin 2005; Steinle 1997; 2002). Using background considerations and material from the case study, I will provide principled reasons to show the need for EE in this research, which drew heavily on non-theory-driven uses of nucleotide and amino acid sequence information and on work in such (inter)disciplines as genomics and proteomics. Currently, no systems for generating general hypotheses and no bodies of fundamental biological or chemical theory, supplemented by appropriate boundary conditions plus general background knowledge, are able to predict both in general and in detail genotype-phenotype relations, or structure-function relations for wide ranges of important biological molecules. A mixture of empirical, specialized theoretical, computational, and “discovery methods” are required for these major tasks.¹ I argue, in part from the ways in which miRNAs regulate cellular functions, that, at every “level” from molecules up, biological entities must be mutually co-adapted to perform their functions. This co-adaptation results in a kind of epigenetic historicity that prevents successful derivation of the functions (in context) of many biological entities by strict analysis of molecular structure and interactive properties. Said differently, part-by-part analysis of relevant biological structures plus the arrangements of their parts does not suffice to reveal their roles in integrated organisms or their functions.² This argument applies, for example, to attempts to construct genotype-phenotype maps. These claims will be supported as I develop the case study and then turn, finally, to a discussion of EE and its importance for biological practice.

Arguments over these matters manifest themselves in contemporary molecular biology in sharp methodological divisions between advocates of “hypothesis driven science” (or the hypothetico-deductive method)

¹ “Discovery methods” as used here generally require major instrumental and computational resources and yield very large quantities of data. Arguably, similar claims pertain to many other fundamental tasks – e.g., establishing rules for protein folding, determining phylogenies, evaluating contributions to fitness, or predicting evolutionary fates. Fundamental theory is employed in dealing with all of these problems, but the multi-dimensional historicity of biological entities (including, as I will argue, molecules) and the openness of biological systems make clean prediction impossible. Evolutionary changes of function and structure and the complexity of the interactions affecting protein structure and structure-function relations contribute to the need for mixed methodologies in dealing with problems like those listed above.

² This is one aspect of the “contingency thesis” maintained by John Beatty and Stephen Jay Gould with respect to evolutionary processes (Beatty 1995; 2002; Gould 1989), brought down, in the present context, to the ontogenetic as well as the evolutionary scale.

and advocates of “data-driven” science (or inductive method).³ It is important to reduce the sharpness of this supposed dichotomy, but it is also to note that the funding apparatus of the grant system has long favored the ideal of “hypothesis driven” science. The success of the human genome project and the spread of data-driven work in genomics and related fields may tip this balance, but the grant system generally requires hypothesis-driven justifications of applications for financial support. Research on miRNA draws heavily on data-driven research, but it also provides a clean argument for the need for flexible ways of combining data-driven experiments and EE with hypothesis- and theory-driven experiments in domains related to molecular biology. All this should help to soften the dichotomy between these two styles and to ask new questions about the relationships between theory and experimental practice.⁴ Indeed, all four of the “roles of theory in [exploratory experimental] activity” suggested by Kevin Elliott in his paper in this issue (see his Fig. 1) can be found at one time or another in the work on miRNA.

Some preliminary comments about EE will help set up the subsequent discussion. As Elliott and O'Malley argue in this issue (Elliott 2007; O'Malley 2007), the term “exploratory experimentation” covers many sorts of experimental work. In general, EE is limited to situations in which experimental outcomes cannot be accurately predicted by available theories together with general background knowledge plus boundary conditions and (as Franklin 2005 adds) “local” theories of the behavior of the specific entities or processes under examination. There can be, of course, many reasons for this. In the classic cases that Friedrich Steinle and I examined in 1997, theoretical background knowledge about the entities in question (electricity in the late-eighteenth and early-nineteenth centuries, and nucleic acids before 1950) simply did not provide a basis for clear expectations about what fundamental units were at stake (i.e., the underlying ontology), the fundamental rules governing their behavior, or the relevant phenomenology (Burian 1997; Steinle 1997).⁵

³ I refer here mainly to methodological arguments among biologists, though these are, of course, influenced by the philosophical quandaries regarding inductive, hypothetico-deductive, and abductive methods, and theory-driven science vs. data-driven science.

⁴ In the version of this paper presented at the ISHPSSB, I treated hypothesis-driven science and EE as opposed to each other. This is a mistake. Ken Waters took me to task for this in his comments as did other discussants. (Waters 2004) is an important paper that forces a fundamental rethinking of the relation of theory to experiment in classical genetics; it has helped me rethink this issue. I am grateful to him and other participants in the ISHPSSB session for persuading me of the need to explore the relationship of EE to available theories and theoretical knowledge in the background.

⁵ The key work on electrostatics examined by Steinle began when it was wholly unclear whether there were one or two kinds of electricity and continued into electrodynamics when the laws of force governing moving charges in magnetic fields and the relationship between electricity and magnetism

The lack of clear theory-based guidance about expectable outcomes is a key motivation for conducting “broad experimentation”, i.e., for varying parameters and circumstances widely to search for regularities, in hope of reconceiving the underlying units, finding rules governing their behavior, or finding phenomenological regularities that call for – or guide – the search for explanation.

This suggests a key issue regarding in-principle dependence on EE in biological systems. The miRNA case will clarify some of the difficulties biologists face in mapping from structure (e.g., the nucleotide sequence and secondary structure of RNA and DNA) to function (e.g., to gene expression and context-dependent differences in the protein products of a given gene).⁶ The context – e.g., what happens or happened in distant parts of an organism or its environment – may fundamentally alter the structure and/or function of a particular biomechanical structure, machine, organ, or process. The resulting contingency or historicity poses serious obstacles to any principled theory of function. This problem is amplified and clarified by coevolutionary changes in structure-function correlations and is exceptionally clearly illustrated in the instance of miRNAs. The general point is that the behavior of biological systems is strongly contingent on the sequence of inputs and structures arising during ontogeny and evolution.⁷ Accordingly, it may not be possible to derive relevantly significant properties of developing *systems* from the properties and relations of the fundamental units out of which they are built at a particular time. miRNAs illustrate this problem clearly enough to ground a principled argument that EE is essential to the characterization

were unknown and entirely open theoretical problems. Similarly, early work on nucleic acids and their distribution in cells and embryos began before it was clear whether the two nucleic acids we now know as RNA and DNA were distinctively plant vs. animal nucleic acids, what their structures or functions were, how they were distributed in cells or synthesized, etc.

⁶ A background issue that will not be pursued seriously in this paper arises because of the hierarchical structure of biological systems. To illustrate the point, consider what is meant by the function of particular (kind of) gene or RNA: The function(s) are not usually given by the immediate products involved. The function of an alcohol dehydrogenase gene is not to produce a particular mRNA, but to produce an enzyme that, perhaps *inter alia*, plays a specific biochemical role in breaking down alcohol. Achieving this function depends on the mobilization of hundreds of enzymes involved in several distinct processes in a properly organized cell and is regulated by numerous controls independent of the gene in question. In general, the function(s) of genes are identified with effects far removed from the RNA transcripts they yield.

⁷ This description raises many issues beyond the scope of this paper. One marker of such issues occurs when “basic” entities (e.g., genes, neurons) interact non-additively because of the configuration or structure of “higher order entities”. E.g., wiring connections in the brain built during development in response to perceptual experience determine the functions of particular neurons. In such cases, EE is required to get at the details of the system and it may be impossible to construct adequate compositional theories of the behavior of the complex system strictly in terms of the behaviors of its components.

of these molecules and their functions. This result is easily extended to an argument for the inescapability of EE in carrying out such tasks as characterizing the functions of developmentally relevant molecules and entities in ontogeny. As will become apparent, the historicity of the roles of various developmentally relevant units (including molecules) in ontogeny guarantees the insufficiency of structural information at a given time for completion of such tasks.

Genomics, “Hypothesis-Driven Science”, and “Discovery Science”

The power of computational and “high-throughput” molecular and cell-biological technologies has increased enormously over the last two decades. Correlatively, experimental practices have moved in the direction of EE (Franklin 2005). Genomics and related fields employ “wholesale” methods, i.e., they use automated processes and computational tools to perform large numbers of exploratory experiments in parallel at once (Dupré 2004). Such work has recently gained great influence over topics that, until recently, were dealt with mainly by sciences using slower moving technologies and “retail” methods, with relatively small numbers of experiments done in series. In this respect, the methods of genetics and earlier molecular biology resembled a craft tradition rather than the industrial methods of mass production industries. The models of good methodology in molecular biology established by such figures as Crick, Delbrück, and Monod were based on “hypothesis-driven” or “theory-driven” science. Even though those models were softened to accommodate the resort to wet biochemistry to solve the genetic code and the use of “grind ’em and find ’em” experiments in numerous contexts (e.g., gel electrophoresis in the study of genetic variation in natural populations), the more-or-less Popperian or hypothetico-deductive ideal of hypothesis-driven research, dominant in many sciences and in philosophy of science for much of the last century, governed molecular biology and allied fields. Nonetheless, EE was common in molecular biology – a “new” science that kept stumbling on such unanticipated novelties as reverse transcription, split genes, and RNA editing, and such novel entities as transposons, restriction enzymes, and prions.

Genomics, proteomics, and related “omic” disciplines represent a break with the ideal of hypothesis-driven science. Though they are hardly novel in this respect (cf., e.g., paleontology and meteorology), molecular biologists have recently debated the virtues and powers of “data-driven” or “discovery” science and the risks that it might displace “hypothesis driven” science (see for example Aebersold *et al.* 2000;

Allen 2001; Elgar 2002; Gerstein *et al.*, 2007; Kell and Oliver 2004). In spite of the difficulty of ensuring the reliability and of digesting and interpreting the enormous quantities of data generated, the new experimental tools have provided access to molecular structures and mechanisms hitherto far beyond reach. Thanks to this, and also to the power and speed of the new tools and their importance in solving hitherto intractable problems, the entry of “discovery” methods into molecular biology is probably irreversible. Furthermore, the new genomic tools and findings have become enmeshed in interdisciplinary projects touching on virtually every biological domain so that they are regularly combined with theoretical and experimental work of many other sorts. Arguably, the combination of methodologies, bringing together tools and results from very different disciplines, has played a major role in altering fundamental views in many biological and applied biological disciplines and greatly increased the power of available experimental tools. To cite but one example relevant to this paper, I suggest that the current consensus that epigenetic and genetic systems are co-responsible for development, largely rejected and seldom taken seriously as recently as twenty years ago, is a product of the interaction of classical (hypothesis-driven) methods with EE using genomic and sequence-based tools. The three papers in this issue of *HPLS* illustrate the rapid increase in the sorts of problem-centered interactions that EE has been made accessible to fine-grained molecular study. In particular, as this paper shows, the contribution of sequencing without respect to particular hypotheses to studies of RNA structure-function relations is beyond question.⁸

⁸ A technical point about RNA sequencing: One of the most important ways of sequencing RNAs is by sequencing cDNAs (DNA copies of RNAs made by reverse transcription from RNA to DNA). This is done “blindly” in the sense that the high-throughput technology takes all available cDNAs from prepared cells and embeds them (or fragments them and then embeds them) in plasmids or other molecules for use in giant sequencing arrays. What is then obtained (by use of extensive biotechnological and computing technology) is an enormous array of data covering all the RNAs recovered from the cells as they were prepared. In general, the sequencing tools of focal interest in this paper generate “broad” arrays of data in this sense. They fall under the rubric of “discovery science” because they let scientists find out which RNAs (within the accessible range) are present without respect to prior expectations. Such tools allow detection of hundreds of RNAs altered by a single change of cell state, where the older techniques could only follow one or a few such changes at a time. Gerstein *et al.* discuss some methodological issues this raises in a recent paper that examined all the RNA transcripts from a significant portion of the human genome:

The advantage of such arrays is that they probe the transcription in an unbiased and detailed way, with no preconceptions as to where to look for activity. On the other hand, the output from a tiling array experiment can be noisy and needs careful interpretation in order to allow the collection of a reliable set of transcribed regions. The amount of detected transcription depends heavily on the thresholds used when calling transcribed regions and to some extent also on the segmentation algorithms used to delineate transcribed regions from nontranscribed regions. Furthermore, since ... experiments were carried out on many different tissues and cell lines, direct comparison between experiments is not trivial, and

We cannot examine the power, scope, and limitations of sequencing technologies here, but it is worth noting that they are adapted specifically to “high throughput” and broad exploratory research (Franklin 2005; see also Baulcombe 2006 on application of such technologies to miRNAs). It is possible to address questions like “which RNAs or proteins, and how many copies of them, are built into one kind of cell (or organism or tissue at a particular ontogenetic stage) in comparison with another?” and “which cells in a particular brain respond to a particular stimulus or signaling molecule?” Answers to such questions often yield previously unpredictable regularities and are key elements in building an account of the phenomena requiring explanation. Although we focus on miRNAs below, parallel points can be made in many other domains.

Sequencing technologies, heavily used in “discovery” studies of the sorts hinted at above can now yield sequences from, literally, hundreds of thousands of properly prepared samples on a microarray in a matter of hours. It is now possible to sequence an entire bacterial genome using the latest “next-generation” sequencing technologies (Margulies *et al.* 2005)⁹ or to specify hundreds of molecular changes triggered by a single event, such as a nutritional change, viral infection, or a step in development, in a day by use of DNA microarray technology (“Focus: Microarray Quality Control” 2006). Where appropriate biochemical or genetic tools are available, there are often multiple avenues of research, using highly independent technologies, for checking, cross-checking, pursuing and (re)characterizing some of the thousands of entities, interactions, or changes in question.¹⁰ Such mixed tools and cross-checks can sometimes establish robust results even when the processes and interactions under study are not well understood.

The robustness and reliability of the findings regarding miRNAs have improved rapidly in the few years since they were discovered. Major experimental and interpretative disputes remain to be resolved, but the outlines of findings provided below appear to be quite secure

the overlap between different transcription maps is sometimes quite low, partly due to the variable biological features of the samples used in the experiments. (Gerstein *et al.* 2007, 675)

⁹In fact, in my university one of the current sequencing machines routinely sequences two bacterial genomes per day – and the next generation of these machines, anticipated in a year or two, is expected to have a capacity about eight times greater (pers. commun., Roderick Jensen).

¹⁰In terms of the three studies in this issue of *HPLS*, miRNAs and proteorhodopsins currently allow better “triangulation” of this sort than toxicological evaluation of nanoparticles. The availability of independent, well understood techniques that employ “local” theories of the entities or processes they measure is important to our questions about the relation of EE to available theories and to experimental technologies.

thanks to the concurrence of numerous methods in the production and characterization of the molecules in question.

Micro RNA: Discovery and Roles in Cells

The term “micro RNA” was introduced in a “Perspectives” comment (Ruvkun 2001) and three simultaneous reports published in *Science* in 2001 (Lagos-Quintana *et al.* 2001; Lau *et al.* 2001; Lee and Ambros 2001). The reports identified about 100 miRNAs in the nematode *Caenorhabditis elegans*, *Drosophila*, and human tissue culture cells. The discovery arose out of a puzzle concerning two genes in *C. elegans* (*lin-4* and *let-7*) that regulated developmental timing but did not yield any protein product (Ruvkun 2001). The *lin-4* gene was relevantly characterized in 1993 (Lee *et al.* 1993). Its expression had long been known to block the expression or effects of another gene, *lin-14*. That block allowed formation of certain late larval and adult proteins and organs of *C. elegans*. Null mutation and other studies showed that *lin-4* expression was required for those stages to form properly. Lee *et al.* showed that two short RNAs are produced from the *lin-4* locus, one 61 nucleotides long, one 22 nucleotides long; the evidence suggested that no protein was made from the product of *lin-4*. Furthermore, the short *lin-4* RNAs partially complemented (base pair matched) part of the untranslated region (UTR) in the mRNA of the *lin-14* gene, so the authors speculated that *lin-4* regulated developmental timing via anti-sense pairing with part of the UTR of *lin-14*, thus blocking or altering the processing of *lin-14* mRNA.¹¹

Similar findings were published for *let-7* in 2000 (Reinhart *et al.* 2000): *let-7* regulates developmental timing in a manner similar to *lin-4* and expression of this gene yields a short (~21 nucleotide) RNA, complementary to another part of the UTR of *lin-14*.¹² Follow-up work

¹¹ In eukaryotes, mRNAs are constructed in the nucleus by splicing together various parts of the material from the “primary transcript” (the RNA produced directly by “transcribing” a gene) and adding some material at the beginning and end of the “mature” mRNA that provide (among other things) signals that it is ready to be transported out of the nucleus to the cytoplasm. Mature mRNAs contain a “leader” (5’ UTR) and a “tail” (3’ UTR) that are not translated, but may contain signals that affect the translation of the protein-encoding material (e.g., the circumstances in which or rate at which it proceeds). These signals may be modulated as described in the text of this paper. Already in 1993 it was shown that the short *lin-4* RNA partially complemented seven distinct locations on the 3’ UTR of *lin-14* and that cooperativity likely played a role in regulating *lin-14* expression (Wightman *et al.* 1993). Follow-up by Moss *et al.* 1997 showed that *lin-4* RNA also complements the UTR of another gene (*lin-28*), also regulating developmental timing in that instance. Their genetic studies revealed complex hierarchically ordered interactions among the genes involved.

¹² A recent paper (Hayes and Ruvkun 2006) reviews decisive evidence that *let-7* not only is required in *C. elegans* for formation of some adult tissues and for terminal differentiation of certain cell types but also that it is highly conserved in animals and its homologs have the same functions very widely,

on *lin-4*, *let-7*, and the other newly discovered miRNAs soon showed that their initial transcripts yield an intermediate product, ~60-70 nucleotide double-stranded RNAs (dsRNAs) with a hairpin turn. These, in turn, are processed by a specific mechanism (reviewed in the sections on miRNA biogenesis in Bartel 2004) to yield the ~22 nucleotide single-stranded RNAs (whence the label “microRNA”) that block or modulate the expression of other RNAs. The cases of developmental effects that had set the original puzzle were thus shown to result from a general mechanism modulating and/or regulating cellular and organismal processes.¹³

A brief outline of miRNA production (“biogenesis”) shows how the great acceleration of work on miRNAs after 2001 was possible. It was already clear by 2001, and is now firmly established, that the primary pathway for miRNA production involves prior production of dsRNAs with distinctive double-stranded stem and loop structures, produced directly from the RNA transcript of the gene encoding the miRNA. Such genes must have extended regions of (approximate) inverted repeat sequences so that the RNA produced will fold into the requisite stem and loop structures, where the stem(s) consist of complementary base pairs, typically with one or a few “bumps”, each produced by an extra nucleotide or two that make a slight outpocketing on one side of the dsRNA or a short stretch of mismatched bases that make an outpocketing on both strands.¹⁴

Studies of cellular mechanisms for handling dsRNA cannot be followed in detail here, but they were one of the keys to following the production and context-dependent effects (“functions”) of miRNAs and related molecules. Investigations of various roles of dsRNAs and of regulatory processes in which they play a role have been greatly facilitated by biochemical studies of the machinery that processes dsRNA and subjecting that machinery to genetic and biochemical analysis.¹⁵ Since

including in humans. Blocked expression of this gene (by miRNAs or other means) is also associated with a number of cancers, presumably because of lack of terminal differentiation in the affected cells.

¹³ This work did not occur in isolation. The work on *lin-4* and *let-7* was closely connected to much other work on *C. elegans* and there was considerable cross-talk with work in plants and other animals that had already found effects already recognized as effects of short regulatory RNAs.

¹⁴ These bumps are critical for recognition of the precursor dsRNA to make miRNA and/or for the processing of that dsRNA to make a miRNA (described briefly below). For a nice illustration of these structures see (Bartel 2004, fig. 1).

¹⁵ Andrew Fire and Craig Mello were awarded the Nobel Prize for just such work. The breakthrough paper was Fire *et al.* 1998, which examined long (~300-1000 kilobase) dsRNAs that interfere with translation of mRNAs. This paper was quickly followed by an enormous amount of work in several directions. By 2000, at least six or seven labs, including Fire’s and Mello’s were using dsRNA to follow the interconnections among short dsRNAs, miRNAs, and another class of RNAs about the same size as miRNAs, called short interfering RNAs (see, e.g., Parrish *et al.* 2000).

a variety of repeatable processes (e.g., viral infection and nutritional switching) involve dsRNA, there are widely shared mechanisms by which cells deal with dsRNA. These devices are more highly elaborated and their functions more diversified in eukaryotes than in prokaryotes.

We can now provide a greatly simplified account of the major steps in typical miRNA biogenesis in animals (overview in Bartel 2004), starting from the dsRNA produced by a prior process from the relevant gene: (1) In the nucleus, a nucleus-specific protein complex, called the microprocessor complex, processes the dsRNA. In animals, it typically truncates dsRNAs that will be made into miRNAs to a ~60-70 nucleotide length, aligning the truncated product so that it can be cut by a protein called dicer to yield a ~22-nucleotide dsRNA.¹⁶ (Each strand then has a ~2 nucleotide overhang at one end of the double stranded material, utilized in further processing.) (2) The resulting dsRNA remains complexed with dicer; together they are exported from the nucleus to the cytoplasm, where (3) the dicer-dsRNA complex typically enters into a cytoplasmic protein complex called RISC (the RNA-Induced Silencing Complex), which may then separate the strands of the dsRNA, converting one of them into an miRNA,¹⁷ and loading it properly into another RNA cutting protein, called slicer. (4) Complexed with RISC, the miRNA now serves as the recognition device by which RISC locates a target RNA via base pairing with its miRNA. RISC employs that base pairing to attach to the target and adds additional attachments, thus locking itself in position.¹⁸ Depending on the cellular context and the signals received, slicer may then cut the target RNA apart or the RISC complex may remain attached to the target (perhaps modified) as a modulator. We will discuss how these findings helped amplify research on miRNA in the next section.

In animals, many (but in plants few) of the known miRNA targets are on the 3' UTR (downstream untranslated region) of mRNAs, thus influencing how the mRNA is processed. Alternatively, miRNAs may target exons (coding regions of mRNAs), where the effect of the miRNA is more typically to trigger the action of slicer to disrupt (cut apart) the target RNA, thus preventing its translation. Again, an miRNA may target a sequence in an intron (a non-coding region that lies between two coding

¹⁶ In plants, long dsRNAs are often diced repeatedly to yield a number of differently sequenced miRNAs.

¹⁷ The other strand may be used in various ways, though it is usually degraded. E.g., it may be used to form a short dsRNA that, in turn, makes a short interfering RNA. This may then complement other copies of the target of the miRNA in the cell and the process may repeat indefinitely, thus greatly amplifying the effects of the miRNA. On the potential importance of this process, see Baulcombe 2006.

¹⁸ The miRNA may complement its targets imperfectly, at least in animals, so the target sequence is not necessarily unique; there may anyhow be some looseness in the attachment of RISC to the target RNA without the additional binding of RISC proteins to the target.

regions in pre-mRNA¹⁹), where it can alter the likelihood that the coding material will be spliced one way rather than another.²⁰ Furthermore, to have a phenotypically detectable effect, several miRNAs may need to act on the target molecule at once, especially when the target is in the 3' UTR (Miska *et al.* 2007).

Thus, by interacting with a gene transcript, an miRNA can destroy the message, alter its content, or alter the conditions in which or the rate at which the message is translated into protein. And miRNAs may interact with non-coding regulatory RNAs as well as coding RNAs, thus altering the effects or availability of other regulatory molecules. To have a significant effect, an miRNA may need to act cooperatively with additional miRNAs that respond to different cues and are present in different, but overlapping conditions. Cases of all these sorts have been experimentally demonstrated in enormous detail.²¹ The effect of a particular miRNA on its target is typically correlated with the cell type, its physiological state, and other circumstances. In short, the result of the interaction of miRNA with a target RNA varies with cellular and molecular conditions in such a way that its effect simply cannot be determined (except post hoc on the basis of specific empirical knowledge) from its nucleotide sequence or those of its potential targets.²²

Intermezzo: Some Preliminary Conclusions about Exploratory Experimentation

Without going further into mechanisms or molecular details, we

¹⁹ A pre-mRNA is a predecessor of a mature RNA before the introns have been excised.

²⁰ As of 2004, about ¼ of known human miRNA genes were embedded in intronic regions of protein-encoding genes (Bartel 2004, 282).

²¹ Still other effects are well documented. E.g., Vasudevan *et al.* 2007 have shown that let-7 can upregulate translation of certain mRNAs (the opposite of the usual effect) in cell-cycle arrest. This was unexpected, since it was thought until recently that miRNAs only downregulated, protected, or disrupted their targets. Such upregulation may play a role in cell differentiation and has been shown to play a role in the production of certain cancers.

²² A clear example of context dependence is an miRNA in zebra fish, miR-430. As reported by Schier and Giraldez, in response to a developmental signal in embryonic zebra fish, this miRNA interacts with the 3' UTR of maternal mRNAs. After the interaction, *in somatic cells* the maternal mRNAs are degraded, with the result that the embryo's mRNAs are more readily able to act coordinately to alter the development of the embryo and a major development step occurs in coordinated fashion. In contrast, *in germ cells*, the interaction of miR-430 with UTR of maternal mRNAs protects them from degradation, with the results that transcripts of certain maternal genes, needed for the earliest stages of embryogenesis, remained preserved in the germ line for use in the next generation (Schier and Giraldez 2006). Another example is Vasudevan *et al.*'s finding of miRNA-induced upregulation of mRNA translation after cell-cycle arrest and repression of translation when cells are proliferating (Vasudevan *et al.* 2007; see also Buchan and Parker 2007).

can draw some intermediate conclusions about experimentation on miRNAs. Some tools for recognizing the importance of RNA-RNA interactions were not available until very recently, and they produced findings that conflicted with orthodox theoretical views in molecular biology prevalent in the 1980s and early 1990s. Though we cannot probe issues about instrumentation and experimental technologies here, some points connecting instrumentation and theory stand out. For one thing, until recently most molecular biologists thought that gene expression is controlled primarily by interactions among proteins and between proteins and DNA or RNA. For another, these expectations meant that investigative tools (both computational and experimental) were built, scaled, and calibrated to seek protein-protein or protein-nucleic acid interactions, or nucleic acid interactions that affected genes, but not to seek networks of regulatory interactions in which RNAs (especially small cytoplasmic RNAs) played a major role. This made it easy, for example, to treat short RNA as junk, much like the “junk DNA” of highly repetitive sequences, and hard to demonstrate that introns contain functional RNAs. Theoretical bias, detection difficulties, choices of instrumentation, and preferred subjects of investigation all favored the idea that most short RNAs are detritus resulting from the interactions that built, spliced or altered functional RNA molecules. For a long time, it seemed safe that, with a few specific exceptions, short RNAs and RNA-RNA interactions do not play a significant role in the lives of cells and organisms.

Against this background and in the absence of tools that could easily follow arbitrary short RNA sequences around in cells (a technologically difficult task!), the experimental and theoretical backgrounds did not encourage the search for short RNA molecules with regulatory roles. Furthermore, it seemed quite implausible that such short molecules could be sufficiently specific and sufficiently powerful to have strong – even selectable – functional effects. To break this logjam, tools that let one follow some processes involving short RNAs proved essential. The cascade of results that are now altering fundamental theoretical views (only a small fraction of which are discussed here) was achieved in good part thanks to the employment of new investigative tools.²³

To illustrate the sorts of tools involved, consider those that were developed to follow dsRNAs. Because dsRNAs often mark the presence of cellular pathogens and because various sorts of strong defenses against them involve highly conserved proteins (e.g., dicers, slicers, the other

²³ Denis Thieffry (pers. commun.) and others, responding to this paragraph, argue that classical (molecular) genetic approaches to gene regulation could have demonstrated the importance of small regulatory RNAs. This highlights the importance of the biases mentioned in the previous paragraph, a topic that merits further exploration.

proteins in RISC, and, in animals, in the microprocessor complex), the technologies of biochemical analysis for proteins were applied to the processing of short dsRNAs and their locations and effects in the cytoplasm. And as the tools for doing this were developed, they provided other new and extended techniques for dealing with other RNAs, e.g., miRNAs and short interfering RNAs that play regulatory roles. Thus the exploratory work was amplified by new experimental tools drawing on well developed experimental and theoretical knowledge – knowledge that in its own right provided little or no guidance about how regulatory RNAs work or what they might do. In particular, it is worth noting that, other than base-pair complementation, nothing in standard theories in molecular biology provided close guidance about the functional roles of miRNAs and the like. As Ken Waters argues is the case for Mendelian genetics (Waters 2004), molecular genetics and biochemistry provide tools for investigating fundamental biological questions. This use of such tools opens up many questions about experimental systems that go beyond the scope of this paper, questions about which Hans-Jörg Rheinberger's work has much to say (e.g., Rheinberger 1997; 1998; 2001).

In addition to the experimental tools of biochemistry, genetics, and molecular biology, computational tools were of great importance. Again, work on dsRNA as a source of regulatory RNAs illustrates the point. Until the nucleotide sequences of miRNAs were closely studied, nothing computationally distinctive was known about them, but dsRNAs with stem and loop structures are computationally distinctive. They must contain near perfect inverted repeats to achieve base pair complementation on the stem. And if the loop is relatively short (e.g., five or six nucleotides), as is typical, there are nice computational constraints on searches for inverted repeat sequences. Thus, as whole genomic sequences became available and the structure of the dsRNA precursors of miRNAs became known, computational tools became enormously important in identifying DNA sequences with the potential for yielding miRNAs. Even better, the dicer mechanism yields short interfering RNAs as well as miRNAs. Accordingly, computational techniques for studying dsRNAs provided access to a variety of short regulatory RNAs. Since biochemical and genetic studies of dicer could check on the computational results and enable one to follow the relevant RNA-RNA interactions even in the absence of knowledge of their importance or of guidance about their functions from molecular biological theory, one could check on the computational results and begin to determine when and where miRNAs were produced and which genetic material was likely to encode miRNAs. This is an ideal set-up for “discovery science”, i.e., for EE.

The value and indispensability of EE in working on miRNAs should

now be clear. Chasing sequences made it possible to determine when and where miRNAs are produced. But what an miRNA does depends on what functional molecules it targets and what effects its interactions with its target molecules have in context. The sequence of the target is independent of the sequence of the rest of the target molecule so its identity and function cannot be determined by the miRNA sequence alone. And the effect of the miRNA on its target molecules depends on the enzyme complex in which it is embedded and numerous properties of the cellular contexts in which the interactions take place. Thus full information about the functions of miRNA cannot be determined from the nucleotide sequence of their targets proper plus structural analysis of the target regions.

Before we return to further conclusions about EE, it will be useful to set some additional background to show the importance of miRNAs and what they can teach us about the use of EE in contemporary molecular biology. In the next two sections I will briefly examine a few of the roles played by regulatory networks of RNA in multicellular eukaryotes and return us, via miRNA, to EE.

RNA Networks, Differentiation, and Development

John Mattick and colleagues argue that the major evolutionary transition to integrated multicellular organization, made only by eukaryotes, depended on the evolution of sophisticated RNA-based regulatory networks. Complex multicellularity, they maintain, requires more coordination and regulation than is feasible with primarily protein-based regulation of the genetic system. In current organisms, an RNA and protein regulatory network is required, among many other things, to facilitate coordination of developmental processes and maintain cooperation among the cells, tissues, and organs of a multicellular organism. I consider here some of Mattick *et al.*'s arguments in support of this view, one of which is illustrated by Fig. 1.²⁴ To a first approximation, prokaryotes and eukaryotes do not exhibit enormous differences in the haploid amount of DNA utilized to encode proteins. The same is true for single-celled vs. multicellular eukaryotes. But there is a striking correlation between multicellular complexity and the proportion of the genome that does not code for protein.²⁵ Multicellularity seems to require

²⁴ Mattick and colleagues present supporting material, too detailed and complex to present here, in other articles, e.g., Mattick and Makunin 2005; 2006; Taft and Mattick 2004; Taft *et al.* 2007.

²⁵ As Mattick put it in an earlier summary, "Prokaryotes have less than 25% non-coding DNA, simple

that at least 50% of the genomic material be dedicated to non-coding

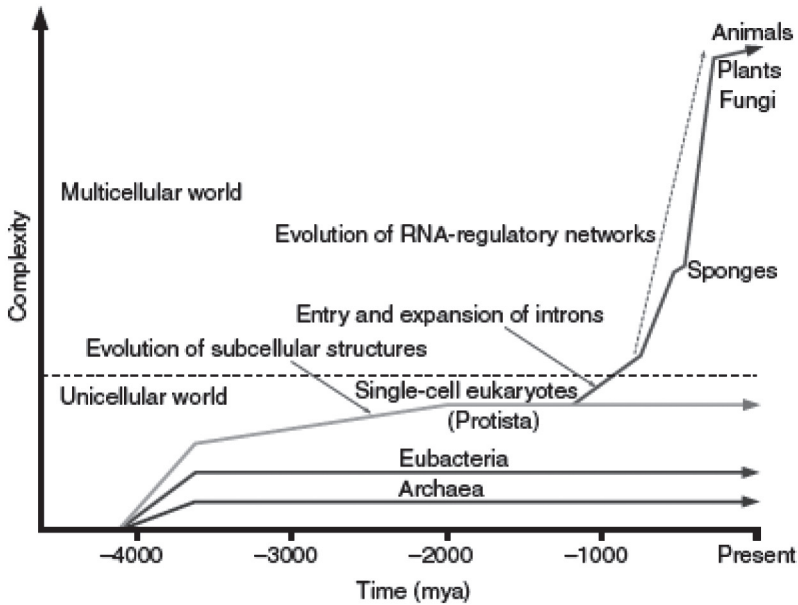


Fig. 1 - A simplified view of the biological history of the Earth.

DNA, and the required percentage of non-coding DNA exceeds 80% for vertebrates. It now appears that the vast majority of the non-coding DNA is transcribed; i.e., it yields RNA. And much evidence has recently accumulated that the additional transcripts yield many kinds of non-coding RNA, some in great quantities, which often interact with one another directly or indirectly, and which regulate or modulate virtually all the processes and products made in bodies and cells.

Let me provide some support for the important claim that an RNA and protein regulatory network is required for coordination of developmental processes and to maintain cooperation among the cells, tissues, and organs of a multicellular organism.²⁶ The concept of cellular heredity, employed in a number of older literatures, is useful in thinking about differentiation and development. Successive differentiation of cells in a cell lineage involves sequentially staged specific “commitments”

eukaryotes have between 25 and 50% non-coding DNA and more complex fungi, plants and animals have more than 50%, rising to approximately 98.5% non-coding DNA in humans — which also have a genome size that is three orders of magnitude larger than prokaryotes” (Mattick 2004, caption for Fig. 1, p. 317).

²⁶The argument that follows was arrived at independently of Mattick 2007, but concurs closely with some arguments in the latter parts of that paper.

of those cells, affecting the chemical processes they perform, their capacities, and their morphologies. Most of the “decisions” in such a chain are extremely hard to reverse, yet to achieve a well differentiated and coordinated body, such “decisions” must be made repeatedly and in orderly succession by cells in numerous cell lineages at different stages of development. Accordingly, although most cell lineages retain some cells at an early pluricompetent stage, lineages of differentiated cells exhibit a strong form of cellular heredity, such that increased differentiation is almost never reversed.

Barring somatic cell mutation, somatic cells of a multicellular organism retain the same genetic content, so the differences between the various sorts of differentiated cells generally do not rest on *genetic* differences (i.e., *differences in nucleotide sequence*). In the contemporary molecular meaning of the term “epigenetic”, which is used for heritable changes that do not involve changes of nucleotide sequences of genes, differentiation requires epigenetic change; i.e., it is a form of *epigenetic cellular heredity*. Orderly development requires yet more – to wit, a system of signals and “local” controls that restrict the timing and locations at which different types of differentiated cells are formed and the interactions among the body’s cells that establish and maintain the correct spatio-temporal distribution of differentiated cells. Such controls enable cells to sum up signals that cross cell boundaries and utilize them so as to act compatibly with the continued functioning of the various organs, tissues, and physiological processes required for maintenance of the body, digestion of food, recognition of foreign substances, etc.

The interactions of regulatory molecules with DNA, RNA, and protein are the key to all of this. Salient features of the regulatory apparatus include the number of copies and distribution of the relevant regulatory molecules, the speed of turnover, transport of signal molecules and regulatory molecules between cells, the rapidity of the regulatory interactions, the combinatorial aspects of the control system, and the extent and length of time of the effects that various interactions have in different contexts. All of these require coevolution and coadaptation of the cells and molecules to each other so that regulatory responses to signals are appropriate for maintenance of key bodily functions and the excursions of the environment with an appropriately wide range.

RNA Regulatory Networks and Exploratory Experimentation

The classification of RNAs is unsettled and some of the classes overlap. Most of these RNAs are involved in regulating or modifying properties

of other RNAs, DNA, or, more rarely, protein (*Regulatory RNAs* 2006; Hannon *et al.* 2006; Mattick and Makunin 2006). The welter of kinds of functional non-coding RNAs (well over two dozen!) is reminiscent of the “fundamental particle zoo” in mid-twentieth century physics, except that there is little reason to suspect that underlying structural principles will allow us to rationalize the classification of the different RNAs in question. We will focus on miRNAs again in this section, concentrating on their roles in the network of regulatory RNAs in animals, but it is important to realize that they are just one of many new kinds of functional non-protein-coding RNA. The results we will examine, some of them controversial, indicate general features of RNA regulatory networks and the need for EE in studying those networks. The argument to this effect extends beyond multicellular eukaryotes and supports the need for EE quite widely in post-genomic biology.

Many miRNAs modulate and coordinate a large number of molecules and processes rather than producing a novel function or regulating a single molecule.²⁷ In some cases, as already indicated, strong forms of cooperativity among miRNAs are important in achieving such modulatory functions. (Recall that some mRNAs have UTRs with multiple miRNA target sites, some of which may be targets for copies of a single miRNA, some of which may be targets for other miRNAs.) Cooperative binding of these sites provides graded combinatorial regulation of the likelihood that the targeted mRNA will be transcribed or disrupted in various circumstances. Recall also that the same binding process can protect the mRNA from degradation in some cell types and increase the likelihood that it will be degraded in other cell types. Since the effects of the regulatory interaction are thus context sensitive, such binding differentially alters the fates of cells in different lineages. Recall too that many regulatory interactions take place at sites (e.g., UTRs) independent of the other functional parts of the regulated molecule (e.g., the protein encoding sequence of an mRNA). It follows that regulatory interactions need not (though they may) change the function(s) of regulated molecules and can evolve independently of other evolutionary changes in the target molecule to affect the timing of interactions, etc.²⁸

²⁷ Here is one of, literally, hundreds of examples: Blocking production of zebra fish miR-430, mentioned in n. 18, allows retention in somatic cells of at least 700 maternal mRNAs that would otherwise be degraded very slowly. About 2/3 of these have been shown to have complementary sequences to miR-430 in their UTRs. The net effect of miR-430 is to speed up degradation of maternal mRNA at a specific developmental stage so that zygotic proteins, rather than maternal proteins are the sole (or predominant?) proteins expressed at the next developmental stage. (See Schier and Giraldez 2006, 197-198.)

²⁸ Denis Thieffry (pers. commun.) responded to this paragraph by pointing out that closely parallel points pertain to DNA transcription factors and their interactions with their target sites, which are now

The available quantities of miRNAs are often strongly regulated. Cellular signals and conditions affect the production, longevity, and availability of miRNAs. In *C. elegans* there are at least three developmentally staged miRNAs that average more than 50,000 copies per adult cell while the average for adult-expressed miRNAs is less than 800 molecules per adult cell (Bartel 2004, 283). Many miRNAs are produced in a cell-specific or tissue-specific manner; to determine their functions one must know which cells to examine. And miRNAs can target from zero, one, or a few mRNAs to hundreds, or perhaps even thousands, of distinct mRNAs (Bartel 2004; Lewis *et al.* 2007; Stark *et al.* 2003). Thus, to assess the function of an miRNA it is typically necessary to determine in which cells and circumstances it is produced, which RNAs it affects, their functions, and the ways in which the miRNA alters their behavior in different contexts. Foreseeably, this can only be done by EE, i.e., by “wide” investigation of miRNAs, their targets, and their behaviors.

One final argument about miRNA regulatory action deserves consideration here. Mattick 2007 suggests that protein regulation is analog since it depends on conformational matching and the like to regulate the occurrence or the rate of regulated interactions, whereas regulatory RNA is digital. This analogy is suspect in general, but there is a core point that should be taken seriously. Short RNAs are virtually freed from conformational constraints in recognizing their targets because the target RNAs are linear molecules generally not highly convoluted – the nucleotides are nearly always accessible to base pairing. Accordingly, short RNAs often operate, effectively, as digital recognition devices, utilizing only sequence information. This sort of sequence information is familiar from Crick’s version of the Central Dogma, according to which linear sequence information in DNA is sufficient to specify amino acid sequence in polypeptides which, in turn, was presumed to be sufficient to specify the three-dimensional structure of proteins and, ultimately, the entire structure of cells and organisms (see Strasser 2006). Although the mechanism of base pairing by itself can accomplish some of the regulatory work of miRNAs, most of that work is done, secondarily, by the many complicated structures with which miRNAs are associated – RISC, slicer, and the like. Thus, on Mattick’s account, *recognition is digital, but much of the regulatory work is done by analog devices attached to the miRNA.*

On an evolutionary scale, the nucleotide sequences of miRNAs can be

studied by use of specialized genomic technologies (SAGE, ChIP-on-chip, DamID, ChIP-seq, etc.). The comparison is useful and needs further exploration. It reinforces my argument about the ubiquity and necessity of EE in post-genomic molecular biology. It also reinforces Thieffry’s suggestion that small interfering RNAs were technologically accessible before their significance was actually recognized.

changed far more nimbly than can protein regulatory devices; alteration of miRNAs and their target sequences is an economical way to modulate which molecules are targeted and what is done to them. And on an ontogenetic scale, short RNAs can be activated and act much more quickly to reach their targets than protein devices – especially if distinct protein devices are required for distinct target molecules. Furthermore, co-evolution of (often distant) short stretches of DNA that regulate a single target is far easier than evolution of protein devices to regulate the many distinct molecules that are the ultimate targets of short RNAs, so modulation of miRNAs and their targets to alter or regulate likelihoods or rates of certain processes is far easier than alteration of separate protein regulatory devices. These are major advantages of the “digital” recognition of RNA targets over protein-protein regulation. The subtlety of the regulatory controls thus made available with minimal evolutionary change is extraordinary and, arguably, unattainable by primarily protein-based regulation.²⁹

It should now be clear why miRNAs and other short interfering RNAs are so effective in combination and can have major regulatory effects while containing so few nucleotides. The mechanics of regulation by miRNA requires far less “information” than would be required for comparable regulation by protein mechanisms and the turnover time to achieve adjustments in regulation via such regulatory tools is considerably shorter than it is with proteins (on both organismal and evolutionary time scales). Short regulatory RNAs respond more rapidly than protein-encoding genes to stress, nutritional change, and other events and sometimes can yield a more flexible regulatory response to environmental change on an evolutionary scale than protein changes, e.g., by regulating developmental timing of otherwise independent gene products. Considerations such as these reinforce and support Mattick *et al.*'s arguments about the connection between RNA regulatory networks and the evolution of multicellularity: the control of development in multicellular organisms, with its requirements of coordination of cells and tissues, is far easier with an RNA regulatory network than with a mainly protein control system. Multicellularity may well have been unattainable with protein-based regulatory control systems.

The findings reviewed to this point show that one cannot delimit the functions of miRNAs by analyzing the sequences of nucleotides to which they are complementary plus the mechanics of their actions plus the rules (such as they are) governing the conformation of RNAs, proteins,

²⁹ The argument of the last two paragraphs has benefited greatly from comments and objections sent me independently by Jean Gayon and Denis Thieffry.

and related molecules. The same sequence of about 22 nucleotides, near enough,³⁰ may occur in hundreds or even thousands of molecules that are potential targets of the miRNA. Whatever proportion of those targets are actually affected by an miRNA, the matches do not significantly constrain the sequences or functions elsewhere in the target molecules. Accordingly, diagnosis of the function(s) of the miRNA is not settled by the mechanics of targeting plus conformational knowledge about its targets. Again, miRNAs may cooperate in affecting their targets, with effects that are generally marginal and decisive only in special circumstances. A much larger range of historical and contextually salient factors determine what cellular and organismal functions are modulated by miRNAs. These factors must be taken into account to evaluate the functions of miRNAs.

On the basis of these results, I argue that theories built on structural molecular formulae (including nucleotide sequence) and structural features of molecules do not contain sufficient information to diagnose the functions of miRNAs. The dimensionality of the problem is greater than that of the body of knowledge going into nucleotide sequence and (secondary) structure of the relevant molecules. To settle systematically which combinatorial possibilities are “programmed” by the devices controlling staging, timing, and coordination of development, one also needs information about the temporal staging of changes in the molecular compositions of the many cells in question; in short, the histories of the altered cellular contents that, in turn, alter the regulatory effects of the interactions of miRNAs and their target molecules. *Prima facie*, this result carries forward rather generally to RNA regulatory networks. It clearly impacts any theory of function that must take account of the contextual differences between different regions within cells and between different kinds of cells. We are on the verge here of very large issues about the status of “higher order” entities, the status of systems biology, and general issues about strong forms of reductionism that cannot be addressed here. It does, however, bring us back one last time to EE.

³⁰ The qualification is due to the fact that, at least in animals, miRNA typically need not complement its target perfectly. In fact, the exactness of complementation between an miRNA and its target may determine whether RISC slices the target apart or represses its translation (Bartel 2004, 288). There is a sequence of seven nucleotides near the beginning of miRNAs that must match the target exactly and is required (or very nearly so) to initiate the binding to the target RNA, but some sloppiness is common in the matching of some of the other nucleotides, especially in the last seven nucleotides of the molecule. How much sloppiness, and whether it has significant effects, depends on poorly understood details of the particular molecules involved and the context.

Conclusion: The Need for Exploratory Experimentation

This examination of recent work on miRNAs and RNA regulatory networks shows that EE has played and continues to play a key role in studying the production and functions of miRNAs. Genomic, gene expression, proteomic, and other databases have been put to good use in determining the locations, behaviors, interactions, and copy numbers of the molecules of interest. This point applies most obviously to studies of multicellular organisms, where RNA regulatory networks play an extensive role in development and in coordination of responses in different tissues and organs, but it also applies to a much greater range of organisms, including prokaryotes.³¹ One must have considerable information about the contents of cells, their histories, and the temporal sequence of their environmental circumstances to analyze relevantly the behavior of regulatory RNA molecules and the molecules they regulate. Current fundamental theories bearing on structure-function relations in molecular biology and genomics, together with boundary conditions regarding the sequences and secondary structures of proteins, RNAs, and DNA in the organism at a given time, do not contain enough information to provide a functional analysis of RNA regulatory networks and, perhaps, RNA regulation more generally. We are dealing here with enchainment and shifting dynamic equilibria.

The claims stated above are justified by dimensional analysis comparing available theories (plus the pertinent boundary conditions) with the determinants of cellular states (though, of course, I did not provide a full-scale dimensional analysis). Not enough information is built into current theories to allow derivation of the behavior or functions of crucial molecules (e.g., miRNAs) from a temporally confined set of boundary conditions. It is highly unlikely that a full table of contingencies for the plausible ranges of novel conditions and combinatorial interactions can be derived from any current or foreseeable theory covering protein interactions with nucleic acids or the like. The theories on the horizon are simply unable to unpack the contingencies that result from shifting spatio-temporal relationships within organisms and the time sequence of changes in the distribution of molecules and molecular constitutions of cells. Developmental construction of molecules, tissues, and organs, however, depends on the sequencing of events and, so far as can be told,

³¹ As O'Malley (2007) demonstrates, we are still far from having adequate general knowledge of the proteins and protein-encoding molecules that prokaryotes contain. It is likely that their regulatory systems will also yield major surprises like those that their proteomes and genomes have already done. And as Elliott (2007) shows, the interactions of molecules at the nano-scale are quite generally unpredictable from highly accurate theories of chemical reactions at a molar scale.

proceeds on the basis of local controls and receipt of signal molecules correlated with distant events.

An alternative formulation may clarify this point. The core difficulty can be stated in terms of epigenetic change. The molecular content and spatial organization of cells and organisms enduring through time is constructed and reconstructed in response to many sorts of inputs. Epigenetic change (and epigenetic cellular heredity) means that full knowledge of the genome and of the conditions pertaining at a particular developmental stage are not sufficient to calculate the organism. This point is not an abstract in-principle concern; it applies concretely and at a practical level to the vast majority of multicellular eukaryotes and quite likely to many single-celled organisms as well. At the same time, the broad investigation of the contexts in which miRNAs are generated and the interactions into which they enter must be integrated with the extensive, hard-won experimental and theoretical knowledge of biological molecules and their interactions, the organization of cells, and the processes that are involved in cellular and organismal functions. Even if such knowledge does not contain sufficient information to solve our problems about regulatory function, without it computational models and sequence information would not be productive. The integration of the experimental and theoretical knowledge already obtained in molecular biology and allied disciplines with data from genomic and other broad investigative technologies is required for understanding the functioning of molecules in the complex regulatory systems of living organisms.

My argument does not depend solely on analysis of regulatory networks. Recall that miRNAs regulate their target molecules in “digital” ways – the regulatory sequences are recognized independently of the functions of the regulated molecules and (by and large) independently of the conformations of the target RNAs. Thus, the functions of regulatory molecules cannot be specified simply by examining the interactions of their regulatory segments with their targets.³² Furthermore, combinatorial modulation of regulatory interactions produces a serious combinatorial

³² This claim is hardly novel. Consider allosteric proteins: these have (at least) two reaction sites, an effector and an active site. Interactions at the effector alter the conformation of the protein so that the active site is activated or deactivated. As Jacques Monod argued, “allosteric interactions do not depend on the structure or the particular chemical reactivity of the ligands themselves, but entirely on the structure of the protein, which acts as a relay” (Monod 1966, 481). In other words, chemical analysis of what happens at the regulatory site does not reveal the function of the ligand or the allosteric protein. Those functions, like the functions of miRNAs, depend on the chain of reactions affected by the active site, which cannot be determined solely by chemical analysis of the interactions between the ligand and the effector site. Parallel comments apply, as well, to the functions of transcription factors, which cannot be determined from their interactions with the DNA binding sites with which they interact, but require a much fuller analysis of the molecular, cellular, and organismal context.

explosion that makes molecule-by-molecule functional analysis extremely difficult. There are so many subtle variations in molecular structure, regulation, and context that exploratory and computational approaches will precede and accompany structural analysis for a long time to come.

For analyses of molecular functions, available technologies limit the requisite information about boundary conditions and relevant cellular states. Recall that full molecular understanding of the regulatory effects of a given miRNA on a given molecule in one kind of cell or cell state may not apply to the regulatory effects of the same miRNA on the same target in another kind of cell or when it is in a different cell state. The distribution of molecules within and between cells and the details of functional responses to “simple” stimuli are rarely tractable analytically. There is no reason to suppose that such limitations are restricted to multicellular eukaryotes; so far as we now know, similar complications pertain to prokaryotes as well (though to some extent these are matters of degree). The enormous experimental and computational armamentarium of biotechnology, genomics, proteomics, and so forth provides extensive, if approximate, knowledge of the details of complex cascades of cellular and organismal responses to local signals in tractable cases and of the effects of particular changes in particular systems. Such findings allow scientists to deploy highly developed instrumentation, “local” theories, and computational tools to follow in rigorous detail some of the cascades and elaborate the consequences that ensue from particular inputs or stimuli. But no general theory of molecular interactions is able to provide across-the-board predictions of such cascades or their consequences.

It should therefore be no surprise that scientists in any number of relevant disciplines seek to resolve many problems by using the rapidly developing “wide” technologies to obtain specific information by tagging and following particular molecules through chains of reactions or by comparing sequences and cascades in different organisms and contexts. The examples covered in this paper illustrate some of the enormously varied problems in divergent domains that depend on “wide” input and EE employing sequencing technologies. Problem solution regarding molecular behavior, of course, still relies on a great variety of powerful theories. EE on miRNAs was followed up by probing their functions using biochemical, genetic, cytological, evolutionary, and many other theories and experimental tools, not to mention the theories required to understand the many experimental technologies employed. And the intensive use of computational technologies in following sequences and in many other aspects of the experimental protocols is beyond question. Nonetheless, the point of my argument remains: without reliance on

broad sequence data, all the other theories (together with the boundary conditions they utilize for explanations and predictions) cannot, in general, predict when and where regulatory reactions will occur. Even with reliance on sequence data it is not, in general, possible to predict in which precise circumstances what, precisely, the immediate effects of a regulatory interaction will be at a fine enough grain to provide structure-function correlations or to trace cellular and developmental reactions from nucleotide sequence to functional result or phenotype. Satisfactory execution of such tasks requires integration of EE with highly specific knowledge, “local” theories, and the like.

In my view, this is exactly as it should be. As best we can tell, the integration of organisms is not under central control, but is the evolutionary result of integrating local controls and signaling cascades in such a way that organisms survive the insults of the environment and interactions with other organisms. If the implicit metaphysics of organisms behind this claim is anywhere near correct, we ought not expect to produce a core theory to handle the integration of organisms and the functions of their (molecular) components. Organisms are historical products, tinkered together over evolutionary time (Beatty 1995; 2002; Gould 1989; Jacob 1977) and developmental time (this paper). They may not be Rube Goldberg devices, but they also were not built by rational engineers working with well defined constraints and resources. Coevolution and cooperation among epigenetically altered entities produce problem shifts and resource shifts and lock in what, from another perspective, are sequence-dependent historical accidents. Such accidents – and the products that they yield – cannot be understood from basic principles.

This last consideration provides a reason to expect post-genomic molecular biology to be forced to integrate EE with classical theoretical disciplines over the long term. Recognition of this need should be of considerable practical importance to biologists and granting agencies alike. It is incumbent on those who seek to understand the methodology of post-genomic biology to help think through just how the integration of EE with more traditional (hypothesis-driven) disciplines and research should proceed. The three papers in this special section of *HPLS* should help orient further work to that end, but there is a long road ahead.

Acknowledgements

Thanks to Kevin Elliott for inviting me to present the predecessor of this paper at the session he organized for the 2007 ISHPSSB meeting. The paper has been improved by discussion at that session and by helpful feedback from Kevin Elliott, Maureen O’Malley, Ken Waters,

two Virginia Tech colleagues, Diya Banerjee and Roderick Jensen, and an anonymous referee. Helpful comments from the following colleagues also led to revisions of the text: Ron Amundson, Michael Dietrich, Jean Gayon, Joram Piatigorsky, Robert Richardson, and Denis Thieffry. I am grateful to all of these good people.

References

- Aebersold R., Hood L.E. and Watts J.D., 2000, "Equipping Scientists for the New Biology", *Nature Biotechnology*, 18(4): 359.
- Allen J.F., 2001, "Bioinformatics and Discovery: Induction Beckons Again", *BioEssays*, 23: 104-107.
- Bartel D.P., 2004, "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function", *Cell*, 116(2): 281-297.
- Baulcombe D.C., 2006, "Short Silencing RNA: The Dark Matter of Genetics", *Cold Spring Harbor Symposia on Quantitative Biology*, 71: 13-20.
- Buchan J.R. and Parker R., 2007, "The Two Faces of miRNA", *Science*, 318: 1877-1878.
- Burian R.M., 1997, "Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952", *History and Philosophy of the Life Sciences*, 19: 27-45.
- Dupré J., 2004, "Understanding Contemporary Genomics", *Perspectives on Science*, 12(3): 320-338.
- Elgar G., 2002, "The High Throughput Revolution", *Briefings in Functional Genomics and Proteomics*, 1: 4-6.
- Elliott K.C., 2007, "Varieties of Exploratory Experimentation in Nanotoxicology", *History and Philosophy of the Life Sciences*, 29(3): 311-334.
- Fire A., Xu S., Montgomery M.K., Kostas S.A., Driver S.E. and Mello C.C., 1998, "Potent and Specific Genetic Interference by Double-stranded RNA in *Caenorhabditis elegans*", *Nature*, 391: 806-811.
- "Focus: Microarray Quality Control", 2006, *Nature Biotechnology*, 24(9): 1039; 1103-1169.
- Franklin L.R., 2005, "Exploratory Experiments", *Philosophy of Science*, 72: 888-899.
- Gerstein M.B., Can B., Rozowsky J.S., Zheng D., Du J., Korbel J.O., Emanuelsson O., Zhang Z.D., Weissman S. and Snyder M., 2007, "What is a Gene, post-ENCODE? History and Updated Definition", *Genomics Research*, 17(6): 669-681.
- Hannon G.J., Rivas F.V., Murchison E.P. and Steitz J.A., 2006, "The Expanding Universe of Noncoding RNAs", *Cold Spring Harbor Symposia on Quantitative Biology*, 71: 551-564.
- Hayes G.D. and Ruvkun G., 2006, "Misexpression of the *Caenorhabditis elegans* miRNA *let-7* is Sufficient to Drive Developmental Programs", *Cold Spring Harbor Symposia on Quantitative Biology*, 71: 21-27.

- Jacob F., 1977, "Evolution and Tinkering", *Science*, 196: 1161-1166.
- Kell D.B. and Oliver S.G., 2004, "Here is the Evidence, Now what is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-driven Science in the Post-genomic Era", *BioEssays*, 26: 99-205.
- Lagos-Quintana M., Rauhut R., Lendeckel W. and Tuschl T., 2001, "Identification of Novel Genes Coding for Small Expressed RNAs", *Science*, 294: 853-858.
- Lau N.C., Lim L.P., Weinstein E.G. and Bartel D.P., 2001, "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*", *Science*, 294: 858-862.
- Lee R.C., Feinbaum R.L. and Ambros V.R., 1993, "The *C. elegans* Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *Lin-14*", *Cell*, 75: 843-854.
- Lee R.C. and Ambros V.R., 2001, "An Extensive Class of Small RNAs in *Caenorhabditis elegans*", *Science*, 294: 862-864.
- Lewis B.P., Burge C.B. and Bartel D.P., 2007, "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets", *Cell*, 120(1): 15-20.
- Margulies M. *et al.*, 2005, "Genome Sequencing in Microfabricated High-density Picolitre Reactors", *Nature*, 437: 376-381.
- Mattick J.S., 2004, "RNA Regulation: A New Genetics?" *Nature Reviews Genetics*, 5: 316-323.
- Mattick J.S., 2007, "A New Paradigm for Developmental Biology", *Journal of Experimental Biology*, 210: 1526-1547.
- Mattick J.S. and Makunin I.V., 2006, "Non-coding RNA", *Human Molecular Genetics*, 15 (Review Issue 1): R17-R29.
- Miska E., Alvarez-Saavedra E., Abbott A.L., Lau N.C., Hellman A.B., McGonagle S.M., Bartel D.P., Ambros V.R. and Horvitz H.R., 2007, "Most *Caenorhabditis elegans* microRNAs are Individually not Essential for Development or Viability", *PLoS Genetics*, 3(12): e215 (pp. 2395-2403).
- Monod J.L., 1966, "From Enzymatic Adaptation to Allosteric Transitions", *Science*, 154: 475-483.
- Moss E.G., Lee R.C. and Ambros V.R., 1997, "The Cold Shock Domain Protein *Lin-28* Controls Developmental Timing in *C. elegans* and is Regulated by the *Lin-4* RNA" *Cell*, 88(5): 637-646.
- O'Malley M.A., 2007, "Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case", *History and Philosophy of the Life Sciences*, 29(3): 335-358.
- Parrish S., Fleenor J., Xu S., Mello C.C. and Fire A., 2000, "Functional Anatomy of a dsRNA Trigger: Differential Requirement for the Two Trigger Strands in RNA Interference", *Molecular Cell*, 6: 1077-1087.
- Regulatory RNAs, 2006, *Cold Spring Harbor Symposia on Quantitative Biology*, 71.
- Reinhart B.J., Slack F.J., Basson M., Pasquinell A.E., Bettinger J.C., Rougvie A.E., Horvitz H.R. and Ruvkun G., 2000, "The 21-nucleotide *let-7* RNA Regulates Developmental Timing in *Caenorhabditis elegans*", *Nature*, 403: 901-906.

- Rheinberger H.-J., 1997, *Towards a History of Epistemic Things: Synthesizing Proteins in the Test Tube*, Stanford, CA: Stanford University Press.
- Rheinberger H.-J., 1998, "Experimental Systems, Graphematic Spaces". In: Lenoir T. (ed.), *Inscribing Science: Scientific Texts and the Materiality of Communication*, Stanford, CA: Stanford University Press, 285-303.
- Rheinberger H.-J., 2001, "Putting Isotopes to Work: Liquid Scintillation Counters, 1950-1970". In: Joerges B. and Shinn T. (eds), *Instrumentation: Between Science, State and Industry*, Dordrecht: Kluwer, 143-174.
- Ruvkun G., 2001, "Glimpses of a Tiny RNA World", *Science*, 294: 797-799.
- Schier A. and Giraldez A.J., 2006, "MicroRNA Function and Mechanism: Insights from Zebra Fish", *Cold Spring Harbor Symposia on Quantitative Biology*, 71: 195-203.
- Stark A., Brennecke J., Russell R.B. and Cohen S.M., 2003, "Identification of Drosophila MicroRNA Targets", *PLoS Biology*, 1(3): 397-409 (E 360).
- Steinle F., 1997, "Entering New Fields: Exploratory Uses of Experimentation", *Philosophy of Science*, 4 Suppl.: S65-S74.
- Steinle F., 2002, "Experiments in History and Philosophy of Science", *Perspectives on Science*, 10: 408-432.
- Strasser B.J., 2006, "A World in One Dimension: Linus Pauling, Francis Crick and the Central Dogma of Molecular Biology", *History and Philosophy of the Life Sciences*, 28: 491-512.
- Vasudevan S., Tong Y. and Steitz J.A., 2007, "Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation", *Science*, 318: 1931-1934.
- Waters C.K., 2004, "What was Classical Genetics?", *Studies in History and Philosophy of Science*, 35: 783-809.
- Wightman B., Ha I. and Ruvkun G., 1993, "Posttranscriptional Regulation of the Heterochronic Gene *lin-14* by *Lin-4* Mediates Temporal Pattern Formation in *C. elegans*", *Cell*, 75: 855-562.

