Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 123–126

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Essay

Life on Two Tracks

Jun Yu *

CAS Key Laboratory of Science and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China

Available online 23 June 2012

There are, I think, several factors that contribute to wisdom. Of these I should put first a sense of proportion: the capacity to take account of all the important factors in a problem and to attach to each its due weight. This has become more difficult than it used to be owing to the extent and complexity of the specialized knowledge required of various kinds of technicians.

—Bertrand Russell, Knowledge and Wisdom

In lieu of an introduction

The second modern synthesis of biology [1] must start with new stratifications and conceptualizations, as well as new ways to connect all of them [2], which are to be perfected through in-depth discussions that may be provoked by seemingly controversial ideas. And new paradigms can thus be built upon the novel thoughts that may eventually converge. As knowledge accumulates based on either newly generated data or frequent visits to old datasets, the new generation of scientists will digest, scrutinize, and cast doubts on the novel thoughts, and all these activities will lead to acceptance or rejection of the new synthesis and its essentials. The knowledge network may either keep growing in multiple dimensions or end temporarily at a single vortex.

Trained as a biochemist in college and then a cell biologist early in my scientific career, I only used molecular biology tools to manipulate genes as "parts" and use them to "assemble" certain "operations" and never worried about the genetic background of the cells and animals we used. Some of the tissues (organs) for primary cultures were actually collected from a local slaughter house as byproducts of steaks and sausages and individual samples were often used as a pool. Working on genetics, genomics, and bioinformatics for a couple of decades, I have encountered problems in extending classic concepts of genetics and evolution to explain biological phenomena, especially those defined based on basic concepts of cell biology and biochemistry. My typical example is how to look for genotype-phenotype relationship for indel selection found in the mammalian minimal introns when the selection not only exists in significant numbers but also is predicted to be weak and lacks definable phenotypes in the context of classic genetics [3–5], although variations of the splice site are often considered as deleterious and attributable to certain inherited diseases [6,7]. While believing that we still have a long way from the understanding of the Rules of Life, I also reckon we better get started now before it is too late as we always say—a journey of a thousand miles begins with the first step.

We now have two tracks of fundamental thinking in biology—geneticists on the one side and cell biologists/biochemists and alike on the other side—informational and operational. The former constructs phenotype-genotype relationship relying on sequence changes in the contexts of populations, species, and lineages, and the latter seeks molecular details leveraging on model systems, such as model organisms and cell lines with less concerns about their generic backgrounds. The future genomics or the modern synthesis of genome biology is to unite the two tracks, of course after some exhaustive discussions. In addition, the differentiation of the two tracks is not to simply draw the boundaries of the two but to seek common grounds for the integration of more information into a knowledge network that connects scientific facts from all fields of biology in a unified realm.

The elements for the origin of life

The modern synthesis is so fundamental that we have to start our thinking with how life originates in two tracks rather than one. To be alive, life has to operate in two basic ways in the operational track: mechanical (or structural)

* Corresponding author.
E-mail: junyu@big.ac.cn (Yu J).

http://dx.doi.org/10.1016/j.gpb.2012.06.001

and chemical (or catalytic) to gain enough mechanistic operations and homeostases for doing different things for its survival and propagation. In an operational point of view, life needs not to be reproducing or replicating but propagating with non-identical offspring if armed with a tool kit for sufficient survival. Initially, the early life forms might be just so flexible in shape and size that the successors of these life forms had also nothing resembling their own predecessors. However, the more complex and dynamic the life forms became, the more they needed ways to keep their surviving and winning skills consistent to avoid being eliminated by competitors. Clearly, life had to create mechanisms to maintain stability of information inheritance. Therefore, life began with the operational track, followed by a slow establishment of the informational track. The two tracks definitely have their own histories and characteristics; one is mechanistic and the other is abstract; one is versatile and the other is stringent; and one implements the other with increasing sophistication and the other guarantees more robust implementation. I believe that the elements of life are multifold even at the beginning, and the operational system and compartmentalization are of essence for life to commence. The operational molecules started with RNAs and later recruited more sophisticated proteins that are more diverse and compact in physiochemical properties, and together they engaged DNAs as the informational molecule.

The RNA world was there and is still here

If there is a RNA world [8], the two tracks must have started from it. RNA is sufficient to provide the right tool kit for the realization of the two tracks. The RNA world hypothesis is essentially based on molecular relics and limited experimental evidence. The former relies on parsimonious principles and logic, whereas the latter often involves new discoveries. Nevertheless, the complexity of the contemporary RNA world being revealed thus far has been so dramatic that it may one day match or even surpass those of DNA-centric and protein-centric mechanisms.

The RNA world starts unified and continues its evolution. There have not been two RNA worlds—modern and contemporary—but one that is both ancient and evolving continuously, as new RNA molecules are constantly in the making [8]. We may one day discover the life forms from the RNA world—some people believe that it is long gone—and regret how stupid we were to have overlooked the most obvious. One recent example is the discovery of the possible chimeric molecules of DNA–RNA in the largely unexploited viral world [9].

The RNA world, however, may have gone through three essential phases. In the first phase, RNA was born of being able to play two indispensable roles: operational and informational, even though there is a possibility that the former might exceed the latter when the latter remained short and fragmented. The operational role is essentially tool-making and physiochemical in nature. It is not only catalytic, usually involving chemical reactions—the break-and-join of chemical bonds—but also non-catalytic such as hybridization that is weak (mainly hydrogen bond) individually but strong when consecutively extended along the nucleotide chain, and even weaker interactive forces, including ionic and the Van der Waals force. The informational role is basically a coding-decoding process between two or more cipher texts that can be both short and a bit lengthy. Together, the two-tracked ribo-protocells were doing quite well; the RNA-made tools might be cumbersome but plenty, and may be imprecise but versatile. Therefore, the offspring of the RNA-based ribo-protocells are vast in number and highly variable in information content.

In the second phase, the operational track in the RNA world became more sophisticated due to the involvement of proteins through a gradual replacement of RNA-based operational molecules. First, many molecular mechanisms were created and perfected in this phase, including RNA splicing, editing, and modification, and of course later protein translations. Since molecular mechanisms are composed of multiple components, they should be more stably inherited than single gene and simple function. We should not be confused by simple, rare, and "big effect" mutations and believe that genes actually act alone. Second, the material flow and signal transduction among the molecular mechanisms are cellular processes, and together they form the modern operational track. Molecular mechanisms are not easily given up once created and the cellular processes are becoming more complex, supporting the sophistication of the complex life forms. Third, the complexity of the operational track necessitates a new round of compartmentalization.

In the final phase, the operational track has compartmentalized to form a ribo-proteo-protocell. Also, in this very phase, the pressure for a more stable informational track increases the needs for the involvement of DNA—a molecule for information. Therefore, compartmentalization has fulfilled two major roles (the third one from unicellular to multicellular organisms consumed nearly a couple billion years): to separate the operational track itself, making RNA and making protein, and to separate the operational track from the informational track, paving a way for the DNA-involved operations and the maturation of the modern informational track and the protocell.

This simple scenario has several implications. First, if life did go through the RNA world via the creation of the two tracks and the two cell types, ribo- and ribo-proteo-protocells, the evolving main-stream early life forms should be eukaryote-like rather than born from bacterium-like and achaeon-like. Only when DNA-based mechanisms were created, did bacteria and archaea start to thrive from the mother cell—the protocell as their "escapees" rather than "founders". Second, the informational track was only truly formed after DNA was actually chosen as the informational molecule. Third, DNA also became a component of the operational track as what RNA does in the RNA world. Fourth, other molecular

mechanisms may evolve with RNA-dominant mechanisms, such as the genetic code [10–13] and the spliceosome.

The two tracks and their intertwinements

The idea of a two-tracked biology needs not to be surprising. First, the two tracks represent two essential schools of biological disciplines in basic sciences. On the one hand, the informational track is largely handled by the fields of genetics (population genetics), evolutionary biology, and genomics. The tools to sequence genomes and to characterize proteomes in large-scale appear ready for their prime time. There have not been any obvious technical obstacles for sequencing everyone's genome and everything's genome for all life forms on earth. On the other hand, the operational track is largely handled by the fields of cell biology, molecular biology, and biochemistry. The tools for figuring out gene/protein function need to be speeded and scaled up, although some are already armed with powerful tools such as monoclonal antibody and transgenics. Second, the data structure for the informational track is largely linear and indirect, albeit sometimes statistical or phenotypically-defined, which is used for building the association between phenotype and genotype. In contrast, the data structure for the operational track is often non-linear and network-like, although the connections may have to be nailed one at a time based on many sporadic activities and efforts. To understand both types of data simultaneously, the more complex becomes the bottle-neck. For instance, cancer is a disease arising from failures in controlling a complex interplay of genetic (largely studied on the informational track) and environmental factors (largely studied on the operational track). Therefore cancer diagnosis, therapy, and prevention require a concerted effort from scientists working on both tracks, let alone dedicated physicians and modern equipments. Obviously, scientists working both tracks need to extend their knowledge bases vigorously; though "digging" independently on their own, "the tunnel of knowledge" needs to be shared and worked on by all members of scientific communities, i.e., to have a final closure with its shortest paths from the two opposite directions. That is to say, the goal of differentiating the two tracks is to find ways to unite them at the end.

As thoughts along the informational track have passed its infancy, the operational track has been lagging behind due to the complexity of its apparatuses and processes. In my opinion, our thoughts on the operational track are still evolving (**Figure 1**). For instance, we can certainly define some of the molecular mechanisms in the operational track in a thorough way but have hard time to characterize cellular processes in the track as they are homeostatic, multi-directional, and extremely intricate. If we believe that the "treasure map" of the informational track is the DNA sequence and its variations within some defined populations, the "treasure hunt" of the operational track has yet to begin—the illustrations for the landscapes need to be worked out first. A simple challenge, for example, is to

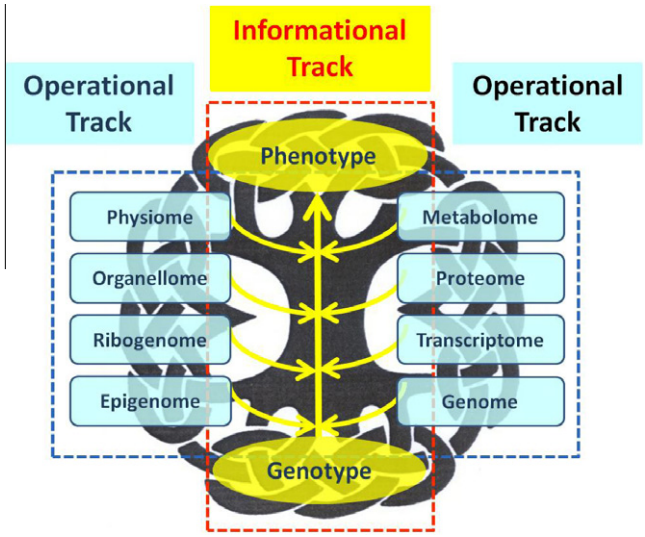

Figure 1 A scheme of the two-track biology in the background of a complex Chinese knob
This scheme emphasizes the pattern-rich and intertwining nature of modern biology and its stratification. The informational track is obviously overwhelmed by a complex network of the operational track that not only has multiple facets but also is rich in real experimental data. The interaction between the two tracks remains to be defined in details.

illustrate all the non-protein-coding genomic sequences including the "junk" and "dark" matters as well as many structural elements. Another obvious challenge is to define "epigenetic allele" (such as a variable CpG site as a broadly defined genotype or an environment-sensitive, incompletely inherited phenotype) where the inheritance is somewhat non-Darwinian. One of the tougher challenges is to distinguish functional genes from their revolutionary "transients" [14] and their true relics in polyploid plant genomes where function-based assays and gene paralogy are both definitely overwhelmed.

In the light of evolutionary paths

In the modern synthesis of biology, we should encourage reconciliation of the Darwinism and the Lamarckism frameworks, albeit in a sense beyond their classic definitions. Although the Darwinian framework suggests that evolution is composed of a contingent series of purposeless and unpredictable events, its footprints are traceable; if a species or lineage of multiple species does not choose, nature does so for it. In other words, organisms on earth as a whole may not have directional evolutionary paths, but they must have distinguishable characteristics that are definable by inheritance and new conception. For instance, the vertebrate lineage has experienced a long-term gain of increasing complexity but the arthropod lineage appeared not, especially in a macro-evolutionary scale. In addition, the unicellular organisms may have established robustness in their relatively compact and yet stable genomes by frequent gain-and-loss of genes. Versatility, variability, and other terms may be used to summarize the essence of other

Table 1 The evolutionary features of genomes in different lineages

Lineages	Genome* duplication	Horizontal transfer	Gene duplication
Mammals	No	Very rare	Rare
Reptiles	Very rare	Rare	Rare
Amphibians	Frequent	Rare	Rare
Fish	Very frequent	Rare	Rare
Insects	Very rare	Rare	Moderate
Plants	Very frequent	Rare	Frequent
Fungi	Frequent	Frequent	Moderate
Bacteria	No	Very frequent	Rare

Note: *Indicates genome duplication and/or polyploidy.

lineage-specific characteristics for plant and arthropod genomes; the former keeps duplicating its genome and the latter never does (**Table 1**) [15,16].

A typical clash of the two frameworks lies in the intertwining zone of the two tracks—defining the causative mechanism of mutations. Aside from the protein-coding property, every nucleotide in a genome is subjected to change; if not functionally selected, they are neutral and free to drift around the four-nucleotide circle. The mutations come from two classes of molecular mechanisms: DNA replication and repair. Replication is rather straightforward, albeit complex in eukaryotic organisms, where the chance to mutate for a given piece of DNA is random—everyone is created equal (well except telomeres so to speak). However, the repair mechanisms are not, especially one of them, the transcription-coupled DNA repair that repairs the transcribed strand rather than the non-transcribed. Furthermore, it repairs the upstream sequence more than the down-stream sequence of a transcript [17]. And even further, it involves multiple molecular mechanisms, including transcription regulation, CpG island density, and R-loop formation [18]. Nevertheless, the sequence signatures of such mechanism show a negative GC-content gradient that is correlated with gene expression [19–21]. In other words, the higher level a gene is expressed, the stronger the gradient there should be: does it not resemble the use-and-disuse principle of Lamarckism for such directional variations when involved in a beneficial function?

> "*A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.*"
> —*Max Planck, quoted in Thomas Kuhn, The Structure of Scientific Revolutions (1970 ed.), p. 150.*

References

[1] Yu J, Wong GK. Genome biology: the second modern synthesis. Genomics Proteomics Bioinformatics 2005;3:3–4.
[2] Yu J, Wang J, He F, Yang H. "Three kingdoms" to romance. Genomics Proteomics Bioinformatics 2003;1:1.
[3] Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK. Minimal introns are not "junk". Genome Res 2002;12:1185–9.
[4] Wang D, Yu J. Both size and GC-content of minimal introns are selected in human population. PLoS One 2011;6:e17945.
[5] Zhu J, He F, Wang D, Liu K, Huang D, Xiao J, et al. A novel role for minimal introns: routing mRNAs to the cytosol. PLoS One 2010;5:e10144.
[6] Tazi J, Bakkoura N, Stamm S. Alternative splicing and disease. Biochim Biophys Acta 2009;1792:14–26.
[7] Wang J, Zhang J, Li K, Zhao W, Cui Q. SpliceDisease database: linking RNA splicing and disease. Nucleic Acids Res 2012;40:D1055–9.
[8] Cech TR. The RNA worlds in context. Cold Spring Harb Perspect Biol 2011, doi: 10.1101/cshperspect.a006742.
[9] Liu W, Zhao Y, Cui P, Lin Q, Ding F, Xin C, et al. Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. Front Genet 2011;2:93.
[10] Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct 2012;7:13.
[11] Yu J. A content-centric organization of the genetic code. Genomics Proteomics Bioinformatics 2007;5:1–6.
[12] Xiao J, Yu J. A scenario on the stepwise evolution of the genetic code. Genomics Proteomics Bioinformatics 2007;5:143–51.
[13] Zhang Z, Yu J. On the organizational dynamics of the genetic code. Genomics Proteomics Bioinformatics 2011;9:1–9.
[14] Wang J, Zhang J, Li R, Zheng H, Li J, Zhang Y, et al. Evolutionary transients in the rice transcriptome. Genomics Proteomics Bioinformatics 2010;8:223–8.
[15] Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) of *Drosophila* and vertebrate genomes. BMC Evol Biol 2009;9:55.
[16] Yu J, Wong GK, Wang J, Yang H. Shotgun sequencing (SGS). In: Meyers RA, editor. Encyclopedia of Molecular Cell Biology and Molecular Medicine, vol 2, 2nd Edition. Wiley-VCH; 2005, p. 71–114.
[17] Yu J. Challenges to the common dogma. Genomics Proteomics Bioinformatics 2012;10:55–7.
[18] Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-Loop formation is a distinctive characteristic of unmethylated human cpg island promoters. Mol Cell 2012;45:814–25.
[19] Wong GK, Wang J, Passey DA, Yu J. Codon-usage gradients in Gramineae genomes. Genome Res 2002;12:851–6.
[20] Cui P, Ding F, Zhang L, Hu S, Yu J. Distinct contributions of replication and transcription to mutation rate variation of human genomes. Genomics Proteomics Bioinformatics 2012;10:4–10.
[21] Cui P, Lin Q, Ding F, Hu S, Yu J. The transcript-centric mutations in human genomes. Genomics Proteomics Bioinformatics 2012;10:11–22.