Divergent Transcription: An Interesting Tale

Terrence P McGarty
White Paper No 106
December, 2013

Divergent Transcription is the transcription of RNA segments not part of the normal process of gene to protein. Many of the RNA segments do not become proteins but do set up a network process enabling the creation of new genes. We examine this process herein.

Notice

This document represents the personal opinion of the author and is not meant to be in any way the offering of medical advice or otherwise. It represents solely an analysis by the author of certain data which is generally available. The author furthermore makes no representations that the data available in the referenced papers is free from error. The Author also does not represent in any manner or fashion that the documents and information contained herein can be used other than for expressing the opinions of the Author. Any use made and actions resulting directly or otherwise from any of the documents, information, analyses, or data or otherwise is the sole responsibility of the user and The Author expressly takes no liability for any direct or indirect losses, harm, damage or otherwise resulting from the use or reliance upon any of the Author's opinions as herein expressed. There is no representation by The Author, express or otherwise, that the materials contained herein are investment advice, business advice, legal advice, medical advice or in any way should be relied upon by anyone for any purpose. The Author does not provide any financial, investment, medical, legal or similar advice in this document or in its publications on any related Internet sites.

Contents

1 INTRODUCTION

In a recent paper by Wu and Sharp the authors discuss the concept of Divergent Transcription. Simply this is a study of all the transcription that generally goes nowhere but from time to time does go somewhere and in this case the development of new genes.

A few decades ago when we looked at the DNA world we thought of it in terms of the Dogma: DNA to RNA to Proteins. Then the proteins did things.

Then we found that we had about 3 Billion base pairs and only about 20,000 genes. That means that we used only about 1-2% of our bases and the other 98-99% were not really used, but not really. That unused DNA was actually used in bits and pieces. There was a ton on non-coding RNA floating all over.

Thus we might think that decades ago the cell was filled with some well-organized proteins, coming from a well-orchestrated RNA process of translation. It was like an airport in the US, with all the people, base pairs, and lining up with the TSA, the promoters, and moving thorough the scanners in order, each being read for proper ID, the scanner being the RNA polymerase and coming out as ticked passengers grouping at each waiting area for the assigned flight. Each waiting area was the proteins composed of the translated bases now nucleic acids. Organized, controlled, and no unverified interlopers.

But now we look again and it really appears like Penn Station in New York. Doors all over, no lining up, people going on Amtrak, LIRR, NJ Transit, subways, no waiting, no seats, no tickets, no security. Then there are vagrants checking out the trash bins, and dozens of other types just wandering and looking. Order may be there but not the type we see at say Newark. There are, if you will, big RNAs and little RNAs, RNAs destined to become proteins, namely passengers on some transport, but there are also just lots of little segments of RNA going nowhere. These are the equivalent of non-coding RNAs just wandering around, crowding up the floor, slowing down the passengers, and at time changing who goes where.

2 THE MODEL

Wu and Sharp conclude:

we propose that divergent transcription at promoters and enhancers results in changes of the transcribed DNA sequences that over evolutionary time drive new gene origination in the transcribed regions. Although the models proposed here are consistent with significant available data, systematic tests of these models await further advances such as in-depth characterization of additional genomes and experiments designed to test specific hypothesis. Over evolutionary times, genes formed through divergent transcription can be shuffled to other locations losing their evolutionary context. We envision future studies will uncover more functional surprises from divergent transcription, and illuminate how intergenic transcription is integrated into the cellular transcriptome.

Divergent Transcription is transcription that follows a different path than the organized transcription that we think of in a highly organized structure. As Seila et al stated:

Transcription initiation by RNA polymerase II (RNAPII) is thought to occur unidirectionally from most genes. Here, we present evidence of widespread divergent transcription at protein-encoding gene promoters. Transcription start site–associated RNAs (TSSa-RNAs) nonrandomly flank active promoters, with peaks of antisense and sense short RNAs at 250 nucleotides upstream and 50 nucleotides downstream of TSSs, respectively. Northern analysis shows that TSSa-RNAs are subsets of an RNA population 20 to 90 nucleotides in length. Promoter-associated RNAPII and H3K4-trimethylated histones, transcription initiation hallmarks, colocalize at sense and antisense TSSa-RNA positions; however, H3K79-dimethylated histones, characteristic of elongating RNAPII, are only present downstream of TSSs. These results suggest that divergent transcription over short distances is common for active promoters and may help promoter regions maintain a state poised for subsequent regulation.

As Wu and Sharp recount the classic transcription we use their description:

In the textbook model of a eukaryotic promoter, the directionality is set by the arrangement of an upstream cis-element region followed by a core promoter. The cis-elements are bound by sequence-specific transcription factors, whereas the core promoter is bound by TATA-binding protein (TBP) and other factors that recruit the core transcription machinery.

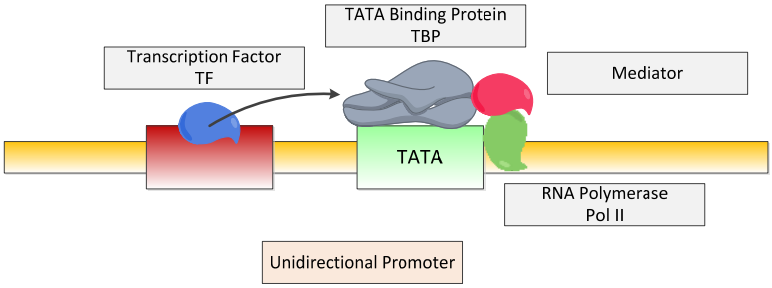
Most mammalian promoters lack a TATA element (TATA-less) and are CpG rich. For these promoters, TBP is recruited through sequence-specific transcription factors such as Sp1 that bind CpG-rich sequences and components of the TFIID complex that have little sequence specificity.

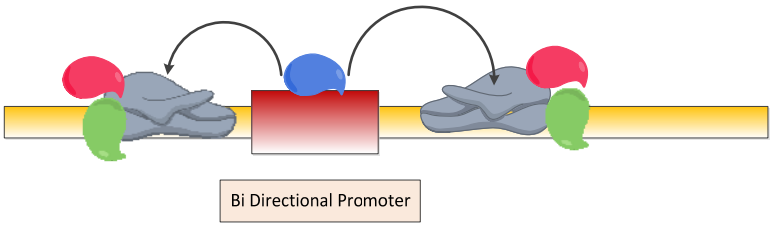
Thus, in the absence of strong TATA elements such as for CpG island promoters, TBP-complexes are recruited on both sides of the transcription factors to form preinitiation complexes in both orientations.

This model is supported by the observation that divergent transcription occurs at most promoters that are associated with CpG islands in mammals, whereas promoters with TATA elements in mammals and worm are associated with unidirectional transcription

We demonstrate the two concepts below using a modified graphic from Wu and Sharp. We show the TATA binding site on the gene and we show the TBP, the TATA binding protein and a mediator and ultimately the RNA Pol II. This is a classic unidirectional process moving across the exons and generating mRNA which is then cleaned and changed to a protein. The related cDNA do not show any of this underlying complexity.

Now below this is a second process, but now we show both forward and backward transcription. This requires a bi-directional promoter which Wu and Sharp discuss.





Wu and Sharp then argue that the model can be characterized by the system below. We have modified their graphic so that we may take a small step further. The Figure below depicts the four processes they consider:

1. Transcription: This is the classic transcription process of taking DNA and changing it to RNA, usually an mRNA.

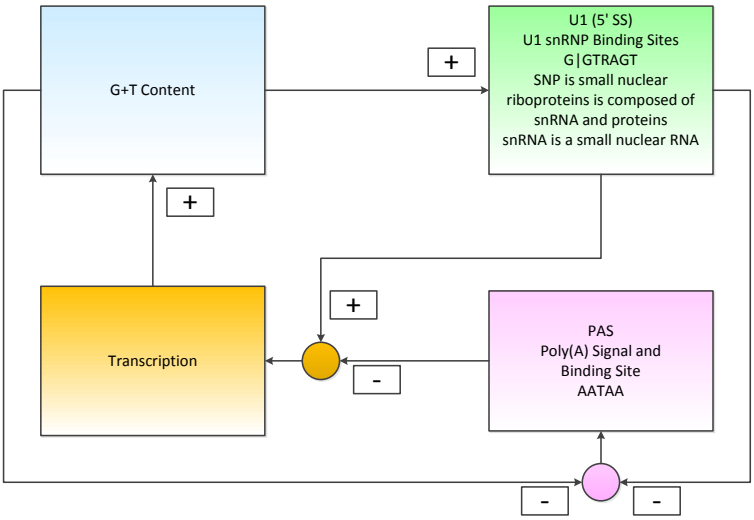
2. G+T Content: This is the G and T content of the intron and the propensity for mutations to occur in that area and thus setting up a region for the introduction of new gene type sequences.

3. U1 Process: There are small nuclear RNAs used to splice RNA segments together and these are called spliceosomes. One of them us the U1 snRNA. As the mRNA segments are produced

they get spliced together by these nuclear RNA segments. They are powerful elements found in the nucleus.

4. PAS Process: The poly(A) is described as, from Baynes & Dominiczak, pp 430-432, as: *At the 3' end of all eukaryotic mRNAs (with the exception of histone mRNAs), a polyadenosine track is added, termed the polyA tail. The adenosine residues are not encoded by the DNA but instead are added by the action of poly(A) polymerase using ATP as a substrate. This polyA tail is frequently >250 nucleotides in length. Although it is still susceptible to the action of exo-RNases, the presence of the polyA tail significantly increases the lifetime of mRNA. The presence of the polyA tail has historically been used to isolate mRNA from eukaryotic cells.*

We now combine these elements into the Wu and Sharp dynamic, as modified, below:



We can then represent this model by the meta-equation below:

$$\frac{dGT(t)}{dt} = \alpha T(t)$$
$$\frac{dT(t)}{dt} = \beta U(t) - \eta PAS(t)$$
$$\frac{dU(t)}{dt} = \mu GT(t)$$
$$\frac{dPAS(t)}{dt} = -\theta GT(t) - \nu U(t)$$

or

$$\frac{dx(t)}{dt} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} x(t)$$

Here we represent GT, T, U and PAS as some measure of each of the four processes represented in the diagram. Admittedly this is at best an ad hoc representation but it does demonstrate that indeed we have some form of dynamical system and in turn this system depending on whatever the constants are can become an unstable and ever growing process.

3 NEW GENES

Out of this process Wu and Sharp argue that new genes can be born. This is an ingenious and compelling argument. The time scale for such a development is not specified but perhaps it may be intuited. Also actual changes have yet to be fully observed from beginning to end. Yet the pieces are logically consistent and are all supported by the evidence.

First a brief summary of the splicesome (from Baynes and Dominiczak, pp 430-432)

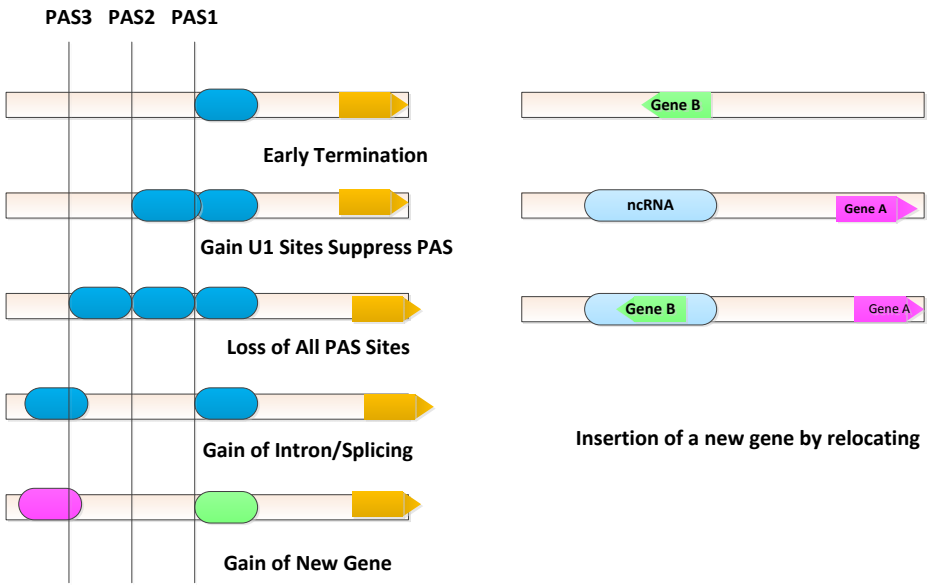
In the more complicated posttranscriptional processing of eukaryotic mRNAs, sequences called introns (intravening sequences) are removed from the primary transcript and the remaining segments, termed exons (expressed sequences), are ligated to form a functional RNA.

This process involves a large complex of proteins and auxiliary RNAs called small nuclear RNAs (snRNAs), which interact to form a spliceosome. The function of the five snRNAs (U1, U2, U4, U5, U6) in the spliceosome is to help position reacting groups within the substrate mRNA molecule, so that the introns can be removed and the appropriate exons can be spliced together precisely. The snRNAs accomplish this task by binding, through base-pairing interactions, with the sites on the mRNA that represent intron/exon boundaries. Accompanying protein factors are responsible for holding the reacting components together to facilitate the reaction.

We summarize the U below:

snRNA	*Size*	*Function*
U1	165 nt	Binds the 5' exon/intron boundary
U2	185 nt	Binds the branch site on the intron
U4	145 nt	Helps assemble the spliceosome
U5	116 nt	Binds the 3' intron/exon boundary
U6	106 nt	Displaces U1 after first rearrangement

We explain how this may work in the Figure below adapted from Wu and Sharp. There is on the left a progression of changes in a segment of DNA which would normally read left to right with the inclusion of a new segment from right to left. The PAS sites, three as shown in the Figure below, are covered by RNA segments ultimately allowing the creation of an Exon and Intron. The process is further elucidated on the right where are gene is putatively relocated from one chromosome to another or even just duplicated.



Let us use Wu and Sharp's text and go through the argument. They proceed as follows:

One consequence of transcription is that it can cause mutations, especially on the coding (nontranscribed) strand.

During transcription, transient R loops can be formed behind the transcribing RNA polymerase II, exposing the coding strand as single-stranded DNA, whereas the noncoding strand is base paired with and thus protected by the nascent RNA.

The lack of splicing signals in the divergent transcript also makes it more vulnerable to R loop formation, as splicing factors have been implicated in suppressing R loop formation.

In addition, divergent transcription generates negative supercoiling at promoters, which facilitates DNA unwinding and promotes R loop formation.

As a consequence of R loop formation, the single-stranded coding strand is vulnerable to mutagenic processes, such as cleavage, deamination, and depurination. Genomics studies have shown that during mammalian evolution, transcribed regions accumulate G and T bases on the coding strand, relative to the noncoding strand or nontranscribed regions.

Evidence suggests that such strand bias may result from passive effects of deamination, transcription-coupled repair, and somatic hypermutation pathways in germ cell-transcribed genes, in the absence of selection.

Accumulation of G and T content on the coding strand will strengthen the U1-PAS axis.

A-rich sequences such as PAS (AATAAA) are likely to be lost when the genomic DNA accumulates G and T.

In contrast, G+T-rich sequences, such as U1 snRNP-binding sites (e.g., resembling 50 splice sites, G|GTAAGT and G|GTGAGT), are likely to emerge in these regions. Since promoter-proximal PAS reduces transcriptional activity, the loss of PAS and gain of U1 sites should contribute to lengthening of the transcribed region as well as its more robust transcription.

The gain of U1 sites could also enhance transcription by recruiting basal transcription initiation factors or elongation factors.

Therefore a positive feedback loop is formed: active transcription causes the coding strand to accumulate sequence changes favoring higher transcription activity.

As noted above, strengthening of the U1-PAS axis also favors extension of the transcribed region. Being longer gives the transcript several advantages: by chance longer RNAs are more likely to contain additional splicing signals such as a 30 splice site to become spliced, or binding sites for splicing-independent nuclear export factors, thus escaping nuclear exosome degradation by packaging and exporting to cytoplasm .

Longer RNAs are also more likely to carry an open reading frame, either generated de novo or by incorporation of gene remnants.

Once in the cytoplasm, the RNA should at some frequency be translated into short polypeptides due to widespread translational activity.

Some of the polypeptides may provide advantage to the organism and become fixed in the population, thereby forming a new gene.

Thus we have seen a mechanism for new gene creation and insertion.

4 OBSERVATIONS

These are a very powerful set if insights and observations. They have significant conclusions as has been articulated by those in Sharp's Lab. The metaphor of a train station with wandering fragments of often "useless" RNA has certain merit. However all too often those fragments are not useless but have ways of interfering and disrupting the normal progress of cellular dynamics.

We now pose a few observation which may have some merit.

1. Somatic vs Germline: These changes seem to be mitotic in nature and thus are reflected in somatic cells. What is the impact in meiosis and germ line cells? Namely can these mutations be carried forward and be selected out in subsequent generations? Or is this process one almost exclusively found in somatic cells and thus may be causes for such diseases as the cancers? I could not find a clear path to follow here.

2. Causation: What causes some of these processes. Many if not most of the links are presented and explained but ultimate causality is missing.

3. Frequency: How frequently do these changes occur? Are they rare or common and at what rate do they occur? What are the overall temporal dynamics of these processes. Can we examine genomes and ascertain where they might occur. We all too often just skip over the Introns, focusing on the Exons and their resultant expression. There also are many regions of the Exons that are not expressed, and are they part of this phenomenon as well?

4. Reaction Dynamics: The actual reaction dynamics could possibly be explained and modelled. We have presented a meta model solely for the visualization of what may happen. It is expected that the model is most likely non-linear and more complex. In fact the actual metrics being measured and modelled are still in question. However not withstanding that we can envision a dynamic model exhibiting not only stability issues but also oscilliatory effects.

5. Methylation and Epigenetic Factors: Clearly the CpG islands play an important factor. Methylation has become a significant area of study over the past decade and the processes described herein rely on many of these CpG islands as well. Is methylation a competing process, an allied process, a controlling or mediating process?

6. What are all these RNA fragments doing?: Ultimately we find that a cell may have not only well understood Dogma based proteins and pathways but also a mass of disconnected non coding RNA spinning about in the nucleus and throughout the cell. Thus we ask; what do these snippets do? Are they just wanderers going nowhere and possibly just bumping into those going somewhere or are the truly entities which have predictable effects on pathways? Are they noise or an aberrant signal?

This is a very compelling paper and it presents in an elegant manner the results of the efforts to date. This effort demands to be followed and examined in detail as it progresses.

5 REFERENCES

1. Baynes Dominiczak: Medical Biochemistry 3E, Mosby (New York) 2013.

2. Seila, A., et al, Divergent Transcription from Active Promoters, Science VOL 322 19 December 2008 1849.

3. Wu, X., P. Sharp, Divergent Transcription: A Driving Force for New Gene Origination? Cell 155, November 21, 2013.