Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 241 (2006) 205–215
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

# Systematic discovery of the grammar of translational inhibition by RNA hairpins

Clark Jeffries[a,*,1], Diana O. Perkins[b,2], Michael Jarstfer[c,3]

[a]Renaissance Computing Institute and School of Pharmacy, Campus Box 7360, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360, USA
[b]Department of Psychiatry, University of North Carolina at Chapel Hill, USA
[c]School of Pharmacy, University of North Carolina at Chapel Hill, USA

Received 19 August 2005; received in revised form 16 October 2005; accepted 14 November 2005
Available online 5 January 2006

**Abstract**

Recent discovery of gene expression mechanisms has propelled molecular genetics to a state of rapid development, a state likely to persist due to continuing advances in understanding control systems of fundamental cellular processes. An algorithm for that advancement starts in this paper with a gene of interest and a characteristic function of that gene. The set of all genes with counteracting function is identified by pathway searches. Also associated with the first gene is the set of the genes which byproducts of its transcription might downregulate, identified relative to searches involving sequence alignments. Our focus is the intersection of the counteracting gene set and the downregulated gene set. The result is hypothesis generation. Examples of and predictions from this approach are given in the context of apoptosis. Also discussed is application of the algorithm to rational drug design from a new development platform.


*Keywords:* Gene regulation; Translational inhibition; Hairpin; miRNA; Alu

## 1. Introduction

This paper addresses the problem of discovering from databases the organizational principles of regulation of gene expression. Text-based searches for gene relationships inherently must be artificially terminated. By contrast, we present a hybrid algorithm of text and biochemical alignment searches that inherently terminates. The product of the algorithm is a gene organization hypothesis.

The hybrid search works, we assert, because some genes are not just "selfish" (seeking above all else their own survival and proliferation in future generations); they are also "devious" in the sense that their transcription byproducts automatically inhibit translation of other genes with counteracting functions (so they quietly undermine opposing genes). In terms of control theory in cell biology, this would be an example of a positive feedforward mechanism, resulting in a "fervid" response as described in Sinclair and Challis (1993). The mechanisms include microRNA (miRNA) hairpins and also the main emphasis of this paper: larger regulatory hairpins from introns. Three well-formed hairpins from one intron of pro-apoptosis gene PAWR (GeneID 5074) are shown in Fig. 3. The first two have stems that resemble Alu repeat consensus sequence on one side and a sequence that is approximately the reverse complement (revcom) of an Alu, that is, an Alurc, on the other. The third hairpin, however,

*Corresponding author. Tel.: +1 919 260 6576; fax: +1 919 843 7624.
*E-mail addresses:* clark_jeffries@med.unc.edu (C. Jeffries), diana_perkins@med.unc.edu (D.O. Perkins), jarstfer@email.unc.edu (M. Jarstfer).

[1]Research supported by 5-P20-RR020751-01-02 from the National Institutes of Health/National Center for Research Resources. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health/National Center for Research Resources.
[2]Research supported by Per USP#SMH01905-01 North Carolina Foundation of Hope, Anonymous donor administered through the San Francisco Foundation.
[3]Elsa U. Pardee Foundation and NSF MCB 0446019.

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.11.027

comes from an intronic portion of PAWR that is first Alu-like, then about 100 other nucleotides (nts), then about 240 nt of an Alurc-like pattern (a full Alu or Alurc is about 285 nt). Thus the stem of the third hairpin in Fig. 3 is not fully Alu/Alurc, but is restricted to only a part of an Alu/Alurc double-stranded RNA (dsRNA).

The predicted hairpins provide feedstock dsRNA for an miRNA pathway that eventually causes inhibition of other genes, as considered in Hannon (2002). In particular, the anti-apoptosis gene BIRC4 (331) is expressed at the same time as PAWR in certain cancer cell lines, as discussed below. According to the output of the hybrid algorithm, PAWR can be hypothesized to inhibit BIRC4 by means of an interference mechanism (and possibly other mechanisms). Thus once apoptosis begins, this double mechanism including PAWR expression and automatic BIRC4 repression might drive the cell efficiently and irreversibly to disintegration (hence fervid, positive feedforward control). Such deviousness by some genes might suggest a logical module that helps researchers to organize concepts of gene expression.

That there is a need to find organizational principles for gene expression including various types of control agents is self-evident. The purpose of this paper is progress in that direction with emphasis on the miRNA pathway applied primarily to larger, non-miRNA hairpins (miRNA hairpins contain ~70 to ~120 nt). Again, a fundamental assumption is that portions of dsRNA of various sources might be exported to the cytosol through an approximately common pathway and participate in translational inhibition of target mRNAs (Hannon, 2002). The dsRNA sources can include, intronic single-stranded RNA (ssRNA) that folds to make dsRNAs, or long dsRNAs generated by bidirectional transcription of repetitive sequences from adjacent promoters (Grewal and Rice, 2004). The subsequent miRNA pathway might lead to binding by chemical affinity to target mRNA and then repression of target gene expression by: inhibition of ribosomal processing; hybridization with and accelerated degradation of mRNA in a processing body (PB); or heterochromatin alteration (Murchison and Hannon, 2004; Pillai et al., 2005; Grewal and Rice, 2004). In this paper the target sequence is assumed to lie in the 3′UTR, but also known are related mechanisms including instances of a functional target in the 5′UTR (Jin et al., 2004) and transcription factor binding site in an intron (Xu et al., 1999; St. Clair et al., 2002). The full range of types of sequence specific pre-mRNA and mRNA targeting is no doubt very large. In addition is the recent finding that half or even more of the human genome might be transcribed, antisense as well as sense; this is a fraction that dwarfs the one or two percent of the genome that contains conventional genes. For a survey of ideas and current research, see Mattick (2005) and other articles on non-coding RNA (ncRNA) in the same special issue of *Science*. Since there are many small and some large sequences in the genome with nearby revcom sequences, there is potentially a vast supply of dsRNAs that could participate in regulatory systems. At any rate, the bipartite graph of such relationships (regulated genes versus blocking agents) would ideally be part of an organized theory or grammar of combinatorial control, the description of which could proceed in terms of its syntax and semantics. By "syntax" we mean sets of rules starting with the list of permitted combinations of neucleotides (nts) that make the allowed sequences (words) engaged as blocking agents in translational inhibition control. In the case of miRNAs, this syntactic description might, for example, include the set of miRNAs and their mature regions with "seeds" (short nt sequences that hybridize with target sequences in an mRNA; Lewis et al., 2003). The next level would include the rules for permitted tables of genes versus control agents (sentences in a relational grammar). Higher levels that make gradual or abrupt control of gene expression possible can be imagined. By "semantics" we define the union of all levels that somehow endows meaning (life) to syntactically correct sets of symbols.

In gene expression theory just as in computer science (Aho et al., 1986), finding formal yet useful definitions of semantics and meaning is difficult, necessitating human intuition and understanding. There are similarities but also extreme differences between gene expression systems and computer systems; for a given set of tasks, humans usually make computers to be as simple and efficient as possible, while nature often allows redundancy, tolerates errors (even fatal), and seems indifferent to inefficiency (e.g. retention of long-disused components and steady production of certain molecules most of which are immediately degraded). The most important human design goals include simplicity and near-optimality, while the goals of nature emphasize robustness and the capacity for self-organization. Still, synthetic and natural control systems are again comparable in that humans believe that they can both be profitably described using two concepts enabled by syntactical structure: *module* and *hierarchy*. Modules are logically interchangeable units, for example, all genes (including isoforms) that produce proteins that are biochemically interchangeable. A hierarchy (of relational syntax) is the organizational pyramid of intra- and intercellular communication that makes multicellular life possible. If gene expression organization is modular and hierarchical, then signaling among genes should occur mainly between genes in the same organizational unit; module-to-module signals likely are relatively infrequent and possibly more important (silencing them is especially problematic to life).

The architecture of gene expression is today studied by use of powerful online search engines, of which more than 100 exist. Some of the engines used in the preparation of this paper are: general resources like NCBI (http://www.ncbi.nlm.nih.gov/), GeneCards (http://www.genecards.org/) (Rebhan et al., 1997), and Ensembl (http://www.ensembl.org/index.html) (EMBL-EBI/Sanger Institute); relationship engines based upon text searches such as