Models of Recombination in Ciliates

Paul Sant* Martyn Amos†

Abstract

In this chapter we describe how certain single-celled organisms "compute" by unscrambling their genetic material. We review extant models of this process, and discuss their implications.

To appear in Computation in Cells and Tissues: Perspectives and Tools of Thought, Paton, R., Bolouri, H., Holcombe, M., Parish, J. H., Tateson, R. (Eds.); Collected volume in the Natural Computing Series, Springer-Verlag, ISBN 3-540-00358-4

1 Introduction

Ciliate is a term applied to any member of a group of around 10,000 different types of single-celled organism that are characterised by two unique features: the possession of hair-like *cilia* for movement, and the presence of two *nuclei* instead of the usual one. One nucleus (the *micronucleus*) is used for sexual exchange of DNA, and the other (the *macronucleus*) is responsible for cell growth and proliferation. Crucially, the micronucleus contains an "encoded" description of the working macronucleus, which is decoded during development. This encoding "scrambles" functional gene elements by both the permutation of *coding* sequences and the inclusion of *non*-coding sequences. A picture of the ciliate *Oxytricha nova* is shown in Figure 1. During development, ciliates reorganise the material in the micronucleus by removing non-coding sequences and placing coding sequences in the correct order. This "unscrambling" may be interpreted as a computational process during which up to 95% of the original sequence is discarded.

The exact mechanism by which genes are unscrambled is not yet fully understood. We first describe experimental observations that have at least suggested possible mechanisms. We then describe two different models of the process. We conclude with a discussion of the computational and biological implications of this work.

*Department of Computer Science, King's College, London WC2R 2LS, UK. E-mail: sant@dcs.kcl.ac.uk

†School of Biological Sciences and Department of Computer Science, University of Liverpool, Liverpool L69 7ZB, UK

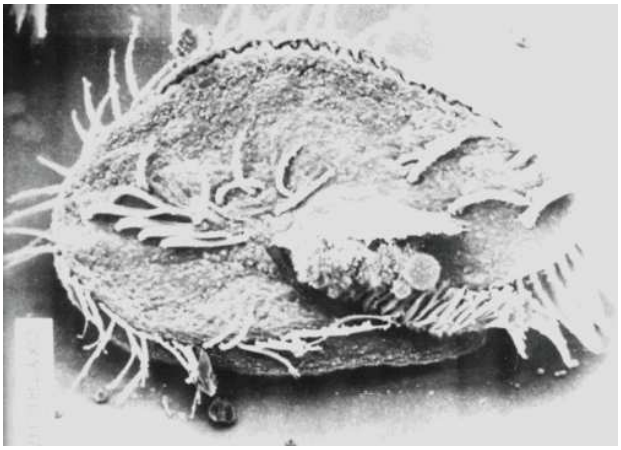

Figure 1: *Oxytricha nova* (Picture courtesy of D.M. Prescott)

2 Biological background

The macronucleus consists of millions of short DNA molecules that are "snipped out" of or *excised* from the micronucleus. Each macronuclear molecule corresponds to an individual gene, varying in size between 400 b.p. (*base pairs*) and 15000 b.p. (the average size is 2000 b.p.) The macronuclear DNA forms a very small proportion of the micronucleus, as up to 95% of micronuclear DNA forms intergenic "spacers", and is eliminated during genetic excision (that is, only 5% of the micronucleus is coding DNA).

During macronuclear development, individual genes are excised from the micronucleus, and are, after the completion of this process, present as individual short molecules in the macronucleus. The formation of the macronucleus triggers the process of *transcription*, whereby genes are read and transcribed into RNA, the "blueprint" for proteins.

2.1 IESs and MDSs

The process of decoding *individual* gene structures is therefore what interests us here. In the simplest case, micronuclear sequences contain many short, noncoding sequences called *internal eliminated sequences*, or IESs. These are short, AT-rich sequences, and, as their name suggests, are removed from genes and destroyed during the development of the macronucleus. They separate the micronuclear version of a gene into *macronuclear destined sequences*, or MDSs (Figure 2(a)). When IESs are removed, the MDSs making up a gene are "glued together", or *ligated* to form the functional macronuclear sequence. In the simplest case, IESs are bordered on either side by pairs of direct repeat sequences of between 2 and 7 b.p. in the ends of the adjacent MDSs (Figure 2(b)).

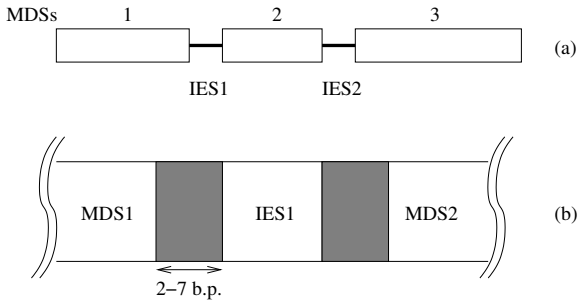

Figure 2: (a) Schematic representation of interruption of MDSs by IESs (b) Repeat sequences flanking an IES

The removal of IESs is a two-stage process; first, the correct pair of repeat sequences bordering the IES must be identified, then the IES must be cut out and the two MDSs ligated together. The molecular mechanisms by which these problems are solved are still poorly-understood. It is thought that, in some cases, staggered cuts are made in the DNA to create "sticky" versions of the repeats plus the ends of the MDSs. The repeats then align and the MDSs are ligated (this process is known as *recombination*). The excision of IESs and splicing together of MDSs can require up to several hundred thousand incredibly precise recombination operations to be carried out over a timescale of a few hours [5].

2.2 Scrambled genes

In some organisms, the gene construction problem is complicated by the "scrambling" of MDSs within a particular gene. In this situation, the correct arrangement of MDSs is present in a permuted form in the micronuclear DNA. For example, the actin I gene in *Oxytricha nova* is made up of 9 MDSs and 8 IESs, the MDSs being present in the micronucleus in the order 3-4-6-5-7-9-2-1-8 [7]. During the development of the macronucleus, the MDSs making up this gene are rearranged into the correct order at the *same* time as IES excision.

Scrambling is often *further* complicated by the fact that some MDSs may be *inverted*. Inverted MDSs are the *reverse complement* of the correct sequence (e.g., the inverse of the sequence CGT is derived as follows; reverse the sequence, giving TGC, then take the Watson-Crick complement, giving a reverse complement of ACG).

2.3 Fundamental questions

Despite exhibiting these seemingly bizarre phenomena, ciliates are remarkably successful organisms. The range of DNA manipulation and re-organization operations they perform has clearly been acquired during millenia of evolution. However, several fundamental questions remain: (1) what are the underlying molecular mechanisms of gene reconstruction, and how did they evolve? (2) how do ciliates "know" which sequences to remove and which to keep?

Concerning the first question, Prescott proposes [5] that the "compression" of a working nucleus from a larger predecessor is part of a strategy to produce a "streamlined" nucleus in which "every sequence counts" (i.e., useless DNA is not present). This efficiency may be further enhanced by the dispersal of genes into individual molecules, rather than their being joined into chromosomes.

In addition, the significance of the existence of MDSs and IESs is still largely a mystery. Subdivision, excision and re-arrangement of their DNA appears to convey no evolutionary advantage upon these organisms.

We may, perhaps, have more success in attempting to answer the second question: how are genes successfully reassembled from an encoded version? In the rest of this paper we address this question from a computational perspective, and describe two extant models that describe the rearrangement process.

3 Models of gene construction

We now present a review of two models that attempt to shed light on the process of macronuclear gene assembly. In [4] Landweber and Kari propose an initial model that was subsequently enhanced in [3], where a formal model for gene rearrangement was presented.

Landweber and Kari propose two main operations that model the process of intra and intermolecular recombination. These can be used to unscramble a micronuclear gene to form a functional copy of the gene in the macronucleus. Both of these operations are based on the concept of repeat sequences "guiding" the unscrambling process.

- Operation 1: $\{uxwxv\} => \{uxv, \bullet wx\}$.
- Operation 2: $\{uxv, \bullet wx\} => \{uxwxv\}$.

The first operation takes as input a single linear DNA strand containing two copies of a repeat sequence x. The operation then "folds" the linear strand into a loop, thus aligning the copies of x. The operation then "cuts" the strands after the first copy of x and before the second copy of x, creating three strands (ux, wx and v). The operation finally recombines ux and v

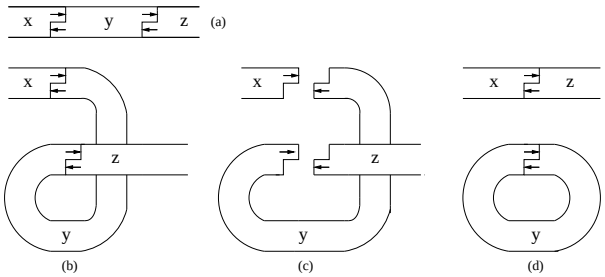

Figure 3: Excision

and wx forms a circular string. The output of the operation is therefore a linear string uxv and a circular string $\bullet wx$. This operation mimics the excision of an IES that occurs between two MDS's that are in the correct (i.e., unscrambled) order. In ciliates the IES is excised as a circular molecule and the two MDS's are "sewn" together to make a single larger MDS.

The second operation takes as input a single linear strand and a separate circular strand. The operation takes two inputs and creates a single linear strand. This allows the insertion of the circular strand within the smaller linear strand and mimics intermolecular recombination.

Intermolecular recombination takes as input a circular molecule and a linear molecule, each of which contain a single copy of a direct repeat sequence x. The result of intermolecular recombination is a single linear molecule containing two copies of the repeat.

A subsequent model, due to Rozenberg *et al.* [2] builds on the work of Landweber and Kari. Repeat sequences are modelled as pointers, connecting one MDS with another. In Figure 3, the two MDSs x and z are separated by the IES y. The pointers are represented by arrows, $\rightarrow$ being the outgoing pointer and $\leftarrow$ being the incoming pointer.

The first operation is the simplest, and is referred to as *loop, direct-repeat excision.* This operation deals with the situation depicted in Figure 3, where two MDSs in the correct (i.e., unscrambled) order are separated by an IES.

The operation proceeds as follows. The strand is folded into a loop with the pointers aligned, and then staggered cuts are made (Figure 3(b) and (c)). The pointers connecting the MDSs then join them together, while the IES self-anneals to yield two molecules, one linear and the other circular (Figure 3(d)).

The second operation is known as *hairpin, inverted repeat excision*, and is used in the situation containing inverted sequences (sequence y in Figure 4(a)). The molecule folds into a hairpin structure (Figure 4(b)), cuts are made (Figure 4(c)) and the inverted sequence is re-inserted (Figure 4(d)),

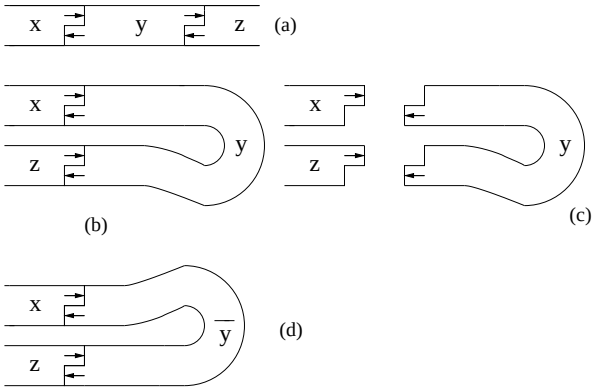

Figure 4: Inversion

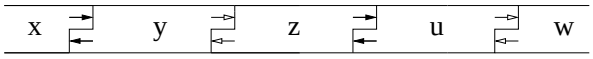

Figure 5: Alternating direct repeat pattern

yielding a single molecule.

The third and final operation in the Rozenberg *et al.* model is *double-loop, alternating direct repeat excision/reinsertion.* This operation is applicable in situations where molecules have an alternating direct repeat pointer pattern, as depicted in Figure 5.

The molecule folds into a double loop, with one loop being aligned by one pointer pair, and the other loop being aligned with the second pointer pair (Figure 6(a)). Cuts are then made ((Figure 6(b)), and the sections representing u and y exchange positions by a process of reinsertion (Figure 6(c)).

The three operations presented are *intramolecular* (as opposed to the *intermolecular* operations of Landweber and Kari) in that a molecule "reacts with" (i.e., folds on) itself. The process by which gene assembly takes place using these operations and the computational properties of the system are discussed in detail in [1].

Although the model presented may appear rather abstract, it has been successfully applied to the assembly of real genes, including the *Actin I* gene of *Urostyla grandis* and *Engelmanniella mobilis*, the gene encoding α telomere binding protein in several stichotrich species and assembly of the gene encoding DNA polymerase α in *Sterkiella nova.* Descriptions of these applications are presented in [6].

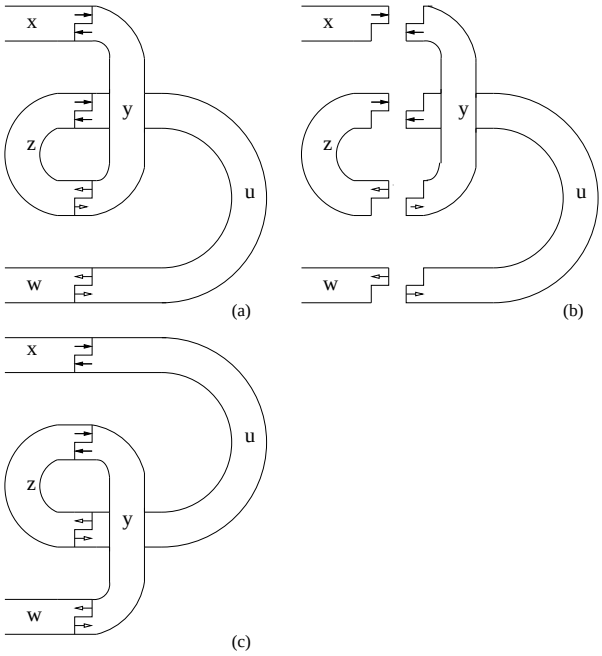

Figure 6: Excision/inversion

4 Discussion

In this paper we have described two models for the assembly of genes in ciliates. Although the fundamental molecular mechanisms underlying the operations within these models are still not well-understood, they do suggest possible areas of experimental enquiry. Looking further ahead, it may well be that in the future these mechanisms may even be exploited by using ciliates as prototype cellular computers.

References

[1] A. Ehrenfeucht, I. Petre, D.M. Prescott, and G. Rozenberg. Formal systems for gene assembly in ciliates. *Theoretical Computer Science*, 2002. In press.

[2] Andrzej Ehrenfeucht, Ion Petre, David M. Precott, and Grzegorz Rozenberg. Universal and simple operations for gene assembly in ciliates. In C. Martin-Vide and V. Mitrana, editors, *Where Mathematics, Computer Science, Linguistics and Biology Meet*, pages 329–342. Kluwer Academic Publishers, 2001.

[3] Lila Kari and Laura F. Landweber. Computational power of gene rearrangement. In Erik Winfree and David K. Gifford, editors, *Proceedings 5th DIMACS Workshop on DNA Based Computers, held at the Massachusetts Institute of Technology, Cambridge, MA, USA June 14 - June 15, 1999*, pages 207–216. American Mathematical Society, 1999.

[4] L.F. Landweber and L. Kari. The evolution of cellular computing: Nature's solution to a computational problem. *BioSystems*, 52(1–3):3–13, 1999.

[5] David M. Prescott. Invention and mystery in hypotrich DNA. *Journal of Eukaryotic Microbiology*, 45(6):575–581, 1998.

[6] David M. Prescott and Grzegorz Rozenberg. Encrypted genes and their assembly in ciliates. In Martyn Amos, editor, *Cellular Computing*. Oxford University Press, New York. To appear.

[7] D.M. Prescott and A.F. Greslin. Scrambled actin I gene in the micronucleus of *oxytricha nova*. *Developmental Genetics*, 13:66–74, 1992.