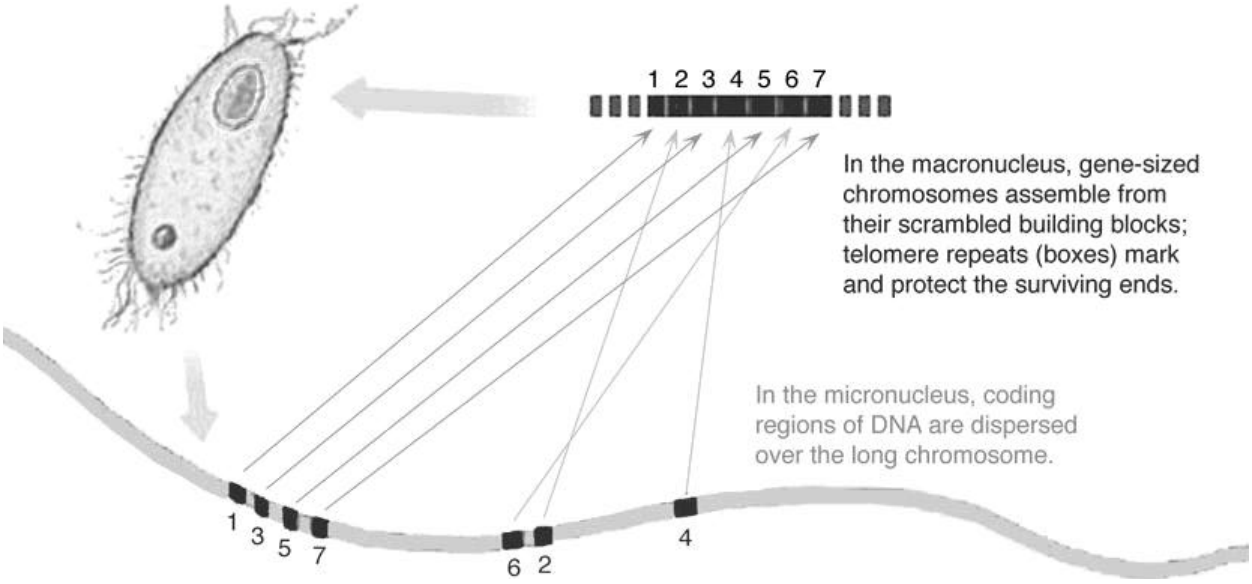

Figure 2: Overview of gene unscrambling. Dispersed coding MDSs 1-7 reassemble during macronuclear development to form the functional gene copy (top), complete with telomere addition to mark and protect both ends of the gene.

This process of unscrambling bears a remarkable resemblance to the DNA algorithm Adleman (1994) used to solve a seven-city instance of the Directed Hamiltonian Path problem. Section 1.4 introduces a formal model of gene unscrambling. (Adleman's algorithm involves the use of edge-encoding sequences as splints to connect city-encoding sequences, allowing the formation of all possible paths through the graph (Figure 1). Afterwards, a screening process eliminates the paths that are not Hamiltonian, i.e. ones which either skip a city, enter a city twice, or do not start and end in the correct origin and final destinations.)

The developing ciliate macronuclear 'computer' (Figures 2-3) apparently relies on the information contained in short repeat sequences to act as guides in a series of homologous recombination events (Table 1). These guide-sequences provide the splints analogous to the edges in Adleman's graph, and the process of recombination results in linking the protein-encoding segments (MDSs, or 'cities') that belong next to each other in the final protein coding sequence ('Hamiltonian path'). As such, the unscrambling of sequences that encode DNA polymerase α accomplishes an astounding feat of cellular computation, especially as 50-city Hamiltonian path problems are often considered hard problems in computer science and present a formidable challenge to a biological computer. Other structural components of the ciliate chromatin presumably play a significant role, but the exact details of the mechanism are still unknown.

1.2. The path towards unscrambling

Typical IES excision in ciliates involves the removal of short (14 - ~600 bp) A–T rich sequences, often released as circular DNA molecules (Tausta and Klobutcher 1989). The choice of which sequences to remove appears to be minimally 'guided' by recombination between direct repeats of only 2 to 14 base pairs (Table 1).

Unscrambling is a particular type of IES removal in which the order of the MDSs in the MIC is often radically different from that in the MAC. For example, in the micronuclear genome of *Oxytricha nova*, the MDSs of α–telomere binding protein (α–TP) are arranged in the cryptic order 1-3-5-7-9-11-2-4-6-8-10-12-13-14 relative to their position in the 'clear' macronuclear sequence 1-2-3-4-5-6-7-8-9-10-11-12-13-14. This particular arrangement predicts a spiral mechanism in the path of unscrambling which links odd and even segments in order (Figure 4; Mitcham et al. 1992).

Homologous recombination between identical short sequences at appropriate MDS–IES junctions is implicated in the mechanism of gene unscrambling, as it could simultaneously remove the

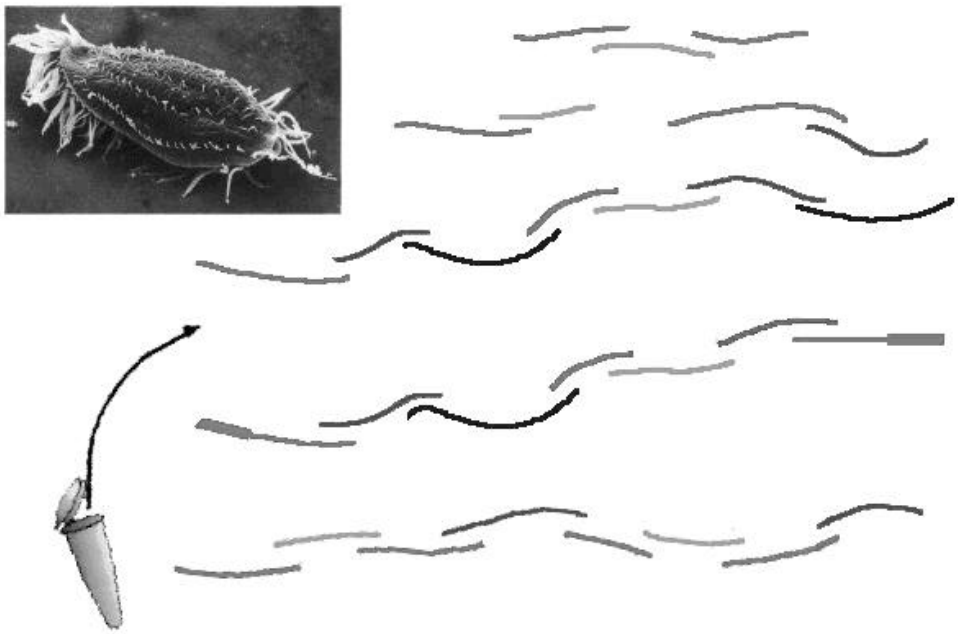

Figure 3: A ciliate molecular computer? Correct gene assembly in *Stylonychia* (inset) requires the joining of many segments of DNA guided by short sequence repeats, only at the ends. Telomeres, indicated by thicker lines, mark the termini of correctly assembled gene-sized chromosomes. Note the striking similarities to DNA computations that specifically rely on pairing of short repeats at the ends of DNA fragments, as in Adleman's experiment (1994).

IESs and reorder the MDSs. For example, the DNA sequence present at the junction between MDS *n* and the downstream IES is generally the-same as the sequence between MDS *n*+1 and its upstream IES, leading to correct ligation of MDS *n* to MDS *n*+1, over a distance (Table 1). However the presence of such short repeats (average length 4 bp between non-scrambled MDSs, 9 bp between scrambled MDSs (Prescott and Dubois 1996)) suggests that although these guides are necessary, they are certainly not sufficient to guide accurate splicing. Hence it is likely that the repeats satisfy more of a structural requirement for MDS splicing, and less of a role in substrate recognition. Otherwise, incorrectly spliced sequences (the results of promiscuous recombination) would dominate, especially in the case of very small (2-4 bp) repeats which would be present thousands of times throughout the genome. This incorrect hybridization could be a driving force in the production of newly scrambled patterns in evolution. However during macronuclear development only unscrambled molecules which contain 5' and 3' telomere addition sequences would be selectively retained in the macronucleus, ensuring that most promiscuously ordered genes would be lost.

1.3. Inversions as catalysts of DNA rearrangements

The micronuclear actin I gene has a scrambled MDS order 3-4-6-5-7-9-2-1-8 in *O. nova*, with MDS 2 inverted (present on the opposite strand

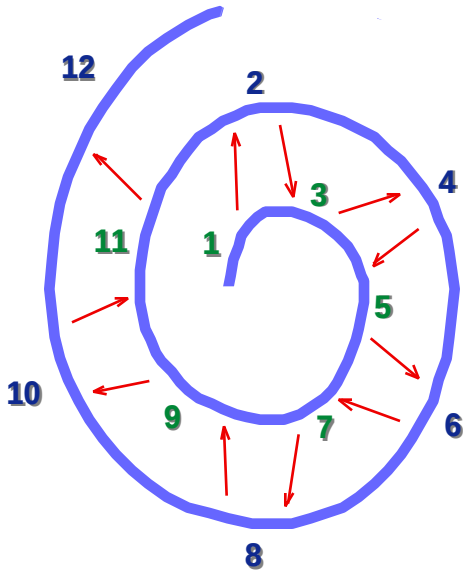

Figure 4: Model for unscrambling in α-TP (adapted from Mitcham et al. 1992).

and in the opposite direction) relative to the others (Dubois and Prescott 1995). DNA polymerase α has at least 44 MDSs in *O. nova* and 51 in *O. trifallax* (Table 1), scrambled in a nonrandom order with an inversion in the middle, and some MDSs located at least several kb away from the main gene (in an unmapped PCR fragment). The resulting hairpin structural constraint predicted in Figure 5 equips the ciliate with a dramatic shortcut to finding the correct solution to its DNA polymerase α unscrambling problem.

Table 1: *O. trifallax* DNA Polymerase α (Data modified from Hoffman and Prescott 1997)

5' MDS/IES junction sequence	MDS	3' MDS/IES junction sequence	No. repeats in Mac sequence*
5' Telomere addition site	1	AGATA	8
AGATA	2	ATT	*
ATT	3	ATA	*
ATA	4	ATGATGAGTGGAAT	1
ATGATGAGTGGAAT	5	AACAGAAC	1
AACAGAAC	6	AGAAATATG	1
AGAAATATG	7	n.d.	
n.d.	9	TTATCATT	2
TTATCATT	10	AAAATAAT	1
AAAATAAT	11	GTTTCTTG	1
GTTTCTTG	12	ATGCAAA	1
ATGCAAA	13	TAAAATGA	1
TAAAATGA	14	AGAGGAG	1
AGAGGAG	15	TAATGATGG	1
TAATGATGG	16	ATGGTGAG	1
ATGGTGAG	17	AAAATCAA	3
AAAATCAA	18	AAAGCATGCTTG	1
AAAGCATGCTTG	19	GATTTCAAGAAAA	1
GATTTTAAGAAAA	20	GTTACTCTTG	1
GTTACTCTTG	21	GCTCAATAAAA	1
GCTCAATAAAA	22	ATCTTG	2
ATCATG	23	AAAACTT	1
AAAACTT	24	GAGAGATAGA	1
GAGAGATAGA	25	TAGTTGCTC	1
TAGTTGCTC	26	AAGCTAGATTTT	1
AAGCTAGATTTT	27	GGAGGATC	1
GGAGGATC	28	CAAGATAA	1
CAAGATAA	29	GTTCAACT	1
GTTCAACT	30	ATAAGACTTTGATGA	1
ATAAGACTTTGATGA	31	CTAATGAA	1
CTAATGAA	32	n.d.	
n.d	36	CTTGAGAT	1
CTTGAGAT	37	AAAGTAGTTTAG	1
AAAGTAGTTTAG	38	CACTTTCAA	1
CACTTTCAA	39	ATGAAAAATAA	1
ATGAAAAATAA	40	CCTTGGATCA	1
CCTTGGATCA	41	AAGAGTGAAT	1
AAGAGTGAAT	42	TGAACAACTTT	1
TGAACAACTTT	43	GTGCTTAG	1
GTGCTTAG	44	n.d.	
n.d	49	ATAAAA	4
ATAAAA	50	AT	*
AT	51	3' Telomere addition site	

*The number of occurrences of di- and tri-nucleotides was not determined since they would be extremely represented in any gene sequence. Note the values shown in this column only represent the number of occurrences of these sequence motifs in the macronuclear copy of the gene. There are also several occurrences of these repeats throughout the noncoding portion of the micronuclear copy of this gene as well as throughout the entire genome; each such occurrence offers the opportunity for incorrect pairing, which would lead to the production of 'dead-end' copies of the gene. These would, however, be unlikely to contain telomere addition sequences at both ends. Junction sequences for MDSs 32-36 and 44-49 are unknown because of missing micronuclear sequence data.

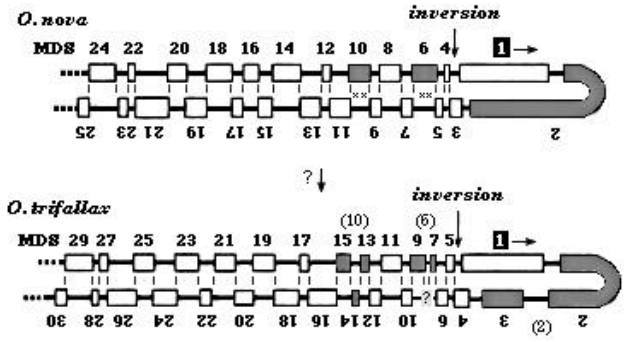

Figure 5: Model for scrambling of DNA pol α. Vertical lines indicate recombination junctions between scrambled MDSs, guided by direct repeats. MDS 1 contains the start of the gene. MDS 10 in *O. nova* can also give rise to three new MDSs (13–15) in *O. trifallax*, one scrambled on the inverted strand, by two spontaneous intramolecular recombination events (×'s) in the folded orientation shown. *O. nova* MDS 6 can give rise to *O. trifallax* MDSs 7–9 (MDS 8, shaded, is only 6 bp and was not identified in (Hoffman and Prescott 1997)). *O. trifallax* non-scrambled MDSs 2 and 3 could be generated by the insertion of an IES in *O. nova* MDS 2 (similar to a model suggested by M. DuBois in Hoffman and Prescott 1997).

Figures 5-6 outline a model for the origin and accumulation of scrambled MDSs. The appearance of an inversion is likely to encourage the formation of new MDSs in a nonrandomly scrambled pattern. By Muller's Ratchet, an inversion makes the addition of new MDSs much more likely, given that the hairpin structure, which juxtaposes coding and noncoding DNA sequences, would promote recombination, possibly between short arbitrary repeats. For example, the arrangement of MDSs 2, 6, and 10 in *O. nova* could have given rise to the arrangement of eight new MDSs in *O. trifallax* (Figure 5).

We have recently discovered scrambling in the gene encoding DNA polymerase α in the micronucleus of a different ciliate, *Stylonychia lemnae*, which enjoys the benefit of a working transfection system (Wen et al. 1996). The scrambled gene in *S. lemnae* appears to share the presence of an inversion with the two *Oxytricha* species. These scrambled genes in ciliates thus offer a unique system in which to study the origin of a complex genetic mechanism and the role of inversions as catalysts of acrobatic DNA rearrangements during evolution (Figure 6). DNA polymerase α's complex scrambling pattern is possibly the best analog equivalent of a hard path finding problem in nature. Alternate splicing at the RNA level, as well as other forms of programmed DNA rearrangements, could also be viewed as solutions to path finding problems in nature. Dynamic processes, such as maturation of the immune response, provide examples of genuine evolutionary computation in cells, whereas the path finding problems here may follow a more deterministic algorithm.

Current effort is directed toward (1) recoding DNA in the laboratory and (2) understanding how cells unscramble DNA, how this process has arisen, and how the 'programs' are written and executed. Do they decode the message by following the shortest unscrambling path or by following a more circuitous but equally effective route, as in the case of RNA editing (below)? Also, how error prone is the unscrambling process? Does it actually search through several plausible unscrambled intermediates or follow a strictly deterministic pathway? The isolation of functional nucleic acid molecules, such as RNA catalysts (ribozymes), from large pools of random sequence offers yet a different paradigm for molecular computation (Bartel and Szostak 1993; Landweber 1997).

1.4. The formal model

Before introducing the formal model, we summarize our notation. An alphabet Σ is a finite, nonempty set. In our case $\Sigma = \{A, C, G, T\}$. A sequence of letters from Σ is called a string (word) over Σ and in our interpretation corresponds to a linear strand. The words are denoted by lowercase letters such as u, v, α_i, x_{ij}. A word that has no letters in it is called an empty word. The set of all possible words consisting of letters from Σ is denoted by Σ^*. We also define circular words over Σ by declaring two words to be equivalent if and only if one is a cyclic permutation of the other. In other words, w is equivalent to w' if and only if they can be factored as w = uv and w' = vu, respectively. We denote by $\cdot w$ a representative of the equivalence class of w. Such a circular word $\cdot w$ refers to any of the circular permutations of the letters in w.

With this notation, we introduce two operations that model the processes that occur during the homologous recombination.

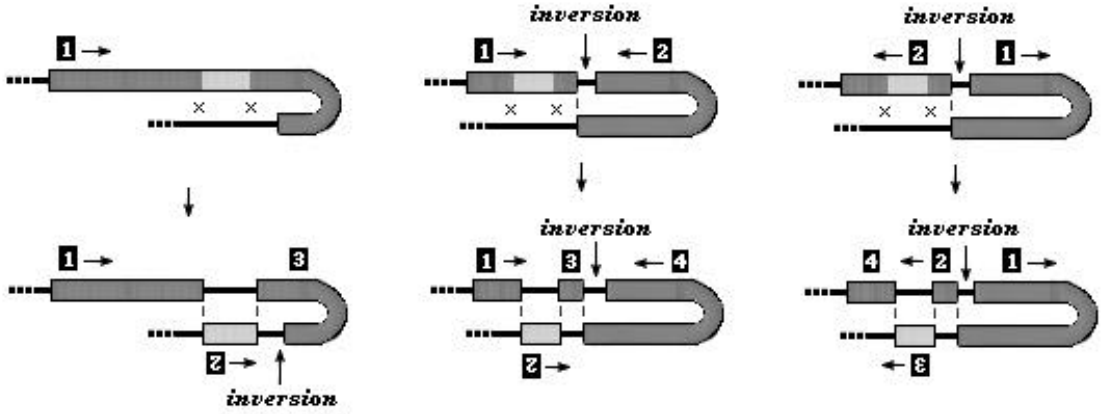

Figure 6: Proposed model for the origin of a scrambled gene. Left: birth of a scrambled gene from a non-scrambled gene by a double recombination with an IES or any noncoding DNA (new MDS order 1-3-2 with an inversion between MDSs 3 and 2). Middle: generation of a scrambled gene with a non-random MDS order, from a non-scrambled gene with an inversion between two MDSs. Right: creation of new scrambled MDSs in a scrambled gene containing an inversion. Inversions may dramatically increase the production of scrambled MDSs, by stabilizing the folded conformation that allows reciprocal recombination across the inversion.

Operation (1), intramolecular recombination, is unary:

$$uxwxv \quad uxv + \cdot wx$$

where u, w, x, and v are words in Σ^*, and x is nonempty. Here "+" is interpreted as the union of the two resulting strands.

Operation (1) is reversible. Note that op1 in the forward direction is formally intramolecular recombination, whereas op1 in the reverse direction is intermolecular recombination.

Thus operation (1) models the process of intramolecular recombination that occurs during unscrambling of the gene. x is a repeated sequence that guides the homologous recombination. After x finds its second occurrence in uxwxv, the molecule undergoes a strand exchange in x that leads to the formation of two new molecules: uxv and a circular DNA molecule $\cdot$wx.

Operation (1) also accomplishes the deletion of either sequence wx or xw from the original molecule uxwxv. The fact that $\cdot$wx is circular implies that we can use any circular permutation of its sequence as an input for a subsequent operation.

Operation (1$\cdot$) is also unary:

$$\cdot uxwxv \quad \cdot uxv + \cdot wx.$$

Operation (1$\cdot$) is similar to op1, the only difference in that the input is circular, which results in the output of two circular strands. Like op1, op1$\cdot$ is reversible.

Operation (2), intermolecular recombination, is binary:

$$uxv + u'xv' \quad uxv' + u'xv$$

where u, x, v, u', v' are words in Σ^*, and x is nonempty. Operation 2 is also reversible.

Operation (2) models most processes of intermolecular recombination. Given two molecules uxv and u'xv', both of which contain a homologous (identical) subsequence x, the molecules undergo a strand exchange (homologous recombination) in x that leads to the formation of molecules uxv' and u'xv. This operation effectively rewrites the input sequences, for example by replacing the suffix v from uxv with v', a process analogous to *trans*-splicing (Sullenger and Cech, 1994).

Note that each operation and its reverse conserves the number of 'ends' (a linear strand has two ends while a circular strand has none). Having defined the operations modeling intra- and inter-molecular recombinations, we now remark that the recombination events predicted to occur during gene unscrambling are capable of generating a variety of products. These include the production of one circular output strand

from two circular inputs, and vice-versa (by op1· and its reverse), the generation of both a linear and a circular output strand from one linear input strand (by op1), the generation of a linear strand by the combination of a linear and circular input (by reversed op1), and finally the generation of two recombined linear output strands from two linear input strands (by op2 and its reverse).

These operations resemble the splicing operation introduced by Head (1987) as a model of DNA recombination and the splicing on circular strands studied by Siromoney *et al.* (1992) and Pixton (1995). Paun (1995) and Csuhaj-Varju et al. (1996) subsequently showed that this model has the computational power of a universal Turing Machine.

The process of gene unscrambling entails a series of successive or possibly simultaneous intra- and inter-molecular homologous recombination events. This is followed by the excision of all sequences $\tau_s y \tau_e$, where the sequence y is marked by the presence of τ_s, a telomere 'start' (at its 5' end), and τ_e, a telomere 'end' (at its 3' end). Thus from a long sequence $u\tau_s y\tau_e v$, this step retains only $\tau_s y\tau_e$ in the macronucleus. Lastly, the enzyme telomerase extends the length of the telomeric sequences (usually double-stranded $\{TTTTGGGG\}_n$ repeats in these organisms) from the 'telomere addition sequences', τ_s and τ_e, to protect the ends of the DNA molecule; however the telomere addition step is not present in our formal model.

We now make the assumption that, by a clever structural alignment, such as the one depicted in Figure 4, and numerous other biological factors, the cell decides which sequences are non-protein-coding (IESs) and which are ultimately protein-coding (MDSs), as well as which sequences x guide homologous recombination. Moreover, such biological shortcuts are presumably essential to bring into proximity the guiding sequences x.

Each MDS, denoted primarily by α_i, $1 \leq i \leq n$ (where n is the number of pieces sparsely present in the micronucleus that assemble to form the functional gene in the macronucleus) is flanked by the guiding sequences $x_{i-1,i}$ and $x_{i,i+1}$. Each guiding sequence points to the MDS that should precede or follow α_i in the final sequence. The only exceptions are α_1, which is preceded by τ_s, and α_n which is followed by τ_e (Table 1). Note that although present generally once in the final macronuclear copy, each $x_{i,i+1}$ occurs at least twice in the micronuclear copy, once after α_i and once before α_{i+1}.

We denote by ε_k an internal sequence that is deleted; ε_k does not occur in the final sequence. Thus, since unscrambling leaves one copy of each $x_{i,i+1}$ between α_i and α_{i+1}, an IES is nondeterministically either $\varepsilon_k x_{i,i+1}$ or $x_{i,i+1}\varepsilon_k$, depending on which guiding sequence $x_{i,i+1}$ is eliminated. Similarly an MDS is technically either $\alpha_i x_{i,i+1}$ or $x_{i-1,i}\alpha_i$. For the purposes of this model, either choice is equivalent.

On a technical note, removal of simple (non-scrambled) IES's in *Euplotes* leaves extra sequences (including a duplication of x_{ij}) at the junctions between ε_k's in the resulting non-protein-coding products. This may result when the x_{ij}'s are as short as two nucleotides (Klobutcher *et al.* 1993). It is unknown whether unscrambling also introduces extra sequences, since it uses considerably longer x_{ij}'s on average. However, since the extra sequences have always been found at junctions between ε_k's, this would not affect our unscrambling model.

The following example models unscrambling of a micronuclear gene that contains MDSs in the scrambled order 2-4-1-3:

$$u\ x_{12}\ \alpha_2\ x_{23}\ \varepsilon_1\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}\ \varepsilon_3\ x_{23}\ \alpha_3\ x_{34}\ v$$

$$u\ x_{12}\ \varepsilon_3\ x_{23}\ \alpha_3\ x_{34}\ v + \cdot\ \alpha_2\ x_{23}\ \varepsilon_1\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}$$

$$=$$

$$u\ x_{12}\ \varepsilon_3\ x_{23}\ \alpha_3\ x_{34}\ v + \cdot\ \varepsilon_1\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}\ \alpha_2\ x_{23}$$

$$u\ x_{12}\ \varepsilon_3\ x_{23}\ \varepsilon_1\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}\ \alpha_2\ x_{23}\ \alpha_3\ x_{34}\ v$$

$$u\ x_{12}\ \varepsilon_3\ x_{23}\ \varepsilon_1\ x_{34}\ v + \cdot\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}\ \alpha_2\ x_{23}\ \alpha_3\ x_{34}$$

$$=$$

$$u\ x_{12}\ \varepsilon_3\ x_{23}\ \varepsilon_1\ x_{34}\ v + \cdot\ \tau_s\ \alpha_1\ x_{12}\ \alpha_2\ x_{23}\ \alpha_3\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2$$

$$\tau_s\ \alpha_1\ x_{12}\ \alpha_2\ x_{23}\ \alpha_3\ x_{34}\ \alpha_4\ \tau_e + \varepsilon_2 + u\ x_{12}\ \varepsilon_3\ x_{23}\ \varepsilon_1\ x_{34}\ v$$

Note that the process is nondeterministic in that, for example, one could start by replacing the first step, which was the recombination between homologous sequences x_{12}, by recombination between the homologous sequences x_{34} instead, obtaining thus

$$u\ x_{12}\ \alpha_2\ x_{23}\ \varepsilon_1\ x_{34}\ \alpha_4\ \tau_e\ \varepsilon_2\ \tau_s\ \alpha_1\ x_{12}\ \varepsilon_3\ x_{23}\ \alpha_3\ x_{34}\ v$$

$$u\, x_{12}\, \alpha_2\, x_{23}\, \varepsilon_1\, x_{34}\, v + \cdot\, \alpha_4\, \tau_e\, \varepsilon_2\, \tau_s\, \alpha_1\, x_{12}\, \varepsilon_3\, x_{23}\, \alpha_3\, x_{34} =$$

$$u\, x_{12}\, \alpha_2\, x_{23}\, \varepsilon_1\, x_{34}\, v + \cdot\, \varepsilon_3\, x_{23}\, \alpha_3\, x_{34}\, \alpha_4\, \tau_e\, \varepsilon_2\, \tau_s\, \alpha_1\, x_{12}$$

$$u\, x_{12}\, \varepsilon_3\, x_{23}\, \alpha_3\, x_{34}\, \alpha_4\, \tau_e\, \varepsilon_2\, \tau_s\, \alpha_1\, x_{12}\, \alpha_2\, x_{23}\, \varepsilon_1\, x_{34}\, v$$

$$\cdot\, \alpha_3\, x_{34}\, \alpha_4\, \tau_e\, \varepsilon_2\, \tau_s\, \alpha_1\, x_{12}\, \alpha_2\, x_{23} + u\, x_{12}\, \varepsilon_3\, x_{23}\, \varepsilon_1\, x_{34}\, v =$$

$$\cdot\, \tau_s\, \alpha_1\, x_{12}\, \alpha_2\, x_{23}\, \alpha_3\, x_{34}\, \alpha_4\, \tau_e\, \varepsilon_2 + u\, x_{12}\, \varepsilon_3\, x_{23}\, \varepsilon_1\, x_{34}\, v$$

$$\tau_s\, \alpha_1\, x_{12}\, \alpha_2\, x_{23}\, \alpha_3\, x_{34}\, \alpha_4\, \tau_e + \varepsilon_2 + u\, x_{12}\, \varepsilon_3\, x_{23}\, \varepsilon_1\, x_{34}\, v$$

in the same number of steps.

In effect, the above examples show that, as the input strand is always linear, albeit scrambled, the intermediate steps will generally produce at most one linear strand as output (apart from telomere addition or other mechanisms that may lead to chromosome breakage or fragmentation). Indeed, in the most basic case, the output of an operation that has only one linear strand as input can never be two linear strands. Consequently, the process involves only iterative application of op1 and op1·. Formally, op2 can only occur after the telomere addition phase (the last step) which we do not define as a separate operation in our model. This does not reduce the generality of the model, as telomere addition happens only once at the end of the process.

Note that, once we assume that the cell has 'decided' which are the α_i's, $x_{i,i+1}$'s and ε_i's, the process that follows is simply sorting, requiring $O(n)$ steps (possibly fewer than n if some of the recombination events take place simultaneously).

However, this 'decision' process, the details of which are still unknown, amounts to finding the correct 'path' linking the pieces of protein-coding-regions in the correct order. Indeed, the occurrence of $\alpha_i\, x_{i,i+1}$ and $x_{i,i+1}\, \alpha_{i+1}$ in the micronuclear sequence provides the link between α_i and α_{i+1}, to indicate that they belong next to each other in the macronuclear sequence. The junction sequences $x_{i,i+1}$ thus serve the role of the 'edge' sequences in Adleman's graph.

A computational difficulty is the presence of multiple copies of the sequences $x_{i,i+1}$ (Table 1) which may direct the formation of incorrect 'paths'. Indeed, throughout the genome, such simple sequences may be present in extremely high redundancy. Some of the $x_{i,i+1}$ even overlap with each other. For example, in Table 1, $x_{24,25}$ = GAGAGATAGA contains $x_{1,2}$ = AGATA as a subsequence.

The search for the proper junction sequences thus amounts to finding the correct 'path' and is potentially the most costly part of the computation. Production of incorrect paths will not necessarily lead to the production of incorrect proteins unless the path sequences start and end with the correct telomere addition sites (τ_s and τ_e), since these ensure survival of the genes in the macronucleus.

2. RNA Editing

RNA editing presents another cellular paradigm for 'biological so f tware'. Driven by small guide RNA sequences, RNA editing also uses sequence pairing, or 'hybridization logic', to convert seemingly disordered sequences into functional coding sequences. Together RNA editing and gene unscrambling provide an array o f potentially usable paradigms for writing and rewriting DNA.

RNA editing is the modification o f RNA sequences by insertion, deletion, or substitution o f bases. At the same time, formal models of contextual insertions/deletions or substitutions have proved that such operations are a powerful computational paradigm with Universal Turing Machine power (Kari and Thierrin 1996; Beaver 1996). Found in a wide variety o f eukaryotes, ranging from parasitic protozoa to humans, this remarkable process alters the sequence o f messenger RNA molecules before translation into protein such that the resulting protein sequence o f ten differs dramatically from the original

DNA G G GTTTTGG AGA G ATTTGG A
RNA **uGuuuuGuuuuuu**GUUUUG**Guuu**AGA**uuuuuuuuGu**AU**GG**uu**Auuu**

Figure 7: RNA Editing. Comparison o f an RNA sequence encoding *H. mariadeanei* COIII with its corresponding region o f the mitochondrial DNA (Landweber and Gilbert 1993). DNA sequences in upper case; uridines in mRNA that are added by RNA editing in lower case (boldface); two encoded thymidines deleted from the mRNA indicated by asterisks.

gene sequence, which sometimes does not encode a protein at all. For example, in *Trypanosoma brucei,* RNA editing by addition and deletion of literally *hundreds* of uridine residues creates initiation and termination codons, alters the structural features of transcripts, and creates over 90% of the amino-acid codons (Landweber and Gilbert 1993). The main features of RNA editing are recognized by differences between RNA and genomic DNA sequences.

Figure 7 illustrates an extreme example of RNA editing of the mitochondrial cytochrome oxidase subunit III (COIII) transcript in *Herpetomonas mariadeanei.* 91% of the 288 amino acid codons in *T. brucei* (Feagin et al. 1988) and *Herpetomonas* (Landweber and Gilbert 1993) are created by editing, with this effect restricted only to uridines (Figure 7). The other bases—A, C, and G—are completely conserved between the DNA and the RNA sequence.

The edited proteins also have a fast molecular clock, which accumulates amino acid substitutions at least twice as fast as the unedited versions. This is due to frameshift mutations that are introduced by fixed changes in the number and position of U's inserted or deleted by editing and then compensated by editing at another site that restores the reading frame (Figure 8). Editing therefore allows the production of combinatorially diverse protein products from a single gene, either within an individual cell (Sommer *et al. 1991*) or over evolutionary time (Landweber and Gilbert 1993). The surprising result that RNA editing provides an additional level of sequence variation, rather than a faithful 'editing' or correcting mechanism, underscores the importance of the question of why is it still used by some organisms to generate a sequence that encodes a single conserved protein (Landweber and Gilbert 1994).

The genetic information for editing is stored in the form of 'guide' RNA molecules (gRNAs), very small (50-70 nt) transcripts that mediate editing by base-pairing with specific regions of the edited transcript, exploiting G:U base-pairs in RNA (Blum et al. 1990). Each gRNA contains the sequence information to edit approximately 30 nt of edited RNA (Landweber et al. 1993) and pairs more efficiently with the final product than with the pre-edited substrate. For every inserted U in the messenger RNA sequence, there is a corresponding A or G in the gRNA which pairs with the fully edited product (Figure 9). Complete editing proceeds 3' to 5' on the mRNA and requires a full set of overlapping guide RNAs. Editing by each guide RNA creates an anchor sequence for binding the next guide RNA (Figure 10, Blum et al. 1990; Maslov and Simpson 1992) leading to an ordered cascade of insertion and deletion editing events. RNA editing is thus a cellular process which uses RNA sequences as guides to convert seemingly disordered RNA sequences into a final messenger RNA molecule: a truly RNA-based computer. Together, the stunning acrobatics of DNA, such as scrambling or editing, give proof to the versatility of nucleic acids and their potential use in solving computational problems that occur in biological systems.

```
Hsa DNA     G AA       A         G G   GG    AGA       G  G      G G      G
Hsa RNA   uuGuAAuuuuuuAuuuuuuuu GuGuuuGGuuuAGAuuuuuuGuuGu  uuGuGuuuuuGu
Hme DNA     G AA          ATTTT G G   GG    AGA       G  GT     G A      G
Hme RNA   uuGuAAuuuuuuuuuuAUUUU GuGuuuGGuuuAGAuuuuuuGuuGU  uuGuAuuuuuGu
Hmm DNA     G AA          ATTTTTG G   GG    AGA       G  GT     G A      G
Hmm RNA   uuGuAAuuuuuuuuuuAUUUU GuGuuuGGuuuAGAuuuuuuGuuGU  uuGuAuuuuuGu
Hma DNA       GAA    G     G       GTTTTGG    AGA        G ATTTGG  A   G
Hma RNA   uuuGAAuuuuGuuuuGuuuu uGUUUUGGuuuAGAuuuuuuuGuAU**GGuuAuuuGuuu
Tbr DNA     G  AA       G    G   GTTTTTGG    AGG        G  GTTTTG  G   G
Tbr RNA   uGuuAAuuuuuuuuGuuuuGu GUUUUUGGuuuAGGuuuuuuuuGuuG**UUGuuGuuuuGu
```

Figure 8: Editing produces and fixes frameshift mutations. Frameshifted nucleotides are in boldface. Hsa, *H. samuelpessoai*; Hme, *H. megaseliae*; Hmm, *H. muscarum muscarum*; Hma, *H. mariadeanei* (Landweber and Gilbert 1993); Tb, *T. brucei* (Feagin et al. 1988). DNA sequences in upper case; uridines in mRNA that are added by RNA editing in lower case; encoded thymidines deleted from the mRNA indicated by asterisks.

Figure 9: Pairing between a guide RNA (top) **and edited mRNA** (bottom). Lowercase u's are added by editing. **U**'s are edited by base pairing to lowercase a's and g's in this gRNA (Maslov and Simpson 1992).

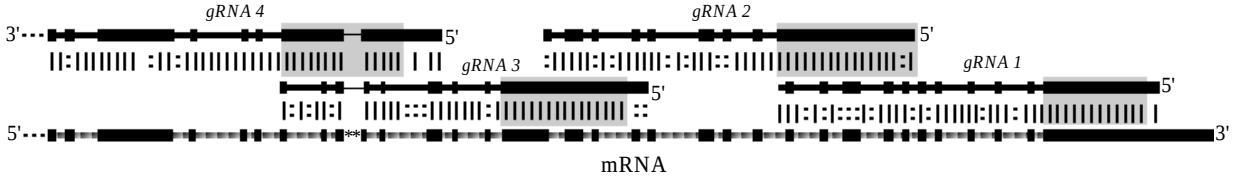

Figure 10: Editing by four overlapping gRNAs. Thick lines in the mRNA are encoded in the mitochondrial DNA. Thin shaded lines are inserted U's; the two asterisks are deleted U's (Maslov and Simpson 1992). Thin lines in the gRNA's are guide nucleotides (A or G) that pair with inserted U's. Vertical lines indicate Watson-Crick base pairs; colons indicate G:U wobble base pairs, illustrating formation of well-paired 'anchors' between the 5' ends of gRNA's and the corresponding region of the mRNA.

Acknowledgements

We thank Richard Lipton, David Prescott, Dirk Faulhammer, and Ed Curtis for discussion and Hans Lipps for DNA. L.F.L. is a Burroughs-Wellcome Fund New Investigator in Molecular Parasitology.

Bibliography

Adleman, L. M. 1994. Molecular computation of solutions to combinatorial problems. *Science* 266: 1021-1024.

Bartel, D.P., and J. W. Szostak. 1993. Isolation of New Ribozymes from a Large Pool of Random Sequences. *Science* 261: 1411-1418.

Beaver, D. 1996. A universal molecular computer. In Lipton, R. J. and E.B. Baum (editors). *DNA Based Computers. DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 27, AMS. Pages 29-36.

Blum, B., Bakalara, N., and L. Simpson. 1990. A model for RNA editing in kinetoplastid mitochondria: 'Guide' RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* 60: 189-198.

Csuhaj-Varju, E., Freund, R., Kari, L., and G. Paun. 1996. DNA computing based on splicing: universality results. In Hunter, L. and T. Klein (editors). *Proceedings of 1st Pacific Symposium on Biocomputing*. World Scientific Publ., Singapore, Pages 179-190.

DuBois, M. and D.M. Prescott. 1995. Scrambling of the actin I gene in two *Oxytricha* species. *Proc. Natl. Acad. Sci., U.S.A.* 92: 3888-3892.

Feagin, J.E., Abraham, J.M., and K. Stuart. 1988. Extensive Editing of the Cytochrome c Oxidase III transcript in *Trypanosoma brucei*. *Cell* 53: 413-422.

Head, T. 1987. Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bull. Math. Biology* 49: 737-759.

Hoffman, D.C., and D.M. Prescott. 1997. Evolution of internal eliminated segments and scrambling in the micronuclear gene encoding DNA polymerase α in two *Oxytricha* species. *Nucl. Acids Res.* 25: 1883-1889.

Kari, L., and G. Thierrin. 1996. Contextual insertion/deletions and computability. *Information and Computation* 131: 47-61.

Klobutcher, L.A., Turner, L.R., and J. LaPlante. 1993. Circular forms of developmentally excised DNA in *Euplotes crassus* have a heteroduplex junction. *Genes Dev.* 7: 84-94.

Landweber, L. F. 1997. RNA Based Computing. DIMACS Technical Report 97-83.

Landweber, L. F. and W. Gilbert. 1993. RNA ed-

iting as a novel source of genetic variation. *Nature* 363: 179-182.

Landweber, L. F. and W. Gilbert. 1994. Phylogenetic analysis of RNA editing: A primitive genetic phenomenon. *Proc. Natl. Acad. Sci. USA* 91: 918-921.

Landweber, L. F., Fiks, A. G., and W. Gilbert. 1993. The Boundaries of Partially Edited Cytochrome c Oxidase III Transcripts are Not Conserved in Kinetoplastids: Implications for the Guide RNA Model of Editing. *Proc. Natl. Acad. Sci. USA* 90: 9242-9246.

Lynn, D.H. and J.O. Corliss. 1991. Ciliophora. In *Microscopic Anatomy of Invertebrates*. Vol. 1: Protozoa. New York: Wiley-Liss, Inc., 333-467.

Maslov, D.A. and L. Simpson. 1992. The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. *Cell* 70: 459-467.

Mitcham, J.L., A.J. Lynn, and D.M. Prescott. 1992. Analysis of a scrambled gene: The gene encoding α-telomere-binding protein in *Oxytricha nova*. *Genes Dev*. 6 : 788-800.

Paun, G. 1995. On the power of the splicing operation. *Int. J. Comp. Math* 59, 27-35.

Pixton, D., 1995. Linear and circular splicing systems. Proceedings of the First International Symposium on Intelligence in Neural and Biological Systems, IEEE Computer Society Press, Los Alamos. Pages 181-188.

Prescott, D.M. and M.L. Dubois. 1996. Internal Eliminated Segments (IESs) of Oxytrichidae. *J. Euk. Microbiol.* 43: 432-441.

Siromoney, R., Subramanian, K.G. and Rajkumar Dare, V. 1992. Circular DNA and splicing systems. In Parallel Image Analysis. Lecture Notes in Computer Science 654, Springer Verlag, Berlin. Pages 260-273.

Smith, W.D. (1996) DNA Computers In Vitro and Vivo. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. 27: 121-185.

Sommer, B., Kohler, M., Sprengel, R. and Seeburg, PH. 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* 67: 11-19.

Sullenger, B.A. and T. R. Cech. 1994. Ribozyme-mediated repair of defective mRNA by targeted, trans-splicing. *Nature*. 371: 619-22.

Tausta, S.L. and L.A. Klobutcher. 1989. Detection of circular forms of eliminated DNA during macronuclear development in *E. crassus*. *Cell* 59 : 1019-1026.

Wen, J., Maercker, C., and Lipps, H.J. 1996. Sequential excision of internal eliminated DNA sequences in the differentiating macronucleus of the hypotrichous ciliate *Stylonychia lemnae*. *Nucl. Acids Res.* 24: 4415-4419.