Gene assembly in ciliates
Part I. Molecular operations

Tero Harju\(^1\), Ion Petre\(^2\), and Grzegorz Rozenberg\(^3\)

\(^1\) Dept. of Mathematics, University of Turku, FIN-20014 Turku, Finland.
\(^2\) Dept. of Computer Science, Abo Akademi University, FIN-20520 Turku, Finland.
\(^3\) Leiden Institute for Advanced Computer Science, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands and
Department of Computer Science, University of Colorado, Boulder,
Co 80309-0347, USA.
emails: harju@utu.fi, ipetre@abo.fi, rozenber@liacs.nl

“If you look inside the cell you find a bunch of amazing little tools.
The cell is a treasure chest.”
Leonard Adleman in CNN Science and Space, August 17, 2003

Natural Computing is concerned with computing taking place in nature and
with human-designed computing inspired by nature. It is an interdisciplinary
research area that has already had a significant effect on the development of
computer science: human-designed computing inspired by nature includes ar-
eas such as neural computing, evolutionary algorithms, DNA computing and
quantum computing.

Also, there is a growing research trend spread throughout the disciplines
of bioinformatics, computational biology and DNA computing (among others)
that investigates computational nature of complex biological phenomena. As a
matter of fact, part of this research considers some of the basic life processes as
computations. One of such processes that has recently attracted the attention is
gene assembly in ciliates (single-celled eukaryotic organisms).

The study of this process was initiated in 1971 (see Prescott et al. [16]) -
the review of subsequent biological studies of this process can be found in
Prescott [14, 15] and Prescott and Rozenberg [20].

The computational nature of gene assembly was first pointed out by Kari
and Landweber [12], [13] - they noticed that the process of assembling genes
resembles the structure of the “molecular solution” of the directed hamiltonian
path problem proposed by Adleman in his seminal paper [1]. Kari and Landweber
have proposed a model based on intermolecular interactions, and subsequently
they have investigated the computational power of this model in the sense of
computation theory. Another model of gene assembly, based on intramolecular
interactions (different parts of the same molecule interacting with each other)
was introduced in [17], and studied in more details in the articles [2]-[11], [18]-
[20]. This research concentrates on the (formal and biological) properties of the
process of gene assembly itself.
This paper surveys the latter line of research (hopefully) making it accessible to a broader audience. It consists of two parts. The current Part I provides the basic biological background and introduces molecular operations involved in the process of gene assembly. Part II reviews theoretical properties of models based on these operations.

A recent monograph by Ehrenfeucht, Harju, Petre, Prescott, Rozenberg [3] provides a broad and in-depth treatment of both the biological and the formal aspects of gene assembly.

1 DNA molecules

The genetic information of a cell is stored in DNA (deoxyribonucleic acid) molecules. The expression of this information is an intricate process, which involves transcription of information coded in DNA into (messenger) RNA (ribonucleic acid) molecules, and then translating RNA molecules into proteins.

DNA molecules are made from nucleotides, and each nucleotide is made up from three components: sugar, phosphate and base. There are four types of bases: adenine, thymine, cytosine, and guanine, abbreviated by A, T, C, and G, respectively. Since nucleotides may differ only in their bases, they are also denoted by A, T, C, and G, according to the sort of base that they contain.

Single stranded DNA molecules are formed by chaining of nucleotides. Two nucleotides in a chain are bound by a strong bond between the sugar of one nucleotide and the phosphate of the other one - in this way the sugar/phosphate backbone of a single stranded DNA molecule is formed with bases “sticking out of it”, see Fig. 1. Such a single stranded DNA molecule has a polarity because it has phosphate available for binding at one end of the chain and the sugar at the other end - the phosphate end is referred to as the 5'-end and the sugar end as the 3'-end. Since there are four types of nucleotides, a single stranded DNA molecule may be represented by a string over the alphabet $\mathcal{N} = \{A, C, G, T\}$ of nucleotides. To get an unambiguous notation one needs to indicate whether the strand is read beginning with the 5'-end or beginning with the 3'-end; the common convention is the reading beginning with the 5'-end. In this way $ATTGCA$ denotes a single stranded DNA molecule consisting of six nucleotides and such that A is its first nucleotide (at the 5′-end), T is its second nucleotide, etc., and A is its sixth nucleotide (or the first one at the 3′-end).

The basic property of the four types of nucleotides is their complementarity: $A$ is complementary to $T$ and $C$ is complementary to $G$. This complementarity plays a crucial role in forming double stranded DNA molecules from single stranded DNA molecules. Two single stranded DNA molecules $\alpha$ and $\beta$ will bind to form a double stranded DNA molecule if they satisfy the following “complementarity and reverse polarity condition”: the first nucleotide at the 5′-end of $\alpha$ is complementary to the first nucleotide at the 3′-end of $\beta$, the second nucleotide at the 5′-end of $\alpha$ is complementary to the second nucleotide at the 3′-end of $\beta$, etc., and the nucleotide at the 3′-end of $\alpha$ is complementary to the nucleotide at the 5′-end of $\beta$. In other words, if we match the 5′-end of $\alpha$ with the 3′-end of $\beta$,
Fig. 1. A segment of a DNA strand. Nucleotides in DNA are joined to form very long strands by connections between the sugar (S) and phosphate (P) groups in each nucleotide. The bases (A, T, G, and C) in the nucleotides project to the side of the strand.

then all nucleotides in these two strands will be complementary. Thus, e.g., the single stranded DNA molecule $\alpha = ATTGCA$ and $\beta = TGCAAT$ (recall that we write single stranded DNA molecules beginning with the 5'-end) will form the double stranded molecule

\[
\begin{align*}
A & T T G C A \\
T & A A C G T
\end{align*}
\] (1)

In this obvious “double string notation” for double stranded DNA molecules the upper string corresponds to the reading of one of the single stranded DNA molecules, in our case $\alpha$, starting with the 5'-end (therefore the lower string corresponds to the reading of the other single stranded DNA molecule, in our case $\beta$, starting with the 3'-end). To close our remarks on the notation for double stranded DNA molecules we want to point out (recall) the well-known fact that although our notation for DNA molecules is linear, the double stranded DNA molecules in living cells form the famous double helix, where the two single stranded DNA molecules twist around each other in a helical fashion.
An inversion of a double stranded molecule is obtained by reversing the direction and exchanging the two strands. For instance, the double strand illustrated in (1) has the following inversion:

\[
\begin{array}{cccc}
A & C & G & T \\
C & T & G & A \\
\end{array}
\]

Proteins are chains of amino acids (there are 20 amino acids used by proteins), and double stranded DNA molecules encode proteins using codons, which are triplets of nucleotides. Since there are 4 (types of) nucleotides there are 64 different codons - many more than 20 amino acids. Thus some amino acids are encoded by many codons, and moreover some codons indicate the beginning and the end of a gene, where a gene is a region in a DNA double helix that specifies the sequence of amino acids of a particular protein (this is a big simplification in defining a gene, but this definition suffices for the purpose of this paper). One strand of each gene begins with the “start” codon ATG, and it ends in one of the “stop” codons TAA, TAG, or TGA. This strand is called the sense strand of the double helix, and the complementary strand is called the antisense strand. Chromosomes are complexes of proteins and amino acids that contain the DNA molecules; hence they contain all the genes in a cell.

2 Nuclear dualism in ciliates

Ciliates form a diverse group of unicellular organisms that are found practically in all environments containing water. They are an ancient group, possibly more than two billion years old. About 8000 different species are known. All ciliates have hair-like cilia using which the organism can move and drive food in its oral apparatus.

A unique feature of ciliates is their nuclear dualism, the presence of two kinds of functionally different nuclei, micronucleus and macronucleus, in a single cell. A micronucleus is a dormant storage of genetic material - no expression of genetic material is ever attempted in a micronucleus. A macronucleus on the other hand is genetically active, i.e., the genes in the macronucleus produce all the RNA transcripts for the expression of proteins, and hence for maintaining the structure and activity of the cell. A micronucleus becomes active only during cell mating when, at some stage, it is transformed into a macronucleus in a process called gene assembly.

Gene assembly is a very complex DNA processing involving extensive elimination of DNA and defragmentation of genetic material. This process is interesting from both biological and computational point of view. The complexity of this process is most significant in the group of ciliates known as stichotrichs. During evolution the genomes of stichotrichs ciliates have experienced remarkable modifications that require intricate manipulations of DNA in the gene assembly process.

All species of stichotrichs have two or more micronuclei and two or more macronuclei, depending on the species. The micronuclei in a stichotrich are genetically and structurally identical, and also the macronuclei are identical.
Micronucleus

Micronuclei in ciliates contain chromosomes, but, as mentioned above, micronuclei are inactive, i.e., they do not produce RNA transcripts for building proteins. The protein synthesis is carried solely by the genes in the macronuclei.

In normal environmental conditions, where there is food available, stichotrichs reproduce by cell division, but if a stichotrich begins to starve, it may mate with another starving stichotrich of the same species. Mutations are essential in evolution. They can be deleterious, or neutral, or, very rarely, they can improve the function of a gene. Favourable mutations can spread through a population by mating. The mating of ciliates proceeds through the following steps (in our example we have assumed that mating ciliates have two micronuclei and two macronuclei) - see Fig. 2:

Fig. 2. Cell mating in stichotrichs.

(1) The ciliate in the pre-mating state.
(2) Two ciliates stick together and a connecting channel between them is formed.

The micronuclei of each cell undergo the process called meiosis so that from
the two diploid micronuclei (ciliates are diploid, i.e., they contain two copies of every chromosome) eight haploid micronuclei are produced (a haploid micronucleus contains one copy of every chromosome). Then one haploid micronucleus from each cell migrates through the connecting channel to the other ciliate.

(3) The received haploid micronucleus fuses with a resident haploid micronucleus, forming a new diploid micronucleus. The two cells separate.

(4) The unused haploid micronuclei and the two macronuclei in each cell degenerate, and the new diploid micronucleus in each cell divides by the process called mitosis which yields two diploid micronuclei.

(5) One of the two daughter micronuclei becomes the new diploid micronucleus, while the other daughter micronucleus develops into a new macronucleus.

(6) The new macronucleus and micronucleus divide yielding the characteristic number of two macronuclei and two micronuclei in the exmated cell and marking the end of the mating process.

**Micronuclear versus macronuclear genes**

The DNA in each micronuclear chromosome consists of one very long molecule. Genes are widely spaced along the chromosome and they are separated by stretches of nongenetic DNA (see Fig. 3). The genes in the micronuclear DNA of stichotrichs have features that have not been observed in any other organisms. They are interrupted by noncoding segments of DNA called *internal eliminated segments*, or IESs. In Fig. 4 a diagram for the micronuclear gene for the βTP protein in *sterkiella nova* is presented. It has six IESs, which divide the gene into seven segments, known as *macronuclear destined segments*, or MDSs. IESs

---

**Fig. 3.** Genes along a segment of chromosomal DNA.

**Fig. 4.** A diagram for the micronuclear gene encoding βTP protein.
are short, usually less than 100 nucleotides long, and they are rich in the nucleic acids A and T. The micronuclear genes that are known in one species of stichotrich indicate that as many as 100,000 IESs are present in all the genes in a haploid genome. During development of micronucleus into macronucleus all these IESs are eliminated, and the MDSs are rearranged and ligated together. Thus, e.g., in the species *styloonychia* around 98% of the DNA in the micronucleus (consisting mostly of spacer DNA and IESs) is eliminated during the formation of macronucleus.

The IESs can be determined by comparing the expressible copy of a gene in the macronucleus, *that has no IESs*, with its micronuclear counterpart. This also gives an indication how IESs are excised during macronuclear development. At the MDS-IES junctions both MDSs contain the same sequence consisting of 3 to 20 nucleotides. Such a repeated sequence is underlined in Fig. 5. The IES is indicated in lower case letters.

![Diagram of IESs and MDSs](image)

**Fig. 5.** A section of the micronuclear gene for βTP in *Sterkiella histriomuscorum*.

This suggests that the repeats are essential for the excision of IESs. When an IES is removed, one copy of the repeat is also removed, and one copy remains at the newly formed junction between the MDSs.

The micronuclear gene structure can be more complex than the one shown in Fig. 4. Figure 6 gives a diagram of the micronuclear gene for the actin protein in *Sterkiella nova*. It has nine MDSs that occur in the scrambled order 3-4-6-5-7-9-2-1-8. Moreover, the coding sequences of nucleotides in the MDSs, excepting MDS 2, are read from left to right. The MDS 2 has been *inverted* with respect to this reading order. The macronuclear development of this gene requires excision of IESs and also rearrangement of the MDSs in the orthodox order 1-2-3-4-5-6-7-8-9 such that also MDS 2 is ordered in the same direction as the other MDSs.

In the final step, gene-sized DNA molecules are created – the shortest DNA molecules occurring in nature. This is an extraordinary event. Indeed, about
25,000 genes in the developing macronucleus are excised, and the spacer DNA is destroyed leaving only short DNA molecules that each contain usually a single gene. A few nucleotides of a specific sequence are then added to the ends of the DNA strands of these short molecules (to protect the ends). Each of these gene-sized molecules is then replicated many times until it is present in hundreds to thousands of identical copies.

3 Homologous recombination

The molecular operations accomplishing gene assembly that we are going to introduce now are based on homologous recombination between pointers. Homologous recombination is a common molecular operation that occurs between two DNA molecules that possess a segment containing an identical sequence of base pairs. These identical sequences are first aligned side-by-side, then an enzyme introduces staggered cuts in the same position in each DNA molecule, then molecules switch parts with each other, and finally an enzyme called ligase repairs the cuts - this is illustrated in Fig. 7 (where $F_2, S_2$ are the switched parts).

4 Molecular operations

The unscrambling and joining of MDSs requires impressive manipulations of DNA during the process of gene assembly when new macronucleus is developed.
from its micronuclear form. These manipulations consist of the excision of at least 100,000 IESs, and rearranging of thousands of MDSs. We shall now introduce the three molecular operations introduced in [17] which account for the gene assembly process in stichotrichs. In the description of these operations the pairs of repeats are called pointers. Clearly, these molecular operations depend on the MDS/IES structure of the genes.

**Loop recombination: ld**

![Diagram of Loop recombination](image)

In the operation *loop, direct repeat*-excision, or *ld* for short, a pair of pointers flanking an IES guides the excision of the IES, as illustrated in Fig. 8 for the IES 2 between MDS 2 and MDS 3 in the βTP gene in *Sterkiella histriomuscorum* (see Fig. 5). The copy of the pointer P3 at the end of MDS 2 is the outgoing pointer (of MDS 2), while the copy of P3 at the beginning of MDS 3 would be the incoming pointer (of MDS 3).
is the incoming pointer (of MDS 3). The two copies of P3 are recombined by homologous recombination. This recombination joins MDS 2 to MDS 3 through one recombined copy of P3, and excises IES 2 as a circular molecule containing the other recombined copy of P3. MDS 2 and MDS 3 form a composite MDS.

**Hairpin recombination: hi**

![Diagram of hairpin recombination]

**Fig. 9.** Hairpin recombination.

The operation (*hairpin, inverted repeat*)-excision/reinsertion, or *hi* for short, is applicable to a molecule containing a pair of pointers where one pointer is the inversion of the other. This is illustrated for the segment of the actin gene containing MDS 1 and MDS 2 (with MDS 2 inverted relative to MDS 1) on Fig. 9. In the recombination of the actin gene MDS 1 is joined to MDS 2 through one copy of P2 by aligning the outgoing copy of P2 (of MDS 1) with the incoming
Fig. 10. Double loop recombination.
copy of P2 (of MDS 2). Note that MDS 2 and its incoming pointer P2 have inverted polarity. By forming a hairpin in the DNA, the two copies of P2 are aligned in the same direction, and homologous recombination can take place.

**Double loop recombination: dlad**

The operation *(double loop, alternating direct repeat)-excision/reinsertion, or dlad* for short applies to a DNA molecule containing two pairs of pointers where the segments encompassed by the pairs of pointers overlap with each other. This is illustrated in Fig. 10 for the micronuclear actin gene in *Sterkiella nova*, see Fig. 6.

The molecule folds into two loops so that the two copies of P5 align with each other in one loop, and the two copies of P6 align with each other in the other loop. Thus, the molecule is in position for two recombinations. Recombination between the two copies of P5 joins MDS 4 to MDS 5 through a copy of P5, making the composite MDS 4 – MDS 5. Recombination between the two copies of P6 joins MDS 5 into a composite MDS with MDS 6 through one copy of P6. As a result of these two recombinations IESs 3, 2, and 4 are joined by the other copy of P5 and the other copy of P6. Thus, the three IESs form a composite segment that now resides after MDS 6.

The instances of the operations *ld, hi, and dlad* may be applied simultaneously or sequentially. It is important to notice that the process of gene assembly accomplished by these operations is in fact a process of removing pointers. The three operations, working singly or in various combinations, account for the excision of IESs and the joining of MDSs.

As shown in [2], [6], and [8] the set of these three molecular operations provides a uniform framework for gene assembly in ciliates.

**Acknowledgements**

The authors are supported by European Union project MolCoNet, IST-2001-32008. T. Harju gratefully acknowledges support of the Academy of Finland, project 39802. G. Rozenberg gratefully acknowledges support by NSF grant 0121422.

**References**


