# 7.31. Semigroups and Biology

**Note.** Recall that a *semigroup* is a set $S$ together with an associative binary operation $\circ$ defined on it (see Definition 7.28.1). If $s \circ t = t \circ s$ for all $s, t \in S$ then $S$ is a *commutative semigroup*. Recall that for $A$ a nonempty set, the set $A_*$ of all finite sequence $(a_1, a_2, \ldots, a_n) = a_1, a_2, \ldots, a_n$ where $n \in \mathbb{N}$ and $a_i \in A$ forms a *free semigroup*, $(A_*, *)$, where the binary operation is concatenation:

$$a_1 a_2 \cdots a_n * a'_1 a'_2 \cdots a'_m = a_1 a_2 \cdots a_n a'_1 a'_2 \cdots a'_m$$

(see Definiton 28.20). In this section we consider applications of semigroups and free semigroups to genetic crossings of organisms and cellular growth.

**Example 7.31.1.** Consider a strain of cattle which can be black or brown, and monochromatic or spotted. Suppose that black is a dominant trait over brown, and that monochromatic is a dominant trait over spotted. This yields four possible types of cattle:

| | |
|---|---|
| $a$: black and monochromatic | $c$: brown and monochromatic |
| $b$: black and spotted | $d$: brown and spotted. |

We treat this differently here from a traditional biological setting (as Lidl and Pilz comment on page 357 when they state "In general, the table for breeding operations is more complicated"). We assume that when an individual displaying a dominant trait can only have offspring which display the dominant trait (so we are not assuming a presence of a dominant allele and a recessive allele for which the phenotype of the dominant trait gives ambiguity of the genotype, since the individual could

either be homozygous dominant or heterozygous). For more details on biologically realistic models, see my online notes for Integrative Biology and Statistics (BIOL 1810) on Evolution Module: 6.1 Hardy-Weinberg (notice the "One-Locus/Two Alleles" model). This simplifying assumption allows us define a binary operation $*$ on the four types of cattle. For example, a type $b$ strain (having the dominant color black and the recessive pattern spotted) when crossed with a type $c$ strain (having the recessive color brown and the dominant pattern monochromatic) will produce progeny that are of type $a$ (black and monochromatic). We indicate this by writing as $b * c = a$. Similarly, we get the following table representing the binary operation:

| $*$ | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|
| $a$ | $a$ | $a$ | $a$ | $a$ |
| $b$ | $a$ | $b$ | $a$ | $b$ |
| $c$ | $a$ | $a$ | $c$ | $c$ |
| $d$ | $a$ | $b$ | $c$ | $d$ |

We still need to check associativity (somewhat tedious), but it holds. So $S = (\{a, b, c, d\}, *)$ is, in fact, a semigroup. Since the table is symmetric with respect to the main diagonal, then $S$ is a commutative semigroup. We also have element $d$ as the identity, so in addition $S$ is commutative monoid.

**Note.** The molecule which carries our genetic information is DNA ("deoxyribonucleic acid"). It is a two-stranded double helix. Each strand of a sequence of nucleotides, A (Adenine), T (Thymine), C (Cytosine), and G (Guanine). These nucleotides pair up in "base pairs" with A and T binding together, and C and G binding together. For more details on structure of DNA, see my online notes

used in Independent Study (MATH 5900) classes on Evolutionary Genetics and Mathematical Biology on Introduction to Molecular Genetics. In this section, we denote the nucleotides as $n_1, n_2, n_3, n_4$, so that a strand of DNA can then be interpreted as a word over the set $\{n_1, n_2, n_3, n_4\}$ (in the sense of free groups and free semigroups). mRNA (messenger ribonucleic acid) binds to DNA to "read" the DNA, decouples from the DNA, and is then used to transcribe protein chains made up of amino acids. There are 21 different amino acids produced in this way which we denote $a_1, a_2, \ldots, a_{21}$. The protein chains can then be thought of as a word over $\{a_1, a_2, \ldots, a_{21}\}$ (in the sense of free groups and free semigroups). We assume that the sequence of amino acids in a protein chain uniquely determine the sequence of nucleotides in the DNA molecule. We denote the free semigroup on set $\{n_1, n_2, n_3, n_4\}$ as $F_4$ and denote the free semigroup on set $\{a_1, a_2, \ldots, a_{21}\}$ as $F_{21}$. The "DNA protein coding problem" is the question: "How many, if any, monomorphisms are there from $F_{21}$ into $F_4$?" The next theorem show that there are infinitely many.

**Theorem 7.31.2.** There are infinitely many monomorphisms form $F_{21}$ into $F_4$. Thus the DNA protein-coding problem has infinitely many solutions.

**Note.** In the proof of Theorem 7.31.2, we showed that there are infinitely many desired monomorphisms by only considering words in $F_4$ of length 3. In fact, DNA protein coding is based on converting triplets of nucleotides (called codons) through mRNA into individual proteins. So the use of words in length 3 in $F_4$ carries biological realism.

*Revised: 11/2/2021*