

DNA computing, insertion of words and left-symmetric algebras*

MURRAY R. BREMNER[†]

Research Unit in Algebra and Logic, and
Department of Mathematics and Statistics
University of Saskatchewan
106 Wiggins Road, Saskatoon, SK, S7N 5E6, Canada

May 16, 2005

Keywords: DNA computing, formal languages, nonassociative algebras, polynomial identities, computational linear algebra

Intended audience: computer algebra users, theoretical computer scientists, pure and applied mathematicians, molecular geneticists

2000 Mathematics Subject Classification: 92D20 (protein sequences, DNA sequences), 68Q70 (algebraic theory of languages and automata), 17A30 (algebras satisfying other identities), 15-04 (explicit machine computation and programs)

1 Introduction

Theoretical computer scientists working on DNA computing have developed models of computation based on the processes of molecular genetics. Much of this work is devoted to the study of biologically motivated operations (“bio-operations”) on formal languages. A detailed study of families of languages defined by bio-operations appears in Daley, Kari and McQuillan [4]. For a general survey of DNA computing see Daley and Kari [5].

This paper discusses linearized versions of the operations of synchronized insertion and simplified insertion on words. It describes Maple programs which search for nonassociative polynomial identities satisfied by these operations. Synchronized insertion satisfies no identity of degree ≤ 5 , but simplified insertion satisfies the left-symmetric identity in degree 3.

Left-symmetric algebras were originally introduced in the early 1960s in connection with differential geometry and Lie group actions (Vinberg [12]). A few years later, they also appeared in the theory of cohomology of associative

*For a longer version see: math.usask.ca/~bremner/research/papers/index.html

[†]bremner@math.usask.ca

algebras (Gerstenhaber [7]). They have also been studied from a purely algebraic point of view (Kleinfeld [9]). More recently, they have received attention in connection with operads, combinatorics of trees, and Hopf algebras (Chapoton and Livernet [3], Dzhumadil'daev and Löffwall [6], Aguiar [1]). Unlike most nonassociative algebras, free algebras in the variety of left-symmetric algebras have a basis which is easily described in terms of nonassociative words (Segal [11]).

Nonassociative algebras have previously been applied to *population* genetics in the theory of *genetic algebras* (Reed [10]). The present paper seems to be the first attempt to investigate nonassociative structures motivated by the processes of *molecular* genetics.

2 Preliminaries on formal languages and free associative algebras

2.1 Languages and monoids

We consider a finite non-empty set S called the **alphabet**. A **word** over S is a finite string $w = a_1 a_2 \cdots a_p$ where $p \geq 0$ and $a_i \in S$ ($1 \leq i \leq p$). For $p = 0$ we have the **empty word** denoted 1. We define the **length** (or **degree**) of w by $|w| = p$. We write $M(S)$ for the set of all words; any subset of $M(S)$ is called a **language**. The operation of **concatenation** sending (x, y) to xy is an associative binary operation on $M(S)$; with this operation, $M(S)$ becomes the **free monoid** generated by S .

2.2 Polynomials and algebras

We consider finite linear combinations of elements of $M(S)$ with coefficients in a semiring R :

$$\sum_{i=1}^n c_i w_i \quad (c_i \in R)$$

For simplicity we assume that $R = \mathbb{Q}$, the field of rational numbers. We write $A = A(S, \mathbb{Q})$ for the set of all such formal noncommutative polynomials. A finite set $L \subset M(S)$ can be represented as an element of A by taking the sum of the elements of S . We can also let L be a multiset, in which case the coefficient of a word is the multiplicity of the word.

We define a distributive multiplication on A by the bilinear extension of concatenation:

$$\left(\sum_{i=1}^m c_i x_i \right) \left(\sum_{j=1}^n d_j y_j \right) = \sum_{i=1}^m \sum_{j=1}^n c_i d_j x_i y_j$$

With this operation A becomes the **free associative algebra** generated by the alphabet S over \mathbb{Q} .

Using elements of the free associative algebra $A(S, \mathbb{Q})$ instead of subsets of the free monoid $M(S)$ means that we are thinking of a finite language L as a formal noncommutative polynomial rather than as a set. This allows us to describe algebraic properties of a language operation (such as the polynomial identities satisfied by the operation) which are not easily expressed in terms of monoids.

3 Synchronized insertion

3.1 Set-theoretic definition

In theoretical computer science, language operations are set-valued operations on words. For any words $x, y \in M(S)$ we define synchronized insertion $x \rightrightarrows y$ as follows:

$$x \rightrightarrows y = \{ y'x'x''x'y'' \mid x = x'x''; y = y'x'y''; x', x'', y', y'' \in M(S); x' \neq 1 \}$$

This operation is the left-right reversal of the operation studied by Daley, Kari and McQuillan [4] (Definition 1, page 54). Note that x' must be nonempty, but that x'', y', y'' may be empty. That is, we search for occurrences of nonempty left factors of x in y , and insert x to the left of each occurrence. Strictly speaking, the product $x \rightrightarrows y$ is a multiset, since the same word $y'x'x''x'y''$ may occur more than once for different factorizations of x and y .

3.2 Linearized definition

When we translate this operation to the free associative algebra, we must be careful to count multiplicities to get the right coefficient. Given two words $x = x_1 \cdots x_p$ and $y = y_1 \cdots y_q$ with $x_i, y_j \in S$, we define

$$s(x, y) = k \text{ when } x_i = y_i \text{ for } 1 \leq i \leq k \text{ but } x_{k+1} \neq y_{k+1} \text{ (or } k+1 > \min(p, q))$$

That is, $s(x, y)$ counts the number of letters to which x and y agree starting from the left. The number of times a nonempty left factor of x occurs starting at position j of y is $s(x, y_j \cdots y_q)$. If we sum the words in the set-theoretic definition of $x \rightrightarrows y$ and count multiplicities then we obtain the linearized version of synchronized insertion:

$$x \rightrightarrows y = \sum_{j=1}^q s(x, y_j \cdots y_q) y_1 \cdots y_{j-1} x y_j \cdots y_q$$

Here is an example: Let $S = \{a, b, c\}$ and set

$$x = abc, \quad y = aababcaba$$

Then we obtain (with the insertions of x underlined)

$$\begin{aligned} x \rightrightarrows y = & \underline{ab}caababcaba + 2 \underline{a}abcbabcaba + 3 \underline{aab}abcabcaba + 2 \underline{aababc}abcaba \\ & + \underline{aababcab}abca \end{aligned}$$

3.3 One generator

The simplest special case occurs when the alphabet consists of a single letter: $S = \{a\}$. We write $x = a^p$ instead of $x = a \cdots a$ (p factors). The formula for synchronized insertion becomes

$$a^p \rightrightarrows a^q = \sum_{j=1}^q s(a^p, a^{q-j+1}) a^{p+q}$$

Proposition 1. *The structure constants for the algebra of synchronized insertion on one generator are:*

$$a^p \rightrightarrows a^q = c(p, q) a^{p+q} \quad \text{where} \quad c(p, q) = \begin{cases} \frac{1}{2}p(2q - p + 1) & \text{if } p < q \\ \frac{1}{2}q(q + 1) & \text{if } p \geq q \end{cases}$$

Proof. For words on a single letter we have

$$s(a^p, a^{q-j+1}) = \min(p, q - j + 1) = \begin{cases} p & \text{if } j \leq q - p \\ q - j + 1 & \text{if } j \geq q - p + 1 \end{cases}$$

So we get

$$\sum_{j=1}^q s(a^p, a^{q-j+1}) = \begin{cases} \sum_{j=1}^{q-p} p + \sum_{j=q-p+1}^q (q - j + 1) & \text{if } p < q \\ \sum_{j=1}^q (q - j + 1) & \text{if } p \geq q \end{cases}$$

which simplifies to the stated result. \square

3.4 Polynomial identities for synchronized insertion on one generator

The operation $x \rightrightarrows y$ is noncommutative and nonassociative. Since the coefficient field \mathbb{Q} has characteristic 0, we may assume that any polynomial identity is equivalent to a finite set of homogeneous multilinear identities (Zhevlakov et al. [13], Chapter 1). Here *homogeneous* means that every term in the identity has the same degree, and *multilinear* means that every letter occurs exactly once in each term.

Lemma 2. *If a nonassociative operation satisfies a nonzero polynomial identity in degree n over \mathbb{Q} , then it satisfies a nonzero polynomial identity in degree n over the field \mathbb{F}_p for any prime p .*

Proof. We can write the polynomial identity of degree n over \mathbb{Q} in the form

$$I = \sum_{i=1}^t c_i m_i(x_1, \dots, x_n)$$

$$\begin{array}{cccccc}
x \rightrightarrows (y \rightrightarrows z) & x \rightrightarrows (z \rightrightarrows y) & y \rightrightarrows (x \rightrightarrows z) & y \rightrightarrows (z \rightrightarrows x) & z \rightrightarrows (x \rightrightarrows y) & z \rightrightarrows (y \rightrightarrows x) \\
(x \rightrightarrows y) \rightrightarrows z & (x \rightrightarrows z) \rightrightarrows y & (y \rightrightarrows x) \rightrightarrows z & (y \rightrightarrows z) \rightrightarrows x & (z \rightrightarrows x) \rightrightarrows y & (z \rightrightarrows y) \rightrightarrows x
\end{array}$$

Table 1: The 12 association types in degree 3

where $m_i(x_1, \dots, x_n)$ are monomials, each consisting of the variables x_1, \dots, x_n in some permutation with some arrangement of parentheses. Each coefficient has the form $c_i = a_i/b_i$ where a_i and b_i are relatively prime integers. Write d for the least common multiple of the denominators b_i . Then dI is a nonzero polynomial identity over \mathbb{Z} (the ring of integers) which is satisfied by the operation. Write f for the greatest common factor of the coefficients dc_i of dI . Then dI/f is a nonzero polynomial identity over \mathbb{Z} which is satisfied by the operation; furthermore, dI/f reduces modulo p to a nonzero identity for any prime p . \square

Theorem 3. *Synchronized insertion on one generator does not satisfy any polynomial identity of degree ≤ 5 .*

Proof. This is a proof by exhaustive computational search. By Lemma 2, it suffices to show that the operation does not satisfy any identity of degree ≤ 5 modulo some convenient prime p . For this we use the Maple package `LinearAlgebra[Modular]` with $p = 101$.

Degrees 1 and 2. It is clear that synchronized insertion on one generator gives neither the zero algebra ($x \equiv 0$) nor a trivial algebra ($x \rightrightarrows y = 0$), and that it does not satisfy either the commutative or anticommutative identity:

$$x \rightrightarrows y - y \rightrightarrows x = 0, \quad x \rightrightarrows y + y \rightrightarrows x = 0$$

These are the only possible identities in degrees 1 and 2.

Degree 3. In degree 3, there are 2 association types:

$$x \rightrightarrows (y \rightrightarrows z), \quad (x \rightrightarrows y) \rightrightarrows z$$

For each association type, we consider all 6 permutations of the variables. This gives a total of 12 nonassociative monomials (Table 1). These 12 monomials form a basis for the vector space of all possible (homogeneous multilinear) nonassociative polynomial identities of degree 3.

We choose a maximum exponent M (for example, $M = 100$). Using the Maple `rand()` function, we generate random numbers e, f, g in the range $0, \dots, M - 1$ corresponding to the elements $x = a^e, y = a^f, z = a^g$. When we evaluate any of the 12 monomials on these random elements using the structure constants of Proposition 1, the result will be a scalar multiple of an element $w = a^h$ with $0 \leq h \leq 3M - 3$.

We initialize a matrix of size $(12 + 3M) \times 12$ using the Maple command `Create`. We think of this ‘‘expansion matrix’’ as consisting of a matrix of size 12×12 stacked on top of a matrix of size $3M \times 12$. We now perform the follow iteration:

1. Generate 3 random elements $x = a^e, y = a^f, z = a^g$.
2. Evaluate each of the 12 monomials on the 3 random elements. For monomial j we obtain a result ca^h where $c \in \mathbb{F}_p$ and $h \geq 0$. Insert the coefficient c into column j and row $13 + h$.
3. After all 12 monomials have been evaluated and the matrix entries stored, each row of the lower $3M \times 12$ matrix contains a linear relation on the coefficients of any polynomial identity (linear combination of monomials) satisfied by synchronized insertion. (Row $13 + h$ of this matrix expresses the constraint that the coefficient of a^h must be 0 when any polynomial identity is fully evaluated.)
4. Compute the row canonical form of the matrix using the Maple command `RowReduce`. The lower $3M \times 12$ matrix is now the zero matrix, so we can repeat the fill and reduce process.

As we repeat the iteration, the rank of the matrix is non-decreasing, and at some point the rank stabilizes. If the rank reaches 12, the nullspace of the matrix is 0, and we have shown that no identity exists. (We have generated enough random counter-examples to prove that no linear combination of the 12 monomials vanishes on all possible choices of words.) If the rank stabilizes at a value below 12, we perform another 100 iterations to be confident that the maximum rank has indeed been reached. At this point, the (nonzero) nullspace of the matrix contains candidates for identities satisfied by synchronized insertion. (We need to verify these identities independently.)

For degree 3, the rank increases by 1 after each iteration until it reaches 12. This shows that there is no (non-trivial) identity of degree 3.

Degree 4. For degree 4, we have 5 association types:

$$w \rightrightarrows (x \rightrightarrows (y \rightrightarrows z)), \quad w \rightrightarrows ((x \rightrightarrows y) \rightrightarrows z), \quad (w \rightrightarrows x) \rightrightarrows (y \rightrightarrows z), \quad (w \rightrightarrows (x \rightrightarrows y)) \rightrightarrows z, \\ ((w \rightrightarrows x) \rightrightarrows y) \rightrightarrows z$$

Since each association type contains 24 permutations, we have a total of 120 monomials. The expansion matrix has size $(120 + 4M) \times 120$. After 125 iterations, the rank reaches 120, and so there is no identity in degree 4.

Degree 5. For degree 5, we have 14 association types and 120 permutations, for a total of 1680 monomials. The matrix has size $(1680 + 5M) \times 1680$. After 1776 iterations, the rank reaches 1680, and so there is no identity in degree 5. \square

Corollary 4. *Synchronized insertion on any number of generators does not satisfy any polynomial identity in degree ≤ 5 .*

Proof. Since the algebra on one generator is contained in the algebra on any number of generators, and a subalgebra satisfies all the identities of the larger algebra, this follows immediately from Theorem 3. \square

4 Simplified insertion

In view of the negative result in Theorem 3, we now consider a simplified insertion operation which has more interesting mathematical properties. Given $x, y \in M(S)$ we consider *all* insertions of x into y :

$$x \rightarrow y = \{ y'xy'' \mid y = y'y'', y', y'' \in M(S) \}$$

Note that we are allowing both y' and y'' to be empty. (This operation is called “normal insertion” in Kari [8].) The linearized form of this operation is the sum of all insertions:

$$x \rightarrow y = \sum_{j=0}^q y_1 \cdots y_j x y_{j+1} \cdots y_q,$$

where $y = y_1 \cdots y_q$ and $y_j \in S$. We call this the **simplified insertion** of x into y ; it extends by linearity to another nonassociative operation on the free associative algebra $A(S, \mathbb{Q})$.

4.1 Polynomial identities of degree 3 for simplified insertion

The **associator** for the simplified insertion operation is defined as usual by

$$(x, y, z) = (x \rightarrow y) \rightarrow z - x \rightarrow (y \rightarrow z)$$

The **left-symmetric identity**

$$(x, y, z) = (y, x, z)$$

expresses the invariance of the associator under the transposition of the left and middle arguments.

Theorem 5. *Simplified insertion satisfies the left-symmetric identity.*

Proof. Gerstenhaber [7], Theorem 2, page 276. □

Proposition 6. *Every identity of degree ≤ 3 satisfied by simplified insertion follows from the left-symmetric identity.*

Proof. It is clear that simplified insertion gives neither the zero algebra nor a trivial algebra, nor a commutative algebra nor an anticommutative algebra.

We now show that every identity of degree 3 satisfied by simplified insertion follows from the left-symmetric identity. We take $S = \{a, b, c, d, e, f\}$ and we consider 3 distinct words of length 2:

$$x = ab, \quad y = cd, \quad z = ef.$$

We use Maple to determine all relations between the 12 possible nonassociative monomials (see (1), but here we are using simplified insertion). When we expand

any of these monomials we obtain a sum over certain shuffles of x, y and z . For example,

$$\begin{aligned}
ab \rightarrow (cd \rightarrow ef) &= ab \rightarrow (cdef + ecdf + efcd) \\
&= abcdef + cabdef + cdabef + cdeabf + cdefab \\
&\quad + abecdf + eabcdf + ecabdf + ecdabf + ecdfab \\
&\quad + abefcd + eabfcd + efabcd + efcabd + efcadb
\end{aligned}$$

There are altogether $6!/(2!)^3 = 90$ shuffles of 3 words of length 2. We create a matrix of size 90×12 in which the columns are labelled by the 12 nonassociative monomials and the rows are labelled by the 90 shuffles. The (i, j) entry of this expansion matrix contains the coefficient of the i -th shuffle in the expansion of the j -th monomial. The nullspace contains the relations satisfied by the expanded monomials. We calculate that the expansion matrix has rank 9 and hence its nullspace has dimension 3. A basis of the nullspace consists of the 3 distinct permutations of the left-symmetric identity:

$$(x, y, z) - (y, x, z), \quad (y, z, x) - (z, y, x), \quad (z, x, y) - (x, z, y)$$

The space of identities satisfied by arbitrary words x, y, z must be a subspace of the identities satisfied by words of length 2. Since we already know that the left-symmetric identity is satisfied by arbitrary words, the proof is complete. \square

5 Free left-symmetric algebras

Now that we know that simplified insertion satisfies the left-symmetric identity, we can reduce the size of the matrices which occur in our computational searches by using only monomials from the free left-symmetric algebra rather than all nonassociative monomials. A basis for the free left-symmetric algebra on a set of generators S over a field \mathbb{F} was constructed by Segal [11]. We briefly review the definitions necessary for the statement of Segal's theorem.

We choose a total ordering on the alphabet S , denoted $<$. The set $W(n)$ of all nonassociative words w of length $|w| = n \geq 1$ is defined inductively:

$$W(1) = S, \quad W(n) = \{ (uv) : |u| + |v| = n \}, \quad W = \bigcup_{n \geq 1} W(n)$$

For example,

$$\begin{aligned}
W(2) &= \{ ab : a, b \in S \}, & W(3) &= \{ a(bc), (ab)c : a, b, c \in S \} \\
W(4) &= \{ a(b(cd)), a((bc)d), (ab)(cd), (a(bc))d, ((ab)c)d : a, b, c, d \in S \}
\end{aligned}$$

The number of words in $W(n)$ is

$$C_n |S|^n \quad \text{where} \quad C_n = \frac{1}{n} \binom{2n-2}{n-1} \quad (\text{Catalan number})$$

The free nonassociative algebra $F(S)$ over \mathbb{F} has basis W and distributive product induced by concatenation of nonassociative words.

In $F(S)$ we consider the T -ideal I generated by the left-symmetric identity. By a T -ideal, we mean an ideal I for which $f(I) \subseteq I$ for every endomorphism $f: F(S) \rightarrow F(S)$. In other words, I is generated (in the usual sense) by the elements obtained by all possible substitutions of words in W for the arguments of the left-symmetric identity. Then the free left-symmetric algebra is the quotient $L(S) = F(S)/I$.

For a word $w \in W$ we define the left and right factors $\lambda(w)$ and $\rho(w)$:

$$\begin{aligned} \lambda(w) &= w, \rho(w) = 1 && \text{if } w \in S, \\ \lambda(w) &= u, \rho(w) = v && \text{if } w = uv \in W-S \end{aligned}$$

We define the parts of a word $w \in W$ as follows:

- if $w \in S$ then the only part of w is w itself;
- if $w \in W-S$ then the parts of w are w itself, the parts of $\lambda(w)$, and the parts of $\rho(w)$.

We extend the total ordering on S to all of W :

$$\begin{aligned} u < v &\text{ if } |u| < |v|, \text{ or} \\ &\text{ if } |u| = |v| \text{ and } \lambda(u) < \lambda(v), \text{ or} \\ &\text{ if } |u| = |v| \text{ and } \lambda(u) = \lambda(v) \text{ and } \rho(u) < \rho(v) \end{aligned}$$

Definition 7. A word w is **bad** if it has the form $w = r(st)$ with $r, s, t \in W$ and $r < s$. A word is **reduced** if it has no bad parts.

We write an overline for the natural surjective homomorphism from the free nonassociative algebra to the free left-symmetric algebra:

$$\overline{}: F(S) \rightarrow F(S)/I = L(S)$$

Let R denote the set of all reduced words.

Theorem 8. (Segal [11]) *The restriction of $\overline{}$ to R is injective, and the free left-symmetric algebra $L(S)$ is an \mathbb{F} -vector space with basis \overline{R} .*

Since we are concerned with multilinear polynomial identities satisfied by left-symmetric algebras, we only need to consider the multilinear subspace of the free left-symmetric algebra $L(S)$. That is, for each n we consider an alphabet S of size n , and the subspace of $L(S)$ spanned by the set $R(n)$ of reduced words of degree n in which each element of S occurs exactly once. This set has a natural graph-theoretical interpretation.

Proposition 9. *In degree n , the number of multilinear reduced words is n^{n-1} .*

Proof. A basis for the free left-symmetric algebra is given in terms of rooted trees by Chapoton and Livernet [3] and Dzhumadil'daev and L\"ofwall [6]. A rooted tree on n vertices is an acyclic connected graph with vertex set $\{1, \dots, n\}$ and with one specified vertex (the root). By Cayley's Theorem, the number of (unrooted) trees with this vertex set is n^{n-2} , and so the number of rooted trees is n^{n-1} (there are n distinct choices for the root). This is therefore also the number of multilinear reduced words in degree n . \square

Corollary 10. *The dimension of the multilinear subspace $R(n)$ of the free left-symmetric algebra on n generators is n^{n-1} .*

6 Polynomial identities of degrees 4 and 5 for simplified insertion

In this section we describe Maple procedures that implement an algorithm for an exhaustive search over all possible polynomial identities of degree 4 and 5 for a left-symmetric operation.

We start by initializing the degree n and the alphabet $S = \{1, \dots, n\}$. The procedures in the first group do basic computations on nonassociative words:

1. Generate recursively the set $W(n)$ of nonassociative words of degree n .
2. Find the degree of a nonassociative word, and its left and right factors.
3. Determine whether two nonassociative words x and y satisfy $x < y$ in Segal's total ordering.
4. Determine if a nonassociative word is "bad" in Segal's sense.
5. Generate recursively all multilinear nonassociative words which are "reduced" in Segal's sense.

The second group of procedures converts among different data structures which represent nonassociative words. For example, the word $((ab)c)(de)$ can be represented either as a nested list $[[[1, 2], 3], [4, 5]]$ or as a flat list $[1, 2, 0, 3, 0, 4, 5, 0, 0]$ (where zero is a postfix operation symbol), or as its underlying permutation together with a number giving the association type: $[[1, 2, 3, 4, 5], 9]$. Here we order the association types in degree 5 as follows:

$$\begin{aligned} & a(b(c(de))), \quad a(b((cd)e)), \quad a((bc)(de)), \quad a((b(cd))e), \quad a(((bc)d)e), \\ & (ab)(c(de)), \quad (ab)((cd)e), \quad (a(bc))(de), \quad ((ab)c)(de), \\ & (a(b(cd)))e, \quad (a((bc)d))e, \quad ((ab)(cd))e, \quad (a(bc))(de), \quad ((ab)c)(de) \end{aligned}$$

The third group of procedures implements the operation of simplified insertion of associative words:

1. Insert one word (or list of words with coefficients) into another word (or list of words with coefficients).

2. Evaluate the association types for nonassociative words of degrees 4 and 5, using the operation of simplified insertion.
3. Compute the expansions of the reduced words when the variables are set equal to words of length 2 in distinct letters.
4. Convert each shuffle appearing in the expansions of the reduced words into a permutation of a multiset.
5. Compute the lexicographical index of each multiset to determine the row of the expansion matrix in which the coefficient will be stored.

The main loop uses the Maple package `LinearAlgebra[Modular]` to compute the row canonical form of the expansion matrix. The nullspace of the expansion matrix contains the linear relations satisfied by the expansions of the reduced words. If this nullspace is zero, we know that no polynomial identity is satisfied by the operation.

Theorem 11. *Every polynomial identity of degree 4 satisfied by simplified insertion follows from the left-symmetric identity.*

Proof. By Corollary 10, in degree 4 the number of reduced words is $4^3 = 64$. These words form a basis for the multilinear subspace of the free left-symmetric algebra. (They are displayed in Table 2 in order of increasing association type with the operation symbols omitted.) Any identity of degree 4 satisfied by simplified insertion will be a linear combination of these 64 monomials. Any identity must be satisfied by words of length 2, so we first look for linear relations on the monomials in the special case

$$w = ab, \quad x = cd, \quad y = ef, \quad z = gh.$$

When any of the 64 monomials is expanded, every term will be a shuffle of the four words of length 2. The number of such shuffles is $8!/(2!)^4 = 2520$. So in this case the expansion matrix has size 2520×64 . The (i, j) entry of this matrix contains the coefficient of the i -th shuffle in the expansion of the j -th left-symmetric monomial. Using Maple, we generate this matrix and determine that it has rank 64; hence its nullspace is $\{0\}$. It follows that there are no identities in degree 4 (except those which are trivial consequences of the left-symmetric identity). \square

Proposition 12. *The nullspace of the expansion matrix for 5 words of length 2 on 10 distinct letters has dimension 5.*

Proof. The basic ideas are the same as in the proof of Theorem 11, although the expansion matrix is larger. In degree 5 the number of reduced words is $5^4 = 625$. Any identity of degree 5 satisfied by insertion will be a linear combination of these 625 monomials. Any identity must be satisfied by words of length 2, so we first look for linear relations on the monomials in the special case

$$v = ab, \quad w = cd, \quad x = ef, \quad y = gh, \quad z = ij.$$

$y(x(wz))$	$z(x(wy))$	$z(y(wx))$	$z(y(xw))$	$(wx)(yz)$	$(wx)(zy)$
$(wy)(xz)$	$(wy)(zx)$	$(wz)(xy)$	$(wz)(yx)$	$(xw)(yz)$	$(xw)(zy)$
$(xy)(wz)$	$(xy)(zw)$	$(xz)(wy)$	$(xz)(yw)$	$(yw)(xz)$	$(yw)(zx)$
$(yx)(wz)$	$(yx)(zw)$	$(yz)(wx)$	$(yz)(xw)$	$(zw)(xy)$	$(zw)(yx)$
$(zx)(wy)$	$(zx)(yw)$	$(zy)(wx)$	$(zy)(xw)$	$(x(wy))z$	$(x(wz))y$
$(y(wx))z$	$(y(wz))x$	$(y(xw))z$	$(y(xz))w$	$(z(wx))y$	$(z(wy))x$
$(z(xw))y$	$(z(xy))w$	$(z(yw))x$	$(z(yx))w$	$((wx)y)z$	$((wx)z)y$
$((wy)x)z$	$((wy)z)x$	$((wz)x)y$	$((wz)y)x$	$((xw)y)z$	$((xw)z)y$
$((xy)w)z$	$((xy)z)w$	$((xz)w)y$	$((xz)y)w$	$((yw)x)z$	$((yw)z)x$
$((yx)w)z$	$((yx)z)w$	$((yz)w)x$	$((yz)x)w$	$((zw)x)y$	$((zw)y)x$
$((zx)w)y$	$((zx)y)w$	$((zy)w)x$	$((zy)x)w$		

Table 2: The 64 multilinear reduced words in degree 4

When any of the 625 monomials is expanded, every term will be a shuffle of the 5 words of length 2. The number of such shuffles is $10!/(2!)^5 = 113400$. So in this case the expansion matrix has size 113400×625 . The (i, j) entry of this matrix contains the coefficient of the i -th shuffle in the expansion of the j -th left-symmetric monomial. In this case the matrix is too large to be processed all at once, so we use the following method to reduce the size of the computations:

1. Generate the expansion of each monomial; this gives a list of shuffles with coefficients.
2. Break down the 113400 shuffles into 945 groups of 120 shuffles. Initialize a matrix of size 745×625 , and regard this as a matrix of size 625×625 stacked on top of a matrix of size 120×625 .
3. For each of the 945 groups of shuffles, perform the following steps:
 - (a) Read the coefficients of the shuffles into the bottom 120 rows of the expansion matrix. That is, for each iteration $i = 1, \dots, 945$ we consider the shuffles numbered $120(i - 1)$ through $120i$; the coefficient of the k -th such shuffle in the expansion of the j -th reduced word is stored in position $(625 + k, j)$ of the matrix.
 - (b) Compute the row-canonical form of the expansion matrix. Now the bottom 120 rows are zero, so we can repeat the fill and reduce process.

The rank stabilizes at 620 after iteration 464; hence the nullspace of the matrix has dimension 5. \square

A basis of the nullspace of the expansion matrix in degree 5 may be described as follows. It consists of a multilinear homogeneous left-symmetric polynomial $I(v, w, x, y, z)$ and the 4 equivalent forms of this polynomial obtained by applying the permutations

$$(vw), \quad (vwx), \quad (vyxw), \quad (vzyxw).$$

$$\begin{aligned}
I(v, w, x, y, z) = & \\
& -z(y(x(wv))) + (xw)(z(yv)) + (yw)(z(xv)) + (yx)(z(wv)) + (zw)(y(xv)) \\
& + (zx)(y(wv)) + (zy)(x(wv)) - (zw)((yx)v) - (zx)((yw)v) - (zy)((xw)v) \\
& + (y(xw))(zv) + (z(xw))(yv) + (z(yw))(xv) + (z(yx))(wv) - ((xw)y)(zv) \\
& - ((xw)z)(yv) - ((yw)x)(zv) - ((yw)z)(xv) - ((yx)w)(zv) - ((yx)z)(wv) \\
& - ((zw)x)(yv) - ((zw)y)(xv) - ((zx)w)(yv) - ((zx)y)(wv) - ((zy)w)(xv) \\
& - ((zy)x)(wv) + (z(y(xw)))v - ((xw)(zy))v - ((yw)(zx))v - ((yx)(zw))v \\
& - ((y(xw)z)v - ((z(xw)y)v - ((z(yw)x)v - ((z(yx)w)v + (((xw)y)z)v \\
& + (((xw)z)y)v + (((yw)x)z)v + (((yw)z)x)v + (((yx)w)z)v + (((yx)z)w)v \\
& + (((zw)x)y)v + (((zw)y)x)v + (((zx)w)y)v + (((zx)y)w)v + (((zy)w)x)v \\
& + (((zy)x)w)v
\end{aligned}$$

Table 3: The 46-term relation in degree 5

The polynomial has 46 terms and is displayed in Table 3. The terms are listed in order of association type, and within each association type by lexicographical order of permutation. There is a certain amount of symmetry in this expression, but I could not find a natural way to write it more compactly. This polynomial is satisfied (at least over the field with 101 elements) by simplified insertion on 5 disjoint words of length 2 in 10 distinct variables. It is an open question whether this is a polynomial *identity* for simplified insertion (that is, $I(v, w, x, y, z) = 0$ holds for arbitrary words) or merely a curious *relation* satisfied by words of length 2.

Some of the author's earlier Maple computations with free left-symmetric algebras appear in Bremner [2].

7 Acknowledgements

I thank Mark Daley, Mark Eramian and Michael Horsch (Department of Computer Science, University of Saskatchewan) for information about DNA computing and formal languages. Preliminary versions of this talk were presented on February 24, 2005 in the seminar *Álgebras de Lie e de Jordan e suas Representações*, Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil (I thank Professor I. P. Shestakov for the invitation to speak) and on April 16, 2005 at the 2005 Saskatchewan Mathematics Mini-Meeting at the University of Regina. This research was supported by a Discovery Grant from NSERC (Natural Sciences and Engineering Research Council of Canada).

References

- [1] M. Aguiar, *Infinitesimal bialgebras, pre-Lie and dendriform algebras*, Hopf algebras, pages 1–33, Lecture Notes in Pure and Applied Mathematics **237**, Marcel Dekker, 2004. MR2051728 (2005c:16053)

- [2] M. R. Bremner, *Additive structure of free left-symmetric and assosymmetric rings*, International Journal of Mathematics, Game Theory and Algebra **12**, 1 (2002) 23–37. MR1904877 (2003b:17004)
- [3] F. Chapoton and Muriel Livernet, *Pre-Lie algebras and the rooted trees operad*, International Mathematics Research Notices **8** (2001) 395–408. MR1827084 (2002e:17003)
- [4] M. Daley, Lila Kari and I. McQuillan, *Families of languages defined by ciliate bio-operations*, Theoretical Computer Science **320**, 1 (2004) 51–69. MR2060183
- [5] M. J. Daley and Lila Kari, *DNA computing: Models and implementations*, Comments on Theoretical Biology **7** (2002) 177–198.
- [6] A. Dzhumadil'daev and C. Löfwall, *Trees, free right-symmetric algebras, free Novikov algebras and identities*, Homology Homotopy and Applications **4**, 2 (2002) 165–190 (electronic). MR1918188 (2003h:17003)
- [7] M. Gerstenhaber, *The cohomology structure of an associative ring*, Annals of Mathematics **78** (1963) 267–288. MR0161898 (28 #5102)
- [8] Lila Kari, *On insertion and deletion in formal languages*, 192 pages, Dissertation, Department of Mathematics, University of Turku, Finland, 1991. MR1219089 (94j:68158)
- [9] E. Kleinfeld, *On rings satisfying $(x, y, z) = (x, z, y)$* , Algebras Groups and Geometries **4**, 2 (1987) 129–138. MR914169 (89a:17001)
- [10] Mary Lynn Reed, *Algebraic structure of genetic inheritance*, Bulletin of the American Mathematical Society **34**, 2 (1997) 107–130. MR1414973 (98e:17043)
- [11] D. Segal, *Free left-symmetric algebras and an analogue of the Poincaré-Birkhoff-Witt theorem*, Journal of Algebra **164** (1994) 750–772. MR1272113 (95f:17008)
- [12] E. B. Vinberg, *The theory of homogeneous convex cones*, Trudy Moskovskogo Matematicheskogo Obshchestva **12** (1963) 303–358, translated in Transactions of the Moscow Mathematical Society. MR0158414 (28 #1637)
- [13] K. A. Zhevlakov, A. M. Slinko, I. P. Shestakov and A. I. Shirshov, *Rings that are nearly associative*, 371 pages, translated from the Russian by Harry F. Smith, Academic Press, 1982. MR0668355 (83i:17001)