AVOIDING SUBSTRINGS IN COMPOSITIONS

Silvia Heubach
Dept. of Mathematics, California State University Los Angeles, Los Angeles, CA 90032, USA
sheubac@calstatela.edu

Sergey Kitaev[1]
The Mathematics Institute, School of Computer Science, Reykjavik University, 103 Reykjavik, Iceland
sergey@ru.is

Abstract

A classical result by Guibas and Odlyzko obtained in 1981 gives the generating function for the number of strings that avoid a given set of substrings with the property that no substring is contained in any of the others. In this paper, we give an analogue of this result for the enumeration of compositions that avoid a given set of prohibited substrings, subject to the compositions' length (number of parts) and weight. We also give examples of families of strings to be avoided that allow for an explicit formula for the generating function. Our results extend recent results by Myers on avoidance of strings in compositions subject to weight, but not length.
Keywords: Compositions, strings, avoidance, (auto)correlation, generating functions
2000 Mathematics Subject Classification: 05A05, 05A15

1. Introduction

In 1981, Guibas and Odlyzko [1] obtained the generating function for the number of strings avoiding a given set of prohibited substrings and then applied this result to non-transitive games. (A string $s = s_1s_2\cdots s_m$ *contains a substring* $b_1b_2\cdots b_k$ of length k if there is an index i such that $s_is_{i+1}\cdots s_{i+k-1} = b_1b_2\cdots b_k$. Otherwise, we say that s *avoids* the substring $b_1b_2\cdots b_k$.) A detailed derivation of this generating function and related results in the binary case was later given by Winterfjord in

[1]The work presented here was supported by grant no. 090038011 from the Icelandic Research Fund.

his Master's thesis [5]. The basic idea in the derivation of the generating function is the notion of the correlation between two strings and being able to enumerate the strings avoiding the set of substrings in two different ways. Let $S_1 = a_0a_1 \dots a_{m-1}$ and $S_2 = b_0b_1 \dots b_{\ell-1}$ be two strings of lengths m and ℓ, respectively, over the alphabet $[n] = \{1, 2, \dots, n\}$. The *correlation* $c_{12} = c_0c_1 \dots c_{m-1}$ is the binary string defined as follows:

$m \le \ell$: For $0 \le j \le m-1$, $c_j = 1$ if $a_i = b_{\ell-m+i+j}$ for $i = 0, 1, \dots, m-j-1$, and $c_j = 0$ otherwise;

$m > \ell$: For $0 \le j \le m-\ell$, $c_j = 1$ if $b_i = a_{m-\ell+i-j}$ for $i = 0, 1, \dots, \ell-1$, and $c_j = 0$ otherwise; for $m-\ell+1 \le j \le m-1$, $c_j = 1$ if $a_i = b_{\ell-m+i+j}$ for $i = 0, 1, \dots, m-j-1$ and $c_j = 0$ otherwise.

In plain English, this means that c_j is equal to 1 if and only if the letters in the overlap of string S_1 and string S_2, shifted (or offset) by j positions to the left, agree, as illustrated in Figure 1.

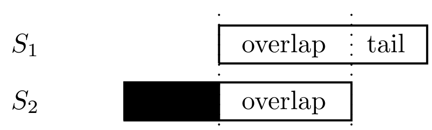

FIGURE 1. Comparing strings S_1 and S_2.

For example, if $S_1 = 112$ and $S_2 = 1211$, then $c_{12} = 011$ and $c_{21} = 0010$, as depicted below:

j							c_j
				1	1	2	
0			1	2	1	1	0
1		1	2	1	1		1
2	1	2	1	1			1

j							c_j
			1	2	1	1	
0				1	1	2	0
1			1	1	2		0
2		1	1	2			1
3	1	1	2				0

In general $c_{12} \ne c_{21}$ and, unless the strings are of the same lengths, the correlations will have different lengths. The *autocorrelation* of a string or word S_1 is just c_{11}, the correlation of S_1 with itself. For instance, if $S_1 = 112$ then $c_{11} = 100$. It is convenient to associate a *correlation polynomial* $C_{12}(q) = c_0 + c_1q + \dots + c_{k-1}q^{k-1}$ with the correlation $c_{12} = c_0c_1 \dots c_{k-1}$. This correlation polynomial is the generating function for the number of letters in the *tail*, the portion that is to the right of the overlap in the substring S_1, as illustrated in Figure 1.

We now state the general result given by Guibas and Odlyzko [1] in the form given (for the special case of binary strings) in Winterfjord [5, Th. 24].

Theorem 1.1. *The generating function $S(q)$ for the number of strings or words on the alphabet $[n]$ that avoid the substrings $S_1, \dots, S_k$ of lengths $\ell_1, \dots, \ell_k$ respectively, none included in any other, is given by*

$$S(q) = \frac{\begin{vmatrix} -C_{11}(q) & \cdots & -C_{1k}(q) \\ \vdots & \ddots & \vdots \\ -C_{k1}(q) & \cdots & -C_{kk}(q) \end{vmatrix}}{\begin{vmatrix} (1-nq) & 1 & \cdots & 1 \\ q^{\ell_1} & -C_{11}(q) & \cdots & -C_{1k}(q) \\ \vdots & \vdots & \ddots & \vdots \\ q^{\ell_k} & -C_{k1}(q) & \cdots & -C_{kk}(q) \end{vmatrix}}, \tag{1.1}$$

where $C_{ij}(q)$ is the correlation polynomial for the substrings S_i and S_j.

Unfortunately, the approach by Guibas and Odlyzko is not applicable to permutations and subpermutations, or when *patterns* (as opposed to strings) are to be avoided. However, the approach generalizes to *compositions* avoiding a set of prohibited substings, and we will derive a formula for the most general case that is an analogue of the formula by Guibas and Odlyzko[2]. This generalization to compositions follows the current interest in compositions which have been studied from different perspectives in the literature, mostly from the viewpoint of pattern avoidance (see [2] and references therein). Our results add a facet to this research.

Let $\mathbb{N}$ be the set of natural numbers. A *composition* $\sigma = \sigma_1 \cdots \sigma_m$ of $n \in \mathbb{N}$ is an ordered collection (or string) of one or more positive integers whose sum, also called the composition's *weight* $w(\sigma)$, is n. The number of *summands* or *letters*, namely m, is called the number of *parts* or *length* of the composition and is denoted by $\ell(\sigma)$. The main result of this paper is the derivation of the generating function

$$G(x,q) = G(S_1, \dots, S_k; x, q) = \sum_{\sigma} x^{w(\sigma)} q^{\ell(\sigma)}$$

where the sum is taken over all compositions with parts in $\mathbb{N}$ simultaneously avoiding the prohibited substrings S_i, $i = 1, \dots, k$, where none of the substrings is included in any other. We state and prove this result in Section 2 and then give applications of our result for families of prohibited substrings in Section 3.

---

[2]As the matter of fact, a recent paper by Myers [4] considers a very similar problem. However, we are able to control both length and weight in compositions, as opposed to just weight, while Myers' result is more general with respect to the alphabet considered.

2. Main result

In order to extend Theorem 1.1 to compositions, we need to adapt the correlation polynomial to also keep track of the weight in addition to the length of the tail. We therefore define the correlation polynomial for a correlation $c_{ij} = c_0 c_1 \dots c_{m-1}$ between $S_i = a_0 a_1 \dots a_{m-1}$ and S_j as

$$C_{ij}(x,q) = c_0 + c_1 x^{w(a_{m-1})} q + \cdots + c_{m-1} x^{w(a_1 a_2 a_3 \dots a_{m-1})} q^{m-1}.$$

For example, for $S_1 = 112$ and $S_2 = 1211$ considered in Section 1, $C_{12}(x,q) = x^2 q + x^3 q^2$, $C_{21}(x,q) = (xq)^2$, $C_{11}(x,q) = 1$, and $C_{22}(x,q) = 1 + x^4 q^3$. Note that since we are considering compositions, all parts are positive and therefore each term but the first one of a correlation polynomial is divisible by xq (the first term is either 0 or 1). We are now ready to state the main result.

Theorem 2.1. *The generating function $G(x,q)$ for the number of compositions that avoid the substrings $S_1, \dots, S_k$ of lengths $\ell(S_1), \dots, \ell(S_k)$ respectively, none included in any other, is given by*

$$G(x,q) = \frac{(1-x) \cdot \begin{vmatrix} -C_{11}(x,q) & \cdots & -C_{1k}(x,q) \\ \vdots & \ddots & \vdots \\ -C_{k1}(x,q) & \cdots & -C_{kk}(x,q) \end{vmatrix}}{\begin{vmatrix} 1-x(1+q) & 1-x & \cdots & 1-x \\ x^{w(S_1)} q^{\ell(S_1)} & -C_{11}(x,q) & \cdots & -C_{1k}(x,q) \\ \vdots & \vdots & \ddots & \vdots \\ x^{w(S_k)} q^{\ell(S_k)} & -C_{k1}(x,q) & \cdots & -C_{kk}(x,q) \end{vmatrix}} \tag{2.1}$$

where $C_{ij}(x,q)$ are the correlation polynomials defined above.

Proof. In finding $G(x,q)$ we adapt the arguments in [1, 5] to compositions. Let $\mathcal{A}$ denote the set of all compositions avoiding the prohibited substrings and let $\mathcal{B}_i$, for $i = 1, \dots, k$, be the set of all compositions ending with S_i but having no other occurrence of any of the prohibited substrings. A composition in $\mathcal{B}_i$ is said to *quasi-avoid* S_i. We denote the generating function corresponding to $\mathcal{B}_i$ by $\mathcal{B}_i(x,q)$ and note that $G(x,q)$ is the generating function of the set $\mathcal{A}$. Furthermore, the sets $\mathcal{A}$ and $\mathcal{B}_i$ are all pairwise disjoint as none of the substrings is included in any of the others.

We now derive recurrences for certain sets of compositions. It is well-known that we can create the unrestricted compositions of $n+1$ recursively from those of $n \geq 1$ by either increasing the last part by 1 or by appending a part 1 at the right end of the composition. However, since we want to avoid the forbidden strings, we need to be careful since each of these two operations can create a substring at the end of the newly created composition. We let $\mathcal{M}^{+1}$ denote the set obtained from a set of compositions

$\mathcal{M}$ by increasing the rightmost part of *each* non-empty composition by 1, and let $\mathcal{M} \times \{1\}$ denote the set obtained from $\mathcal{M}$ by adjoining the new rightmost part 1 to *each* composition in $\mathcal{M}$. Letting ϵ denote the empty composition, we obtain the following equation for the set of compositions that either avoid or quasi-avoid the forbidden substrings:

$$\mathcal{A} \cup \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_k = \{\epsilon\} \cup (\mathcal{A} \cup \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_k - \{\epsilon\})^{+1} \cup (\mathcal{A} \times \{1\}). \tag{2.2}$$

The expression on the right hand side follows as increasing the last part of a composition of n that either avoids all substrings or quasi-avoids one of the substrings creates a composition of $n+1$ that either avoids the forbidden substrings or quasi-avoids one of the forbidden substrings. Appending the part 1 to a composition that avoids the forbidden set either results in a composition that still avoids S or quasi-avoids some S_i. However, the same operation for a composition in $\mathcal{B}_i$ creates a composition that contains S_i and therefore no longer avoids the forbidden set of substings, so that operation is not allowed for a composition in $\mathcal{B}_i$. Increasing the last part of a composition results in an increase in the weight of the composition by 1 but no increase in the number of parts, while appending the part 1 increases both the weight and the length of the composition. Thus (2.2) can be expressed in terms of generating functions as

$$(1 - x - xq)G(x, q) + (1 - x)(B_1(x, q) + \cdots + B_k(x, q)) = 1 - x, \tag{2.3}$$

where we have used that the generating function of the union of disjoint sets is the sum of the respective generating functions, and the generating function of a Cartesian product is the product of the respective generating functions.

We now create an alternative connection between the sets $\mathcal{A}$ and $\mathcal{B}_i$. Naively, one might create the compositions in $\mathcal{B}_i$ by adjoining the string S_i to all the compositions in $\mathcal{A}$. However, this operation my create the occurrence of another substring S_j for some $j \neq i$ across the connection point, as illustrated in Figure 2.

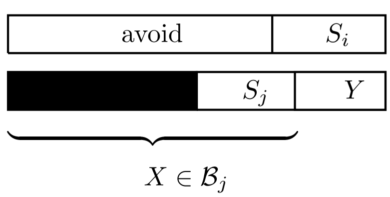

FIGURE 2. The ij-tail Y.

Note that the string Y at the end of the composition must agree with the end of the string S_i. Thus, for a composition (or string) X from $\mathcal{B}_j$, we call a string Y with $\ell(Y) \leq \ell(S_i) - 1$ a *possible ij-tail* if XY ends with the substring S_i. With this definition, we obtain that

$$A \times \{S_i\} = \cup_{1\leq j\leq k} \left(\mathcal{B}_j \times \{\text{possible } ij\text{-tails}\}\right). \tag{2.4}$$

Comparing Figure 2 to Figure 1, we see that the generating function for the possible ij-tails is given by $C_{ij}(x,q)$. Therefore, we can translate (2.4) into a set of equations in terms of generating functions for $i = 1, \ldots, k$:

$$G(x,q)x^{w(S_i)}q^{\ell(S_i)} - \sum_{j=1}^{k} B_j(x,q)C_{ij}(x,q) = 0. \tag{2.5}$$

Equation (2.4) is identical to the corresponding statement for strings that can be found in [1, 5] (it does not matter whether we deal with strings or compositions in this case), while for the generating functions, the difference is that we also keep track of the weight in the compositions using the variable x. Combining (2.3) and (2.5) results in the following set of $k+1$ equations in $k+1$ unkonwns:

$$\begin{pmatrix} 1-x(1+q) & 1-x & \cdots & 1-x \\ x^{w(S_1)}q^{\ell(S_1)} & -C_{11}(x,q) & \cdots & -C_{1k}(x,q) \\ \vdots & \vdots & \ddots & \vdots \\ x^{w(S_k)}q^{\ell(S_k)} & -C_{k1}(x,q) & \cdots & -C_{kk}(x,q) \end{pmatrix} \begin{pmatrix} G(x,q) \\ B_1(x,q) \\ \vdots \\ B_k(x,q) \end{pmatrix} = \begin{pmatrix} 1-x \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Using Cramer's rule to solve for $G(x,q)$ gives formula (2.1). □

3. Applications of Theorem 2.1

Even though Theorem 2.1 provides an explicit solution to the enumerative problem, it involves the evaluation of determinants which may not be a simple thing to do. While one can easily find explicit formulas for the generating function that do not involve determinants when there are just a few prohibited substrings, it is interesting to know in which cases the determinants can be evaluated for families of prohibited substrings. In this section, we evaluate the determinants for a family of prohibited substrings which generalizes the *well-based sets* used in [3] to count independent sets in certain graphs called *path-schemes*.

Let 1^i denote the string consisting of i 1s, $1 \leq a_1 < a_2 < \cdots < a_k$, and let $V = \cup_{1\leq i\leq k}\{21^{a_i-1}2\}$ be the set of substrings to be avoided. Note that none of the substrings in V is included in any other. Thus we can apply formula (2.1) to find the generating function for the number of compositions avoiding all the substrings in V simultaneously.

Corollary 3.1. *The generating function $G(V;x,q)$ for the number of compositions that avoid the family of substrings V defined above is given* *by*

$$\frac{(1-x)(1+x\sum_{i=1}^{k}(xq)^{a_i})}{(1-x(1+q)+(1-x)x^2q)(1+x\sum_{i=1}^{k}(xq)^{a_i})-(1-x)x^2q}. \tag{3.1}$$

Proof. It is easy to see that the correlation polynomial for the two strings $21^{a_i-1}2$ and $21^{a_j-1}2$ is $C_{ij}(x,q)=\delta_{ij}+x(xq)^{a_i}$, where δ_{ij} is the Kronecker delta. Also, $x^{w(21^{a_i-1}2\})}q^{\ell(21^{a_i-1}2)}=x^{a_i+3}q^{a_i+1}$. From Theorem 2.1 it follows that $G(V;x,q)$ is given by

$$\frac{(1-x)\cdot\begin{vmatrix}-1-x(xq)^{a_1} & -x(xq)^{a_1} & \cdots & -x(xq)^{a_1}\\ -x(xq)^{a_2} & -1-x(xq)^{a_2} & \cdots & -x(xq)^{a_2}\\ \vdots & \vdots & \ddots & \vdots\\ -x(xq)^{a_k} & -x(xq)^{a_k} & \cdots & -1-x(xq)^{a_k}\end{vmatrix}}{\begin{vmatrix}1-x(1+q) & 1-x & 1-x & \cdots & 1-x\\ x^{a_1+3}q^{a_1+1} & -1-x(xq)^{a_1} & -x(xq)^{a_1} & \cdots & -x(xq)^{a_1}\\ x^{a_2+3}q^{a_2+1} & -x(xq)^{a_2} & -1-x(xq)^{a_2} & \cdots & -x(xq)^{a_2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ x^{a_k+3}q^{a_k+1} & -x(xq)^{a_k} & -x(xq)^{a_k} & \cdots & -1-x(xq)^{a_k}\end{vmatrix}}.$$

To compute the determinant in the numerator, replace row 1 by the sum of all rows and then factor out the common factor $(-1-x\sum_{i=1}^{k}(xq)^{a_i})$. Next subtract column 1 from columns $2,3,\dots,k$ to obtain

$$-(1+x\sum_{i=1}^{k}(xq)^{a_i})\cdot\begin{vmatrix}1 & 0 & 0 & \cdots & 0\\ -x(xq)^{a_2} & -1 & 0 & \cdots & 0\\ -x(xq)^{a_3} & 0 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ x(xq)^{a_k} & 0 & 0 & \cdots & -1\end{vmatrix}$$

which simplifies to $(-1)^k\cdot(1+x\sum_{i=1}^{k}(xq)^{a_i})$.

To compute the determinant in the denominator, replace column 1 by the sum of column 1 and $x^2q\cdot$(column $(k+1)$) and for $i=2,3,\dots,k$, replace column i by the difference of column i and (column $(k+1)$) to yield

$$\begin{vmatrix}1-x(1+q)+(1-x)x^2q & 0 & \cdots & 0 & 1-x\\ 0 & -1 & \cdots & 0 & -x(xq)^{a_1}\\ 0 & 0 & \cdots & 0 & -x(xq)^{a_2}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & -1 & -x(xq)^{a_{k-1}}\\ -x^2q & 1 & \cdots & 1 & -1-x(xq)^{a_k}\end{vmatrix}.$$

To obtain an upper triangular matrix we replace the last row in this determinant by

$$\frac{x^2q(\text{row }1)}{1-x(1+q)+(1-x)x^2q}+(\text{row }2)+(\text{row }3)+\cdots+(\text{row }(k+1))$$

which yields that the determinant of the denominator is equal to

$$(-1)^k\Big[(1-x)x^2q-(1-x(1+q)+(1-x)x^2q)(1+x\sum_{i=1}^{k}(xq)^{a_i})\Big],$$

completing the proof. □

Further simplifications of $G(V;x,q)$ are possible whenever $\sum_{i=1}^{k}(xq)^{a_i}$ can be simplified. We provide three examples here. For compactness of display, we define $f(x,q)=1-x(1+q)+(1-x)x^2q$.

Example 3.2. *The set of prohibited substrings* $V_k=\{22,212,\dots,21^{k-1}2\}$ *corresponds to* $a_i=i$ *for* $i=1,\dots,k$. *Then (3.1) implies that* $G(V_k;x,q)$ *is given by*

$$\frac{(1-x)(1-xq+x^2q(1-(xq)^k))}{f(x,q)(1-xq+x^2q(1-(xq)^k))-(1-x)(1-xq)x^2q}.$$

The initial values of $G(V_2;x,q)$ (avoiding 22 and 212) are

$$\begin{aligned}&1+qx+\big(q+q^2\big)x^2+\big(q+2q^2+q^3\big)x^3+\big(q+2q^2+3q^3+q^4\big)x^4\\&\quad+\big(q+4q^2+3q^3+4q^4+q^5\big)x^5+\big(q+5q^2+9q^3+5q^4+5q^5+q^6\big)x^6+\cdots\end{aligned}$$

Example 3.3. *The set* $V_k^e=\{22,2112,\dots,21^{2k}2\}$ *of prohibited substrings that have an even number of 1s is represented by the set* $\{a_1,a_2,\dots\}=\{1,3,5,\dots,2k+1\}$. *Then (3.1) implies that* $G(V_k^e;x,q)$ *is given by*

$$\frac{(1-x)\left(1-(xq)^2+x^2q\left(1-(xq)^{2k+1}\right)\right)}{f(x,q)\left(1-(xq)^2+xq^2\left(1-(xq)^{2k+1}\right)\right)-(1-x)\left(1-(xq)^2\right)x^2q}.$$

The initial values of $G(V_2^e;x,q)$ (avoiding $\{22,2112,211112\}$) are

$$\begin{aligned}&1+xq+(q+q^2)x^2+(q+2q^2+q^3)x^3+(q+2q^2+3q^3+q^4)x^4\\&\quad+(q+4q^2+4q^3+4q^4+q^5)x^5+(q+5q^2+9q^3+6q^4+5q^5+q^6)x^6\\&\quad+(q+6q^2+13q^3+16q^4+9q^5+6q^6+q^7)x^7\\&\quad+(q+7q^2+19q^3+28q^4+26q^5+12q^6+7q^7+q^8)x^8+\cdots\end{aligned}$$

Example 3.4. *The set* $V_k^o = \{212, 21112, \ldots, 21^{2k-1}2\}$ *of prohibited substrings that have an odd number of 1s is represented by the set* $\{a_1, a_2, \ldots\} = \{2, 4, 6, \ldots, 2k\}$. *Then (3.1) implies that* $G(V_k^o; x, q)$ *is given by*

$$\frac{(1-x)\left(1-(xq)^2+x^3q^2\left(1-(xq)^{2k}\right)\right)}{f(x,q)\,(1-(xq)^2+x^3q^2\,(1-(xq)^{2k}))-(1-x)\,(1-(xq)^2)\,x^2q}.$$

The initial values of $G(V_2^o; x, q)$ (avoiding $\{212, 21112\}$) are as follows:

$$\begin{aligned}1 &+ xq + \left(q+q^2\right)x^2 + \left(q+2q^2+q^3\right)x^3 + \left(q+3q^2+3q^3+q^4\right)x^4\\ &+ \left(q+4q^2+5q^3+4q^4+q^5\right)x^5 + \left(q+5q^2+10q^3+8q^4+5q^5+q^6\right)x^6\\ &+ \left(q+6q^2+15q^3+18q^4+11q^5+6q^6+q^7\right)x^7\\ &+ \left(q+7q^2+21q^3+33q^4+30q^5+15q^6+7q^7+q^8\right)x^8 + \cdots\end{aligned}$$

Clearly, other families of substrings can be created that allow for similar simplification of the generating function.

References

[1] L. J. Guibas and A. M. Odlyzko, String overlaps, pattern matching, and nontransitive games, *Journal Comb. Theory Series A* **30** (1981), 19–42.

[2] S. Heubach and T. Mansour, *Combinatorics of Compositions and Words*, CRC Press, Boca Raton, 2009.

[3] S. Kitaev, Counting independent sets on path-schemes, *Journal of Integer Sequences* **9**, no. 2 (2006), Article 06.2.2, 8pp.

[4] A.N. Myers, Forbidden substrings on weighted alphabets, *The Australasian Journal of Combinatorics* **45** (2009), 59–65.

[5] B. Winterfjord, Binary strings and substring avoidance, Master thesis, CTH and Göteborg University (2002).